

# Proceedings of 1st ICSU-WDS Conference

Global Data for Global Science



3 – 6 September 2011  
Kyoto University, Kyoto, Japan



WORLD DATA SYSTEM

## The 1st ICSU World Data System Conference - Global Data for Global Science -

September 3-6, 2011

Kyoto University, Kyoto, Japan



Prof. Huadong Guo  
(President of CODATA)



Prof. Hideo Miyahara  
(President of NICT)



Prof. Jean-Bernard Minster  
(Chair of WDS-SC)



Prof. Kiyoshi Yoshikawa  
(Executive Director of  
Kyoto University)

### WDS-SC Committee members (2009-2012)

- Jean-Bernard Minster, **Chair**. (Scripps Institution of Oceanography/UCSD)
- David Clark (National Geophysical Data Center/NOAA)
- Pierre Cilliers (Hermanus Magnetic Observatory)
- Michael Diepenbroek, **Vice Chair**. (Institute for Marine Environmental Sciences, University Bremen)
- Françoise Genova (Strasbourg Astronomical Data Centre)
- Luiz Horta (Large Scale Biosphere-Atmosphere Experiment in Amazonia /INPE)
- Wim Hugo (South African Earth Observation System)
- Ruth Neilan (Central Bureau of the International GNSS Service/JPL)
- Lesley Rickards (Permanent Service for Mean Sea Level/BODC)
- Takashi Watanabe (Solar-Terrestrial Environment Laboratory, Nagoya University)
- Baoping Yan (Computer Network Information Center/CAS)
- Michael Zgurovsky (National Technical University, Kiev Polytechnic Institute)

### *Ex officio*

- Ray Harris (ICSU ad-hoc Strategic Coordinating Committee for Information and Data)
- Yasuhiro Murayama (National Institute for Information and Communications Technology)
- Mustapha Mokrane (Science and Information Technology Officer, ICSU)

**Proceedings of  
1st ICSU-WDS Conference  
Global Data for Global Science**

**3–6 September 2011  
Kyoto University, Kyoto, Japan**

**Published by  
ICSU-WDS International Programme Office**

**October 2012**

## **On this issue**

This volume includes reports and preprints of presented papers at the 1st ICSU-WDS Conference—*Global Data for Global Science*—held on 3–6 September 2011 in Kyoto, Japan. These papers have been submitted to a special issue of the Data Science Journal, which is the online journal of CODATA\*. Although most papers in this issue were refereed, minor revisions may be applied during the editorial process, before posting finally in the journal. Owing to this reason, all papers in this volume should be treated as ‘preprints’.

In total, 46 papers are published in this volume, including 8 invited papers and 38 contributed papers. The online version of this issue will be posted soon on the WDS website\*\*. Enquiries on this publication should be addressed to Takashi Watanabe.

This volume is published by the ICSU-WDS International Programme Office (WDS IPO).

On behalf of the Editorial Committee

Takashi Watanabe (Managing Editor)  
takashi.watanabe@icsu-wds.org

## **Editorial Committee**

J.-B. Minster (*Chair*), T. Watanabe (*Managing Editor*), D. Clark, M. Diepenbroek, F. Genova, L. Horta, W. Hugo, T. Iyemori, M. Mokrane, Y. Murayama, R. Neilan, L. Rickards, B. Yan, M. Zgurovsky

\* [www.codata.org/](http://www.codata.org/)

\*\* [www.icsu-wds.org/](http://www.icsu-wds.org/)

## Contents

Global Data for Global Science	J.-B. Minster	ix
Summary and Shared Understandings		xii
Schedule of the Conference		xiv
<b>Partnership and Coordination</b>		
ICSU and the Challenges of Data and Information Management for International Science	P. Fox, and R. Harris	2
<b>Activity Report from WDS Members</b>		
The Global Observing Systems Information Center (GOSIC): A Comprehensive Portal for Global Climate Data and Information	H. Diamond	15
Development of WDS Russian-Ukrainian Segment	M. Shaimardanov, A. Gvishiani, M. Zgurovsky, A. Sterin, A. Kuznetsov, N. Sergeyeva, E. Kharin, and K. Yefremov	19
The Contribution of A Geophysical Data Service: The International Service of Geomagnetic Indices	M. Menvielle	29
Operations of The World Data Centre for Geomagnetism, Edinburgh	S. J. Reay, E. Clarke, E. Dawson, and S. Macmillan	33
Activities and Plan of Center for Geophysics (Beijing) from WDC to WDS	F. Peng, M. Ma, L. Peng, J. Zhang, G. Chen, Y. Li, B. Sun, and Y. Zhang	38
Strasbourg Astronomical Data Centre (CDS)	F. Genova	42
Experience and Strategy of Biodiversity Data Integration in Taiwan	K. T. Shao	46
Development of Global Soil Information Facilities	N. H. Batjes, H. I. Reuter, P. Tempel, T. Heng, J. G. B. Leenaars, and P. S. Bindraban	55
The British Geological Survey's New Geomagnetic Data Web Service	E. Dawson, J. Lownde, and P. Reddy	60
Data Handling within the International VLBI Service	D. Behrend	66
The Activities at World Data Center for Geomagnetism Mumbai, India	M. Doiphode, R. Nimje, and S. Alex	70

Japanese Contribution to the World Data Center for Oceanography	A. Seta, S. Wakamatsu, T. Miyake, and Y. Iwabuchi	74
Data and Information Activities of ICSWSE, Kyushu University, Japan	S. Abe, K. Yumoto, A. Ikeda, T. Uozumi, and G. Maeda	77
Information about the World Data Centers for Solar-Terrestrial Physics and Solid Earth Physics; Regional Multidisciplinary Initiatives of Russian-Ukrainian World Data Centers Segment for Occurrence in the World Data System	N. Sergeyeva, E. Kharin, L. Zabarinskaya, A. Rodnikov, I. Shestopalov, T. Krylova, and M. Nisilevich	82
The Application of an Online Data Visualization Tool, PTPLOT, in The World Data Centre (WDC) for Solar-terrestrial Science (STS) in IPS Radio and Space Services, Australia	K. Wang, and C. Yuile	86
<b>Data Intensive Multidisciplinary Science</b>		
Lessons Learned from Data Management Activities after Great East Japan Earthquake in March 2011	A. Kitamoto	91
Multi-disciplinary Approaches to Intelligently Sharing Large-volumes of Real-time Sensor Data during Natural Disasters	S. E. Middleton, Z. A. Sabeur, P. Löwe, M. Hammitzsch, S. Tavakoli, and S. Poslad	95
Mathematical Tools for Geomagnetic Data Monitoring and INTERMAGNET Russian Segment	A. Soloviev, S. Bogoutdinov, A. Gvishiani, R. Kulchinskiy, A. Chulliat, and J. Zlotnicki	99
A New Approach to Research Data Archiving for WDS Sustainable Data Integration in China	W. Juanle, S. Jiulin, Y. Yaping, S. Jia, and Y. Xiafang	104
The State of IPY Data Management: The Japanese Contribution and Legacy	M. Kanao, A. Kadokura, M. Okada, T. Yamnouchi, K. Shiraishi, N. Sato, and M. A. Parsons	108
Beyond Data Regulation: Finding Solution to a Persistent Problem of Marine Debris and Sea Surface Temperature Measurement along Coastline of Lagos, Nigeria	O. A. Ediang, and A. A. Ediang	113
Visualization of Flux Rope Generation Process Using Large Quantities of MHD Simulation Data	Y. Kubota, K. Yamamoto, K. Fukazawa, and K. T. Murata	117

## Application of Information Technologies to Data Systems

A Science Cloud for Data Intensive Sciences	K. Murata, S. Watari, T. Nagatsuma, M. Kunitake, H. Watanabe, K. Yamamoto, Y. Kubota, H. Kato, T. Tsugawa, K. Ukawa, K. Muranaga, E. Kimura, O. Tatebe, K. Fukazawa, and Y. Murayama	122
SPASE: The Connection Among Solar and Space Physics Data Centers	J. R. Thieman, D. A. Roberts, and T. A. King	130
Cell Based GIS as Cellular Automata for Disaster Spreading Predictions and Required Data Systems	K. Arai	137
Data Mining Approaches for Habitats and Stopovers Discovery of Migratory Birds	Q. Xu, Z. Luo, and B. Yan	142
Metadata Modelling of IPv6 Wireless Sensor Network in Heihe River Watershed	W. Luo, and B. Yan	153
An Integrated Management System of Multipoint Space Weather Observation	H. Watanabe, K. Yamamoto, T. Tsugawa, T. Nagatsuma, S. Watari, Y. Murayama, and K. T. Murata	158
Inter-university Upper Atmosphere Global Observation Network (IUGONET)	H. Hayashi, Y. Koyama, T. Hori, Y. Tanaka, S. Abe, A. Shinbori, M. Kagitani, T. Kouno, D. Yoshida, S. Ueno, N. Kaneda, M. Yoneda, N. Umemura, H. Tadokoro, T. Motoba, and IUGONET project team	162
Innovations for the Curation and Sharing of African Social Survey Data	H. L. Woolfrey	168
A Maturity Model for Digital Data Centres	W. Hugo	172
Geophysical Data Stewardship in the 21 <sup>st</sup> Century at the National Geophysical Data Center (NGDC)	E. A. Kihn, and C. G. Fox	176
Application and Metadata Format of Cryosphere Data Archive Partnership (CRDAP)	H. Yabuki	180
Toward a Normalized XML Schema for the GGP Data Archives	A. Gabillon, J.-P. Barriot, Y. Verschelle, and B. Ducarme	184
Research Environment and Information Service of Space Weather Cloud	S. Watari, H. Kato, K. T. Murata, K. Yamamoto, H. Watanabe, Y. Kubota, and M. Kunitake	192
Digital Database of Long-term Solar Chromospheric Variation	R. Kitai, S. Ueno, H. Maehara, S. Shirakawa, M. Katoda, Y. Hada, Y. Tomita, H. Hayashi, A. Asai, H. Isobe, H. Goto, and S. Yamashita	196

A State-space Approach to Explore the Strain Behavior Before and After the 2003 Tokachi-Oki Earthquake (M8)	T. Takanami, G. Kitagawa, H. Peng, A. T. Linde, and I. S. Sacks	199
Metadata Publication and Search System in JAMSTEC	Y. Hanafusa, H. Saito, and Y. Abe	203
Solar-terrestrial Data Analysis and Reference System (STARS) - Its High Potentiality for Collaborative Research	M. Kunitake, K. Yamamoto, S. Watari, K. Ukawa, H. Kato, E. Kimura, Y. Murayama, and K. T. Murata	207
<b>Data Publication</b>		
The World Ocean Database	S. Levitus, J. I. Antonov, O. K. Baranova, T. P. Boyer, C. L. Coleman, H. E. Garcia, A. I. Grodsky, D. R. Johnson, R. A. Locarnini, A. V. Mishonov, J. R. Reagan, C. L. Sazama, D. Seidov, I. Smolyar, E. S. Yarosh, and M. M. Zweng	212
Is Data Publication the Right Metaphor?	M. Parsons, and P. Fox	218
Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing	IJ. J. Aalbersberg, J. Dunham, and H. Koers	233
Re-evaluation of Geomagnetic Field Observation Data at Syowa Station, Antarctica	K. Takahashi, Y. Minamoto, S. Arita, I. Tomofumi, and A. Kadokura	241
A Data-driven Method for Selecting Optimal Models Based on Graphical Visualisation of Differences in Sequentially Fitted ROC Model Parameters	K. M. Mwitondi, R. E. Moustafa, and A. S. Hadi	245
Digitization of Bromide Paper Records to Extract One-minute Geomagnetic Data	N. Mashiko, T. Yamamoto, M. Akutagawa, and Y. Minamoto	251



## Global Data for Global Science

The Proceedings of the 1st ICSU World Data System Conference, held 3–6 September 2011 at Kyoto University, Japan, contains the papers submitted to the *Data Science Journal*, CODATA. With the theme ‘Global Data for Global Science’, this conference was remarkably consonant with the long-term ICSU vision of

‘...a world where science is used for the benefit of all, excellence in science is valued and scientific knowledge is effectively linked to policy-making. In such a world, universal and equitable access to high quality scientific data and information is a reality and all countries have the scientific capacity to use these and to contribute to generating the new knowledge that is necessary to establish their own development pathways in a sustainable manner.’

The importance attached by ICSU to equitable access to data and information is not new. The history detailed in the report of the *ad hoc Strategic Committee on Information and Data* (SCID) to the 2008 ICSU General Assembly (GA) highlights the creation of the *Federation of Astronomy and Geophysics data analysis Services* (FAGS) and the *World Data Centres* (WDCs) prior to the 1958 International Geophysical Year. At the beginning of the 21st century, about a dozen services and over 50 data centres worldwide provided the international scientific community with access to data and information focussed on the natural sciences. This level of international participation clearly demonstrated the success of these endeavours.

However, self-reviews by FAGS and WDCs—conducted after five decades of operations, and incorporated in the SCID report—highlighted the need to restructure ICSU’s data and information activities into a global system. The initial motivations, rooted in the desire of the scientific community to mitigate the fragmentation associated with the Cold War, had evaporated. Instead the focus had shifted to what was described in ICSU Committee on Data for Science and Technology (CODATA) documents, as the ‘digital divide’. So instead of facilitating an East-West exchange of scientific data and information, the need had become to develop more effective North-South exchange channels. As a result, on October 23, 2008, the ICSU GA voted to establish the ICSU *World Data System*, with the following charge:

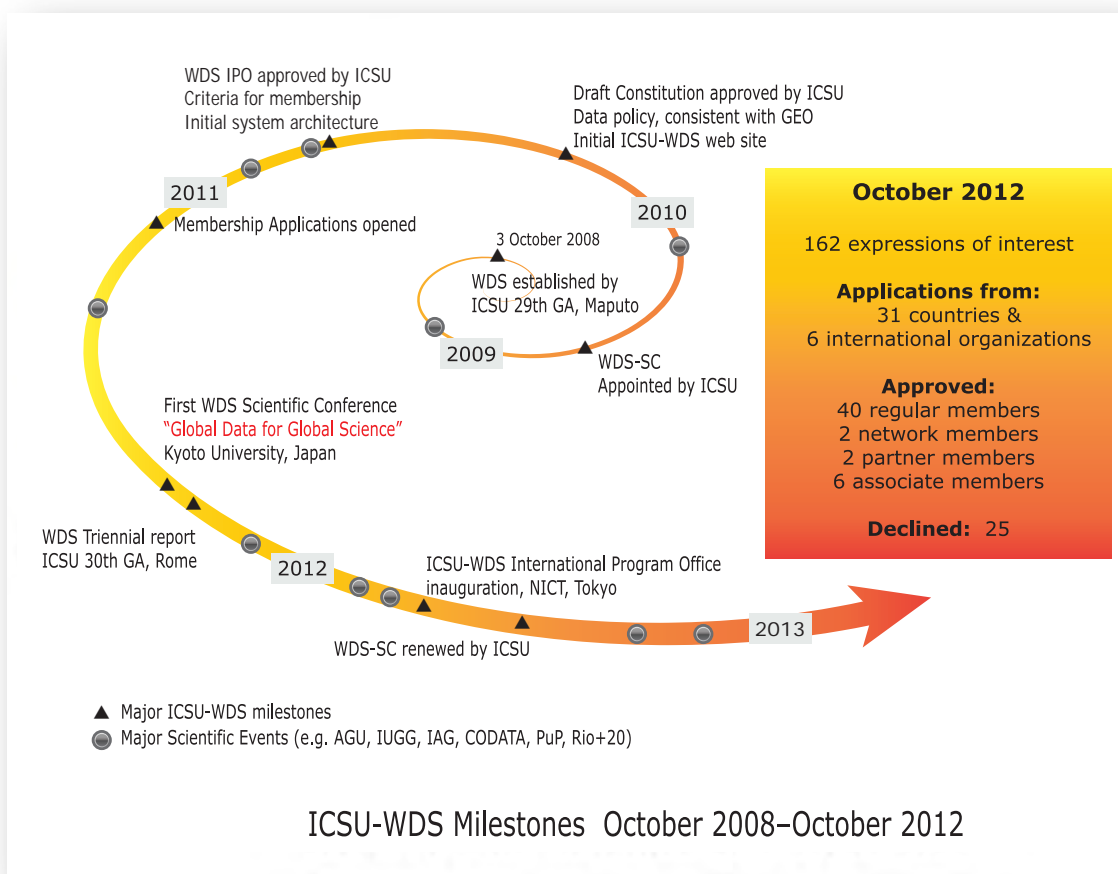
‘To emphasize the critical importance of global data for global science activities; to further ICSU strategic scientific outcomes by addressing pressing societal needs (e.g. sustainable development, digital divide); to highlight the very positive impact of universal and equitable access to data and information; to support services for data

and information long-term stewardship, and to promote and support data publication and citation.’

The first WDS Scientific Committee (SC) was appointed by ICSU in June, 2009. Its broadly diverse membership—both disciplinary and geographical—enabled it to address its tasks quite effectively. The figure below shows the various milestones met by the WDS-SC, which led the ICSU 30th GA held in Rome to approve a decision to ‘consolidate and expand the ICSU World Data System’.

Since its inception, the WDS-SC agreed on an ambitious mission for WDS: *to provide an International framework for (1) long-term stewardship of science-driven data and products; (2) access to, and dissemination of, quality-assessed data and information and quality-verified global data analysis services; (3) interdisciplinary integration of data (including human sciences); (4) harmonization and interoperability across disciplines, geography, and technology over time; and (5) implementation of a global community of excellence for the dissemination of data and information.*

During its first term, the WDS-SC drafted a WDS constitution that was approved by ICSU in 2010; adopted a data policy consistent with the *Group on Earth Observations*



data sharing principles; and defined several categories of membership—Regular, Network, Partner, and Associate—depending on the level of data activity. Applications for membership are subjected to a formal review procedure (initially by the WDS-SC, and now also by current WDS Members). Relevant documents are all to be found on the ICSU-WDS website, and were initially developed through collaboration between the WDC for Geoinformatics and Sustainable Development (Kyiv, Ukraine) and the Centre for Marine Environmental Sciences (Bremen, Germany). ([www.icsu-wds.org](http://www.icsu-wds.org))

In 2010, a major development was approval by ICSU of the creation of a WDS International Program Office (IPO). Following a worldwide call for proposals, the WDS IPO, hosted by the National Institute of Information and Communication Technology in Tokyo, Japan was inaugurated in May 2012. The WDS IPO is now fully staffed and enjoys substantial support from the Government of Japan.

As the reader will see in this special issue, WDS includes data archive centres, data analysis centres, service providers, data producers and developers, observing systems and networks, and virtual observatories at scales ranging from regional, to national, to global. Recently, ICSU strongly urged WDS to expand their membership beyond the natural sciences, and to recruit members from other domains such as biodiversity and the health sciences. This confirms the long-held opinion of the WDS-SC that the architecture of WDS should evolve into a ‘*scale-free* system of data systems’.

Also illustrated in the figure are special efforts by the WDS-SC to raise the visibility of WDS in various global scientific meetings and symposia. Particularly notable is WDS participation in the ICSU-CODATA biennial conferences, and ICSU-sponsored symposia such as *IPY 2012* in Montréal, Canada; *Planet Under Pressure 2012* in London, UK; and the *Forum on Science, Technology and Innovation for Sustainable Development* at *Rio+20* in Rio, Brazil. Indeed, collaborations with CODATA, ICSU’s Integrated Research on Disaster Risk programme, and ICSU National and Union Members were strongly recommended by SCID and the subsequent ICSU *ad hoc Strategic Coordinating Committee on Information and Data*.

ICSU-WDS now enjoys a momentum that builds on decades of expertise developed by its predecessors. The historical distinction between data repositories and data analysis services is rapidly being blurred. The WDS-SC is of the opinion that this trend serves global science well.

On behalf of the ICSU-WDS Scientific Committee

Jean-Bernard Minster (Chair)

## **Schedule of the Conference**

The 1st ICSU-WDS Conference was originally scheduled to be held on 3–6 September 2011. Owing to a typhoon hitting the Kyoto area unexpectedly, activities scheduled on the first day of the Conference (3 September) were unfortunately cancelled. Invited talks initially scheduled on the first day (Sessions 1, 2, and 3) were therefore moved to the second day (4 September), and the invited talks in Sessions 4, 5A, and 5B were also presented on that day. Furthermore, the oral presentations of contributed papers for these sessions were replaced by posters. From the third day (5 September), we returned to the prearranged timetable. The revised schedule is shown below, while the original schedule and programme are posted on the Conference page of the WDS website ([www.ICSU-WDS.org](http://www.ICSU-WDS.org)).

### **4 September 2011 (SUN)**

Session 1: Opening Talks

Session 2: What is ICSU-WDS?

Session 3: Partnership and Coordination

Session 4: Activity Reports from WDS Members

Session 5-A: Data Intensive Multidisciplinary Science (Disasters Data Management)

Session 5-B: Data Intensive Multidisciplinary Science (Multidisciplinary Data Systems for Earth Science)

### **5 September 2011 (MON)**

Session 6: Application of Information Technologies to Data Systems

Session 7: Data Publication

### **6 September 2011 (TUE)**

Open Session

Session 8: ICSU WDS Members' and Partners' Open Forum

Summary Session

## **Summary and Shared Understandings of 1st ICSU-WDS Conference**

### **Summary**

Around 155 participants (including 86 local participants) from over 22 countries attended the 1st ICSU World Data System (WDS) Conference—*Global Data for Global Science*—held at Kyoto University, Kyoto, Japan, from 3 to 6 September 2011. Participants included representatives of data centres and data services covering a wide range of scientific disciplines, data scientists, and engineers working in a variety of fields such as natural sciences, social sciences, and information technologies, as well as data publishers.

It should be noted that the Conference was held a mere six months after the 2011 Tohoku earthquake and subsequent tsunami, and the failure of the Fukushima Daiichi nuclear power plant. In this exceptional context, a session on disaster data was organized in collaboration with the ICSU-cosponsored Integrated Risk and Disaster Research (IRDR) programme and the Disaster Prevention Research Institute at Kyoto University.

Furthermore, the passage of typhoon Talas (T201112) forced the organizers to cancel events planned for 3 September and restructure the conference agenda with little warning.

The 23 invited talks, 36 contributed talks, 70-plus poster papers, and 5 exhibits enabled the nascent WDS community to engage in effective scientific collaboration, and provided a constructive forum for lively exchanges of views and ideas. Important feedback was also provided to the WDS Scientific Committee during a Members' and Partners' open forum, which will certainly influence and help shape WDS in the future.

The Conference was also an opportunity to initiate a dialogue with WDS stakeholders and important partners such as the Committee on Data for Science and Technology (CODATA) and the International Oceanographic Data and Information Exchange.

The Conference was sponsored by the ICSU-WDS International Program Office—hosted by the National Institute of Information and Communications Technology—and by the Science Council of Japan and the Graduate School of Science, Kyoto University.

The Conference proceedings will be published in a special issue of the CODATA Data Science Journal (<http://www.codata.org/dsj/index.html>). More details on the Conference can be found on the WDS website (<http://icsu-wds.org>).

### **Shared Understandings**

The ICSU World Data System—by supporting the missions and objectives of ICSU—upholds the Principle of Universality of Science. It seeks to ensure the long-term stewardship and

provision of quality-assessed data and data services to the international science community and other stakeholders. WDS should also aim at integration of multidisciplinary scientific data and information to address the needs of ICSU programmes including the Earth System Sustainability Initiative.

Scientists are not only users but also providers of data. The WDS policy of open and full access to data will benefit the international scientific community and ultimately society at large. The concepts of data publication and data citation should be adopted and promoted by WDS to facilitate the release of data, with proper recognition of providers.

The series of disasters that befell Japan in March 2011, recent events in New Zealand, as well as other events have highlighted the importance of timely and reliable acquisition, management, quality control, and dissemination of disaster-related data. On-going collaboration between ICSU programmes—WDS, CODATA, and IRDR—will be critical for building interoperable data systems that guarantee full and open access to data in support of better disaster prediction, understanding, and mitigation, and hence of more effective response to devastating events.

6 September 2011

Kyoto, Japan

# Partnership and Coordination



# ICSU AND THE CHALLENGES OF DATA AND INFORMATION MANAGEMENT FOR INTERNATIONAL SCIENCE

*Peter Fox<sup>1\*</sup>, and Ray Harris<sup>2</sup>*

<sup>\*1</sup>*Rensselaer Polytechnic Institute (RPI), 110 8<sup>th</sup> St, Troy, NY 12180, United States*

*Email: [pfox@cs.rpi.edu](mailto:pfox@cs.rpi.edu)*

<sup>2</sup>*University College London, Gower St, London WC1E 6BT, UK*

*Email: [ray.harris@ucl.ac.uk](mailto:ray.harris@ucl.ac.uk)*

## ABSTRACT

*The International Council for Science (ICSU) vision explicitly recognises the value of data and information to science and particularly emphasises the urgent requirement for universal and equitable access to high quality scientific data and information. A universal public domain for scientific data and information will be transformative for both science and society. Over the last several years, two ad-hoc ICSU committees, the Strategic Committee on Information and Data (SCID) and the Strategic Coordinating Committee on Information and Data (SCCID), produced key reports that make 5 and 14 recommendations respectively aimed at improving universal and equitable access to data and information for science, and providing direction for key international scientific bodies such as the Committee on Data for Science and Technology (CODATA) as well as a newly ratified (by ICSU in 2008) formation of the World Data System. This contribution outlines the framing context for both committees based on the changed world scene for scientific data conduct in the 21<sup>st</sup> century. We include details on the relevant recommendations and important consequences for the worldwide community of data providers and consumers ultimately leading to a conclusion and avenues for advancement that must be carried to the many thousands of data scientists world-wide.*

**Keywords;** Data Science; Data Access; Data Release; Data Management; Interdisciplinary Science.

## 1 INTRODUCTION

There is no doubt that scientific data and information<sup>1</sup> have made significant impacts on our society. The understanding of contemporary climate change is dependent upon high quality data and information, the major advances in our understanding of the origins and evolution of the universe are built upon the solid foundation of high quality astronomy data, and in the last decade the major steps taken in understanding the human genome have been dependent on high quality data in the life sciences.

The challenges facing science as far as managing scientific data and information are concerned fall into two contrasting camps. First, there are the enormous volumes of data that are being and will be produced in science sectors such as astronomy, biomedicine, environmental science, Earth observation and particle physics, often termed the data deluge (Hey and Trefethen 2003), the data tsunami or the fire hose of data. Second, there is the reluctance of some scientists to share their data because of the overheads incurred in preparing the data so that they can be shared (Nelson 2009). Several recent reviews have implicitly or explicitly noted these two major challenges, including reports produced by the European Commission (EC 2010, GRDI 2011), the Organisation for Economic Cooperation and Development (OECD 2007) and an alliance of German science organisations (Anon 2008). The size of the problem can be quickly gauged in Table 1. Hilbert and Lopez (2011) have compiled estimates through modelling of the world's technological capacity to store, communicate and compute information, and the table presents a selection of their results for storage, telecommunications and general purpose computation. It is no surprise to see that the annual growth rates for data and data processing are very high, with the estimates for the year 2007 being far in excess of those for the year 1986. An interesting analogy on data storage is that if all the data used in the world were written to CD-ROMs and the CD-ROMs piled up in a single stack, the stack thereby created would stretch from the Earth to the Moon and a quarter of the way back again. The explosion in the quantity of data and information available to science continues apace. Whilst the absolute size of this explosion varies across disciplines, the general trend is for rapid growth in all disciplines from the social sciences to seismology, from the humanities and social sciences to high energy physics. By the end of 2011 it was estimated that 30,000 human genome sequences will have been completed (Nature 2010b),

<sup>1</sup>For the purposes of this paper a definition of data and information is given in Appendix A.



creating information about billions of bases and requiring petabytes of data storage. A study by the International Data Corporation (IDC 2010) in 2010 estimated that by the year 2020 there will be 35 zettabytes (ZB) of digital data created per annum. The IDC estimate of the total digital storage capacity in the world to be available in 2020 is 15 ZB, less than half the amount of digital data produced by then. When the Square Kilometre Array radio telescope in astronomy is fully functional in 2024 it will be able to produce more digital data than is capable of being processed in all the world's computers put together.

**Table 1.** Estimates of the world's capacity to store, communicate and compute data and information. Source: Hilbert and Lopez (2011) who give detailed descriptions of the variables plus links to back-up tables for each variable.

		1986	1993	2000	2007	Percent annual rate of change 1986-2007
Storage	MB optimal compression per capita (installed capacity)	539	2,866	8,988	44,716	23
	Approximate CD-ROM equivalent per capita	<1	4	12	61	
	Percent digital	0.8	3	25	94	
Telecommunications	MB optimal compression per capita per day (effective capacity)	0.16	0.23	1.01	27	28
General purpose computation	MIPS per capita (installed capacity)	0.06	0.8	48	968	58

While the recognition of the data deluge has been relatively recent, the International Council for Science (ICSU) has been actively involved in the management of data and information since the 1950s when the World Data Centres were established as part of the International Geophysical Year of 1957-58. In its vision statement (ICSU 2006, 2011) ICSU explicitly recognises the value of data and information to science:

*[the vision of ICSU is] ... a world where science is used for the benefit of all, excellence in science is valued and scientific knowledge is effectively linked to policy making. In such a world, universal and equitable access to high quality scientific data and information is a reality ...*

Within the ICSU family there are organisations actively exploring how to implement the ICSU vision for universal and equitable access to high quality scientific data and information against the backdrop of the major challenges noted earlier. These organisations include the following.

- Committee on Data for Science and Technology (CODATA)
- International Network for the Availability of Scientific Publications (INASP)
- International Council for Scientific and Technical Information (ICSTI)
- World Data System (WDS)

The purpose of this paper is to explore the challenges of data and information management for international science by focussing on the ways in which the International Council for Science has reviewed and acted upon these challenges. The paper is based on discussions and reports from two *ad-hoc* ICSU committees, the Strategic Committee on Information and Data (SCID, ICSU 2008) and the Strategic Coordinating Committee on Information and Data (SCCID, ICSU 2011). These committees produced reports that make 5 and 14 recommendations respectively aimed at improving universal and equitable access to data and information for science, and providing direction for key international scientific bodies such as CODATA and the World Data System.

## 2 DATA AND INFORMATION CHALLENGES

### 2.1 The Fourth Paradigm

The volume and complexity of data and information available to science has given rise to what some call the Fourth Paradigm of science (Hey et al 2009). This fourth paradigm puts data-intensive science into the context of its three main predecessors, namely (Bell et al 2009):

- First Paradigm. Observation, descriptions of natural phenomena and experimentation.
- Second Paradigm. Theoretical science such as Newton's laws of motion and Maxwell's equations.
- Third Paradigm. Simulation and modelling, such as in astronomy.
- Fourth Paradigm. Data-intensive science that exploits the large volumes of data in new ways for scientific exploration, such as the International Virtual Observatory Alliance in astronomy.

The Fourth Paradigm is certainly characterised by massive data volumes, but also by complexity of data sets and by the potential for extensive cross-fertilisation of data, information, information technology and publishing. The Fourth Paradigm acknowledges the central role played by data in science and in some ways reflects the empirical but computationally-limited First Paradigm.

### 2.2 Data overload

The last decade has seen substantial change in the creation, use and management of scientific data and information, not least amongst scientists, data managers, libraries and publishers. The traditional reward mechanisms for scientists have been in grants, publications, citations, prizes and promotion. There is now a strong interest in publishing data and for such publication reward or recognition systems do not commonly exist. When scientists use data they must often now be concerned with the conditions of access to the data, for example copyright, onward distribution and use licences, as well as with the data themselves, and they must also enter the arena of standards and interoperability so that they can read the digital data needed for their work and produce outputs that are accessible to other scientists. Data managers are now often in charge of very large data repositories, for example in astronomy, and they need to provide tools to help scientists use data.

In the last decade there has been a rapid expansion of the responsibilities of libraries to encompass digital repositories, including data repositories, alongside traditional books and journals. This means in particular that there is a need for knowledge of deposit and access conditions, digital rights such as Creative Commons licences and the use of standards, metadata schemes and persistent identifiers, such as those promoted by DataCite (2011) to ensure correct data citation. In parallel, publishers have also made major changes to encompass digital data. Some publishers encourage, or even require, the submission of data to either their own journals as supplemental material or to recommended data centres. As an illustration, the journal *Nature* makes it mandatory for certain types of human genome data that are associated with accepted publications to be submitted to a community-endorsed, public repository: for example, DNA and RNA sequence data have to be submitted to the Protein DataBank or UniProt or to GenBank/EMBL/DDBJ nucleotide sequence database.

Open access journals have been changing the landscape of journal publishing away from the traditional model of payment by subscribers and libraries. More than 6,000 titles are currently registered in the Directory of Open Access Journals. Traditional publishers are experimenting with new business models and increasingly offering open access options, for example the relationship between the publisher Elsevier and the PANGAEA data centre in Germany.

### 2.3 Data complexity

There are four important characteristics of complex data; high dimensionality, multimodality, multi-scale and heterogeneity. Multimode data appears in fields ranging from neuroscience to astronomy and while its origins are in imagery, it is now appearing in application areas such as air quality where the modes of measurement are very different. In a range of fields from environment and climate to biomedicine, crossing scales has emerged as a key need. For example, crossing the scales from molecular to cell to tissue, and then to organ and organism scales in animals, each of which has different measurement and structural data representations. While there are promising approaches to reduce complexity, further complications such as dependency among dimensions may

result in redundancy and inaccuracy in semantics. However, progress using a variety of means (algorithmic, representational and computational) is beginning to occur in some fields. In the present context, this is all dependent on data and information and the application of data science.

## **2.4 Changing expectations**

Expectations on scientists in the area of data and information management have evolved and increased over the past decades as science itself has moved into the data-intensive era. The main drivers of these changing expectations are the changing nature of science, science funders, policy makers and governments as well as society at large. Science is more than ever a globalized international activity with a strong collaborative component. To carry out their research, scientists are not only expected to manage, share and archive their data professionally but also to use cutting-edge information and communication technologies for data and information discovery and analysis. Unfortunately, the vast majority of scientists who work with data are neither well equipped nor trained to meet these high expectations. On the other hand, data scientists are working at the forefront of information technology and have the knowledge to develop the tools and training in this important area of data management.

Scientists have to respond and adapt to new expectations coming from governments and funding agencies, such as the National Science Foundation in the United States, which are increasingly requiring a full data management plan to be submitted with applications for research funding. Scientists are also facing new expectations from society at large as the outcome of their research is used by policy makers in designing public policies that affect society directly and by applied users from both the public and private sectors. Scientists need not only to communicate honestly and openly their research, but also to share and open their data to public use and media scrutiny, as illustrated in the field of climate change by the so-called Climategate scandal (Nature 2010a).

## **2.5 Digital divide**

While there are rapid advances in data capture technologies and the ability to handle the data deluge, there is still a digital divide with those scientists in the less economically developed countries (LEDCs) who lack access to both data and technology. The meetings of the World Summit on the Information Society in both 2003 and 2005 identified the digital divide as a major concern for society. Data on computer availability (UN 2008) show that while the countries in the North have better than one computer for every two people, the LEDC countries have about one computer for every 10-20 people. Broadband penetration in LEDCs lags similarly behind the provision of computers, although undersea cables are set to have a major impact on connectivity in African countries. Data from the International Telecommunication Union for 2009 (ITU 2009) show that there is only one fixed broadband subscriber for every 1,000 people in Africa compared to one for every 200 in Europe. In the countries of the North, National Research and Education Networks (NRENs), such as GEANT2, SINET and AARNet, have developed alongside commercial broadband capacity to provide dedicated services and support to research and education. NRENs are either absent from or only recently emerging in the LEDCs, particularly in Sub-Saharan Africa. Whilst there have been significant recent connectivity developments in South Africa and Kenya, the picture in the rest of Africa is still very much one of limited or poor connectivity.

## **3 ICSU STRATEGIC REVIEWS OF INFORMATION AND DATA**

During the last decade ICSU has launched several strategic reviews of the capability of international science to handle the growing volume and complexity of data and information. In 2004 ICSU's Panel Area Assessment (PAA) on Information and Data (ICSU 2004) identified three main requirements for improved data and information management in science. First, to ensure universal and equitable access to data and information. Better access will lead to better science. Second, to develop an improved capability to manage data professionally, vital both for access to good quality data now and to ensure that future scientists will have access to historical data. Third, to consider the question of who pays for data and for professional data management because reliable funding is always required for the creation and management of data and information: no funding means no data. The PAA report was extensive in its recommendations, but kept returning to ways to encourage and enable the scientific community to improve its strategic capability to think about and then take action on data and information management, both within the ICSU family of national members and scientific unions and in relation to other organisations.

Following the PAA report, ICSU established a Strategic Committee on Information and Data (SCID) to examine how in practice to facilitate a new, coordinated global approach to scientific data and information that ensures equitable access to quality data and information for research, education and informed decision-making. The SCID report (ICSU 2008) recommended the creation of a new World Data System (WDS) based on the former World Data Centres (WDCs) and the Federation of Astronomical and Geophysical data analysis Services (FAGS). The purpose of the World Data System is to provide a coordinated, professional approach to the management of scientific data and the production of services based on the data. The World Data System is described in the next section of this paper. In addition, the SCID report encouraged CODATA to become more prominent in science by having clearer strategic goals that are linked with ICSU's vision for science. ICSU has national members and scientific union members, and these members provide a vital means of communication throughout the scientific community. ICSU members were also encouraged by the SCID report to engage proactively in professional data management for science.

The most recent ICSU examination of strategy for scientific data and information management has been the Strategic Coordinating Committee on Information and Data (SCCID, ICSU 2011). SCCID continued the process of developing strategic priorities for the improvement of professional data management in science.

## **4 PROGRESS WITH THE WORLD DATA SYSTEM**

### **4.1 WDS objectives from the ICSU perspective**

In exploring desirable attributes of an 'ideal' system, the SCID process included a separation of these attributes into three categories: Mission, Coordination and Execution (SCID Report, pp. 13-14.). The aim of this distinction was to match certain functions with existing international organizations (e.g. ICSU, CODATA) but importantly to identify functionality gaps in both national and international (i.e. world) data centres. In essence, an ideal system became a combination of existing activities as well as the new World Data System. For example, for the Mission of an ideal system, a) Enable and encourage the advancement of science through the open provision of high quality data and information services, b) Increase global knowledge and reduce the knowledge divide between richer and poorer countries by providing universal and equitable access to scientific data and products, c) Identify structural gaps in data and information provision and seek solutions to fill these gaps, and d) Develop further the structure for long term stewardship of scientific data, including in the form of formal public libraries for data. Coordination included: a) Fostering multi-disciplinary, large scale, complex science, b) Leading and championing professional data management, and c) Informing discussions on data policy from a science perspective. Finally Execution emphasized: a) Taking a lead role in developing, testing and implementing standards for data access to provide services for all scientists, b) Promoting the publication of data and data products, with the associated recognition and accreditation that are common to peer-reviewed science publications, c) Providing reliable and trustworthy science-reviewed data and derived products, d) Serving discipline-based science communities with exemplary data repositories and data products, e) Integrating data sets using community-consensus algorithms, and lastly, f) Enabling seamless access to data.

The WDS adopted the vision provided in the SCID (2008) report and has articulated the following goals (Minster et al. this volume, <http://www.icsu-wds.org/>): a) Enabling universal and equitable access to scientific data and information, b) Facilitate better access to data, c) Strive for simpler access to data, d) Work to provide quality assured data and information, e) Promote improved data stewardship, f) Work to reduce the digital divide and g) Ultimately, provide data for better science. These goals represent an initial amalgam of SCID's Mission and Execution suggestions with a focus on near and medium term activities.

### **4.2 Criteria for professional data management in the WDS**

The SCCID report (ICSU 2011) took to task the issue for professionalization, both for data science and data management, and the task for the WDS is challenging. In particular SCCID, in their 'Recommendation 6' stated: *We recommend the development of education at university and college level in the new and vital field of data science. The example curriculum included in appendix D [of the SCCID report] can be used as a starting point for course development.* Also, in Recommendation 7: *We recommend that both the CODATA and the World Data System biennial conferences include forums for data professionals, including data librarians, to share experiences across a range of science disciplines.*

WDS has an opportunity to share experience with the broader community on their experience with *what are the required credentials, knowledge and skills (technical, scientific, personal, user needs, etc.) to train and give data professionals more explicit recognition* (SCCID report, p. 21). The WDS must participate in the identification and definition of a community of data professional peers and provide a variety of fora for them to meet and exchange ideas, experiences and solutions.

### **4.3 Active participation in the WDS**

The WDS, via its Scientific Committee (WDS-SC) is responsible for soliciting active participation in the WDS. The forms of membership include: Regular, Network, Partner and Associate. At the time of writing there were 29 Regular members and 1 for each of the remaining categories (see <http://www.icsu-wds.org/wds-members/wds-members> for current statistics). As a mark of the new diversity in the WDS, the active member list comprises: past World Data Centres (WDCs), Federation of Astronomical and Geophysical data analysis Services (FAGS), research institutes, international scientific unions, consortia and commercial publishers, all of which bring a welcome diversity to the WDS. The future seems bright for a revitalized and synergistic World Data System.

### **4.4 Future plans for the WDS**

The WDS-Transition Team (WDS-TT) and subsequently the WDS-SC took responsibility for SCID's second recommendation to 'work closely with CODATA and with the new ICSU ad hoc Strategic Coordinating Committee.' In particular the WDS-SC efforts to develop and begin implementation of a strategic plan for the WDS will be key to its future. Now that the WDS is forming and the SCCID term has ended, additional consideration must be given to exactly which entities, beyond CODATA, the WDS-SC and the WDS, to work closely with as well as how they interact. For example, the structure for making scientific data and information management effective and efficient will need to be re-conceived because the SCCID role of coordination has ended and the interaction and coordination roles need to be re-defined. SCCID recommendations 2 (open access), 4 (beyond the science community), and 7 (new forums for data professionals) are significant for ICSU and require adequate resource allocation and attention for the desired outcomes to be achieved. SCCID recommendation 8 (explicit visibility enhancement of data and science engagement) appears well within the current WDS-SC strategy and commensurate with the emerging set of WDS members. However, the later SCCID recommendations including 10 (exploitation of standards expertise), 12 (multi-way organizational engagement for closing the digital divide), and 14 (in-reach, raising the profile of data science) are expected to place a strain on the current WDS capability and capacity to participate beyond what many institutional hosts for World Data System nodes may consider their core mission. Experience from both SCID and SCCID deliberations suggest that deliberate and frequent communication between ICSU, WDS-SC, CODATA, other relevant ICSU inter-disciplinary bodies, International Scientific Unions and ICSU National Members will be required.

## **5 FUTURE IMPROVEMENTS IN DATA AND INFORMATION MANAGEMENT FOR SCIENCE**

As noted earlier, after the production of the SCID report and the stimulation for the creation of the World Data System, ICSU took a wider view of other data and information challenges in its Strategic Coordinating Committee on Information and Data (SCCID). The purpose of this section of the paper is to present some of the key recommendations of the SCCID report.

### **5.1 Best practice in data management**

Advice and guidance on the principles of best practice in data and information management is needed, both for the members of the ICSU family and for all of science. Every sector of science can learn from previous experience in professional data management, and improvements in data management will lead to better science by improving access to data and information. A short, practical guide to best practice for professional data management in science has been produced and it is included in Appendix B of this paper. The guide draws on experience from (amongst others) the Protein Data Bank, the International Polar Year, the Intergovernmental Panel on Climate Change and the International Virtual Observatory Alliance in astronomy.

## 5.2 Open access

The Open Access movement emerged in the new era of electronic information and the concept was initially introduced and formalised in the field of access to publications through the “3B” declarations listed below.

- Budapest Open Access Initiative, <http://www.soros.org/openaccess/read.shtml>
- Bethesda Statement on Open Access Publishing, <http://www.earlham.edu/~peters/fos/bethesda.htm>
- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, <http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/>

The 34 members of the Organisation for Economic Co-operation and Development (OECD) have agreed at ministerial level a statement on *OECD Guidelines and Principles for Access to Research Data from Public Funding* (OECD 2007). On open access the OECD principles state:

*Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.*

The OECD principles cover 13 topics in total, including transparency, legal conformity, interoperability and quality. The principles are regarded by the OECD as “soft law”, that is they have a moral authority and strong support by ministers but they are not legally binding on OECD member states.

The open access notion has been extended to various degrees of unlimited access to online data and information and, in the domain of research data, is clearly related to the needs and practices of data sharing and re-use. As open access has a generally positive impact on scientific progress it is increasingly supported, via formal statements and policies, by research institutions, scientific unions, government bodies and funding agencies. However, the terminology used in open access is uncertain and at times confusing. Uncertainty has been created by the use of different ideas such as full and open access, free access, public access, universal and equitable access and by the (somewhat artificial) distinction between access to data and access to publications. At the same time some initiatives have been trying to formalise ‘open’ beyond the initial access definitions, for example open data, open archives, open content, open knowledge and open notebook science. There is certainly merit in establishing a forum for the exploration and eventual agreement in relation to science of all the terms used under the broad umbrella of Open Access. Without agreement it seems likely that uncertainty will grow.

## 5.3 Data as a publication

The evolution of data analysis and the publication of scientific research are parallel to the development of science paradigms described earlier in this paper. In the first to third paradigms data were contained within scientific papers in that the results from scientific theories, models and experiments were presented within the paper, at least in summary form. The fourth paradigm (Hey et al 2009) implicitly or explicitly disconnects the results of research from the data that were used to prepare the research findings in that the data are too voluminous to publish in a conventional form. This evolution, closely linked both to scientific progress and technological advances, calls for a fresh view of the concept of “publishing” data sets in trustworthy repositories with long-term sustainability prospects. The concomitant recognition of, and credit accorded to, such activities are viewed increasingly as essential to such endeavours. As a result, the roles of libraries and publishers are changing in regard to data and information. Most countries have a national deposit library which has a legal responsibility to hold a copy of any publication and such legal deposit libraries might provide a valuable vehicle for ensuring long-term data stewardship. As an example, the national deposit library in The Netherlands is one of the world leaders on considering data as a publication and so requiring easy access and long term stewardship. At the same time, methods for citing data have evolved rapidly, such as the Digital Object Identifier concept. This provides a ready link with peer-recognition of the work of scientists who produce high quality data and with the many emerging mechanisms to communicate science.

## 5.4 The role of education

The conduct of science research is increasingly data driven: from data assimilation through modelling, simulation and visualisation to long-term time series of data. It is now well established that data have an intrinsic value that outlast current science foci. Unfortunately there is only part time attention given by most scientists to data science. Perhaps most important is the need to give a new value to science in the form of data citation, attribution and data publication. It is essential to identify the required credentials, knowledge and skills (technical, scientific, personal, user needs, etc.) to acquire and to give data professionals more explicit recognition. Formal training in the key cognitive and skill areas of data science will enable graduates to become key participants in eScience collaborations. The need is to teach key methodologies in application areas based on real research experience and build a skill-set.

## **6 CONCLUSIONS**

### **6.1 Momentum**

In science and the popular press today, barely a week goes by without an article, blog or social media dissemination appearing, focusing on 'data'. Big data is hot news: "floods", "tsunamis", "tidal waves", exascale, etc., but so are other elements of complex data such as dimensionality, scale, modality, source heterogeneity and inter-disciplinarity. Importantly, these factors are now being placed in plain view of the scientific community and the responses are not just coming from science, but from funding and operational agencies, governments, commercial sectors and the private sector. Some truly inspiring opportunities lay ahead. However, it is very important to point out that while many consider this new news, ICSU has been paying very careful (though perhaps without being highly visible) attention since well before 2004 when the PAA (on Information and Data) report was released. The visibility was substantially increased with the SCID and SCCID activities, which substantially took on the role of strategic examination of sectors, and needs to fulfil the ICSU vision, as well as stimulating active coordination among key organizations. Around 2005-2006 several "International Year" activities were being planned as 50<sup>th</sup> year celebrations of the International Geophysical year (1957-1958). The Electronic Geophysical Year, the International Polar Year, the International Heliospheric Year and the International Year of Planet Earth ran approximately from 2007-2008 and all had data at the forefront. Of note, was the link back to IGY and further back to previous IPYs and that for all our modern advances and technology, effect management and use of data was still a tremendous challenge. Looking back, it was most likely the confluence of national and international attention with these celebratory IGY community efforts that truly allowed data science and informatics to fully emerge in their respective areas, and reinforce that ICSU was clearly paying attention and stepping up for its role in strategic coordination.

As an aggregate, the early 'results' arising out of much greater awareness among international organizations around data, willingness to broaden the conversation, and a much more inclusive trend are very promising in advancing the ICSU vision for data and information. Exemplars include the last three ICSTI conferences devoted to advanced aspects of large data and visualization, the formation of a CODATA Task Group on Data Publication, the aforementioned WDS membership composition and response, and leading efforts such as the Polar Information Commons (PIC; <http://www.polarcommons.org/>) and their PIC 'badging' efforts are truly advancing the discussion around data as a first class science object, and in turn challenging more traditional approaches to data.

It is not possible, however, to present a uniformly glowing report of responses. Both the SCID and SCCID reports strongly emphasized the role for ICSU national members and ICSU Scientific Unions as being essential stakeholders and participants in a cohesive future. After all, the universality of science lies at the heart of the ICSU vision, and the direct resources and the attention of ICSU national members are required to implement such a vision. Several nations, or aggregates of nations, as well as scientific unions have provided substantial (and in some cases, long standing) responses to the presently articulated data agendas. Unfortunately, as a whole, the response of nations and scientific unions is poor and remains a challenge for ICSU but more-so for the scientific communities. Such under-served outcomes, may ultimately lead to more and undesirable digital divides.

On the matter of extant digital divides, the EDC-LEDC distinction and the need to erase or at least dramatically reduce the inherent disadvantages faced in LEDCs in the contemporary digital information world, opens up significant opportunities for scientists with minimal resources (data, computation) to become data scientists and

thus be thrust well into a small but growing cadre of such career professionals. The opportunity is also fraught with cultural challenges but is clearly on the agenda for organizations such as the WDS and CODATA.

As we bring this paper to a close, one trend in the levels of discussions and participation reported herein is notable. ICSU activities such as the PAA, SCID, SCCID, WDS-SC involve approximately tens of people, and this is true for executive/ committee leads for related organizations; ICSTI, INASP, PIC, etc. Their greater activities such as conferences and workshops in turn reach often hundreds of participants (e.g. the WDS Science Conference featured in this volume). These numbers fall far short of the much greater audience penetration that is needed. This means, penetration through to scientific unions, professional societies and the working ranks where the numbers are in the thousands. In other words, scaling is needed to move from tens to hundreds to thousands, otherwise the universality of science and the data and information that underpins it will be incomplete.

## 6.2 Avenues for action

In conclusion, we see several avenues for action for a variety of stakeholders.

We call on present and prospective new WDS members of all types, to be forward looking and conversant with the greater goals and vision embraced by ICSU on behalf of the entire scientific community. The extant reports articulate many attributes and details of these goals, including new roles and new partners.

New communities are entering the conversation regarding data. Cultural, organizational and economic barriers for publishers, librarians, and technical solutions from commercial sectors are ever present. We suggest that the required ensuing conversations and the explorations and demonstrations of mutual benefit, often measured by very different means, are an initial step worth pursuing.

Coordination of inevitably overlapping roles and responsibilities (and often authorities) around data, information and its generation, access and use is essential. The tendency to ignore the reality that true coordination and collaboration is actually carried out among *individuals* in organizations is yet one more barrier to progress (especially noting that an individual may participate in many organizations). Even so, while the coordination opportunities abound often the resources and attention assigned to them are discordant. We encourage that strategic attention in each organization be paid to timely and valuable coordination activities, retaining a level of agility to respond to new and changing needs.

To begin the immense task of increasing participation, or in turn addressing the scale of penetration of science communities' attention to data, an immediate avenue is effective engagement at the professional society and scientific union level. Motivated and knowledgeable scientists and managers can introduce discussions of data policies, access, management, etc. at any and every turn. A new group of peers is emerging: those with a career approach to data and information.

Data scientists are here to stay but their explicit numbers are small as are the programs for educational preparation. A clear call to action is the introduction of initially graduate level courses, leading to the wide spread establishment of data science curricula, and degree and career paths for data scientists. In the longer term, undergraduate majors in data science are inevitable. The future will tell the remainder of the story.

## 7 ACKNOWLEDGEMENTS

The authors were both members of the ICSU Strategic Committee for Information and Data (SCID), 2007 - 2008 and the Strategic Coordinating Committee for Information and Data (SCCID), 2009 – 2011. We are very grateful to all our colleagues on the two committees for the extensive and detailed discussions that have informed this paper. The SCID and SCCID reports including the members of the committees can be accessed at the ICSU web site [www.icsu.org](http://www.icsu.org).

## 8 REFERENCES

Anon (2008) *Priority Initiative Digital Information*, Alliance of German Science Organisations, Berlin 11 June 2008.



- Bell G, Hey, T. & Szalay, A. (2009) Beyond the data deluge, *Science* 323, 6 March 2009, 1297-1298.
- DataCite (2011) *DataCite*, available at <http://www.datacite.org/>, last accessed 29 November 2011.
- EC (2010) *Riding the wave. How Europe can gain from the rising tide of scientific data*, Final report of the High Level Expert Group on Scientific Data, European Commission, Brussels, 2010, 36pp.
- GRDI (2011) *Global Research Data Infrastructures: The GRDI2020 Vision*, project funded by the European Commission, 7th Framework Programme for Research and Technological Development, [www.grdi2020.eu](http://www.grdi2020.eu)
- Hey T, Tansley, S. & Tolle, K. (2009) *The Fourth Paradigm. Data-intensive scientific discovery*, Microsoft Research, Redmond, Washington, 252pp.
- Hey, A. J. G. & Trefethen, A.E (2003) The data deluge: an e-science perspective, in F Berman, G Fox and A J G Hey (eds) *Grid Computing - Making the Global Infrastructure a Reality*, Wiley and Sons, Chichester, 809 – 824.
- Hilbert M & Lopez, P. (2011) The world's technological capacity to store, communicate and compute information, *Science* 332, 1 April 2011, 60-65.
- ICSU (2004) *Panel Area Assessment on Information and Data*, International Council for Science, Paris.
- ICSU (2006) *ICSU Strategic Plan 2006-2011*, International Council for Science, Paris, 64pp
- ICSU (2008) *Ad hoc Strategic Committee on Information and Data*, Final Report to the ICSU Committee on Scientific Planning and Review, International Council for Science, Paris, 36pp.
- ICSU (2011) *Ad hoc Strategic Coordinating Committee on Information and Data*, Final Report to the ICSU Committee on Scientific Planning and Review, International Council for Science, Paris, 35pp. ICSU (2011) *ICSU Strategic Plan II, 2012-2017*, International Council for Science, Paris, 56pp.
- IDC (2010) *IDC Digital Universe Study*, sponsored by EMC, May 2010, available at <http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm>
- ITU (2009) *The World in 2009, ICT facts and figures*, [http://www.itu.int/ITU-D/ict/material/Telecom09\\_flyer.pdf](http://www.itu.int/ITU-D/ict/material/Telecom09_flyer.pdf)
- Nature (2010a) Closing the Climategate, *Nature* 468, 18 November 2010, 345.
- Nature (2010b) Genomes by the thousand, *Nature* 476, 28 October 2010, 1026-1027.
- Nelson B (2009) Empty archives, *Nature* 461, 10 September 2009, 160-163.
- OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding, <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- UN (2008) *United Nations Global Development Goals Indicators 2008*, <http://mdgs.un.org/unsd/mdg/Default.aspx>

## 9 Appendix A

### Definition of data and information

Data and information can be considered as a continuum ranging from raw research data through to published papers. “Data” includes at a minimum digital observation, scientific monitoring, data from sensors, metadata, model output and scenarios, qualitative or observed behavioural data, visualizations, and statistical data collected for administrative or commercial purposes. Data are generally viewed as input to the research process. “Information” generally refers to conclusions obtained from analysis of data and the results of research. But the

distinction between data and information is flexible and will vary according to the situation. Increasingly, the output of research (traditionally viewed as “information”) includes data and has become input to other research, rendering the output-input distinction between data and information meaningless.

## 10 Appendix B

### Principles of best practice for data and information management

#### 1. Policy

- Document early the reason(s) for the data policy and the policy itself, and make documents available online.
- Articulate the desired outcomes of the data policy.
- Identify and be explicit about the benefit/cost ratio of professional data management.
- Ensure that guidelines for participation are easily accessible by encouraging open access to data policies, practices and experiences.

##### *Examples*

- ICSU World Data System data policy, available at <http://www.icsu-wds.org/organization/data-policy>
- International Polar Year data policy, available at [http://classic.ipy.org/Subcommittees/final\\_ipy\\_data\\_policy.pdf](http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf)
- OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007, available at <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- Panton Principles for open data in science, see <http://pantonprinciples.org/>
- Creative Commons licences, available at <http://creativecommons.org/choose/>

#### 2. Governance

- Ensure that data management is an integral and funded part of project planning, approval and performance measurement.
- Appoint expert advisory groups where necessary and charge them with defined tasks.
- Exploit major international science conferences and events as dates/locations to hold meetings, and use these meetings to encourage interactions between scientists and data/information professionals.
- Acknowledge the different skills and roles required in professional data and information management.
- Ensure open, online access to all minutes of meetings and decisions taken.

##### *Examples*

- The core agreement for the Worldwide Protein Data Bank, 2003, available at [http://www wwpdb.org/wwpdb\\_charter.html](http://www wwpdb.org/wwpdb_charter.html)
- The Intergovernmental Panel on Climate Change structure and working groups, see [http://www.ipcc.ch/working\\_groups/working\\_groups.htm](http://www.ipcc.ch/working_groups/working_groups.htm)

#### 3. Planning and organisation

- Consider the advantages and disadvantages of distributed versus centralised data repository models in the light of user needs.
- Use service-based data access methods.
- Exploit what already exists for data management.
- Data infrastructure should be completed, ready and available in time for its use by scientists in research projects.
- Incorporate user feedback into all aspects of the data management lifecycle.

##### *Example*

- GenBank, the annotated collection of all publicly available DNA sequences, see <http://www.ncbi.nlm.nih.gov/genbank/GenbankOverview.html>

#### 4. Standards and tools

- Use international standards (e.g. ISO, OGC, XML, GML) where possible, and if not possible then base domain-specific standards on international standards.
- Provide tools to support the implementation of the standards used, including documentation on how to use the project data.

##### *Examples*

- Dublin Core Metadata Initiative, available at <http://dublincore.org/documents/dces/>
- ISO 19115 for geographical information and services, available at [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=26020](http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020)
- Open Geospatial Consortium standards and specifications, see <http://www.opengeospatial.org/standards>
- International Virtual Observatory Alliance, documents and standards, available at <http://www.ivoa.net/Documents/>

#### 5. Data management and stewardship

- Minimise uncertainty at all phases of the data lifecycle, including for example working with manufacturers to avoid device dependency for data and information.
- Embrace science-programme and project-level data management planning.
- Ensure that documented plans for long term stewardship of data exist.
- Implement a plan for formal process for data and information selection and appraisal.
- Produce a plan for data stewardship at the outset of a project or programme, not as the last item in the plan.

##### *Examples*

- International Polar Year Data and Information Service, see <http://ipydis.org/index.html>
- Research Information Network, stewardship of digital research data – principles and guidelines, 2008, <http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines>

#### 6. Data access

- Minimise the burden on the providers of data.
- Provide a single portal for user discovery from distributed sources of information.
- Implement open access policies where appropriate.

##### *Examples*

- GEO portal, see [http://www.geoportal.org/web/guest/geo\\_home](http://www.geoportal.org/web/guest/geo_home)
- Ocean Data Portal, see <http://www.oceandataportal.org/>

# Activity Reports from WDS Members



# THE GLOBAL OBSERVING SYSTEMS INFORMATION CENTER (GOSIC): A COMPREHENSIVE PORTAL FOR GLOBAL CLIMATE DATA AND INFORMATION

*H Diamond<sup>1\*</sup>*

*<sup>1</sup>NOAA's National Climatic Data Center, 1100 Wayne Avenue, Silver Spring, Maryland USA 20910  
Email: howard.diamond@noaa.gov*

## ABSTRACT

*The Global Observing Systems Information Center (GOSIC), which was initiated in 1997 at the request of the Global Climate Observing System (GCOS) Steering Committee, responds to a need identified by the global climate observing community for easier and more effective access to observational climate data and information. GOSIC manages an online portal providing an entry point for users of climate-related global observing systems data and information systems and also helps serve the needs of the World Data Center for Meteorology, Asheville.*

**Keywords:** Climate, Observations, Data, GCOS, ECV, NCDC

## 1 INTRODUCTION

The Global Observing Systems Information Center (GOSIC) [see <http://gotic.org>] was established by the Global Climate Observing System (GCOS) program as a way to provide better and more centralized access to an extremely diverse array of climate related datasets that cross atmospheric, oceanic, and terrestrial domains that are collected from non-satellite, in-situ and satellite observing platforms. To aid in improving access to climate observing datasets, the GOSIC staff located at NOAA's National Climatic Data Center (NCDC) has developed an Essential Climate Variables (ECV) Data Access Matrix. The basic intent of the ECV matrix is to provide users with a centralized resource to access climate observing datasets from trusted sources for each of the defined atmospheric, oceanic and terrestrial variables as well as metadata and reference documentation. The GOSIC staff is constantly adding new climate datasets identified by the global observing systems and the world data centers as the best available collection of data for a particular variable. Information on spatial and temporal coverage, data gaps, quality control and additional data needs is also available in the matrix. The ECV Matrix is meant to be a "one-stop-shop" to access trusted ECV datasets and information and can be accessed online at: <http://gotic.org/ios/MATRICES/ECV/ECV-matrix.htm> body text of the paper should be Times New Roman, 10pt and justified. The introduction should outline the motivation for the research and contain a review of the relevant literature. The paper should be organized into logical parts or sections and include a description of the problem, the means of solution, results and conclusions.

## 2 MISSION AND BACKGROUND OF THE GOSIC

The GOSIC's mission is to provide a broad spectrum of users with a centralized resource to aid in finding worldwide climate observing system datasets and related information from the Global Climate Observing System (GCOS), the Global Ocean Observing System (GOOS) and the Global Terrestrial Observing System (GTOS). However, the GOSIC does not limit itself to these three systems, and also incorporates access to climate data from other partner programs such as the Global Atmosphere Watch (GAW) and the Atmospheric Circulation Reconstructions over the Earth (ACRE), and does so in a consistent fashion across a diverse array of international data centers.

The GOSIC was established in 1997 by the GCOS Steering Committee to develop methods for easy on-line access to the comprehensive base of Global Observing Systems data and information. Under an initial joint National Oceanic and Atmospheric Administration (NOAA) and National Aeronautics and Space Administration

(NASA) grant, the GOSIC was developed at the University of Delaware, College of Marine Studies, building on experience with information systems for international climate research programs. Guidance and evaluation of the GOSIC is provided by the various scientific and steering committees of the three observing systems. The GOSIC reports at these meetings and receives directions for further development. Formal performance reviews were conducted in 2001 and 2003 by groups appointed by each observing system, and the results of these were extremely helpful in shaping the form and function of the GOSIC. Since 2007 the GOSIC has become operational and is a facility operated under the auspices of the U.S. Global Climate Observing System (USGCOS) program based at NOAA's NCDC, and is run on behalf of the international observing community (Diamond, 2009). In addition, the GOSIC has become a key part of the infrastructure in helping to serve data via the World Data Center for Meteorology, Asheville [see <http://wdca-meteorology.org>].

### 3 THE ESSENTIAL CLIMATE VARIABLES (ECV) DATA ACCESS MATRIX

The ECV Matrix was developed by the GOSIC staff to provide users with a centralized "one-stop-shop" resource to access climate observing datasets from trusted sources for each of the defined atmospheric, oceanic and terrestrial essential climate variables.

GCOS first defined a list of the ECVs in 2003 that were identified as "feasible for global implementation and have a high impact" on the requirements of the United Nations Framework Convention on Climate Change (UNFCCC) and the Intergovernmental Panel on Climate Change (IPCC). Originally there were 44 ECVs that included such variables as air temperature, precipitation, sea surface temperature, salinity, snow cover, and albedo. By the end of August 2010, GCOS published an updated set of ECVs that built on the original set by adding variables such as soil moisture, soil carbon, ocean oxygen content, and also recognizing the role of precursors in forming ozone and aerosols. The updated list of 50 ECVs is as follows:

#### **Atmospheric (over Land, Sea & Ice):**

**Surface:** Pressure, Air Temperature, Precipitation, Surface Radiation Budget, Water Vapor, Wind Speed and Direction.

**Upper-Air:** Cloud properties, Earth Radiation Budget, Temperature, Water Vapor, Wind Speed and Direction.

**Composition:** Carbon Dioxide, Methane and other Long-Lived Green House Gases (Nitrous Oxide (N<sub>2</sub>O), Chlorofluorocarbons (CFCs), Hydrochlorofluorocarbons (HCFCs), Hydrofluorocarbons (HFCs), Sulfur Hexafluoride (SF<sub>6</sub>), and Perfluorocarbons (PFCs)), Ozone, Aerosol Properties, Precursors supporting the Aerosols and Ozone ECVs (NO<sub>2</sub>, SO<sub>2</sub>, HCHO, CO).

#### **Oceanic:**

**Surface:** Carbon Dioxide Partial Pressure, Current, Ocean Color (for Biological Activity), Sea Ice, Sea Level, Sea State, Sea Surface Salinity (SSS), Sea Surface Temperature (SST), Ocean Acidity, Phytoplankton.

**Sub-surface:** Temperature, Salinity, Current, Nutrients, Carbon, Ocean Tracers, Phytoplankton, Ocean Acidity, Oxygen.

**Terrestrial:** River Discharge, Water Use, Ground Water, Lakes Levels, Snow Cover, Glaciers and Ice Caps, Permafrost and Seasonally-Frozen Ground, Albedo, Land Cover (including Vegetation Type), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Leaf Area Index (LAI), Above Ground Biomass, Fire disturbance, Soil moisture, Ice Sheets, Soil Carbon.

The design of the ECV Matrix provides users with a quick overview of the ECVs identified by domain and easy access to data download, metadata and other information with a minimal amount of clicks. See Figure 1 below for a view of the ECV Matrix main page on the GOSIC Portal.

Each of the 50 ECVs has a dedicated web page that provides a list of trusted data sets identified by the global observing systems and world data centers and divided into two categories: 1) Non-Satellite or in-situ 2) Satellite (see example in Fig. 2). The data sets for each of the ECVs are identified based on documentation such as the Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC (Update 2010) August 2010, GCOS-138, GTOS-184, GTOS-76, WMO-TD/No.1523 publication I and the GTOS Assessments of the Status of the Development of Standards for the Terrestrial Essential Climate Variables. II Each data set has a description of content, data download link, metadata, data documentation, list of variables, program information, and more. These web pages also include additional information such as definitions, spatial and temporal coverage, data gaps, quality control, contributing and status of networks, additional data needs, and reference documentation.

The GOSIC Portal provides convenient, central, one-stop access to data and information identified by the Global Climate Observing System (GCOS), the Global Ocean Observing System (GOOS) and the Global Terrestrial Observing System (GTOS) and their partner programs, such as the Global Atmosphere Watch (GAW) and regional observing systems, such as the GOOS Regional Alliances (GRA). [More About the GOSIC](#) - [What's New on the GOSIC Portal](#)

<p><b>How do I find Climate Datasets Quickly?</b></p>	<ul style="list-style-type: none"> <li>Search Data by GCOS Essential Climate Variable (ECV)</li> <li>Search Global Observing Data</li> <li>Search using Data Access Matrices</li> <li>Text Search</li> <li>Metadata Search</li> <li>Find and plot data with the Climate Explorer (CNMI)</li> <li>NASA Satellite Data Access (Mirador)</li> <li>NCAR DSI Research Data Archive</li> <li>Global Surface Databank</li> <li>Asia Pacific Data Research Center</li> <li>JAMSTEC Data Search Portal</li> </ul>
<p><b>Access to Observing System Data, Metadata &amp; Information</b></p>	<ul style="list-style-type: none"> <li>GCOS - The Global Climate Observing System</li> <li>GAW - The Global Atmosphere Watch</li> <li>GTOS - The Global Terrestrial Observing System</li> <li>GOOS - The Global Ocean Observing System             <ul style="list-style-type: none"> <li>National Activities Summaries of Operational &amp; Planned Observation Programs</li> <li>Overview of the GOOS Observation Programs' Growth</li> </ul> </li> <li>GRA - The GOOS Regional Alliances - News</li> <li>Metadata</li> <li>Maps and Google Earth™ Products</li> <li>Publications</li> <li>Observational Datasets in Support of the IPCC</li> <li>GEO - <small>Group on Earth Observations</small> - GEO Group on Earth Observations:             <ul style="list-style-type: none"> <li>GEO Portal</li> <li>GOSIC on the GEO Portal</li> <li>GOSIC Works with GEO Portal to Provide Climate Information</li> </ul> </li> </ul>
<p><b>Related Observing System Information</b></p>	<ul style="list-style-type: none"> <li>What's New</li> <li>Data Flow Diagrams</li> <li>Meeting Calendars: GCOS - GOOS - GTOS</li> <li>Publications/Documents: GCOS - GOOS - GTOS - GAW - GOSIC</li> <li>Data Management Plans: GCOS-GOOS-GTOS - GCOS - GOOS - GTOS - GAW</li> <li>Strategic Plans: GCOS - GAW - GCOS-GOOS-GTOS</li> <li>Review the scientific and technical basis for the design of GCOS - GOOS - GTOS - GCOS-GOOS-GTOS</li> <li>WHO Integrated Global Observing System (WIGOS)</li> <li>Scientific Panels (AGPC, OOPC &amp; TOPC)</li> <li>Frequently Asked Questions (FAQ)</li> <li><small>World Meteorological Organization and Atmospheric Administration (NOAA)</small></li> </ul>

Figure 1. Main View of the ECV Matrix

**GCOS Atmospheric Surface ECV<sup>®</sup> Pressure**

*\*over land, sea and ice*

- Definition
- Introduction
- Atmospheric Surface Domain ECVs
- Contributing Networks & Status
- Observations over Land and Ocean

**Data, Product, Metadata and Information Access**

[ECV Matrix Main Page](#) | [About the ECV Matrix](#) | [Reference Documents](#) | [Contact](#) | [Updated June 3, 2011](#)

Non-Satellite or in-situ	Satellite
<ul style="list-style-type: none"> <li><b>GCOS Surface Network (GSN) Monthly Data (GSHM01)</b> (NOAA/NCDC) is a global network of approximately 1000 stations selected from the network of many thousands of existing meteorological stations. The GSN is intended to comprise the best possible set of land stations with a spacing of 2.5 to 5 degrees of latitude, thereby allowing coarse-mesh horizontal analyses for some basic parameters. These data are archived at NCDC after QC of the CLIMAT temperature and precipitation data have been completed (in general about 2 months after the end of the observation month) (see GSNMC data sets below) (data access) (metadata) (data documentation) (GSN program information) (contact)</li> <li><b>GCOS Surface Network (GSN) Monitoring Centre Monthly Data (GSMBC) (DWD)</b> In order to offer an earlier access to the GSN data, a so called "quick" GSN data set is provided as soon as it becomes available (about day 2) of the following month. This data set is marked by a "Q" in the table (see data access below) and does NOT include any information about data quality. The quality flags included are all set by default to "1". The final GSNMC data set, marked as "F", is sent on a monthly basis to the <a href="#">WGA for Meteorology in Asheville, NC, USA</a> (url: <a href="#">http://www.wga.gov</a>)</li> </ul>	

Figure 2. Example of Individual ECV web page - Partial view of the Atmospheric Surface Pressure ECV page

Since its inception, the success of the GOSIC has been in its ability to incorporate new search methodologies as well as to work in a synergistic way with other facilities to ensure maximum exposure of the GOSIC to as wide an audience as possible. The GOSIC has worked with the Group on Earth Observations (GEO)<sup>1</sup>, the World

<sup>1</sup> See <http://www.geowebportal.org/>

Meteorological Organization Information System (WIS)<sup>2</sup>, as well as the NOAA Climate Portal<sup>3</sup>, to ensure that work already undertaken by the GOSIC does not have to be duplicated by these systems. As such, the consistent look and feel of the GOSIC that many users have become accustomed to can also now be found as powering the climate data search engines of these major systems. NCDC has been nominated by the U.S. to be a formal WIS Data Collection and Production Center (DCPC)<sup>4</sup>, and the GOSIC will form a key part of that DCPC function for NCDC; this nomination was confirmed at the 16th WMO Congress in May 2011.

## 4 CONCLUSION

The unique value that the GOSIC offers its users is the ability to search, using a variety of tools, and quickly link users to a wide range of downloadable data sets that reside at multiple data centers around the world via a consistent and user friendly interface. The goal of the GOSIC Portal is to provide access to global observing system data with the fewest number of clicks as well as provide tailored search capabilities through a variety of search tools such as matrices, registries and search by key word, global observing system, data center, program and joint programs, theme, variable, and more.

The newly developed ECV Matrix provides another data access tool that allows users to search for data sets based on the 50 ECVs. Using this tool, the users can efficiently view all of the ECVs by domain (atmosphere, ocean and land), and with the fewest number of clicks access individual web pages for each of the ECVs listing trusted data sets with links to data download, metadata, and other relevant information. The GOSIC staff is constantly adding new climate datasets identified by the global observing systems and the world data centers as the best available collection of data for the ECVs. The GOSIC staff invites persons to become actively involved in the site by providing us feedback at <gosic@noaa.gov>; the staff is quite responsive and their goal is to provide the easiest and most convenient access to global climate observing datasets from the atmospheric, oceanic, and terrestrial observing domains. Finally, the GOSIC has a very small staff and depends on users to help us identify datasets for inclusion. The staff is quite flexible and looks forward to input and suggestion from users to assist us. Therefore, people should feel welcome to provide feedback and rest assured that it will be addressed in a timely fashion.

## 5 ACKNOWLEDGEMENTS

The Director of the World Data Center for Meteorology would first like to acknowledge the work of Ms. Christina Lief who serves as the day-to-day manager of the GOSIC and is responsible for its easy to use interface and up-to-date contact. Second, the Director would like to recognize NOAA's National Climatic Data Center for hosting the WDC and for making the GOSIC utility possible on an operational basis.

## 6 REFERENCES

Diamond, H.J. & C. J. Lief (2009). A Comprehensive Data Portal for Global Climate Information, EOS Trans. AGU, 90(39), doi:10.1029/2009EO390001

<sup>2</sup> See <http://gis.ncdc.noaa.gov/geoportal/catalog/main/home.page>

<sup>3</sup> See <http://climate.gov>

<sup>4</sup> See [http://www.wmo.int/pages/prog/www/WIS/centres\\_en.html](http://www.wmo.int/pages/prog/www/WIS/centres_en.html)



# DEVELOPMENT OF WDS RUSSIAN-UKRAINIAN SEGMENT

*Marsel Shaimardanov<sup>1\*</sup>, Alexei Gvishiani<sup>2</sup>, Michael Zgurovsky<sup>3</sup>, Alexander Sterin<sup>4</sup>, Alexander Kuznetsov<sup>5</sup>, Natalia Sergeyeva<sup>6</sup>, Evgeny Kharin<sup>7</sup>, and Kostiantyn Yefremov<sup>8\*</sup>*

<sup>1</sup>*All-Russian Scientific and Research Institute of Hydrometeorological Information – World Data Center, WDC for Meteorology, 6, Koroleva St., Obninsk, Kaluga Region, 249035 Russian Federation  
Email: marsel@meteo.ru*

<sup>2</sup>*Geophysical Centre of Russian Academy of Science, 3, Molodezhnaya St., 119296 Moscow, Russian Federation  
Email: a.gvishiani@gcras.ru*

<sup>3</sup>*National Technical University of Ukraine “Kyiv Polytechnic Institute”, 37, Peremohy ave., 03056 Kyiv, Ukraine  
Email: zgur@zgurov@kiev.ua*

<sup>4</sup>*All-Russian Scientific and Research Institute of Hydrometeorological Information – World Data Center, WDC for Rockets and Satellites and Rotation of the Earth, 6, Koroleva St., Obninsk, Kaluga Region, 249035 Russian Federation  
Email: sterin@meteo.ru*

<sup>5</sup>*All-Russian Scientific and Research Institute of Hydrometeorological Information – World Data Center, WDC for Oceanography, 6, Koroleva St., Obninsk, Kaluga Region, 249035 Russian Federation  
Email: kuznet@meteo.ru*

<sup>6</sup>*Geophysical Centre of Russian Academy of Science, WDC for Solid Earth Physics, 3, Molodezhnaya St., 119296 Moscow, Russian Federation  
Email: nata@wpcb.ru*

<sup>7</sup>*Geophysical Centre of Russian Academy of Science, WDC for Solar-Terrestrial Physics, 3, Molodezhnaya St., 119296 Moscow, Russian Federation  
Email: kharin@wpcb.ru*

<sup>8</sup>*National Technical University of Ukraine “Kyiv Polytechnic Institute”, WDC for Geoinformatics and Sustainable Development, 37, Peremohy ave., 03056 Kyiv, Ukraine  
Email: k.yefremov@wdc.org.ua*

## ABSTRACT

*Establishment of the Russian-Ukrainian WDS Segment, its state of the art, main priorities and research activities are described. One of the high priority tasks for Segment members is development of common information space – transition from Legacy Systems and individual Services to a common globally interoperable distributed data system that incorporates emerging technologies and new scientific data activities. The new system will build on the potential and added value offered by advanced interconnections between data management, data processing components for disciplinary and multidisciplinary applications. Thereby the principles of architectural organization of intelligent data processing systems are determined in this paper.*

**Keywords:** World Data System, ICSU, Intelligent data processing, Interdisciplinary research, Heterogeneous data sources

## 1 INTRODUCTION

Modern scientific researches connected with the search of answers on global challenges arising in the beginning of XXI century are interdisciplinary and focused on solving of bad structured tasks. The example of such researches is the analysis of sustainable development processes in global and regional context (Zgurovsky, Stratukha, Melnichenko, Voitko, Boldak, Yefremov et al., 2010).

When we talk about interdisciplinary (Somervill, & Rapport, 2000) we mean the usage of data which is quantitative or qualitative assessments that characterize different phenomena or objects. On the basis of this data and models developed in different scientific spheres the generalized (interdisciplinary) models of complete presentation of object of research are developed. As a rule in terms of such researches the formal models are

absent, but at the same time there is a possibility to use the results of objective measurements. Notably the tasks of such researches can be described as bad structured (Newell, & Simon, 1972) for deciding which of the results of objective measurements and subjective expert assessment are used.

To deal with such tasks the methods of scientific calculations (Yang, 2008) that are the essence of intelligent data processing concept (Yang, 2008) are used more and more often. The concept consists of organization of the detection process in “raw” data of unknown nontrivial practically useful knowledge which can be interpreted and may be useful for decision making in different spheres of human activity.

It is clear that concept of intelligent data processing is oriented not only on the usage of special program tools but also on special information-communicative infrastructure, which allows to use huge volumes of data of different origin and its processing for search of solutions of interdisciplinary tasks.

To harmonize the actions for establishment of such infrastructure and organizing a common information space to maintain acquisition, handling, and exchange of data and solving of fundamental and applied interdisciplinary problems in 2008 Russian and Ukrainian World Data Centers united into a Segment (Zgurovsky, Gvishiani, Yefremov, & Pasichny, 2010).

The aim of this paper is to describe state-of-the-art of Russian-Ukrainian WDS Segment and to determine the principles of architectural organization of intelligent data processing system on the basis of which the components of open program system can be created and integrated, which give the user the data and tools to solve the challenges of interdisciplinary researches.

## 2 STATE OF THE ART

Five Russian ICSU World Data Centers (WDC) for Oceanography, Meteorology, Rockets, Satellites and Rotation of the Earth, Solar-Terrestrial Physics and Solid Earth Physics more than 50 years collect, analyze, archive and distribute data for the broad spectrum of observatory types. The Centers provide open and convenient access to great volumes of data, permanently increase the information resources in the Internet. At 2006 Ukrainian WDC for Geoinformatics and Sustainable Development was formed, it is one of leaders in various fields of sustainable development research in Ukraine and it collects, processes, analyses and disseminates global and national data necessary for sustainable development research.

Several meetings of the ICSU World Data Centers in Russia and Ukraine were held in Obninsk, Moscow and Kyiv in 2008-2010. This activity has resulted in establishment of Scientific Council of Russian and Ukrainian World Data Centers and forming of the Russian-Ukrainian segment of WDCs (Segment).

The top priorities for our Segment are following:

- Integration to new WDS and effective cooperation with WDS members
- Providing safety of data from non-digital data carriers and its digitization
- Establishing data quality policy
- Development of data providers' infrastructure
- Formation of common information space for Russian-Ukrainian WDS segment

Creation of common information space for our Segment is a high priority task. It is important to provide unified data formats and develop unified tools for efficient data exchange. Such approach would make possible to create a single access point to all data and services of Russian-Ukrainian WDS Segment. That would also provide a flexible framework for unified data processing toolkit development. The main features of this common information space:

- Flexible and scalable cross-platform open source-based architecture (e.g. SOA)
- Centralized data & services registry
- Easy integration with existing systems (using SOAP, WSDL, UDDI, etc.)
- Single access point
- Easily created and customized UI based on existing services
- Common approach for acquiring data from various data sources

All Segment members expressed their interests to join the new ICSU World Data System and have successfully completed the necessary stages of certification. Today all Russian and Ukrainian WDCs have obtained status of regular WDS members.

## 2.1 WDCs in Obninsk

There are three World Data Centers established on the basis of All-Russian Scientific and Research Institute of Hydrometeorological Information – World Data Center (RIHMI-WDC):

- World Data Center for Meteorology;
- World Data Center for Oceanography;
- World Data Center for Rockets, Satellites and Rotation of the Earth.

WDC for Meteorology has commitment to provide long term secure preservation and dissemination of meteorological data and products both for global and regional scales. Data holding contains observed meteorological data for the long period from XVIII century to the present days. Data are collected from the Russian network of meteorological stations and from the other parts of the globe by means of telecommunication system of WMO. Data are validated, checked and updated in continuous way. Specialized high quality data sets for climate study are of particular importance and are provided for online access ([http://meteo.ru/english/climate/cl\\_data](http://meteo.ru/english/climate/cl_data)). Climate surveys are published on a regular basis (<http://meteo.ru/english/climate/bulletin/>).

WDC for oceanography collects data and products of national and international projects in the field of physical and chemical oceanography. Data holding contains observed oceanographic data collected from single observational platforms (research vessels, buoys and other devices) and coastal stations for the long period from XIX century to the present days. WDC also provide additional products including data analyses, maps of data distributions, and data summaries. Online access to the data and products is made available through the <http://www.meteo.ru/mcd/ewdcoce.html>.

WDC for rockets, satellites and rotation of the Earth collects meteorological data of national and international rockets and satellites, data on Earth rotation. Necessary quality assessment and quality control procedures are applied when required. Data holding contains observed rockets and satellites meteorological data for the period from 1966 up to present days. Online access to the data and products is provided through the entry point <http://www.meteo.ru/mcd/ewdcroc.html>.

Internal data management is based on original data description language (DDL) and hierarchical DBMS AISORI developed in RIHMI-WDC. This ensure data and metadata identity, integrity and efficient archiving and usage. All data and information submitted to the WDCs are classified and registered in catalogues and directories of metadata base. The metadata elements enable to identify type and origin of data, their spatial and temporal coverage, other characteristics necessary to ensure authenticity of data sets. Metadata are used to prepare published and electronic Data Catalogues. These are posted on RIHMI-WDC Web site ([http://meteo.ru/english/data\\_b/](http://meteo.ru/english/data_b/)).

All data are checked and validated by means of a set of visual and automatic QC procedures in accordance with WMO and IOC Manuals and Guides. The numerical criteria used in QC procedures are being updated on the base of continuous climatological research. The enhanced procedure for duplicates check have been developed and applied to avoid data duplicates within global data sets. Also to ensure metadata integrity the directories of organizations, maritime research projects, research vessels are regularly validated to be consistent with international counterpart directories of IODE, GCMD, ICES, WMO, EDMO, EDMERP.

To achieve a high rate of data processing and provide effective online user access to the data they are loaded into relational Oracle DBMS which is a component of integrated web-technology. The last one enables user to discover the data of user's interest, to retrieve, browse and view them in a table form or as a plot, map or diagram.

For exposing on the Web metadata consistent with ISO 19115 standard there are two international metadata profile being in use within RIHMI-WDC: WMO core profile for meteorological data (<http://www.wmo.int/pages/prog/www/WDM/Metadata/documents.html>) and CDI for oceanographic data (<http://www.seadatanet.org/Standards-Software/Metadata-formats>). WDCs also strictly adhere to international standards for data exchange, in particular BUFR/CREX data format developed under WMO umbrella, and

widely used NetCDF and ODV data formats.

Gathering up and dissemination of international publications (atlases, gazetteers, reference books, manuals and guidelines) also are the matter of responsibility of the World data centers hosted by RIHMI-WDC. Catalogs of international publications in the field of meteorology, oceanography, rockets and satellites collected by WDC are available online (<http://meteo.ru/english/publish/>).

Each World data center uses RIHMI-WDC technical infrastructure and software utilities for long-term data preservation and dissemination. All procedures used in RIHMI-WDC for the long-term data storage (holdings with appropriate conditions, periodical check and recovery as necessary) are applied to the WDCs data and information along with national data and information.

To perform processing functions necessary for data archiving and provision of user access to the data the IBM z9 BC mainframe is employed. IBM System Storage DS8300 offers high performance, higher capacity storage up to 512 Tb. Full Disk Encryption with local key management provides relentless data security. The firewall system also is maintained to secure user access via Internet to the operational data base and other information resources.

To ensure long-term and safe storage all data are archived within two robotic IBM System Storage 3500 Tape Libraries. One of these is the Main library and another one is a Mirror library used for data backup and recovery. The libraries are located in two buildings detached. The direct access of external user to the library is impossible. A set of information service systems provide data stewardship and preservation. In particular technological schema of backup and long-term storage is based on IBM Tivoli Storage Manager software and Content Manager on Demand. This solution provides cost-effective functionality, scalability and ease of use for the entry-level storage user.

All of these developments serve to ensure long-term data preservation and free timely access to the WDCs data.

## **2.2 WDCs in Moscow**

WDC for Solar-Terrestrial Physics and WDC for Solid Earth Physics, Moscow are the parts of the Geophysical Center of the Russian Academy of Sciences (GC RAS).

WDC for Solar-Terrestrial Physics (WDC for STP) was one of the original data centers established in the USSR by the Academy of Sciences of the USSR in 1957 to support the IGY. The WDC for STP holds data sets relating to solar activity and interplanetary phenomena, ionosphere, geomagnetic variations, and cosmic rays. The Center maintains and provides services for the archive of historical and modern results of geophysical observations on global networks of observatories. The data are available in different traditional forms, e.g., paper, microfilms and microfiches, and electronic form. WDC for STP converts old analog data into digital form in order to preserve data to provide them to scientists in convenient ready-to-use form. All data are registered in the computer database and listed in the data availability catalog. Digital data in non-standard formats, metadata and data availability catalogs are available at free access on the WDC for STP's web site <http://www.wdcb.ru/stp/index.en.html>. The World Data Center for Solar-Terrestrial Physics together with Laboratory of Network Information Technologies of GC RAS and National Geophysical Data Center of USA supports databases "Space Physics Interactive Data Resource" (SPIDR). SPIDR is a distributed network of databases and service programs that are synchronized in real time and allow the user concurrent access to a network of thematic databases; interactive visualization of time series, maps and images and sampling of the multidisciplinary data; search for specific events in "space weather" in terms of a natural language with fuzzy logic application. The WDC for STP is the participant of the international project in the field of information technologies and geophysics INTERMAGNET.

WDC for Solid Earth Physics (WDC for SEP) maintains extensive archives of data on seismology, geomagnetism (the main magnetic field), archeo- and paleo-magnetism, gravimetry, geothermy, recent movements. Data stored in the Center are obtained during the International Geophysical Year (1957-1958) and subsequent international projects, such as the "Upper Mantle", "Geodynamics Project", "International Polar Year 2007-2008" et al. They are results of land and sea expeditions, launches of satellites, special experiments, results of geophysical observations (seismological, geomagnetic etc.) on global networks of observatories. The Center accepts data according to a long term relationship with producers of the data: separate stationary observatories, regional, national and international observatory networks data-processing and analytical centers. The data provided by scientific institutes and other organizations is accepted only after they have passed examination in these organizations and have received the status of the data intended for the international exchange in the

decision of a commission of experts. The Center provides access to all these data and also serves as information and reference center. Some parts of the data are stored as publications on paper and microfilm, but considerable part are also available in digital electronic form on the WDC for SEP's web site <http://www.wdcb.ru/sep/>. Data, metadata, thematic databases, inventory catalogs are available on line on the web site. The WDC for SEP is the participant of the international and interdisciplinary project in the field of geology and geophysics InterMARGINS, concerned all aspects of continental margin research.

The WDC for STP and WDC for SEP give data, metadata and other products to scientists all over the world without restrictions and are free-of-charge. The Centers provide all conditions for long-term secure preservation and dissemination of these data. The Centers continuously improve implementation of new technologies of data maintenance, software and hardware. They aspire to have the modern level of technologies for collection, handling, transmission and storage of data and information and to consider new scientific requirements. The Centers have modern technical facilities, use modern technologies for data processing and providing permanent online access to data. The user interface is developed for convenient search, browsing, visualization and retrieval of the data on the WDCs web sites.

All data received by WDCs for STP and SEP are analyzed: a discipline, type of observations, period of observations, geographical territory, to which the observations are related etc. are defined. Data are registered and are placed in an appropriate section of the archive. The archive is structured. Reserve copies are made for electronic data. Data analysis and data quality control is carried out with special QC software. Descriptions of data, data formats and metadata of GCMD DIF standard are prepared.

## 2.3 WDC in Kyiv

World Data Center for Geoinformatics and Sustainable Development (WDC-Ukraine) is situated in the Institute for Applied System Analysis (IASA) of the National Academy of Sciences (NAS) of Ukraine and Ministry of Education and Science, Youth and Sport (MESYS) of Ukraine in the structure of the National Technical University of Ukraine the "Kyiv Polytechnic Institute" (NTUU "KPI"), Kyiv, Ukraine. WDC-Ukraine was created by the decision of Presidium of the National Academy of Sciences of Ukraine (NAS), Ministry of education and science of Ukraine and Geophysical Center of the Russian Academy of Sciences (GC RAS) from April, 3, 2006 as a subdivision of the Russian World Data Centers "WDC for Solar-Terrestrial Physics" and "WDC for Solid Earth Physics" (Moscow). Afterwards in the 2008 Ukrainian branch of WDC obtained an independent status as World Data Center for Geoinformatics and Sustainable Development.

The World Data Center in Ukraine should afford access to global information resources of the ICSU on Earth sciences, planetary and space physics, and related subjects for the Ukrainian scientific community and to provide acquisition and storage of national scientific data on the above disciplines and their presentation to the world community. WDC-Ukraine collects, processes, analyses and disseminates global and national data necessary for sustainable development research. Center is one of leaders in various fields of sustainable development research in Ukraine. Scientific and technical staff of the WDC-Ukraine performs fundamental and applied research and analysis for solving interdisciplinary problems with system nature, particularly quantitative measurement and modelling of sustainable development processes and evaluation of the impact of the global threats set on sustainability in global and regional contexts, it also develops and implements information technologies for solving of wide range of tasks connected with the collection, exchange, processing and analysis of interdisciplinary data and solving different tasks of the applied system analysis.

A deep study of World Data Centers in other countries, the interdisciplinary orientation of the IASA, and system approach have allowed proposing a unique (for the World Data System) network model of functioning of the WDC-Ukraine as a unified interdisciplinary national data center. According to this model, each research area is supervised by one or several scientific organizations of the National Academy of Sciences of Ukraine. Here are some of them:

- Institute for Applied Systems Analysis NAS of Ukraine and MESYS of Ukraine (system coordination of interdisciplinary data, sustainable development);
- S. I. Subbotin Institute of Geophysics NAS of Ukraine (data on seismology, gravimetry, heat flow, archeo- and paleo-magnetism, and magnetic measurements);
- Scientific Center for Aerospace Research of the Earth, Institute of Geosciences NAS of Ukraine (aerospace pictures to be used in geology, ecology, agriculture, forestry, and water industry, to predict risks of natural and technogenic processes, global environmental changes, and catastrophic processes);

- Main Astronomical Observatory NAS of Ukraine (space geodesy and geodynamics; cosmic rays);
- Marine Hydrophysical Institute NAS of Ukraine (oceanology and hydrometeorology);
- Institute of Geography NAS of Ukraine (cartography);
- Chernobyl Center for Nuclear Safety, Radioactive Waste and Radioecology.

The network model was first presented on October 7, 2009 at the special WDC session “Emerging Technologies and Opportunities for Global Data Management and Exchange” within the framework of the CODATA-2008 conference (October 5–8, 2008, Kyiv, Ukraine), where it was approved and conventionally called “Network of networks” (Starostenko, Yatskiv, Lyalko, Ivanov, Rudenko, & Yefremov, 2008). At the session of WDS Scientific Committee on October 13–14, 2009 in Paris, this model was taken as a sample for other WDCs.

In Ukraine, such an approach, on the one hand, allows efficient use of the technological capabilities of the Ukrainian Research and Academic Network (URAN), which unites Ukrainian scientific, educational organizations, and the high-performance computing cluster of the NTUU “KPI” (cluster is a part of the national GRID-infrastructure so in case of need tasks can be distributed among all partners of this network), and on the other hand, focusing of the efforts of the WDC staff on solving interdisciplinary system problems important for all the WDC partners:

- provision of all main phases of data management (collection, quality assurance, storage, processing, sharing, reporting and long-term stewardship) for scientific data of various nature;
- development of mathematical models, methods and tools for assessment and decision-making in complex systems;
- development and support of information systems and services focused on data analysis and processing.

### 3 JOINT PROJECTS

To realize the aims of Segment centers fulfill a number of joint projects aimed at development of World Data System and its Russian-Ukrainian Segment. These projects are supported by Basic Research Foundations and Academies of Sciences from both countries:

- 2008 – 2009 –«Development of set of databases and processing algorithms aimed to system prevision of complex anthropogenic and natural systems’ behavior»;
- 2009 – 2010 –«Development World Data Centers network for investigation of basics of complex natural and anthropogenic systems’ global modeling»;
- 2010 – 2011 –«Development of fundamentals analytical methods of multidisciplinary data for creation integrated access system to information resources of the World Data Centers in Russia and Ukraine»;
- 2011 – 2012 –«Development of indices and indicators of Ukrainian and Russian regions sustainable development based on combined usage of causality and stochastic semantics»;
- 2012 – 2013 –«Development of the general approach and methods for system adjustment of data of various nature in the distributed multidisciplinary databases infrastructure of the World Data System Russian-Ukrainian Segment for solving of fundamental interdisciplinary tasks of processes correlation in the geospheres system».

According to basic principles of the WDS development, one of which is a transition from existing stand-alone WDCs and individual Services to common globally interoperable distributed data system, some projects were targeted on creation of Common Data Catalogue (single access point) that gave an access to heterogeneous data sources using agent-oriented and ontology-based approaches. Pilot version of Catalogue was developed and given an access for testing purpose via Internet. Such approach will allow uniting practically unlimited quantity of diverse data sources into the single heterogeneous environment that would be transparent for the user and organize intelligent selection of the data sets according to the user query. Now we continue working on the implementation of the services in this system to organize intelligent data processing with the use of adaptive approach. User can formulate his or her query in the subset of the natural language.

Besides the development of fundamental basis and methods for interdisciplinary data analysis and integration of access to information resources of Russian and Ukrainian WDCs, there are also projects aimed on the development of an intelligence GIS "Russia-Ukraine" for support of fundamental and applied interdisciplinary studies of complex systems of different nature and on the creation of interregional information node for collecting and processing of data from Russian-Ukrainian segment of INTERMAGNET.

To present WDS in Internet and providing an access to data and metadata of WDS a new web-site was designed and developed by World Data Center for Geoinformatics and Sustainable Development. It is located on the address [www.icsu-wds.org](http://www.icsu-wds.org) and supported by WDC-Ukraine staff. The Web-portal is constructed on the basis of flexible and easily scaled platform allowing, if necessary easily and quickly revise not only material but also the portal structure. The site has its proper member zone, organized by integrating MediaWiki and GoogleDocs, in the framework of which members of WDS-SC actively exchange materials.

## 4 COMMON INTELLIGENT DATA PROCESSING SYSTEM

Russian-Ukrainian WDS Segment provides wide range of data for wide diversity of disciplines: Seismology, Gravimetry, Heat Flow, Magnetic Measurements, Archeo- & Paleo-magnetism, Solar Activity and Interplanetary Medium, Cosmic Rays, Ionospheric Phenomena, Geomagnetic Variations, Space Geodesy and Geodynamics, Oceanography, Meteorology, Cartography, Remote Sensing, Sustainable Development etc.

Segment members are involved increasingly frequently to solve fundamental and applied interdisciplinary problems that need system adjustment of data of various nature and using intelligent data processing technologies. For this purpose we start to unite tools for intelligent data processing in special framework.

### 4.1 Organization of the intelligent data processing

As it is seen from Figure 1 which shows a typical process of intelligent data processing (Zgurovsky, 2010) it can be divided into stages of preliminary processing, analysis and post-processing of data.

Each of these stages solves special tasks with the usage of methods of multidimensional statistic analysis, statistic-expert methods and other methods.

You can also notice that the same methods can be used to solve different tasks and the same task can be solved by different methods.

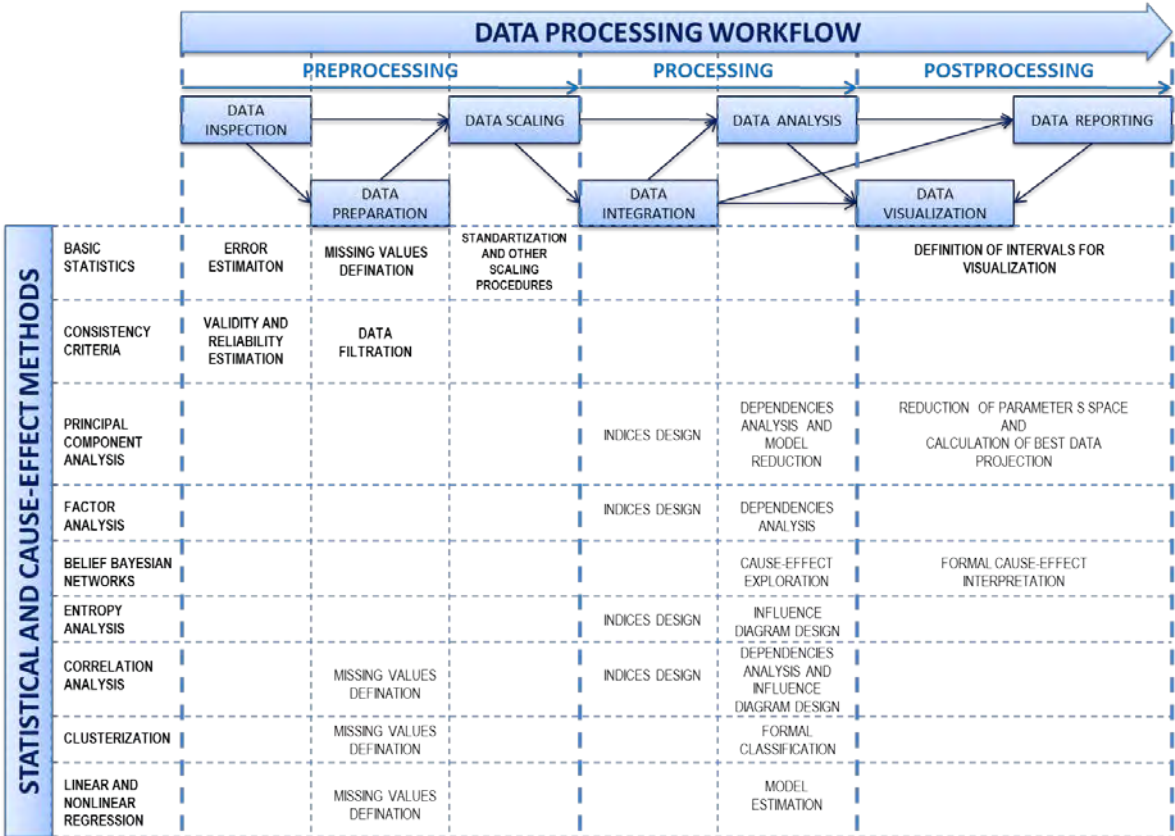


Figure 1. Process on intelligent data processing

That's why we can point out software-based means which realize statistic and expert methods regardless of context of their usage to solve particular tasks of data processing. These software-based means can be organized as universal software-based libraries used for realization of that programs responsible for dealing with the task of data processing in terms of general process, shown on Figure 1.

It should be noticed that the possibility of organization of separate programs and libraries in terms of general calculating process depends on presence of data compatibility and results. Also there should be determined general models of data that will be used for the specification of parameters and results of work of particular programs.

Also it should be pointed out that data which are used can be kept in different storages. That is why it is suitable to develop special unified software-based mechanisms of access to them and provide data with structures of metadata, determined in a result of their diversity on another level from sources of data. Metadata is the combination of some regularity in particular sphere and they are suitable for further combination of different output data. For interdisciplinary researches the problem of development of program methods information transfer to metadata is actual (Shapovalova, Yefremov, & Glukhanik, 2011).

At last, every process of intelligent data processing should be performed as a tool for solution of particular application accessible for user.

Paying attention to the mentioned above the module organization of software-based methods of system of intelligent data processing, divided into four levels is proposed:

- level of data model presentation used by software methods of higher levels for data and results specification;
- level of software implementation of data processing methods that can be used for solving of different tasks;
- level of software implementation of stages and the overall process of data processing for a particular application task;
- level of software implementation of joint mechanisms of access to data, data visualization, reports generating etc.

The described organization under condition of sufficient spectrum of software methods on each level actually allows combining accessible software-based components for solving different tasks of intelligent data processing supporting the concept of "quick" development.

## **4.2 Architectural principles of organization of intelligent data processing system**

Choice of architectural principles of building of intelligent data processing system must consider general and specific requirements which are advanced for such systems from side of all interested people. Regardless of list of particular applications, which the intelligent data processing system deals with, it must satisfy the requirements conditionally divided on two categories. One from them is connected with supplying requirements which are advanced by user. The second one is determined by the possibility of quick development of a whole system as well as its components.

From user's view the system must correspond to such requirements:

- system must allow user to process data of such volume which exceeds the possibilities of its equipment;
- system must process raw data format;
- system must be accessible to the wide range of users;
- system must be simple in use even for users who do not know programming.

Systems requirements are as follows:

- system must provide increasing of calculating sources and volumes of data storage due to its workload;
- internal mechanisms of data processing must be hidden from user. They can include algorithms which are author's now-how or use the intermediate information, which is not free in access;
- architecture and software-based mechanisms of system must predict the possibility of adding of new function and (or) system reconfiguration.



These requirements determine the principles of building of intelligent data processing system on architectural and structural levels of their providing.

Possible increasing of sources, needed for data processing, over possibilities of user's equipment forces to use the system division into client and server components in terms of one of architectural principles of system organization. Such client-server organization means that the data processing may be held on some distant servers (Korzhov, 1997).

In this case the client just enquires and gets the results of counting, notably the user equipment is used as the second client. For organization of relation between client and server net protocols of appendixes can be used and the Internet itself can be used as communication environment. In this way the access to system is provided for wide range of users.

Modern net protocols of appendixes level can be divided into two categories: protocols, based on usage of access to information sources and protocols, based on distant call of procedure.

The first category is called RESET (Representational State Transfer) protocols. In case of usage of such architecture the user agents have the possibility to correlate with heterogeneous sources. This correlation is provided with the help of unique interface of standard commands HTTP (GET, POST, PUT and DELETE). In this case the resource contains all information needed for its processing (Flanders, 2009).

Protocols of second categories use mechanisms of distant call of procedures, among which protocol SOAP (Simple Object Access Protocol) is the most widespread. It provides the transfer structural messages on the base of XML. As opposed to REST, SOAP is protected protocol with guaranteed transfer of messages (MacVittie, 2007).

Each of described protocols has its advantages and disadvantages. To provide the possibility of integration of intelligent data processing system with other systems, the target is to realize not one but several protocols with the usage of which the message transfer is held.

The usage of client-server architecture with communication environment on the base of net Internet allows to examine the intelligent data processing system with the point of view of concept of cloud computing, in the terms of which the information is placed and stored on the distant servers, which are accessible with the help of the Internet and is only temporary storage on the client base (Mell, & Grance, 2009). The selection of cloud computing concept hides the mechanisms of operation of data processing tools which is one of the requirements for developing system.

Among such views on the system from the point of view of user, from renting him the hardware devices or virtual calculating environment and up to providing the access to the functionally finished software-based product, for the intelligent data processing system the most appropriate is its presentation as Software as a Service (SaaS) (Kolesov, 2008).

Concept of cloud computing can be realized in systems designed on different architectural principles. In terms of one of them which has the title service-oriented architecture (SOA), the module method is used in the development of software, based on services usage (service) with standard interfaces, when system is examined as combination of autonomic services combined by general communication mechanism.

It should be pointed out that it predicts the necessity of inputting into system architecture of interim chain (strip) which organizes the relation between connections to it services. The methods of services connection to strip are standardized. For user the system strip, its functions provided by dispatcher (he is responsible for calls processing and results delivery (Ferguson, & Stockton, 2005)), realizes the concept of one point of access to the application SaaS.

Thereby, from the user point of view the intelligent data processing system must be a finished application SaaS which allows to store and treat data with the help of server sources. It is reasonable to use principles of service-oriented architecture with dispatcher that realizes the functionality of system strip creating preconditions for possible scaling and system expansion.

## **5 CONCLUSION**

Creation of Russian-Ukrainian WDS Segment and its successful activities became the important step for development of Russian-Ukrainian Data Community and establishment of scientific data infrastructure for Russian and Ukrainian scientific organizations that become possible thanks largely to the implementation of joint bilateral projects with financial support from the Russian and Ukrainian Academies of Sciences and Basic Research Foundations of both countries.

Intelligent data processing is one of stages of interdisciplinary holistic process of work with scientific data. Methods proposed by authors became the basis for organization of information-communication infrastructure for World Data Center of Geoinformatics and Sustainable Development and its partners, part of functional modules implemented as web-services within the scopes of WDC-Ukraine portal (<http://wdc.org.ua>). Also proposed approaches were successfully used for building intelligent data processing subsystem for Segment common distributed data system – Common Data Catalogue.

## 6 REFERENCES

Ferguson, D. & Stockton, M. (2005) *SOA programming model for implementing Web services, Part 1: Introduction to the IBM SOA programming model*. Retrieved October 20, 2012 from the World Wide Web: <http://www.ibm.com/developerworks/library/ws-soa-progmodel/index.html>

Flanders, J. (2009) An Introduction To RESTful Services With WCF. *MSDN Magazine*, January 2009. Retrieved October 23, 2012 from the World Wide Web: <http://msdn.microsoft.com/en-us/magazine/dd315413.aspx>

Kolesov, A. (2008) SaaS Model — in World and in Russia. *Byte*, 10(119). Retrieved November 12, 2012 from the World Wide Web: <http://www.bytemag.ru/articles/detail.php?ID=12825>

Korzhev, V. (1997) Multilevel client-server system. *Networks*, 06. Retrieved October 17, 2012 from the World Wide Web: [http://www.osp.ru/nets/1997/06/142618/#part\\_1](http://www.osp.ru/nets/1997/06/142618/#part_1) (in Russian)

MacVittie, L. (2007) REST as alternative for SOAP. *Networks and Communication Systems*, 1. Retrieved October 22, 2012 from the World Wide Web: [http://www.ccc.ru/magazine/depot/07\\_01/read.html?0502.htm](http://www.ccc.ru/magazine/depot/07_01/read.html?0502.htm).

Mell, P. & Grance, T. (2011) The NIST Definition of Cloud Computing. *National Institute of Standards and Technology, Information Technology Laboratory*, SP 800-145.

Newell, A. & Simon, H. (1972) *Human problem solving*, Englewood Cliffs, NJ: Prentice-Hall

Shapovalova, S.I., Yefremov, K.V. & Glukhanik, A.I. (2011) Organization of integrated access to information resources. *Proceedings of the XI International Conference “Intelligent Analysis of Information”*. Kyiv, Ukraine (in Russian)

Somervill, M. & Rapport, D. (2000) *Transdisciplinarity: Recreating Integrated Knowledge*. Oxford, UK: EOLSS Publishers Co. Ltd.

Starostenko, V., Yatskiv, Ya., Lyalko, V., Ivanov, V., Rudenko, L., & Yefremov, K. (2008) Ukrainian science data: mutual goals and approaches, *Proc. 21st Int. CODATA Conf.*, Kyiv, Ukraine

Yang, X. S. (2008) *Introduction to Computational Mathematics*, World Scientific Publishing, 2008. — 245 p.

Zgurovsky, M.Z. (2010) System Adjustment of Various nature DATA for Global Modelling of Sustainable development. *Proc. 22nd Int. CODATA Conf.*, Cape Town, South Africa

Zgurovsky, M.Z., Stratukha, G.A., Melnichenko, A.A., Voitko, S.V., Boldak, A.A., Yefremov, K.V. et al. (2010) *Sustainable development analysis – global and regional contexts. P.1. Global analysis of quality and security of life*, Kyiv, Ukraine: NTUU “KPI”

Zgurovsky, M.Z., Stratukha, G.A., Melnichenko, A.A., Voitko, S.V., Boldak, A.A., Yefremov, K.V. et al. (2010) *Sustainable development analysis – global and regional contexts. P.2. Ukraine in the sustainable development indicator analysis*, Kyiv, Ukraine: NTUU “KPI”

Zgurovsky, M.Z., Gvishiani, A. D., Yefremov, K.V. & Pasichny, A.M. (2010) Integration of the Ukrainian science into the world data system. *Cybernetics and Systems Analysis*, Vol. 46, No 2, pp 211-219

# THE CONTRIBUTION OF A GEOPHYSICAL DATA SERVICE: THE INTERNATIONAL SERVICE OF GEOMAGNETIC INDICES

*M Menvielle*<sup>1,2\*</sup>

<sup>\*1</sup>Université Versailles St-Quentin; CNRS/INSU, LATMOS-IPSL, Guyancourt, France

<sup>2</sup>Département des Sciences de la Terre, univ. Paris Sud, Orsay, France

Email: michel.menvielle@latmos.ipsl.fr

## ABSTRACT

*Geomagnetic indices are basic data in Solar-Terrestrial physics, and for operational Space Weather activities. The International Service of Geomagnetic Indices (ISGI) is in charge of the derivation and dissemination of geomagnetic indices that are acknowledged by the International Association of Geomagnetism and Aeronomy (IAGA, an IUGG association). Institutes which are not part of ISGI started early in the internet age to circulate on-line preliminary values of geomagnetic indices. In absence of quality stamping, this resulted in a very confusing situation. The ISGI label was found to be the simplest and the safest way to insure quality stamping of circulated geomagnetic indices.*

**Keywords:** Geomagnetic indices, On-line dissemination, Real-time, ISGI

## 1 INTRODUCTION

The geomagnetic activity corresponds to that part of the transient variations of the geomagnetic field observed at the surface of the Earth which bears the magnetic signature of currents flowing in the ionized environment of the Earth, the dynamics of which is driven by the solar wind. Since the beginning of systematic geomagnetic field observations more than one hundred years ago, efforts to characterize the geomagnetic activity led to the definition of few geomagnetic indices that are widely used by the scientific community, and officially recognized by the International Association of Geomagnetism and Aeronomy (IAGA). These quantities are hereafter referred to as IAGA geomagnetic indices.

Since the dynamics of currents flowing in the ionized environment of the Earth is driven by the solar wind, geomagnetic indices are basic data in Solar-Terrestrial physics, and there is accordingly a need for long homogeneous data series. It is worth being noted that, although geomagnetic indices are thus widely used in magnetospheric physics, as e.g. indicators of the magnetosphere-ionosphere system activity, or as input parameters in the models, ignorance of their real physical meaning often leads to use them as black boxes. The International Service of Geomagnetic Indices (ISGI) has the mission to make available reference values for the IAGA geomagnetic indices to the scientific community, and to maintain archives of these data series.

Magnetospheric activity may otherwise cause disturbance - which can be severe - in the functioning and behaviour of satellites and space systems: geomagnetic indices are also basic data for operational Space Weather activities, and this requires the shortest possible dissemination delay. Early in the Internet age, many institutions which are not part of ISGI started to derive and disseminate online preliminary values (also called quick-look values) of IAGA geomagnetic indices, without any control of IAGA bodies on the quality of their products. This resulted in a very confusing situation, and could have resulted in a dramatic loss of confidence of users in the quality of geomagnetic indices made available through Internet.

ISGI had then to face up two challenges:

- establish a policy for on-line dissemination of indices;
- ensure a quasi real time on-line dissemination of reliable quick-look values of IAGA geomagnetic indices.

ISGI decided to routinely make available on-line preliminary values for almost all IAGA geomagnetic indices. ISGI also drove discussions in the frame of IAGA. They resulted in an IAGA resolution urging the producers of estimated indices to make clear by a specific label that they are not official IAGA indices. The ISGI label thus appeared to be the simplest and the safest way to ensure the quality stamping of circulated geomagnetic indices, and the ISGI Collaborating Institutes therefore decided to quote their ISGI membership in all their publications and data bases. This ISGI contribution is used to illustrate the role that a WDS service could play as a provider of

reference values for given quantities, and to highlight the need for such reference data services.

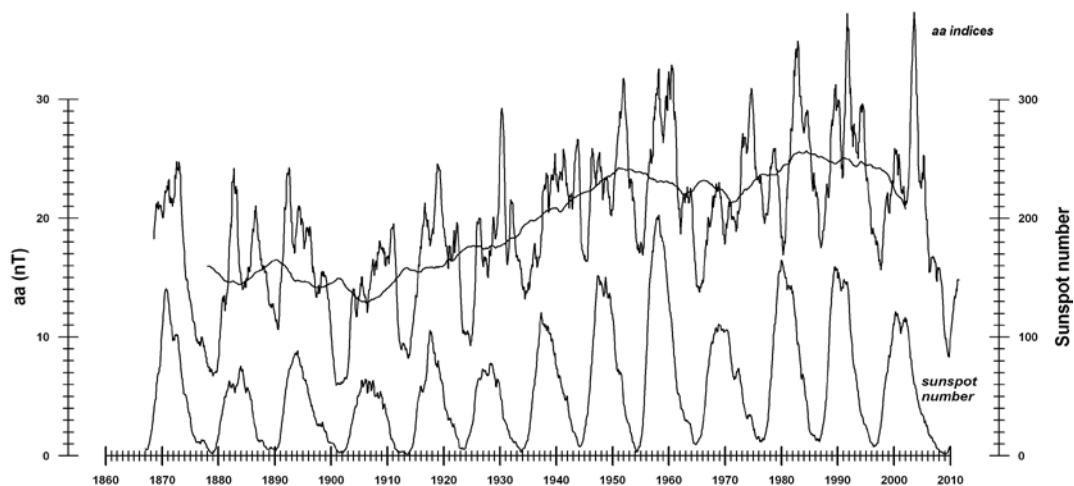
## 2 GEOMAGNETIC INDICES

The Sun emits a variable supersonic plasma flow: the solar wind. The Sun's magnetic field is driven by the solar wind in the interplanetary medium, giving rise to the interplanetary magnetic field. In the vicinity of the Earth, the solar wind is thermalized and flows around the obstacle that constitutes the Earth magnetic field. This results in compression of the force lines of the geomagnetic field on the day side and stretching of these lines into a long tail on the nightside, which gives rise to the magnetospheric cavity.

The solar wind is the energy source of the magnetosphere. The solar wind characteristics directly influence the shape and size of the magnetosphere, the amount of energy that is transferred into the magnetosphere, and how this energy is dissipated. The rapid and wide variations of the solar wind characteristics result in a very dynamic behaviour of the magnetosphere. The transient variations of the geomagnetic field are a signature of the plasma convection inside the magnetosphere, and thus of the magnetospheric status that can be observed at the surface of the Earth. The great complexity and the large dynamics of magnetospheric plasma convection result in complex magnetic signatures at the Earth's surface.

The first systematic geomagnetic field observations date from the middle of nineteenth century, and they are currently made at more than two hundred magnetic observatories distributed on the five continents. Summarizing quantities that aim to characterize the magnetic activity have soon been introduced to describe in a simple although relevant way the transient magnetic variations at the Earth's surface, and their changes with time.

Geomagnetic indices that aim to characterize the magnetic activity at a global scale, or the ionospheric auroral electrojets in the auroral zones, or the behaviour of the magnetospheric ring currents, as well as the lists of *storm sudden commencements* and of *solar flare effects* are currently recognized as IAGA geomagnetic indices. The longest homogeneous data series are more than 140 years long (they begin in 1868), and most of geomagnetic indices data series are more than 50 years long. The reader is referred to, e.g., Menvielle and Berthelier (1991), Menvielle et al. (2011), and to the ISGI Internet homepage (<http://isgi.cetp.ipsl.fr>) for a complete description of geomagnetic activity indices.



**Figure 1.** Long-term variations of aa indices (12-month and 20-year running averages; scale on the left) and of sunspot numbers (12-month running averages; scale on the right) from 1868 until 2011. The aa index monitors geomagnetic activity at a planetary level and the sunspot number monitors the solar activity.

Figure 1 shows the evolution of planetary geomagnetic activity as described by the aa planetary geomagnetic index, together with that of the solar activity as described by the Sunspot number. The 11-year solar cycle is clearly depicted by the sunspot number data series. It also modulates the geomagnetic activity, although in a more complex way. The geomagnetic activity is minimum during the periods of minimum solar activity. Then it increases during the ascending phase of the solar cycle and reaches a maximum during the years where the sunspot number is maximum. A secondary maximum of geomagnetic activity is observed during the descending phase of the solar cycle: this secondary maximum corresponds to solar activity related to the onset of the following cycle

(see, e.g., Legrand and Simon, 1991 and references therein). The variations of the 20-years running average of geomagnetic activity show that long-term variations are superimposed to those related to the solar cycle: the average level of activity remained quite low between 1868 and about 1925, while it was quite high since the beginning of the 1950's. The low activity level observed in 2009-2010, during the last solar minimum may indicate that this high activity period is coming to its end.

Geomagnetic activity indices bear information on the evolution during more than 140 years of the solar wind (see, e.g., Rouillard et al., 2007), of its coupling with the magnetosphere and on the state of the ionized environment of the Earth (see, e.g., Ouattara et al, 2009). Geomagnetic indices have many other applications. For instance, they are unique tools for the separation of magnetic variations according to whether their source is located in the ionized environment of the Earth or inside the planet (see, e.g. Thomson and Lesur, 2007)

Although geomagnetic indices of magnetic activity are fundamental data in solar-terrestrial physics and space weather, they are most often used as "black boxes". This situation, which results from the very complex morphology of geomagnetic activity, implies the existence of an organization, the International Service of Geomagnetic Indices (ISGI), with the mission to calculate, to make available to the scientific community, and to maintain archives of reference values for the IAGA geomagnetic indices, and to take advantage of its expertise in this area to serve the scientific community

### **3 THE INTERNATIONAL SERVICE OF GEOMAGNETIC INDICES**

#### **3.1 History and current organization**

Founded in 1906 as the Central Bureau of Terrestrial Magnetism for the calculation of the "*International Magnetic Character*", it was hosted by the *Koninklijk Nederlands Instituut Meteorologisch* (De Bilt, The Netherlands) until 1987. At that date, it was transferred to France, where it was hosted by the *Institut de Physique du Globe de Paris* until 1990, then by the *Centre d'études des Environnements Terrestre et Planétaires (CETP, Saint Maur)* until the date (2009) when *CETP* was replaced by *LATMOS (Laboratoire Atmosphères, Milieux, Observations Spatiales, Guyancourt, France)*. ISGI headquarters are currently hosted by *LATMOS*.

ISGI is a kind of network that brings together the activities of the four institutions – called ISGI Collaborating Institutes – that are responsible for calculating and disseminating IAGA Geomagnetic Indices: the *LATMOS*, the *Data Analysis Center for geomagnetism and Space Magnetism* (Kyoto, Japan), the *GeoForschungZentrum Potsdam* (GFZ, Potsdam, Germany) and the *Observatori de l'Ebre* (Roquetes, Spain). ISGI is under the supervision of the Union of Geodesy and Geophysics (IUGG). ISGI activities are supervised by a scientific board appointed by the IAGA Executive Committee. The Council meets each second year, during the General and Scientific IAGA Assemblies.

ISGI used to be a service of FAGS (Federation of Astronomical and Geophysical Data Services); it was accepted for membership in the World Data System (WDS) in August 2011

#### **3.2 Current ISGI activities**

The primary responsibility of ISGI is the calculation, distribution and archiving of reference values for IAGA geomagnetic indices. As already mentioned, each ISGI Collaborating Institutes is responsible for calculating, disseminating and archiving one or several IAGA Geomagnetic Indices. ISGI headquarters are, as such, responsible for:

- maintaining and updating of the Internet ISGI homepage. This site is designed as a portal to all sites of collaborating institutes. Together with the these sites, it provides reference values of IAGA geomagnetic indices;
- promoting the participation of ISGI to networks which aims to facilitate user access to all necessary data in the physics of solar-terrestrial relations;
- publishing and distributing a Monthly Bulletin that disseminates provisional values of IAGA geomagnetic indices.

ISGI expertise is widely recognized by academic (Solar Terrestrial Physics) and operational (Space Weather) communities. ISGI has thus a role of expertise and advice regarding the characterization of geomagnetic activity by mean of indices. In the frame of IAGA, ISGI is responsible, through the IAGA ad-hoc committee on geomagnetic indices currently chaired by the ISGI Director, for making proposals for all matters relating to geomagnetic indices: dissemination policy, definition and endorsement of new indices

### 3.3 A noteworthy ISGI contribution

During the last decades, Internet and computer developments resulted in a revolution in data handling, and – as a consequence of Internet facilities – users strongly required to have preliminary values of geomagnetic indices available on line within delays as short as possible.

Early in the Internet age, many institutions which are not part of ISGI started to derive and disseminate online preliminary values of IAGA geomagnetic indices, using derivation schemes different from those endorsed by IAGA. In addition, the used derivation schemes were generally not clearly published, and they varied from one institution to another. It was thus common to find differing estimated preliminary values for a given IAGA geomagnetic index during a given time interval. Such situation was very confusing for the users since there was no longer clear quality stamping. From the user's point of view, the crucial question became: how to be sure of the quality of geomagnetic indices that I find on Internet? This could have resulted in a dramatic loss of confidence of users in the quality of geomagnetic indices made available through Internet. Such a loss of confidence would have in turn challenged the possibility to continue the production of high quality long term data series. This made it clear the necessity for a policy for on-line dissemination of indices, and for quasi real time on-line dissemination of reliable preliminary values of IAGA geomagnetic indices.

ISGI Collaborating Institutes then decided to routinely make available on-line state of the art preliminary values (few hours to two days delay for all IAGA geomagnetic indices, since 1996 for most of them; 30 min delay since 2004 for the aa index). In parallel, ISGI led discussions within IAGA. IAGA resolved that the reference values of geomagnetic indices are those produced by the ISGI Collaborating Institutes, and that they must be posted first at the ISGI and ISGI Collaborating Institutes homepages. IAGA also took a resolution urging the “producers of the estimated indices to clearly label them with “est” at the end of each index name to distinguish them from the official IAGA indices” (Resolution 5, IAGA News 38 1998, p. 42). This policy succeeded both in fitting with this new environment and in preserving the historical heritage, namely the high quality, homogeneous long term data series. This policy is still in force

## 4 CONCLUSION

The evolution in the geomagnetic indices activity dissemination during this transition period makes clear the need (i) for reference(s) scientific organization(s) – IAGA in the present case – in charge of the definition of the policy dissemination of the “added value products” (geomagnetic indices in the present case) and of the labelling of the reference places, and (ii) for reference places where users are sure to find reliable “added value products” – ISGI and ISGI Collaborating Institutes in the present case. Finally, it is important to note that success in defining a new policy for on-line dissemination of IAGA geomagnetic indices was made possible by the deep involvement of the producers in related research activities: their understanding of user needs helped them to propose effective solutions tailored to these needs.

## 5 REFERENCES

- Legrand, J.P. & Simon, P. (1991) A two components solar cycle. *Sol. Phys.*, 131, 187.
- Menvielle, M. & Berthelier, A. (1991) The K-derived planetary indices: description and availability. *Reviews Geophys. Space Phys.*, 29, 415-432; erratum: 30, 91, 1992.
- Menvielle, M., Iyemori, T., Marchaudon, A. & Nose, M. (2011) Geomagnetic indices. In Manda, M. and Korte, M., (Eds.), *Geomagnetic Observations and Models*, IAGA Special Sopron Book Series 5, DOI 10.1007/978-90-481-9858-0\_8, Springer
- Ouattara, F., Amory-Mazaudier, C., Menvielle, M., Simon, P. & Legrand, J.-P. (2009) On the long term change in the geomagnetic activity during the 20th century, *Ann. Geophys.*, 27, 2045–2051.
- Rouillard, A.P., Lockwood, M. & Finch, I. (2007) Centennial changes in the solar wind speed and in the open solar flux, *J. Geophys. Res.*, 112, A05103, doi:10.1029/2006JA012130.
- Thomson, A.W.P. & Lesur, V. (2007) An improved geomagnetic data selection algorithm for global geomagnetic field modelling, *Geophys. J. Int.*, 169, 951–963, doi: 10.1111/j.1365-246X.2007.03354.

# OPERATIONS OF THE WORLD DATA CENTRE FOR GEOMAGNETISM, EDINBURGH

*S J Reay\*, E Clarke, E Dawson, S Macmillan*

*\*British Geological Survey, Murchison House, West Mains Road, Edinburgh, EH9 3LA, United Kingdom  
Email: [sjr@bgs.ac.uk](mailto:sjr@bgs.ac.uk)*

## ABSTRACT

*The British Geological Survey has operated a World Data Centre for Geomagnetism since 1966. Geomagnetic time-series data from around 280 observatories worldwide at a number of time resolutions are held along with various magnetic survey, model and activity index data. The operation of this data centre provides a valuable resource for the geomagnetic research community.*

*The operation of the WDC and details of the range of data held are presented. The quality control procedures that are applied to incoming data are described, as is the work to collaborate with other data centres to distribute and improve the overall consistency of data held worldwide. The development of standards for metadata associated with datasets is demonstrated and current efforts to digitally preserve the BGS analogue holdings of magnetograms and observatory yearbooks are described.*

**Keywords:** Geomagnetism, World Data Centre, World Data System, Metadata, Digital data capture

## 1 INTRODUCTION

A World Data Centre (WDC) for Geomagnetism was established in the United Kingdom in 1966. This was operated by the Institute of Geological Science, which later became the British Geological Survey (BGS), in Herstmonceux, Sussex. The WDC moved to its current location in Edinburgh in 1977. BGS is part of the Natural Environment Research Council (NERC), a research centre funded by UK government. In the past this WDC focused its attention primarily on gathering data for use in global magnetic field modelling, particularly geomagnetic observatory annual means. The WDC for Geomagnetism, Copenhagen, then hosted by the Danish Meteorological Institute, gathered geomagnetic observatory one-minute and hourly mean values. In 2007 BGS agreed to take over responsibility for the digital datasets held at WDC Copenhagen. These were transferred to Edinburgh and a copy made of the data catalogue web pages to ensure the data were available in an identical form from the BGS website ([www.wdc.bgs.ac.uk](http://www.wdc.bgs.ac.uk)).

The WDC operations in Edinburgh are carried out by the Geomagnetism Team of BGS. There are several staff members involved in WDC operations to various degrees, although all have additional duties within the team. BGS staff have experience in processing and delivering geomagnetic data, since the team operate eight geomagnetic observatories worldwide and conduct repeat station observations across the UK. They are also experienced in using geomagnetic data, as the team's research scientists work in the field of global geomagnetic field modelling and space weather science.

In August 2011 the WDC for Geomagnetism, Edinburgh applied to become part of the newly established ICSU World Data System (WDS) (<http://www.icsu-wds.org/>). At the time of writing the application awaits approval by the WDC Scientific Committee.

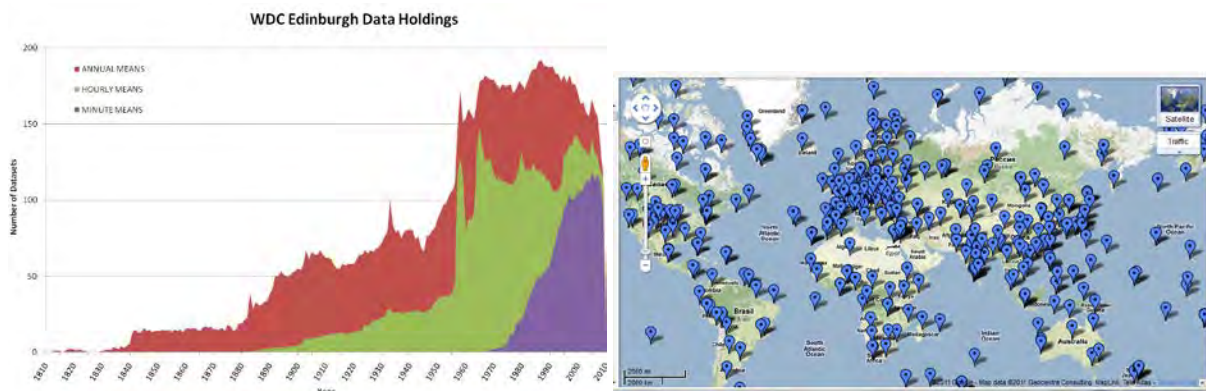
## 2 DATA HOLDINGS

WDC for Geomagnetism, Edinburgh is an active and growing data centre keenly seeking out new data. An individually-tailored email is sent annually to more than 100 geomagnetic observatory operators requesting any new data and highlighting current gaps in the data holdings for their observatories. Reciprocal arrangements with other WDCs to share data directly submitted to us is in place and an agreement with INTERMAGNET<sup>1</sup> has been

<sup>1</sup> <http://www.intermagnet.org>

established to collect data from this network and distribute via the WDC. BGS hope to increase links with other WDCs in the future and widen this data distribution network.

The WDC holds an extensive range of geomagnetic data including digital data from around 280 geomagnetic observatories worldwide containing one-minute mean values from 1969; hourly values from 1883; and annual means from 1813 (Figure 1). Data from land, marine, satellite and aeromagnetic surveys and repeat stations worldwide from 1900 onwards are available, as well as charts and computations from main field models, including the World Magnetic Model (WMM)<sup>2</sup> and International Geomagnetic Reference Field (IGRF)<sup>3</sup> model. Further digital data include a complete set of the definitive magnetic activity indices ( $K$ ,  $K_p$ ,  $ap$ ,  $Ap$ ,  $aa$ ,  $Aa$ ,  $Cp$  and  $C9$ ) and solar activity indices (*International Sunspot Number* and  $F10.7$ ). An archive of historical and analogue records is also maintained. These include geomagnetic data collected during NERC funded university projects; magnetograms from all UK observatories from 1850 (the digital capture of which is described in section 5); and a library containing observatory yearbooks from around the world, expedition memoirs, original survey observations and other miscellaneous items.



**Figure 1.** Data holdings of the WDC; Left – the number of observatory data holdings by time resolution and year, Right – the locations of current & past observatories for which we hold annual, hourly or minute data. (Google map is © Google 2011, map data © Europa Technologies, MapIT, Tele Atlas)

### 3 QUALITY CONTROL WITHIN OUR WDC

All data submitted to the WDC are subjected to quality assurance checks before ingestion into the database. When problems are encountered BGS work with the data originators to try to resolve the issues and improve the dataset. It is known that errors do exist within the historical data holdings; in some cases data quality could be improved or there are inconsistencies in the data held at the various WDCs for Geomagnetism. This is of great concern to the scientific community and is the main focus of BGS's current WDC activities.

The problem of disagreement between datasets held at different WDCs is not a straightforward one. Geomagnetic observatory data are commonly published annually for the preceding calendar year. Finalised data are classed as 'definitive'. However, data may then have corrections applied to produce a second, improved 'definitive' dataset. In the past, subsequent corrections may have been made by the data originator and not passed on to all data centres. Also, some data centres may have made their own corrections to the data and not documented or distributed the changes. The corrections may take the form of a change to the absolute ('baseline') level of the data, the removal of obviously erroneous data points (such as cases where there is evidence of significant environmental noise), or the correction of simple typographical and formatting errors. In all cases, because of the lack of version information within the geomagnetism community's file format standards, the history of these changes may be lost.

Since 2007 BGS has worked to improve the quality of the data to aid the scientific community. Simple typographical or formatting errors within the hourly observatory data holdings were sought and corrected

<sup>2</sup> <http://www.geomag.bgs.ac.uk/research/modelling/WorldMagneticModel.html>

<sup>3</sup> <http://www.ngdc.noaa.gov/IGAG/vmod/igrf.html>



(Dawson *et al.*, 2009). Errors of this type, such as the use of an incorrect “missing sample flag” value, can result in gross effects in any subsequent data analysis. Furthermore, holdings of hourly and minute mean values have been compared with those held in other WDCs for Geomagnetism in order to identify gaps in the respective holdings.

BGS are working with the WDC for Geomagnetism, Kyoto in particular to harmonise the data holdings. In theory, the datasets common to different WDCs should be identical (given that they are holding ‘definitive’ data). However, in a recent analysis by Dawson *et al.* (2011), it was discovered that almost 20% of the datasets have some level of disagreement between them. The WDCs at Edinburgh and Kyoto have agreed to work together to resolve these issues. This is not a simple undertaking; it is not always clear why some data are in disagreement and, without records of the corrections made, the task of identifying the authoritative ‘definitive’ dataset is made more difficult. This work will need to be carried out in partnership with the observatory operators who will have the final say on which version should be considered definitive. However in some cases, where observatories are no longer operational or the host institute no longer exists, absolute certainty may not be possible.

## 4 METADATA

The issues caused by the lack of documentation, with regard to quality control procedures carried out on data held at the WDCs, demonstrate the importance of metadata. There have been discussions between the WDCs and among the wider geomagnetic community to address this matter and establish a metadata standard for geomagnetic observatory data (e.g. Fischman *et al.*, 2009 or Reay *et al.*, 2011). Current metadata standards for geospatial data may act as a guide in establishing a standard but as geomagnetic observatory data are time-series, and every aspect of metadata associated with an observatory (up to and including its location) could change with time, it is difficult to determine what to assign a metadata record to.

Within the WDC for Geomagnetism, Edinburgh the decision was made to focus on metadata gathering, collecting useful information which can be re-formatted according to an agreed metadata standard at a later date. As part of the annual ‘call-for-data’, basic metadata such as observatory name, location, dates of operation, contact information, instruments used, and other similar details are requested. Historically, observatory yearbooks were produced in a reasonably similar format around the world and can be considered the unofficial metadata standard for observatory data. The continued production of yearbooks is actively encouraged and electronic copies of these are made available for download via the WDC website wherever possible. These annual reports thoroughly describe the operations carried out at an observatory for a given year and usually contain all the information, such as processing history, instrumentation etc. that would be required for a complete metadata record.

The metadata records collated to date are available online for users to examine. These records are being populated with information on known QC issues associated with the data, including notes of the version history of the data if it has been modified. To help populate these records, feedback from both the providers of the original datasets and the WDC users are welcome.

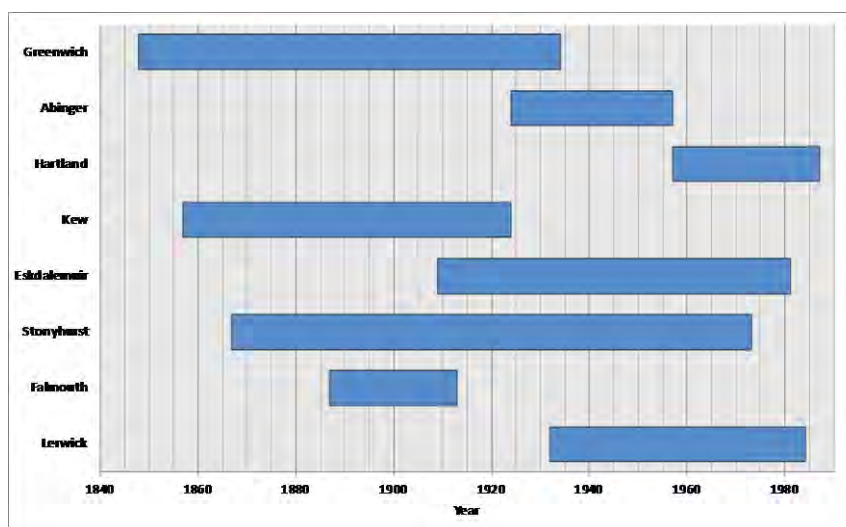
## 5 PRESERVING HISTORICAL DATA

In the middle of the 19<sup>th</sup> century regular systematic measurements of the Earth’s magnetic field began in earnest. Analogue data, in various forms, were produced until the late 20<sup>th</sup> century when the majority of observatories introduced digital data recording. BGS are custodians of the original records and results from historical UK observatories. These consist of magnetogram records on photographic paper between 1848 - 1987 and a range of yearbooks produced to summarise the results and describe the operations. The operational timeline of each observatory is shown in Figure 2. These records provide a unique and continuous record of magnetic variations across the UK. Records for historic British colonial observatories are also held and copies of various yearbooks from observatories worldwide are stored as part of WDC operations.

In the past it was realised that some of these records were in a poor physical state caused by storage in unsuitable conditions. There was also a risk of further deterioration. In 2008 a programme commenced to create digital back-up copies of each paper magnetogram and make these available online (Clarke *et al.*, 2009). The records were also moved to a secure, environmentally controlled archive room conforming to BS5454 (the British

Standard for the preservation of archival material). The digital capture was carried out using a 21-megapixel camera at a fixed focal length. The images are stored in a database and an online magnetogram archive is available from <http://www.bgs.ac.uk/data/Magnetograms/home.html>. This work is ongoing; as of December 2011 approximately 90% of the 472-observatory years of data have been digitally captured.

A collection of 378 yearbooks from current and historical UK and British colonial observatories are held, plus many articles and survey data records. Work to digitise the historical observatory yearbooks, in conjunction with the digitisation of the magnetograms from the same observatories, was started early in 2011. Some articles related to observatory results and operations were also included. The yearbooks contain the published results for each observatory and information about the observatory operations and measurements - a critical source of metadata. A 'Bookeye' scanner was used to capture good quality images of each page. This work is now complete and the electronic copies will soon be available to view online.



**Figure 2.** The temporal coverage of the analogue UK magnetogram records for eight UK observatories.

## 6 FUTURE DEVELOPMENTS

This paper has discussed the status of the World Data Centre for Geomagnetism, Edinburgh and the work currently undertaken there. In the future it is hoped that this WDC can join the WDS and contribute to this new global initiative. New data will continue to be ingested and in the future the collection of additional data types, such as one-second magnetic observatory data, will be considered. The aim to make progress on quality control issues in collaboration with the WDC for Geomagnetism, Kyoto is high on the agenda. Datasets will also be better documented by increasing the quality and range of metadata held. The programme to digitise BGS holdings of analogue UK magnetograms will continue until all records are captured and available online. Furthermore, a catalogue of all other analogue records held will be completed, with unique and at-risk documents identified for digitisation. Finally, the current WDC online interface will be replaced to improve how users can access and analyse data in the future. Web services technologies, such as that described in Dawson *et al* (2012), will be used to better deliver data to the scientific community.

## 7 ACKNOWLEDGEMENTS

This paper is published with the permission of the Director of the British Geological Survey (Natural Environment Research Council).

## 8 REFERENCES

Clarke, E., Flower, S., Humphries, T., McIntosh, R., McTaggart, F., McIntyre, B., Owenson, N., Henderson, K., Mann, E., MacKenzie, K., Piper, S., Wilson, L. & Gillanders, R. (2009) The digitization of observatory

magnetograms. Poster presented at: 11th IAGA Scientific Assembly, 23-30 Aug 2009, Sopron, Hungary. Available from: <http://nora.nerc.ac.uk/12494/>

Dawson, E., Lowndes, J. & Reddy, P. (2012) The British Geological Survey's new geomagnetic data web service. Manuscript submitted for publication to Data Science Journal (*proceedings of 1<sup>st</sup> ICSU World Data System conference*)

Dawson, E., Reay, S., Macmillan, S., Flower, S. & Shanahan, T. (2009) Quality Control Procedures at the World Data for Geomagnetism (Edinburgh). Poster presented at: 11th IAGA Scientific Assembly, 23-30 Aug 2009, Sopron, Hungary. Available from: <http://nora.nerc.ac.uk/11740/>

Dawson, E., Macmillan, S., Humphries, T. & Beggan, C. (2011) A comparison of data holdings at World Data Centres for geomagnetism in Edinburgh and Kyoto. Poster presented at: IUGG XXV General Assembly, 28 June - 7 July 2011, Melbourne, Australia. Available from: <http://nora.nerc.ac.uk/15365/>

Fischman, D., Denig, W.F. & Herzog, D. (2009) A Proposed Metadata Implementation for Magnetic Observatories. Proceedings of the XIIIth IAGA Workshop on Geomagnetic Observatory Instruments, Data Acquisition, and Processing: U.S. Geological Survey Open-File Report 2009-1226, 82-85 pp

Reay, S.J., Herzog, D.C., Alex, S., Kharin, E., McLean, S., Nosé, M. & Sergeyeva, N. (2011) Magnetic Observatory Data and Metadata: Types and Availability. In: Manda, M. & Korte, M. (eds) *Geomagnetic Observations and Models*, IAGA Special Sopron Book Series 5, DOI 10.1007/978-90-481-9858-0\_7 pp149-181

# ACTIVITIES AND PLAN OF CENTER FOR GEOPHYSICS (BEIJING) FROM WDC TO WDS

*Fenglin PENG<sup>1,2\*</sup>, Maining MA<sup>3</sup>, Le PENG<sup>2,5</sup>, Jian ZHANG<sup>3</sup>, Gengxiong CHEN<sup>1,2</sup>, Yufang LI<sup>4</sup>, Bo SUN<sup>1</sup>, and Yunfei ZHANG<sup>2</sup>*

<sup>1</sup>*Institute of Geology and Geophysics, Chinese Academy of Sciences, P. O. Box 9825, Beijing, 100029, P. R. China, Email: pengfy@yahoo. cn*

<sup>2</sup>*Center for Geophysics, World Data System, P. O. Box 9825, 100029, Beijing, P. R. China*

<sup>3</sup>*Earth Science College, Graduate University of Chinese Academy of Sciences, 100039, Beijing, P. R. China*

<sup>4</sup>*Department of Electricity & Information, Normal College, Beijing Union University, Beijing, 100011, P. R. China*

<sup>5</sup>*Department of Physics, City University of New York, New York, USA*

## ABSTRACT

*In this report we introduce the development of WDC for Geophysics, Beijing including our activities in the eGY and in the transition period from WDC to WDS. We also present our future plan. We engaged in the development of geophysical informatics and related data science. We started the data visualization of geomagnetic field in the GIS system. Our database has been expanded from the geomagnetic data to the data of solid geophysics including geothermal data, gravity data, and the records of aurora sightings in the ancient China. We also joined the study on the history of the development of geophysics in China organized by the Chinese Geophysical Society (CGS).*

**Keywords:** WDC, WDS, Geophysical Informatics, eGY.

## 1. INTRODUCTION

The establishment of the World Data Center system was one of the achievements of International Geophysical Year (IGY, 1957-1958). China is one of the first participants of the IGY. A committee for IGY was founded by Chinese Academy of Sciences (CAS). Academician Kezhen Zhu was the chair of the committee. The committee of IGY promoted the establishment and development of the branches of the geophysics in China. The Shenshan Geomagnetic Observatory was awarded a gold medal by International Association of Geomagnetism and Aeronomy. Since IGY, Chinese geophysicists and other scientists working in the adjunct areas have made great achievements.

In late 1980s, the Chinese scholars on Earth science proposed to establish data centers as WDC in China. This proposal was supported by Academician Duzheng Ye and Honglie Sun, the vice presidents of Chinese Academy of Sciences. With their effort, the Chinese WDC Coordination Committee and nine data centers in the WDC system were founded in 1988. The chair of the committee was Academician Honglie Sun, and the secretary of the committee was Academician Jiulin Sun.

The WDC for Geophysics, Beijing (original name: the WDC-D for Geophysics) is one of the first centers of the WDC system in China. The WDC is located in the Institute of Geophysics, CAS. It has got a lot of help from the institute of Geodesy and Geophysics in Wuhan, CAS. In 1999, the Institute of Geophysics, CAS, and Institute of Geology, CAS, combined as the Institute of Geology and Geophysics, CAS. The Laboratory of Space Physics (ionosphere and upper atmosphere physics) also move into this institute then. The WDC for Geophysics, Beijing, has got more and more support of core data resources in geophysical research fields ranging from solid geophysics to geospace physics.

## 2. WORKS OF WDC FOR GEOPHYSICS, BEIJING, BEFORE eGY

The precious magnetograms measured in the Shanghai Sheshan Observatory, which was founded in 1870's, played an important role in the foundation of the WDC. The Sheshan Observatory which was named initially as

the Xujiahui Observatory, has run for more than one hundred years. It is one of the world oldest geomagnetic observatories.

Beginning with the data from the Sheshan Observatory, WDC for Geophysics, Beijing started the data work of geophysics in China. The center have done data work both on traditional media (such as photopapers and films) and electronic media. The center has built a group of databases: Beginning with the database of geomagnetic field and then covered most branches of geophysics. A lot of results of geophysical observations was digitalized and shared online in the data in this data center.

The first project of the WDC was the recovery of the magnetograms which were obtained at the Sheshan Observatory since 136 years ago, and were damaged due to aging. These magnetograms were of significant importance for the studies of geomagnetic field. The broken magnetograms were recovered by specialists. Magnetic plots which were recorded on 36mm photographic microfilms, were copied on microfilms. Our center also started to study the construction of China Geophysics Database (Gao et al., 1992).

During middle 1990's, the WDC attempted to forecast the peak of each solar activity by making use of the sunspot data from the Sheshan Observatory as well as the world-wide geomagnetic data. The result of this research project was presented in the International Conference of Solar-Terrestrial Prediction in 1996 (Peng, Chang, Tschu. and Wang, 1997).

In late 1990s, there was an increasing demand of the scientific data service online. According to the task of the pilot project of the Ministry of Science and Technology, the WDC began to utilize a computer network to provide data service. Senior engineer Hongfei Chen helped the center to set up the network server, and the website of the WDC was posted online. The geomagnetic data was first uploaded on the server. Then the data measured at the Zhongshan Observatory, Antarctica, and at the newly-founded geomagnetic observatory of CAS were added to the database.

The online database was initially built with ASP language and SyBase environment in 2000. In 2003, the online database of geomagnetic-field 1-minute mean values was accomplished by using JSP and SQL server. In the new system, a lot of hard drive space was saved and the search speed of the database was increased significantly. The data sets of daily measurement were uploaded to the online database automatically, and the site could generate the magnetograms of arbitrary time-interval dynamically according to users need (Peng, Wang, Zheng, Xing *et.al.*2007)

Besides the data of geomagnetic field, the WDC for Geophysics, Beijing, also started to collect the data of other branches in geophysics. With the critical support of Xiangru Kong, the former deputy director of the Institute of Geophysics, CAS, the data sets of magnetotelluric sounding in Inner-Mongolia and Tsinghai-Tibet were added to the website.

In 2005, Academician Jiyang Wang and Prof. Shengbiao Hu provided their data collection of the geothermic measurements in the mainland of China during the last 50 years. The WDC also collected the records of gravity anomaly in different areas of China, which were measured by R. P. Lejay, Gongshu Gu, Rongsheng Zeng, and Zhongyin Zhang in 1940s. This set of data, which was measured during the war years, is of great significance. In 2006, the data of solid geophysics in Taiwan, the data of oceanic geothermic and the data of deep seismic sounding in North China were added to the online database.

### **3. ACTIVITIES FROM eGY TO WDC-WDS TRANSECTION PERIOD**

During 2006-2008, the WDC for Geophysics, Beijing, put a lot of effort into the promotion of eGY activities in China. According to the instruction of Huadong Guo, the deputy general secretary of CAS, Peng F. suggested to the secretary general of Chinese Geophysical Society, Rixiang Zhu, to join the eGY. By the support of presidents of CGS and directors of institutes on geophysics and space science in China, The Chinese government decided to found a national committee of eGY in China. Chinese Geophysical Society, Chinese Society of Space Research, Seismological Society of China, Chinese Meteorological Society and Chinese Society of Oceanography joined the foundation of the committee. Liu Guangding, honorary president of CGS, was appointed as the honorary president of the eGY national committee of China; WANG Shui, president of CGS, was appointed as the president of the eGY national committee of China. PENG Fenglin, the director of WDC for Geophysics, Beijing, was appointed as deputy secretary general of this committee.

The WDC for Geophysics, Beijing, promoted the data sharing between the academic organizations. A national

data-sharing platform of geosciences was formed. This is the base of the further cooperation between the academic organizations. And the development of the WDC for Geophysics also benefited from the point of view of data sharing.

In 2007 May, the conference of the centers in the World Data Center organization was held in Bremen, Germany. 19 representatives from 9 data centers in China attended the meeting. In this conference, the organization reached the agreement that the interoperation of the data should be developed among World Data Centers. A beta portal was first set up between 10 centers in the WDC system including WDC for Geophysics, Beijing.

From 2007, a virtual geophysics platform has been constructed. The virtual platform for Geophysics (<http://www.geophys.cn>) and concerted work environment platform for virtual organization on geophysics (<http://e.geophys.cn> & <http://vp.geophys.cn>) were included in the whole platform (Peng, Wu, Guo, Zhang, Chen, Zhu, et al., 2008).

In May 2008, collaborating with UNECO SOC Global Alliance for ICT and Development, we hold a session on the sharing of scientific data and knowledge in the seventh Asia-Pacific City Informatization Forum at Shanghai, held in the same year. In October, we hold a special session in the annual conference of CGS: Session 1 - eGY and the advance of geophysical informatics.

In 2008-2009, we started the data visualization analyzing on geomagnetic field in GIS system. Cooperating with the Graduate University of CAS and Peking University, WDC for Geophysics, Beijing, made the data of IGRF 10 (The 10th generation of International Geomagnetic Reference Field) visualized in the Google Earth system. We use a new toolbox in Matlab-Google Earth Toolbox, which provides various plotting/drawing functions that can be saved as KML output, and loaded in Google Earth. With the functions in this toolbox, we can display spatially and temporally distributed data within Google Earth (Wang, Shen, Peng, Yuan, Tang, Xing, et al., 2008; Wang, Peng, Ma, Yuan, Bai, and Sun, 2009).

In 2010, the information committee of CGS and the Computer Network & Information Center, CAS, organized an e-science salon on Earth-planet-space physics and engineering. Many senior scientists on geophysics, geology, astrophysics, space physics and information science/technology joined the activity eagerly. The holding of this salon will be kept on in the future. It will take an active role required to push the development of the data-sharing, data science and informatics of geophysics.

During 2010-2011, WDC for Geophysics, Beijing, studied the data and events recorded in ancient Chinese documents. The Institute of Science History organized the ancient records of aurora-sightings in China. These records and the data of the geomagnetic field variations during the eclipses in 1930s were uploaded to the server and shared online.

#### **4. ON-GOING WORK AND PLAN OF OUR CENTER**

As a partner of Data-sharing Infrastructure of Earth System Science built by the Ministry of Sciences and Technology of P. R. China, the Center for Geophysics can get more financial support every year from 2011. Our center plans to organize all resources of data including the data of geomagnetics, geothermic, gravity, Earth electricity, and Earth deep structure. 9 main theme databases on geomagnetics, geothermics, gravity, geophysics in Taiwan, earth deep structure, planetary geophysics and urban environment geophysics will be constructed in the near years. The most critical ones of them are listed below:

- (1) Basic parameters of geophysics;
- (2) Geomagnetic data, including:
  - the data of geomagnetic observatories,
  - the field of geomagnetic survey in China,
  - paleomagnetism data
- (3) Gravity data in China since 1900s.
- (4) Geothermics;
- (5) Geodynamics;
- (6) Structure of deep Earth, including data of magnetotelluric sounding and deep seismic sounding;
- (7) Earth electromagnetism
- (8) Space physics
  - Database of Ionosphere and upper-atmosphere physics: Observation data of Ionosphere and upper-atmosphere Physics, Data of national Ionosphere GPS observation net, which is the biggest data set of ionosphere GPS

observation in China  
(9) Historical and scholars' information of Chinese geophysics

We plan to organize a series of salons and conferences to promote the development of data science and geophysical informatics.

## 5. SUMMARY

In the more than 20 years, starting with the magnetograms of Sheshan Observatory, WDC for Geophysics, Beijing have collected the data from various branches of geophysics such as geomagnetic field, gravity and geothermics. The WDC for Geophysics, Beijing, also put a lot of effort into the digitalization and data visualization of the results of all kinds of observations on geophysics. With the contribution of the scholars and researchers from the universities and research organizations in China, WDC for Geophysics, Beijing, have built a group of high-quality databases. The online databases were maintained and updated by making use of the information technology developed during last twenty years. And an online platform was founded for data-sharing between the research organizations for geophysics in China.

In the past 6 years, WDC for geophysics, Beijing also promoted the activities of eGY as an important role for the eGY in China. Several conferences and academic activities about eGY and the development of geophysical informatics, e-science for geophysics and data science in geophysics have been hold.

Center for Geophysics and the National Laboratories collaborates on Data-Sharing Infrastructure of Earth System Science. These works are the base of the further development of geophysical informatics.

## 6. ACKNOWLEDGEMENTS

Thanks for the financial support from the Ministry of Sciences and Technology of P. R. China by through the Data-sharing Infrastructure of Earth System Science and Chinese Academy of Sciences by the key project of knowledge creation to WDC system.

## 7. REFERENCES

- Gao, M. Q., Lu, W., Wen, X. and Gao, M., (1992) China Geophysics Database in WDC-D for Geophysics, *CODATA Bulletin*, 24(6), 2.
- Peng, F., Chang, H., Tschu, K. K. & Wang, J. L., (1997) The Use of Geomagnetic AQD near Sq Focus for Predicting the Magnitude of Sunspot Maximum, *Solar-Terr. Predict.*, 5, pp 99-102.
- Peng, F., Wang, D., Zheng, X., Xing L., Tang, K., Yue, B., Shen, X., Zhang, J., Peng, L. & Huang, Q. (2007) The Internet Databases of the World Data Center for Geophysics, Beijing. *Data Science J.* 6(S), 879-882.
- Peng, F., Shen, X., Tang, K., Zhang, J., Huang, Q., Yue, B., & Wang, D. (2007) Data-sharing Work of the World Data Center for Geophysics, Beijing, *Data Science J.* 6(S), 404-407.
- Peng, F., Wu, Z., Guo, J., Zhang, J. Chen, X., Zhu, R., Wang, S. & Liu, G. (2008) eGY in China: From IT to geophysical informatics and its outreach. *21st International CODATA Conference* (pp139), Kyiv, Ukraine
- Wang, D., Shen, X., Peng, F., Yuan, X., Tang, K., Xing, L. & Peng, L. (2008) Basic Research of IGRF 10 Model on Web Visualization. *21st International CODATA Conference* (pp51), Kyiv, Ukraine
- Wang, D., Peng, F., Ma, M., Yuan, X., Bai, C., & Sun, L. (2009) Visualization Research of IGRF Model. *Seismological and Geomagnetic Observation and Research* 30(4), 7-11.
- Yang, L., Peng, L., & Peng, F. (2008). The Weather, Geologic and Geophysical Environment as a Factor in Transportation Information System. *21st International CODATA Conference* (pp157), Kyiv, Ukraine

# STRASBOURG ASTRONOMICAL DATA CENTRE (CDS)

*F Genova*<sup>1\*</sup>

<sup>1</sup>*Observatoire Astronomique de Strasbourg (UMR7550 UNISTRA/CNRS), 11 rue de l'Université, 67000 Strasbourg, France  
Email: francoise.genova@astro.unistra.fr*

## ABSTRACT

*The Centre de Données astronomiques de Strasbourg (CDS), created in 1972, has been a pioneer in the dissemination of digital scientific data. Ensuring sustainability on several ten years has been a major issue since science and technology evolve continuously and the data flow increases endlessly. The paper briefly describes CDS activities and main services, and its R&D strategy to take advantage of new technologies. The next frontiers for CDS are the new Web 2.0/3.0 paradigm, and at a more general level global interoperability of astronomical on-line resources in the Virtual Observatory framework.*

**Keywords:** Data Centre, Astronomy, Web 2.0/3.0, Interoperability, Virtual Observatory

## 1 INTRODUCTION

The *Centre de Données astronomiques de Strasbourg* (CDS, <http://cdsweb.u-strasbg.fr/>, Genova, Egret, Bienaymé, Bonnarel, Dubois, Fernique et al., 2000) has been a very early player in the dissemination of digital scientific data: it was created in 1972 by the French astronomy agency, INAG (National Institute for Astronomy and Geophysics), which is now CNRS/INSU (National Institute for Universe Sciences), in agreement with the University Louis Pasteur, now University of Strasbourg. The mandate it has been given showed a far-seeing vision, since it included, in these early times, the collection of “useful” data on astronomical objects, in electronic form, their improvement by critical evaluation and combination, the distribution of the results to the international community, and also conducting research using the data. The whole idea of electronic data collection, curation, dissemination and scientific re-use, which is the guideline of current policies about scientific data, has thus been present from the very beginning at CDS. The data centre had originally been created as the *Centre de Données Stellaires* (Stellar Data Centre), with the initial aim of gathering stellar data for studying the galactic structure, but it was renamed to its current name, keeping the already well known acronym, in 1983 when its domain of action was extended to all astronomical objects (outside the solar system).

CDS main role is to support the international community in its research tasks, not only to collect and to curate information. Its core task is to provide highly used value-added services (Section 2), and the main keywords of the activities are quality, scientific and technical relevance, collaboration with other actors or the field and networking of expertise and resources. Its strategy, including its R&D strategy (Section 3), is *user* and *science driven*, not technology driven. The CDS has built along the years a unique expertise on scientific data, data dissemination and exchange standards. It plays a major role in the astronomical Virtual Observatory, which aims at providing seamless access to the wealth of astronomical on-line resources.

## 2 CDS SERVICES

CDS has developed highly successful added-value services: SIMBAD, which summarizes information about astronomical objects – it reached 5,000,000 million objects in 2011; VizieR, the reference service for astronomical catalogues and tables published in academic journals; and Aladin, an image visualizer, conceived to access images and catalogues stored locally or remotely. These services are daily used by the international astronomical community, and their usage is constantly increasing (500,000 queries/day on average in 2010).

SIMBAD (see Figure 1) is the reference database for identification and bibliography of astronomical objects, providing a homogenized view across astronomy sub-disciplines (Wenger, Ochsenbein, Egret, Dubois, Bonnarel,

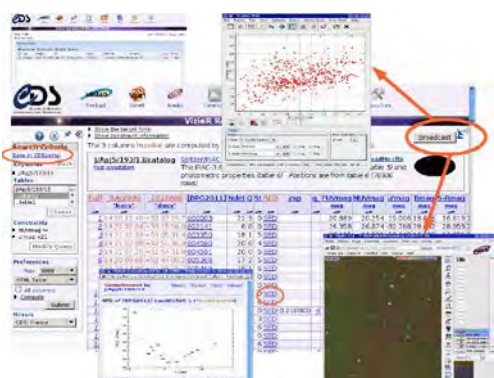


Borde et al., 2000). SIMBAD data is selected from articles published in academic journals and astronomical catalogues. The first version of SIMBAD was developed at the beginning of the seventies, and the current version of the software, the fourth major update since 1972, has been operational since 2006. This last update to date has been from a home-made, object oriented database to the open source platform PostgreSQL (Wenger & Oberto, 2007). Another interesting trend has been the inclusion of “less controlled” information in addition to measurements, bibliography and data: notes from the CDS team (2002) and more recently (2010) the possibility for users to post annotations, the first CDS Web2.0 implementation. The database content is built by a team of highly qualified librarians, working closely with CDS scientists. SIMBAD contains in December 2011 5,400,000 objects, 15,200,000 object identifiers, 250,000 bibliographic references (respectively 3,000,000, 8,300,000 and 140,000 in 2003). To cope with the rapidly increasing flux of data, methods have been developed for semi-automated entry of information from the article texts (Lesteven, Bonnin, Derriere, Dubois, Genova, Oberto et al., 2010) and tables, but the results are validated by a specialist to keep a high level of quality.



**Figure 1.** SIMBAD user interface (top left), and usage of SIMBAD in other services, providing object types to ADS and “name resolving” to coordinates to various telescope observation archives.

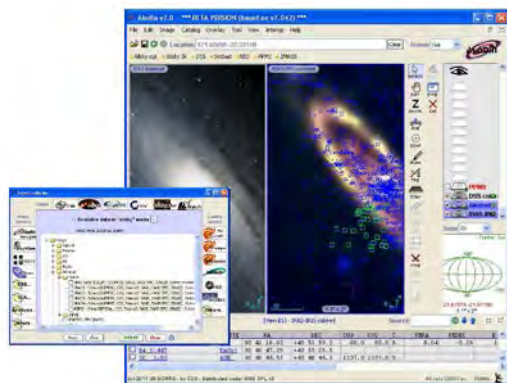
VizieR (see Figure 2) is the reference database for tabular data from astronomical catalogues and tables published in scientific papers – CDS is the data curator of these tables at the international level. Tables are completed by their description, which links their physical and astronomical content (Ochsenbein, Bauer & Marcout, 2000). Tables are available through a ftp service, and are also stored in a relational database system which allows users to browse them, to discover and extract information. The master database is a Sybase system, and queries are distributed on a local cluster. Some of the seven mirror copies are using PostgreSQL. A specific system allowing very efficient queries by position is implemented for very large catalogues (more than a few 10 million objects, the largest ones contain more than  $10^9$  objects). The catalogue and table collection has been built in close collaboration with several of the major astronomy academic journals, beginning with the “on-line only” publication of “long” tables from *Astronomy & Astrophysics* by CDS as early as 1993, as explained in Section 3. VizieR contains 9.500 catalogues in December 2011, compared to 3.800 in 2003. More and more tables now come with “attached data” such as images, spectra, time series, etc, well in line with the increasing requirement from the funding agencies to make data produced by research available for checking the research process and for re-use.



**Figure 2.** VizieR user interface (centre), with display of data through VO tools, Topcat and Aladin, and an object spectral energy distribution built from the newly implemented photometric metadata.

Aladin (see Figure 3) is a reference software dedicated to the integration, visualization and manipulation of images and catalogues provided by CDS, or the user, or remotely by astronomical data centres around the world. It has been continuously evolving, with functionalities allowing users to manipulate huge images and data cubes, to deal with photometry, to convolve images, to carry out cross-match, to use the software in scripts, etc. Aladin is used by ESA, the Space Telescope Science Institute, the NASA Extragalactic Database NED, the Canadian Astronomical Data Centre, to provide visualization of their images. Aladin is also the astronomical Virtual Observatory image portal, giving access to all data provided in VO-enabled services and able to interact with the other VO tools thanks to the VO interoperability framework. A major recent evolution is the usage of Healpix sky tessellation (Gorski, Hivon, Banday, Wandelt, Hansen, Reinicke et al., 2005), which provides a hierarchical view of data with fast

zooming capabilities and is used by the Planck and Gaia projects. This allows a new way of using the tool, also adapted to building views of the full or part of the sky from data obtained by individual projects, which are offered the possibility to build a local Healpix database which they can open for usage by all Aladin users, or by their collaborators, or keep for themselves (Fernique, Oberto, Boch & Bonnarel, 2009). The CDS reference image database is rapidly growing by the addition of “reference skys” in different wavelengths using this method.



**Figure 3.** Aladin displaying the list of available All Sky views (left) and images of a galaxy at optical and infrared wavelengths, with objects from SIMBAD and a catalogue displayed on the right view.

### 3 RESEARCH & DEVELOPMENT AT CDS

Ensuring sustainability on several ten years whereas technology can evolve very quickly is not an easy task. It requires in particular a continuous and significant effort on technological and methodological watch. E.g., very soon after the advent of the WWW, CDS has been at the forefront for networking of on-line astronomical resource, in close collaboration with the academic journals, the ADS bibliographic service and observatory archives. For instance, the CDS and *Astronomy & Astrophysics*, which was then a European journal (it now includes also several South American countries among its partners), agreed to publish long tables electronically only at CDS, instead of printing them, as early as 1993. This was a true change in paradigm, since information which since then had been available in print only became usable and searchable data, and the system also allows navigation between publication and databases (Ochsenbein, Bertout, Lequeux & Genova, 2003).

R&D is thus a fundamental activity for medium/long term sustainability, and it has to be maintained in spite of the heavy constraints linked to the core data centre role (inclusion of the ever increasing data flow in the databases, software development and maintenance, operations). At CDS it is an in-house activity, which takes a significant fraction of the time of engineers and “instrumentalist” researchers.. R&D actions have to be properly focussed: they are driven by the data centre needs, and not technology driven, i.e. new technologies are assessed only when there is a serious promise that they could improve CDS services or functioning, and not because they are trendy. Relevant technologies have to be implemented early enough to fulfil users’ expectations, but one critical requirement is that they are “sustainable enough” for usage for a certain number of years, in a technology landscape where buzz and bandwagon effect tend to be dominant and highly praised technologies can disappear within a few years.

One current frontier is the implementation of the so-called Web 2.0 user-centric approach, and of the Web 3.0 framework, with the usage of the semantic web, mobility and universality. This is mandatory since users are expecting to find in their work environment the kind of functionalities they are using in their everyday life. CDS has already implemented the first steps, with a portal which provides “mash-up” of its services (Boch & Derriere, 2010), and the possibility for users to post annotations in SIMBAD and VizieR. The implementation of a personal user space opens the way for personalized customization of the user interface, and allows one e.g. to store preferences or results. On the other hand, the evolution towards a CDS Web 3.0 will require a deep evolution of the user interface towards a more intuitive human-machine interaction. A first version of the CDS portal for mobile phones is available.

Another frontier for astronomy service providers is global interoperability of on-line resources, the so-called Astronomical Virtual Observatory. The CDS has been a precursor of the VO in many respects. It has also been a major player of the astronomical Virtual Observatory endeavour since the emergence of the project circa 2000. It has been participating actively in the definition of interoperability standards under the auspices of the International Virtual Observatory Alliance, and implements the standards in its services, which are important building block of the astronomical information system. The CDS services are thus seamlessly available to the VO tools, and Aladin has become the image portal of the Virtual Observatory, able to interact with other tools to manage images, tabular data, spectra or data cubes, to fully explore astronomical data.

## 4 CONCLUSION

Since its creation 1972, CDS has successfully fulfilled its mission, to provide support without borders to the astronomical research community, and has played an important role in the networking of on-line astronomical resources, with observation archives, academic journals, and other data centres. This has required over the years an agile strategy, to deal with the constant evolutions of astronomy, users' expectations, and technology, and with the endless data flow that the data centre has to manage. Among lessons learnt is that quality, relevance to user needs and partnership with other actors are critical to ensure long term sustainability.

## 5 REFERENCES

Boch, T., & Derriere, S. (2010). The CDS Portal: a Unified Way to Access CDS Services. *Astronomical Data Analysis Software and Systems XIX ASP Conference Series* 434 (pp. 221-224). Sapporo, Japan.

Fernique, P., Oberto, A., Boch, T., & Bonnarel, F. (2010). Another Way to Explore the Sky: HEALPix Usage in Aladin Full Sky Mode. *Astronomical Data Analysis Software and Systems XIX ASP Conference Series* 434 (pp. 163-166). Sapporo, Japan.

Genova, F., Egret, D., Bienaymé, O., Bonnarel, F., Dubois, P., Fernique, P. et al. (2000). The CDS information hub. On-line services and links at the Centre de Données astronomiques de Strasbourg. *Astron.Astrophys. Suppl.* 143, pp. 1-7.

Gorski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M. et al. (2005). HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophys.J.* 622, pp. 759-771.

Lesteven, S., Bonnin, C., Derriere, S., Dubois, P., Genova, F., Oberto, A., et al. (2010). DJIN: Detection in Journals of Identifiers and Names . *Library and Information Services in Astronomy VI: 21<sup>st</sup> Century Astronomy Librarianship, From New Ideas to Action. ASP Conference Series* 433 (pp. 317-323). Pune, Maharashtra, India.

Ochsenbein, F., Bauer, P., & Marcout, J. (2000) The VizieR database of astronomical catalogues. *Astronomy & Astrophysics Suppl.* 143, pp. 23-32.

Ochsenbein, F., Bertout, C., Lequeux, J., & Genova, F. (2003). *Navigating from Publications to Astronomical Databases. Library and Information Services in Astronomy IV: Emerging and Preserving: Providing Astronomical Information in the Digital Age* (pp. 257-262). Prague, Czech Republic.

Wenger, M., & Oberto, A (2007) SIMBAD4: Experiences Gained from the Development. *Astronomical Data Analysis Software and Systems XVI ASP Conference Serie* 376 (pp. 527-530). Tucson, Arizona, USA.

Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S. et al. (2000). The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astronomy & Astrophysics Suppl.* 143, pp. 9-22.

# EXPERIENCE AND STRATEGY OF BIODIVERSITY DATA INTEGRATION IN TAIWAN

*K T Shao\*, K C Lai, Y C Lin, L S Chen, H Y Li, C H Hsu, H Lee, H W Hsu, and G S Mai*

*\*Biodiversity Research Center, Academia Sinica, 11529 Taipei, Taiwan  
Email: zoskt@gate.sinica.edu.tw*

## ABSTRACT

*The integration of Taiwan's biodiversity databases started in 2001, the same year that Taiwan joined GBIF as an Associate Participant. Taiwan, hence, embarked on a decade of integrating biodiversity data. Under the support of NSC and COA, the database and websites of TaiBIF, TaiBNET (=TaiCOL), TaiBOL and TaiEOL have been established separately and collaborated with the GBIF, COL, BOL and EOL respectively. A cross-agency committee was thus established in Academia Sinica in 2008 to formulate policies on data collection and integration, as well as the mechanism to make data available to the public. Any commissioned project will hereafter be asked to include these policy requirements in the contract. So far, TaiBIF has gained recognition in Taiwan and abroad for its effort in the past several years. It can provide its experience and insights for others to reference or replicate.*

**Keywords:** Biodiversity informatics, GBIF, TaiBNET, TaiBIF, Database

## 1 2001: THE YEAR DATA INTEGRATION BEGAN

Before 2001, the biodiversity databases in Taiwan were scattered in various government agencies, private organizations and academic institutions. There was no real horizontal integration; these databases, at most, provide on their websites links to other sites or the home pages of relevant databases. The agencies and institutions may have departments or research units under them, each in turn may have its own websites and databases. For example, under the Council of Agriculture (COA), there are The Fisheries Agency, Forestry Bureau, and Taiwan Endemic Species Research Institute; under Construction and Planning Agency, there are many National Parks. As for the biodiversity-related private organizations, more than 30 of them have established databases and websites. The large-scale or integrated research projects promoted by the government also have their own websites, such as Forestry Bureau's National Survey and Mapping of Floral Diversity Project, Bureau of Animal and Plant Health Inspection and Quarantine's invasive species project, Council for Economic Planning and Development's National Geographic Information Systems, and National Science Council's Long-Term Ecological Research Network. However, these sites usually cover project introductions, research reports, literature, news articles, and policy and regulation guidance, but lack metadata, raw data, or primary data of the research projects. Moreover, the reports are often in paper form and kept at the funding agencies, making it difficult to achieve the goal of sharing and integrating research data across agencies.

Taiwan began to integrate biodiversity data in 2001. The new National Digital Archives Program aimed to archive not only data in the field of humanities and social sciences, but also data in biological and natural sciences such as specimens and species information. The Executive Yuan approved the Biodiversity Promotion Plan in the same year. One of the projects under the Promotion Plan is for the National Science Council (NSC), leading nine co-organizers, to start to collect and integrate biodiversity data and exchange them with global organizations. The data collected cover expert list, species checklist, specimen information, geographical distribution, spatial and temporal distribution, invasive species, species description, literature and biological resources. Also in 2001, the Global Biodiversity Information Facility (GBIF) was formally established and Taiwan joined it as an Associate Participant. As a result, Taiwan can apply the technologies and standards of GBIF's metadata and exchange platform to promote the integration of its biodiversity information and the exchange with GBIF.

In 2002, NSC began to provide funding for Biodiversity Research Center, Academia Sinica to create the website of Taiwan Species Checklist "TaiBNET" (<http://taibnet.sinica.edu.tw>). More than a hundred taxonomists are

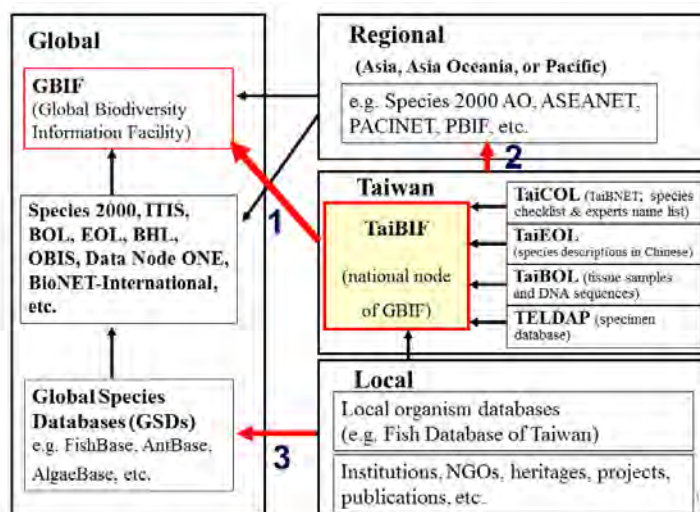
involved in the work. In 2004, GBIF's Taiwan Portal, TaiBIF (Taiwan Biodiversity Information Facility) was built. Its tasks are to consolidate Taiwan's biodiversity data, develop software tools, and hold educational training workshops in related metadata standards and technologies. In 2007, the second phase (2007-2012) of the National Digital Archives Program was transformed to Taiwan e-Learning and Digital Archives Program (TELDAP), and an International Collaboration and Promotion Division was added. One of the tasks of this Division is to promote international cooperation of biodiversity information. TaiBIF was then incorporated into TELDAP and its tasks were re-defined (Shao et al., 2008a). TaiBIF is to collect the ecological distribution data and data in *Biota Taiwanica* (in English), which is commissioned by NSC. It is also in charge of the integration of cross-agency and institution biodiversity data. On the other hand, TELDAP is to integrate data from its Institutional Projects, Request-for-Proposals Project, and Biosphere and Nature Thematic Group. It concentrates on the collection of species descriptions (in Chinese), specimens and literature information. Its task of collaboration with international organizations was assisted by TaiBIF.

As to the issue of intellectual property rights (IPR), TELDAP originally employed the "Creative Commons (CC)" licenses approach. Under a combination of Attribution, Non-Commercial, No Derivative Works, and Share-Alike requirements, data are to be put on the Internet to be browsed by the public. However, for a great deal of cultural content, the IPR issues were not clarified. In addition, the nature of some information makes it unsuitable for CC. For example, the IPR of Taiwanese aborigines is covered by separate laws. Therefore, TELDAP in 2008 had to conduct an inventory of IPR for every item and require project directors to re-sign license agreement; hence solved the problem of TaiBIF not being able to receive data from Union Catalog and then submit them to GBIF.

## 2 INTEGRATE BIODIVERSITY DATA IN TAIWAN AND EXCHANGE THEM INTERNATIONALLY

Currently TaiBIF/TELDAP integrates cross-agency, cross-institution, cross-project, NGO, and private biodiversity databases through the use of species name and GPS (latitude and longitude). The most critical element is to build a species checklist since scientific name is the keyword or linker to the integration of biodiversity data. Via the scientific name of a species, its specimen information, DNA barcode (Barcode of Life, BOL), phylogeny (Tree of Life, TOL), ecological distributional data (in EML format), and other information (Encyclopedia of Life, EOL) can all be accessed.

There are three different ways to exchange and share Taiwan biodiversity data internationally (Shao, 2006; Shao et al., 2007b). The first path is to link to GBIF directly through TaiBIF. The second path is to go through regional networks which are GBIF's Associate Participants themselves. The third path is for the local organism databases to send Taiwanese native or endemic species information to global species databases (GSDs) such as FishBase and AntBase. They in turn can link to GBIF through large-scale international cooperation projects such as COL, BOL, EOL, BHL, OBIS, BioNET-International, etc. (Fig. 1).



**Figure 1.** Integrate biodiversity databases in Taiwan and link them globally

In order to more effectively and broadly employ the integration framework and advancement of global biodiversity information, Taiwan will (1) cooperate with Catalogue of Life (COL) of Species 2000 to continue maintaining and operating Taiwan's COL, i.e., the database and website of TaiCOL (renamed from TaiBNET); (2) cooperate with CBOL and iBOL to continue maintaining and operating Taiwan's BOL, and change the name of the present “Cryobanking Program for Wildlife Genetic Material in Taiwan”(http://cryobank.sinica.edu.tw) to TaiBOL; (3) cooperate with EOL and build Taiwan's EOL, TaiEOL.

All the databases mentioned above will be integrated into TaiBIF (<http://taibif.org.tw>) which is linked to global databases such as GBIF. The old URLs will remain accessible. The new URLs and their website contents are listed below:

- (1) TaiCOL (<http://col.taibif.tw>) = TaiBNET: 54,000 native species, 1,351 alien species, along with conservation species and fossil species.
- (2) TaiBOL (<http://bol.taibif.tw>): 10,457 tissue samples from 2,981 species, 1,255 DNA sequences from 844 animal species.
- (3) TaiEOL (<http://eol.taibif.tw>): established in 2011 and will have information in Chinese on 16,000 species by the end of 2013.
- (4) TaiBIF (<http://taibif.tw>): In addition to the information on the three websites mentioned above, there is *Biota Taiwanica* with English information on 11,000 species and videos, literatures, specimens, etc. Other than database integration work, TaiBIF provides biodiversity communities in Taiwan with information technology assistance to accelerate the integration and sharing of data.

### 3 STARTING WITH SPECIES CHECKLIST—TaiCOL (TaiBNET)

The scientific name of a species is the most important keyword in linking biological information; hence, the most basic task of integrating biodiversity data is to first establish an accurate, authoritative, and complete species checklist. TaiBNET was established in 2002 to integrate and update the checklist of valid species in Taiwan. So far a total of 54,417 native species in eight kingdoms (Virus, Bacteria, Archaea, Protozoa, Chromista, Fungi, Plantae, Animalia), 59 phyla, 143 classes, 662 orders, 3,128 families and 17,706 genera have been compiled, including more than a thousand cultivars, alien species, and fossil species. TaiBNET offers users handy search functions such as search by partial scientific name, Chinese name, common name, keyword or string. Users can also browse by the taxonomy tree. Each of the tree's classification level displays the numbers of recorded organisms in its sub-levels. When a species is chosen, a description page with its classification, synonyms, and literature is shown. For some organisms, links to other databases such as Fish Database of Taiwan, FishBase, Discovery Life, and EOL are given. A mechanism is provided letting users submit related information and images to the site. A “2008 Workshop: Research and Status of Taiwan Species Diversity” was held at National Museum of Natural Science in Taichung on August 15-16, 2008. Half a year later, the books of “2008 Taiwan Species Diversity I. Research and Status” and “2008 Taiwan Species Diversity II. Species Checklist” were published along with a DVD. The book and DVD of “Taiwan Species Checklist 2010” were published in 2010 and can be freely downloaded as well (Shao et al., 2008b, 2008c, 2010).

The database of TaiBNET is currently maintained by a full-time assistant. Authorized taxonomists and their doctoral students can update data online. Afterward, the data are reviewed by top ranking scholars in each taxon. Another way to check the validity of data is to compare the checklist with existing databases such as Sp2000 or the electronic version of *Biota Taiwanica*. When discrepancies are found, experts in that taxon are invited to double check. Additionally, the fungus can be matched against CABI's Index of Fungi and the marine organisms can be matched against WoRMS.

One advantage of international cooperation is to obtain a wealth of other information. For example, Taiwan has 3,086 valid fish species as of 2011; from FishBase, 14,000 synonyms of its native fishes were obtained. The collection of synonyms is important because scientific names and classification systems are constantly revised. To fish, 1/10 of fish names are changed approximately every 10 years (Pauly & Froese, 2000) (Shao et al., 2007a). Even the taxonomists will not be able to track and remember the changes and can only rely on automatic database comparison to detect them. When a user enters an invalid name of a species, the system automatically finds the valid name and links to the right webpages. Otherwise, much of the valuable specimen information in herbaria and museums will be lost if users only know synonyms of a species.

## 4 ESTABLISHING TAIWAN ENCYCLOPEDIA OF LIFE WITH COMMUNITY ENGAGEMENT

Advocated by biologist E. O. Wilson, the Encyclopedia of Life project began in 2007. Its goal is, through the joint efforts of scientists around the world, to gather and share scientific knowledge about the 1.9 million known organisms in a single online resource. The TaiEOL team started to communicate and exchange information technology with the EOL team in 2009. In 2011, the team began to work on Taiwan Encyclopedia of Life (TaiEOL; in Chinese) and hope to complete in three years the online information content of 20,000 species, including 8,000+ endemic species.

### 4.1 Status and analysis of species data in traditional Chinese characters

TaiEOL from the start utilizes its own Biodiversity Bibliography System to record the species with Chinese descriptions from the biological publications in Taiwan. There are now more than 400 books written in traditional Chinese and a cumulative total of 45,000 scientific names in the system. Using Insecta as an example, there are 21,000 recorded insect species in Taiwan; yet only 4,700 species have Chinese descriptions. Furthermore, only 1,600 out of 6,000 endemic insects have information written in Chinese. These statistics help with the decision on which taxon to invite researchers to fill in with information. They also provide users with the bibliographies of a species when they retrieve information on that species.

### 4.2 Open source software platform of TaiEOL portal

The information collaboration and user-participation concept of Web 2.0 has become an international trend in sharing and integrating biodiversity data. LifeDesks, the modules developed by the EOL project, and Scratchpads, provided by the European Distributed Institute of Taxonomy (EDIT), are two of the most popular participatory platforms. Based on these two types of platforms and using open source software, TaiEOL began to develop its Chinese version with species information as its core.

The TaiEOL portal will build on the existing species checklist of Taiwan. For the 5.4 million species, it will, in staggered phases, invite the taxonomists and citizen scientists to join the effort and contribute to its content. Taiwan's endemic species especially need descriptions, images, and popular science material so that the public can easily learn about them and become knowledgeable in Taiwan's biodiversity. Moreover, TaiEOL can provide the information on endemic species to EOL to be shared by the global community. Through the portal's biological content management system and user interaction mechanisms, the biodiversity information of Taiwan can be directly and effectively shared and exchanged.

## 5 INTEGRATING DATABASES AND LINKING GLOBALLY THROUGH TaiBIF

TaiBIF, the Taiwan node of GBIF, uses the metadata format and tools recommended by GBIF to integrate Taiwan's biodiversity information and exchange them with global community. The homepage of TaiBIF is shown in Figure 1. TaiBIF consolidates various databases in Taiwan, including species checklists, *Biota Taiwanica*, species occurrence data, Taiwan Encyclopedia of Life, etc. Using biodiversity informatics and tools, TaiBIF enables the general public, academic scholars, and government agencies to gain access to the data on its platform. In addition, TaiBIF introduces international data standards and protocols so that it can meet the two objectives of "deepening Taiwan people's biodiversity knowledge" and "helping to make the world's biodiversity picture more complete."



Figure 2. Homepage of TaiBIF

## **5.1 Information currently available online at TaiBIF portal (Shao et al., 2009)**

### **5.1.1 Species checklist:**

The Catalogue of Life in Taiwan project (TaiBNET) offers species names (scientific names and Chinese names) and classification hierarchy, and keeps track of taxonomic work. Being the backbone of the project, these data provide means to access all the biodiversity information.

### **5.1.2 Primary species occurrence data (specimen and observational data):**

There are specimen data from TELDAP and observational data from various collaborating domestic databases. A total of 1.59 million digital specimen and observational records are now available. The data come from institutions such as Academia Sinica, National Museum of Natural Science, National Taiwan University, National Taiwan Museum, Taiwan Forestry Research Institute, Fisheries Research Institute, Agricultural Research Institute, Taiwan Endemic Species Research Institute, National Museum of Marine Biology and Aquarium, National Taiwan Ocean University, National Tsing Hua University, National Sun Yat-sen University, etc.

### **5.1.3 *Biota Taiwanica*:**

NSC's Division of Life Sciences, in order to fulfill the requirement of the Biodiversity Promotion Plan, assists domestic taxonomists in composing contents for *Biota Taiwanica*. So far, information on more than 10,000 species has been uploaded online and released to the public.

### **5.1.4 Literature:**

Relevant research papers on biodiversity are provided by the Science and Technology Policy Research and Information Center of the non-profit National Applied Research Laboratories. Currently, there are research projects (8,079 articles), research reports (3,987 articles), periodical papers (10,611 articles), conference papers (4,767 articles), Master's theses and Doctoral dissertation (4,171 articles), English journal papers (2,270 articles), and publications (499 articles).

### **5.1.5 Eco-Photo**

Using Cooliris software, TaiBIF presents ecological photos and videos it collected. It will include the images gathered by TaiEOL in the future.

### **5.1.6 Species description**

The text and images of native Taiwan species as well as popular science material from TaiEOL will be incorporated into TaiBIF in the future.

## **5.2 Web service and tool development**

After years of research and development effort, the TaiBIF team has made it available many biodiversity information tools to accelerate the integration and sharing of biodiversity data. For example, there are tools to check scientific names and geological coordinates:

- i. The scientific names are matched against TaiBNET and Sp2000, and any errors in spelling are highlighted and reported back to the users for reference. Additional information on the taxon level,



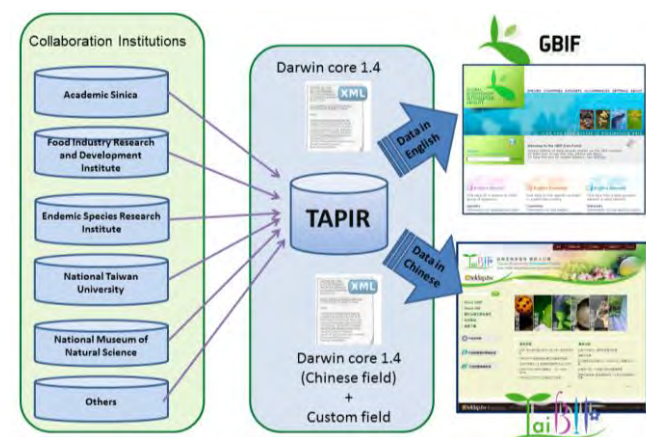
author, and publication year of the species are provided in the correct sequence so that users can easily standardize the format and increase the accuracy of data.

- ii. TaiBIF converts geological coordinates to the format commonly used in Taiwan so that users can update old ecological distribution data and conduct further analysis and research work. TaiBIF also verifies the coordinates. Its tools will check spatial data for errors; e.g. based on the description of a location and latitude and longitude, the tools can determine if the coordinates fall outside the range. TaiBIF will pass the information back to the users (data providers) to make necessary changes. In doing so, TaiBIF increases the quality of the species occurrence data and accomplish the goal of integrating and accessing biodiversity data through geographic information.

## 6 INFORMATION TECHNOLOGY FOLLOWING INTERNATIONAL TREND

In terms of information technology, GBIF uses DiGIR, BioCASE, and TAPIR with Darwin Core as its metadata standard to integrate global species occurrence data (Hill et al., 2009; Chapman, 2005), and has accumulated 312 million occurrence data to date. TaiBIF in 2004 started to use DiGIR (Distributed Generic Information Retrieval) as the protocol to exchange data with GBIF Secretariat. Due to the limitations of certain fields, CRIA (Centro de Referência em Informação Ambiental) developed TAPIR (TDWG Access Protocol for Information Retrieval) to try to solve the problem. TAPIR is a REST style distributed data exchange protocol transmitted by HTTP. It combines and reinforces the strengths of BioCASE and DiGIR and allows users to communicate with data providers in a simple and direct way. It also provides more choices in data distribution standard, including different versions of Darwin Core and the capacity to define data by custom XML, making it more flexible to share data. TAPIR was officially adopted by GBIF and became one of the official TDWG standards since October 2009.

TAPIR is chosen by the TaiBIF team to be its data exchange protocol, integrating data (in English) into the original information architecture. However, the majority of Taiwan's species occurrence data also contains Chinese characters. TaiBIF, therefore, made use of TAPIR's customization feature and created Chinese-language XML extension. Consequently, not only the English data can be shared with GBIF, the Chinese data can also be shared within the original Darwin Core framework. Furthermore, a retrieval platform was established at TaiBIF (Fig. 3). This framework constantly assists with system deployment at the biodiversity-related institutions in Taiwan. Various institutions and museums have now accumulated over 1.59 million species occurrence data.



**Figure 3.** Flow of data to TaiBIF (Chinese) and GBIF (English) using TAPIR

## 7 DIFFICULTIES AND SOLUTIONS OF BIODIVERSITY DATA INTEGRATION

Comparing to specimen, literature, and species checklist, the collection and integration of the observational raw data are much more difficult. It is due to the fact that, once a species is identified and its distribution is recorded by taxonomists or ecologists, the information can be analyzed and written into papers if the information is posted online. On the contrary, specimens can only be borrowed with the agreement of their original collectors or

managers. As a result, most of the researchers perhaps are willing to provide analyzed charts or tables, but hesitate to disclose and share data before they have a chance to publish their papers. It is fairly common, and a regrettable happening, for research data to get lost, destroyed, or buried somewhere (Shao et al., 2007c).

The reluctance to submit raw data is only one of the reasons why, in the past 20 years, there is not much progress in the information integration process among or cross government agencies in Taiwan. Other reasons are: the complicated issues of IPR, the lack of information collection and management unit in the agencies, the absence of clear, executable policies for information, etc. Hence, entrusted by National Science Council, Academia Sinica in 2008 set up the National Committee for GBIF (GBIF-ROC), with committee members being either the representatives from biodiversity-related agencies or researchers who are in charge of databases. The Committee began to establish data exchange standard, study and develop viable information policies, and request all agencies to include a term on the contracts demanding all their commissioned projects to submit raw data when the projects end. After a certain period of time, agreed upon beforehand, the data will be open to the public.

## **7.1 Principles for submitting ecological distribution data**

After several discussions, GBIF-ROC reached a conclusion on July 20, 2009 entitled “The principles of government-funded ecological distribution data collection and archiving, promoted by GBIF-ROC Committee” to be considered by the Sustainable Development Research Committee and the Biodiversity Promotion Plan.

- i. Type and scope of the submitted data: In the first stage, only the digital raw data of the ecological distribution (species occurrence data) need to be submitted. Other related data will be added later.
- ii. Collection and submission of survey and monitor data: All government agencies are requested to include a term on the contracts demanding all their commissioned projects to submit raw data when the projects end.
- iii. Format, methodology, and evaluation of data submitted: (i) Submitted information should include survey or monitor methodology, definitions, and data. The format should use the international customary standards such as Darwin Core, Ecological Metadata Language (EML) (Jones et al., 2006), and ISO19115. (ii) Each agency should designate its data collection (storage) unit based on its technology capability. (iii) Each agency should define an evaluation system; e.g. further funding is provided only after data are submitted.
- iv. When the data are made public (IPR issue): Each agency should determine when to make the data accessible by the public.

## **7.2 Format of ecological distribution data submitted**

Consensus has been reached about the format of the information which the commissioned projects are required to submit. There are 12 items of metadata and one item of raw data.

- i. Title of the project;
- ii. Data Owner information, at least including name, organization, position title, and contact information;
- iii. Associated Parties information when applicable, at least including name, organization, position title, and contact information;
- iv. Research Summary (Abstract);
- v. Keywords;
- vi. Usage Rights;
- vii. Contact Window information, a least including name, organization, position title, and contact information;
- viii. Content Description, including Materials and Methods, Temporal Coverage, Geographic Coverage, and Taxonomic Coverage;
- ix. Variable Names;
- x. Variable Labels;
- xi. Variable Definition;
- xii. Variable Measurement Definition, including detailed definition of measurement categories. If it is Nominal or Ordinal, it should include description and definition of the values. If it is Interval or Ratio,

- it should include Unit, Precision, and Number Type. If it is Date-Time, it should include Format and Precision.
- xiii. Raw data of items ix-xii mentioned above.

### 7.3 Future challenges and strategies

In order to achieve its goals of collecting and integrating biodiversity data, TaiBIF has officially proposed to higher-level governmental decision makers such as the National Science and Technology Conference and the National Council for Sustainable Development to adopt a top-down approach. However, the database integration work in many agencies still faces some common problems and challenges which are listed below. The difficulties can only be overcome with the support and determination of the managers in each agency. The managers need to commit more manpower and material resources to the work, and also bestow recognition and encouragement on their employees.

- i. Without professional IT personnel to serve on the staff, the researchers are forced to outsource system development since they usually don't have the time to manage or maintain the databases on their own. When open source software is not used, it is difficult to maintain a database sustainably. The database typically ceases to exist when a project ends.
- ii. Database management in general is regular work. Consequently, its funding is often cut back year after year.
- iii. The accomplishment of the researchers is not recognized. The evaluation of a researcher's performance is based exclusively on the published papers (SCI scores). Hence, most researchers are reluctant to spend their time on databases or share their data; making it difficult to conduct further work on the data. Recently GBIF has started to promote the publication of "data papers" in scientific or SCI journals; a good incentive strategy for scientists to publish their data and give public open access to them.

Other methods and strategies promoted by TaiBIF are as follows:

- i. Cooperating with other websites to increase page views through mutual links.
- ii. Sharing budget and accomplishment. Data integrators need to establish credibility by distributing funds equitably, making their own data public, and attributing achievement to team members and data providers.
- iii. Meeting the needs of users. The convenience of users and data providers should take precedence over that of information technology or data integration.
- iv. In addition to the number of SCI (Science Citation Index) papers and their impact factors, digitized material such as the number of records (including specimen collected, ecological distribution data, and DNA sequences) archived and uploaded to the Internet, the so-called "Repository Impact Factor," should be taken into account when assessing research performance. Data compilation and submission in the form of "data papers" to academic journals should also be encouraged. This mechanism facilitates the publishing of biodiversity data resources (Chavan et al., 2011).
- v. Implementing performance evaluation effectively in order to receive data effortlessly.
- vi. Nurturing talents. Provide opportunities for young scholars to organize conferences or attend international meetings so that they can learn and exchange ideas.
- vii. Advocating the benefits of data integration
  - ✓ Offer offsite backup (data stored in different places)
  - ✓ Help with data validation to improve data quality
  - ✓ Contribute to the society; academic services; the taxpayers' rights
  - ✓ Increase page views of the data
  - ✓ Assist government in making sure it receives concrete results from the huge investment of funds it allocates to scientific researches and surveys. Raw data are archived and preserved so that when a project ends, these data, rather than summarized reports, are available to the public.
  - ✓ Be able to reanalyze and re-simulate models with the data, using newer statistical software such as cluster analysis, ecological model, etc.
  - ✓ Can be used as an important tool, e.g. a quantitative indicator of biodiversity, for resource conservation, sustainable use, and management policies.

## 8 CONCLUSION

In order to meet the tasks stipulated by the Convention on Biological Diversity, it is elemental to integrate biodiversity data and make it accessible to the public. Only through sufficient information sharing can the multi-dimension objective of biodiversity conservation, research, education, and resource sustainability be achieved. The promotion of data integration involves establishing work flow and overcoming difficulties occurred in the process. The difficulties cover many aspects such as intellectual property rights, data policy, data standards, data quality concept, standard operating procedure of collecting data, and software/hardware development technologies. In this paper, we present the results of the ten-year experience of Biodiversity Research Center, Academia Sinica in promoting biodiversity data integration. Future work will consider the addition of environmental and hydrological conditions to analyze in order to biodiversity sustainable development or policymaking.

## 9 REFERENCES

- Chapman A.D. (2005) *Uses of Primary Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. GBIF, 1–117.
- Hill A.W., Guralnick R., Flemons P., Beaman R., Wieczorek J., Ranipeta A., et al. (2009) Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics*, 10, S3.
- Jones M.B., Schildhauer M.P. & Reichman O.J. (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution and Systematics*, 37, 519–544.
- Pauly D., & Froese R. (2000) *FishBase 2000: concepts, design and data sources*. ICLARM, Los Banos, Laguna, Philippines, 1–344.
- Shao K.T., Lin Y.C. & Lin H.H. (2007a) Linking the Taiwan fish database to the global database. *Data Science Journal*, 6, 164–171.
- Shao K.T., Peng C.I., Yen E, Lai K.C. , Wang M.C., Lin Y.C. , Lee H, Yang A & Chen X.Y. (2007b) Integration of biodiversity database in Taiwan and linkage to global databases. *Data Science Journal*, 6, 2–10.
- Shao K.T., Lai K.C., Lin Y.C. & Lee H (2007c) Progress and impediment of the integration of biodiversity database. *Handbook of Open & Free: New Enterprise in the Information Age — An International Workshop*, 17–18, Institute of Information Science, Academia Sinica, Taipei.
- Shao K.T., Huang S.C. , Chen S , Lin Y.C. , Lai K.C. , Ko C.J. , Chen L.S. & Yang J.L. (2008a) Establishing a Taiwan biodiversity information network and its integration with germplasm databanks. TARI, COA. *APEC-ATCWG Workshop on Capacity Building for Risk Management Systems on Genetic Resources*, 71–78. Hsin-Hua, Taiwan.
- Shao K.T., Peng C.I. & Wu W.J. (2008b) *2008 Taiwan Species Diversity I. Research and Status*, 1–373, Taipei: Forest Bureau, Council of Agriculture and Biodiversity Research Center, Academia Sinica. (in Chinese)
- Shao K.T., Peng C.I. & Wu W.J. (2008c) *2008 Taiwan Species Diversity II. Species Checklist*, 1–796, Taipei: Forest Bureau, Council of Agriculture and Biodiversity Research Center, Academia Sinica. (in Chinese)
- Shao K.T., Peng C.I. & Wu W.J. (2010) *Taiwan Species Checklist 2010*, 1–840, Taipei: Forest Bureau, Council of Agriculture and Biodiversity Research Center, Academia Sinica. (in Chinese)
- Chavan V. & Penev, L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science, *BMC Bioinformatics*, 12 (Suppl 15):S2

# DEVELOPMENT OF GLOBAL SOIL INFORMATION FACILITIES

*N H Batjes<sup>1</sup>\*, H I Reuter, P Tempel, T Hengl, J G B Leenaars, and P S Bindraban*

*\*<sup>1</sup>ISRIC - World Soil Information, P.O. Box 353, 6700 AJ Wageningen, The Netherlands*

*Email: [niels.batjes@wur.nl](mailto:niels.batjes@wur.nl)*

## ABSTRACT

*ISRIC - World Soil Information has a mandate to serve the international community as custodian of global soil information and to increase awareness and understanding of the role soils in major global issues. To adapt to the current demand for soil information, ISRIC is updating its enterprise data management system, including procedures for registering acquired data such as lineage, versioning, quality assessment and control. Data can be submitted, queried and analysed using a growing range of web-based services—ultimately aiming at full and open exchange of data, metadata and products—through the ICSU-accredited World Data Centre for Soils.*

**Keywords:** Soils, Data curation, Analysis services, Web services, ICSU World Data System

## 1 INTRODUCTION

Human activities are a major driver of current changes in the Earth's atmosphere, oceans and landscapes (Watson *et al.*, 2000; UNEP, 2007) and increasingly affecting key environmental, social and economic functions of soil (Blum, 2005). Such changes need to be measured, monitored and modelled to allow governments, civil society and other sectors to take informed decisions about climate change mitigation and adaptation, world-food security, production of biofuels, combating environmental degradation, and other global issues. In this context, ISRIC – World Soil Information is strengthening its capabilities in collecting, storing, processing and disseminating global soil and terrain information for research and development of sustainable and productive land use (Bindraban *et al.*, 2010).

ISRIC has a mandate to serve the international community as custodian of global soil information and to increase awareness and understanding of the role of soils in major global issues. It was created in 1966, following a request by the 1964 UNESCO General Council, and obtained the status of *ICSU World Data Centre* in 1989, initially named *WDC for Soil Geography and Classification*, later *WDC for Soils*. In August 2011, ISRIC was accredited as regular member of the new *ICSU World Data System (WDS)*. In that capacity, it deals directly with data curation and data analysis services.

This paper discusses recent developments at ISRIC, focusing on the on-going development and implementation of distributed global soil information facilities to ultimately provide quality-assured soil information for better science and development.

## 2 DATA CURATION

These days, new soil data are collected less and less while the same old (legacy) data and derived information are repeatedly used. Hence, it is vital to secure, maintain and expand data that support current information, including soil reference collections, soil maps and reports, and this ideally for public, free access (see 3.1).

Various collections are being maintained and regularly expanded by ISRIC. Since its establishment, a primary responsibility has been to create and maintain a world soil reference collection. The collection comprises over 1000 soil monoliths sampled to be representative of the soil units of the *FAO-UNESCO Soil Map of the World (FAO, 1971-1981; FAO et al., 2009)*, forming a unique educational, cultural and scientific resource. A selection of the monoliths is on display in the World Soil Museum in Wageningen, and the whole collection (e.g., monoliths, reference samples, soil morphological and analytical data, thin sections, imagery) has been documented in a relational database, queryable on-line. This information is supported by a large collection of, often unique, country reports, books and maps that originate largely from third world countries. The map collection contains mainly small-scale (1:250,000 or smaller) maps, part of which have been digitised in the framework of the *EuDASM* project (Panagos *et al.*, 2011). So far, some 25% out of some 15,400 of the historic books and country reports has been scanned, while this is some 72% (out of ~ 7600) for the map collection

(<http://www.isric.org/services/world-soil-library-and-maps>). This wealth of soil geographic and attribute information has been used to develop a range of broad scale, GIS databases (e.g., Batjes, 2009; FAO et al., 2009; Nachtergaele *et al.*, 2011) that were generally distributed on-line as stand-alone, compressed files both through the ISRIC website and Global Change Master Directory (GCMD). The information is being used for a wide range of assessments, including agro-ecological zoning, global crop production, soil vulnerability to pollution, soil carbon stocks and changes, and soil gaseous emission potentials (e.g., Milne *et al.*, 2007; Gibbs *et al.*, 2008; Fischer *et al.*, 2010; Mekonnen & Hoekstra, 2011; Wang *et al.*, 2011).

### 3 WEB-BASED INFORMATION FACILITIES

#### 3.1 Aims

To better manage the current demand for soil information, ISRIC is in the process of implementing an enterprise data management system, including procedures for registering acquired data using versioning, quality assessment and control. The centralized enterprise database will contain quality-controlled and authorized data with defined and registered accuracy and quality (e.g., documented data lineage; detailed metadata). Subsequent to international peer-review and adjustments, the spatially-enabled RDMS will go into production and be accessible through different facilities (e.g., Data Entry, MetaDataService, WMS, and WPS).

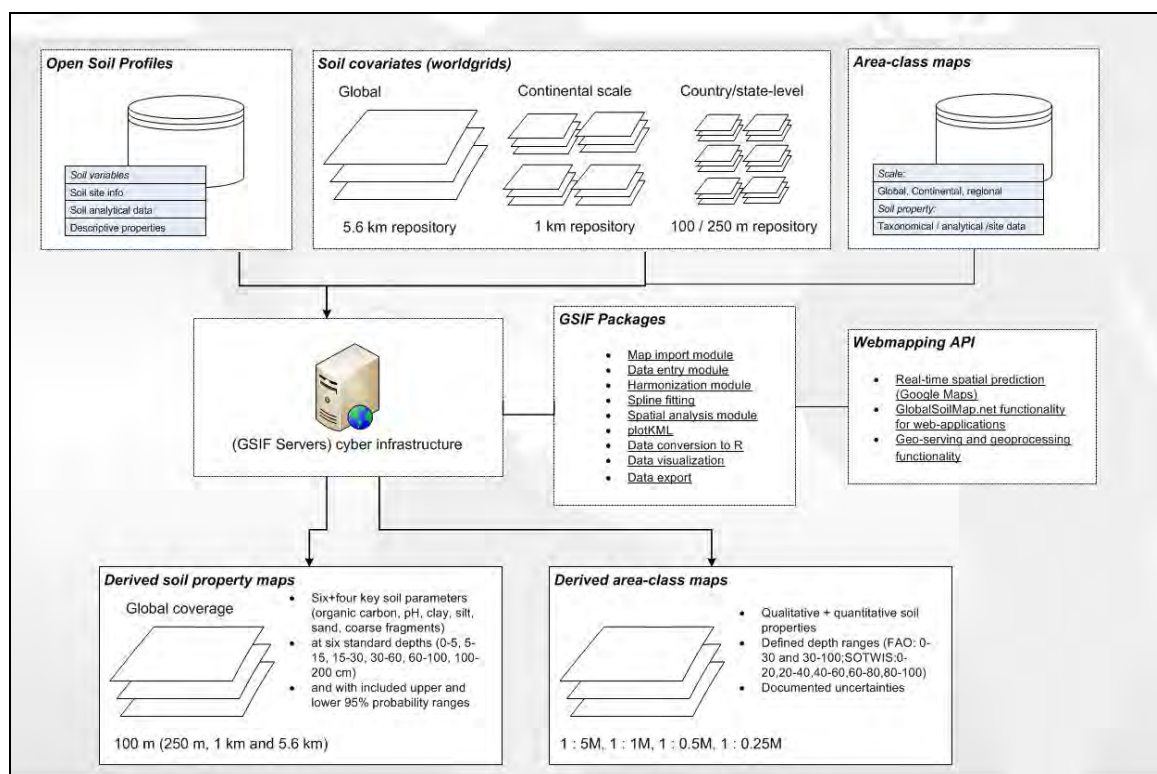
ISRIC's data policy follows the same principles for data sharing as those of ICSU-WDS, aiming at "full and open exchange of data, metadata and products shared within WDS, recognizing relevant international instruments and national policies and legislation" (see Uhlir *et al.*, 2009). When registering, the "contributor should present any existing restriction that is applicable to the exchange of the data, metadata and products submitted" (GEO, 2010).

#### 3.2 Main components

Since June 2011, ISRIC's website provides access to a range of new facilities that are in various stages of development:

- ISRIC enterprise database: All soil-related data managed or maintained by ISRIC will be made available online from one central database environment in one uniform format (Tempel & van Kraalingen, 2011). The spatially-enabled relational database management system (RDBMS) will be used to hold and manipulate a variety of data types, scales and sources. Version 1.0 is being populated using a range of disparate, stand-alone soil databases developed by ISRIC and its partners (e.g., Batjes, 2009; Van Engelen, 2011); for the future, other data(sets) will be added using both on-line and off-line facilities for data entry (<http://www.isric.org/data/wosis>). In conjunction with this, a range of web-based tools is being developed to extract user-defined data from the system using agreed data-exchange formats (e.g., SoilML, GeoSciML). The system is implemented using open standards (e.g., OGC, PostGresQL).
- Vocabulary service: This multilingual, soil-environmental thesaurus is being developed as a reference indexing and retrieval tool. It will provide ontologies of interest in the Soil Science domain and was created at the request of the Working Group on Soil Information Standardisation of the International Union of Soil Sciences (IUSS). Currently, the service is based on selected databases (FAO - Agrovoc; EEA - Gemet). Over the next 2-4 year, the service will be developed further in collaboration with colleagues from Europe (GS-Soil project), Australia (CSIRO SEEGrid) and similar projects around the world. Currently, it uses SEEGrid Vocabulary Service implementation as a frontend while editing is based on FAO's VocBench (<http://aims.fao.org/tools/vocbench-2>), which is open source.
- Metadata service (operational): Implemented as a multilingual catalogue application to enter, search and retrieve metadata records and manage spatially referenced resources. Currently, the service harvests soil-related metadata from various sources worldwide, including ISRIC's Library, FAO, and the Consultative Group on International Agricultural Research (CGIAR). The service is based on *GeoNetwork open source* (<http://geonetwork-opensource.org/>) and follows the principles of Free and Open Source Software (FOSS) and International and Open Standards for services and protocols (e.g., ISO/TC211 and OGC, Open Geospatial Consortium); see <http://www.isric.org/data/metadata-service>.
- OneSoil Map Viewer: Provides visualisations of qualitative and quantitative soil properties worldwide. Maps are accessed from various data sources and providers in a distributed way around the world. The viewer will create a dynamic soil map for different soil properties with standardized legends and data exchange formats (<http://www.isric.org/data/mapviewer>).
- Web mapping service (WMS): This service provides OGC-conform maps for different layers and allows users to visualize datasets hosted at the WDC for Soils (<http://www.isric.org/data/web-map-service>).

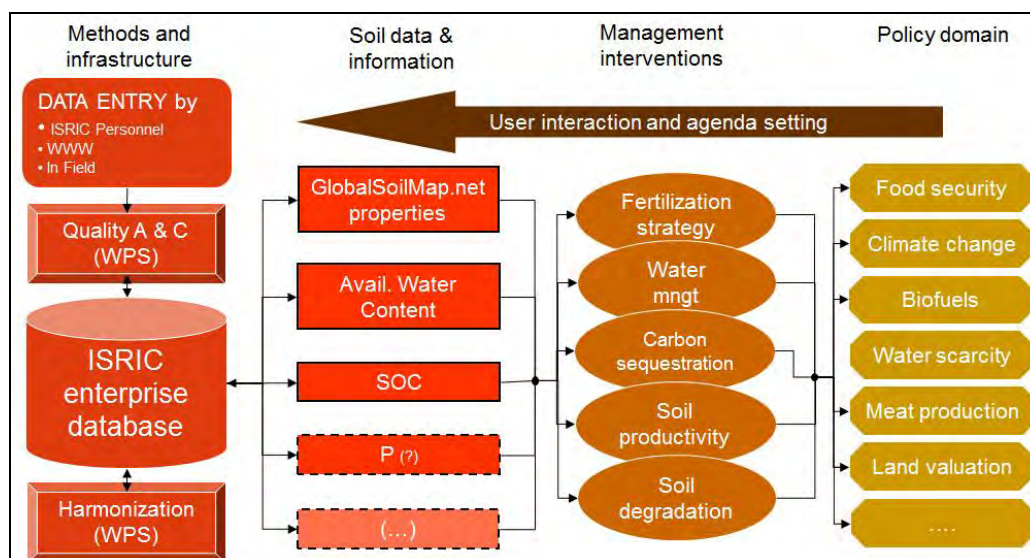
- Web processing service (WPS): Capability to compute soil functions or parameters of interest in the Soil Science domain. Four different services are currently available: a) interpolation over depth of quantitative soil profile properties using spline function; b) extraction of point information of soil information or auxiliary data sources for a given location on the planet, and soil information or auxiliary data sources for a given area on the earth specified by the corner coordinates; c) overlay of two raster data sources to provide statistical information for a predefined zone-grid (e.g., watershed, countries) for soil information or auxiliary data sources. The WPS delivers OGC conforming XML files (<http://www.isric.org/data/wprocessservice>).
- Worldgrids: Provides a repository of gridded predictors with global or at least partial global coverage. As WDC- Soils, ISRIC facilitates the collation and use of multi-thematic gridded repositories of Digital Soil Mapping covariates. Worldgrids already contains over 100 grid maps at a 5.6 km resolution and several 1 km resolution maps. Functionalities include querying, extraction, and creation of an overlay of a user-specified zone grid and a defined covariate map (see: <http://www.isric.org/data/worldgrids>).
- Global Soil Information Facility: GSIF will provide an overarching methodological framework for the production of consistent global soil information products of various types and scales/resolutions (Hengl *et al.*, 2011). Key principles for the design of GSIF, for production of open soil information, are outlined elsewhere (<http://www.isric.org/projects/global-soil-information-facilities-gsif>). Major components, as schematised in Figure 1, are expected to be fully operational by mid-2012. Where appropriate, algorithms and model states will be documented and preserved – recommended procedures are being developed and tested to align with the frameworks of on-going, large international projects such as *GlobalSoilMap.net* (Sanchez *et al.*, 2009) and *e-SOTER* (Van Engelen, 2011).



**Figure 1.** Proposed components of ISRIC's Global Soil Information Facilities and their relations

### 3.3 Applications

Upon their completion and testing, the ISRIC data and web facilities may be used freely —subject to any existing restrictions, see 3.1— to increase awareness and understanding of soils in major global issues; this is illustrated in Figure 2 under “policy domain”. Area-class and soil-property maps derived from the evolving web-processing services will also serve the global soil observing system as part of the *Global Earth Observing System of Systems (GEOSS)* and support collaborative activities with partner organisations such as the Food and Agricultural Organization (UN-FAO) and the Joint Research Centre (EU-JRC) in the broader framework of the *Global Soil Partnership* ([http://www.fao.org/nr/water/landandwater\\_gsp.html](http://www.fao.org/nr/water/landandwater_gsp.html)).



**Figure 2.** Management and use of soil information in support of research and informed decision making

## 4 CONCLUSIONS

The importance of judicious soil use and management in addressing major global issues, such as climate change, food security, energy, water conservation, and preserving biodiversity is increasingly recognised. As a result, there is a growing demand for quality-assessed soil data and information, of various types and scales, by other scientific disciplines, practitioners, policy and society. ISRIC is developing, testing and implementing a new enterprise data management system and a range of WMS and WPS services using international standards. Typically, new procedures and approaches are developed within the framework of on-going large international projects. Within the new ICSU World Data System, ISRIC will continue to ensure the long-term stewardship and provision of quality-assessed data and supporting web facilities to the international science community and other stakeholders.

## 5 REFERENCES

- Batjes, N.H. (2009) Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. *Soil Use and Management* 25, 124-127
- Bindraban, P.S., Batjes, N.H., Leenaars, J.G.B. & Bai, Z. (2010) Relevance of soil and terrain information in studies of major global issues. In: *Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World*. eds R. J. Gilkes & N. Prakongkep). International Union of Soil Sciences, 1-6 August, Brisbane, Australia, pp. 38-41.
- Blum, W. (2005) Functions of soil for society and the environment. *Reviews in Environmental Science and Biotechnology* 4, 75-79.
- FAO (1971-1981) *FAO-Unesco Soil Map of the World, 1:5,000,000 (Vol. 1 to 10)*. United Nations Educational, Scientific, and Cultural Organization, Paris.
- FAO, IIASA, ISRIC, ISSCAS & JRC (2009) Harmonized World Soil Database (version 1.1). Food and Agriculture Organization of the United Nations (FAO), International Institute for Applied Systems Analysis (IIASA), ISRIC - World Soil Information, Institute of Soil Science - Chinese Academy of Sciences (ISSCAS), Joint Research Centre of the European Commission (JRC), Laxenburg, Austria
- Fischer, G., Prieler, S., van Velthuisen, H., Lensink, S.M., Londo, M. & de Wit, M. (2010) Biofuel production potentials in Europe: Sustainable use of cultivated land and pastures. Part I: Land productivity potentials. *Biomass and Bioenergy* 34, 159-172.



- GEO (2010) GEOSS data sharing action plan. Group on Earth Observations, Geneva, pp. 8 ([http://www.earthobservations.org/documents/geo\\_vi/07\\_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf](http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf))
- Gibbs, H.K., Johnston, M., Foley, J.A., Holloway, T., Monfreda, C., Ramankutty, N. & Zaks, D. (2008) Carbon payback times for crop-based biofuel expansion in the tropics: the effects of changing yield and technology. *Environ. Res. Lett.* 3, doi: 10.1088/1748-9326/1083/1083/034001.
- Hengl, T., MacMillan, R.A., Walsh, M.G. & Reuter, H.I. (2011) Global Soil Information Facilities: A methodological framework for open soil information. ISRIC - World Soil Information, Wageningen, pp. 231 (*in the press*)
- Mekonnen, M.M. & Hoekstra, A.Y. (2011) The green, blue and grey water footprint of crops and derived crop products. *Hydrology and Earth System Sciences* 15, 1577-1600.
- Milne, E., Adamat, R.A., Batjes, N.H., Bernoux, M., Bhattacharyya, T., Cerri, C.C., Cerri, C.E.P., Coleman, K., Easter, M., Falloon, P., Feller, C., Gicheru, P., Kamoni, P., Killian, K., Pal, D.K., Paustian, K., Powlson, D.S., Rawajfih, Z., Sessay, M., Williams, S. & Wokabi, S. (2007) National and sub-national assessments of soil organic carbon stocks and changes: The GEFSOC modelling system. *Agriculture, Ecosystems & Environment* 112, 3-12.
- Nachtergaele, F.O., Van Engelen, V.W.P. & Batjes, N.H. (2011) Qualitative and quantitative aspects of world and regional soil databases and maps. In: *Handbook of Soil Sciences (2nd ed.)*. eds Pan Ming Huang, Yuncong Li & M. E. Sumner). Taylor and Francis Group, (*in the press*).
- Panagos, P., Jones, A., Bosco, C. & Kumar, P.S.S. (2011) European digital archive on soil maps (EuDASM): preserving important soil data for public free access. *International Journal of Digital Earth* 4, 434-443.
- Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A. & Zhang, G.-L. (2009) Digital Soil Map of the World. *Science* 325, 680-681.
- Tempel, P. & van Kraalingen, D. (2011) Global Soil Information Facility: Database component. ISRIC - World Soil Information, ISRIC Report 2011/03. Wageningen, pp. 260
- Uhlir, P.F., Chen, R.S., Gabrynowicz, J.I. & Janssen, K. (2009) Toward implementation of the Global Earth Observation System of Systems data sharing principles. *Data Science Journal* 8, GEO1-GEO91.
- UNEP (2007) *Global Environmental Outlook: Environment for development (GEO-4)*. United Nations Environment Programme, Nairobi.
- Van Engelen, V.W.P. (2011) Standardizing soil data (*e-SOTER* regional pilot platform as EU contribution to a Global Soil Information System). *International Innovation* June, 48-49.
- Wang, M., Li, Y., Ye, W., Bornman, J.F. & Yan, X. (2011) Effects of climate change on maize production, and potential adaptation measures: A case study in Jilin province, China. *Climate Research* 46, 223-242.
- Watson, R.T., Noble, I.R., Bolin, B., Ravindramath, N.H., Verardo, D.J. & Dokken, D.J. (2000) *Land Use, Land-Use Change, and Forestry (a special Report of the IPCC)*. Cambridge University Press, Cambridge.

# THE BRITISH GEOLOGICAL SURVEY'S NEW GEOMAGNETIC DATA WEB SERVICE

*E Dawson<sup>1\*</sup>, J Lowndes<sup>1</sup> and P Reddy<sup>2</sup>*

*<sup>1</sup>British Geological Survey, West Mains Road, Edinburgh EH3 9LA, United Kingdom*

*Email: ewan@bgs.ac.uk*

*<sup>2</sup>School of Computing, Robert Gordon University, Schoolhill, Aberdeen AB10 1FR, United Kingdom*

## ABSTRACT

*Increasing demand within the geomagnetism community for high quality real-time or near-real-time observatory data means there is a requirement for data producers to have a robust and scalable data processing infrastructure capable of delivering geomagnetic data products over the internet in a variety of formats. We describe a new software system, developed at BGS, which will allow access to our geomagnetic data products both within our organisation's intranet and over the internet. We demonstrate how the system is designed to afford easy access to the data from a wide range of software clients, and allows rapid development of software utilizing our observatory data.*

**Keywords:** Web Services, REST, Geomagnetism, Data Visualisation, Java, Restlet

## 1 INTRODUCTION

The British Geological Survey (BGS) operates a network of observatories that continuously measure the geomagnetic field at eight locations around the world. Each of our geomagnetic observatories has one or more GDAS (Geomagnetic Data Acquisition Systems) instruments that sample the field at a rate of 1Hz. These data are transmitted to our data processing centre in Edinburgh over the internet in near-real-time. The data are stored on our Storage Area Network (SAN) in fixed-width format plain-text files in a hierarchical file system.

These continuous 1Hz time-series data comprise our raw base data, but our primary data product is a time-series of one-minute mean values for each observatory. As well as our primary data product, we also generate a number of derivative data products for use by the public, academics and industry. In addition, we also create other derivative data products for the purpose of detecting environmental interference and other issues in the data as part of our quality control procedures. These derivative data products are also stored in text files on our SAN.

Definitive data from geomagnetic observatories are available from third-party archives such as INTERMAGNET<sup>1</sup> and the World Data Centres for Geomagnetism<sup>2</sup>. However, due to the post-processing involved in deriving definitive geomagnetic field values from the observatory data (see Macmillan, 2007 for details), these definitive data only become available many months after recording. Some preliminary or quasi-definitive data (as defined in Peltier & Chulliat, 2010) are available much sooner from the BGS public website, although this is quite limited in terms of the range of data available, and is not easily accessible in a standard format.

In 2010, we started to look into using web services technology in order to facilitate easier access to our data products for both external and internal users. The result, which we describe here, is a web service providing timely access to a wide range of BGS geomagnetic observatory data products.

## 2 WEB SERVICE REQUIREMENTS

<sup>1</sup> <http://www.intermagnet.org>

<sup>2</sup> For example, <http://www.wdc.bgs.ac.uk>

The web service will be used by both internal and external users who require access to our geomagnetic observatory data products. It will also be used within BGS to provide data for visualisation tools to assist in observatory quality control procedures, and to display observatory magnetograms on our public website. From these use-cases we derived a number of requirements that the web service must satisfy. These are summarised below.

The web service should:

1. Allow the user to retrieve a time-series of geomagnetic field mean values and derivative data products using criteria such as: *observatory*, *interval* (start and end date of time-series) and *cadence* (second, minute or hour).
2. Allow the user to specify the format in which the data should be returned. Supported formats should include:
  - XML - for easy integration with third-party tools
  - JSON - to allow easy integration with browser-based web applications
  - IAGA-2002 - a plain-text data exchange format commonly used in the geomagnetism community (see McLean, 2011)
3. Respond to common queries in less than one second; the service must perform fast enough to be used by observatory operations staff in interactive data processing applications.
4. Be easily accessible by human users and programmatic clients alike.
5. Allow for easy integration with a wide range of commonly used tools, such as web-browsers, MATLAB, R, Excel, etc.
6. Allow for users to be authenticated, and restrict access to certain datasets depending on authorisation.

### 3 WEB SERVICE DESIGN

#### 3.1 Data transfer protocol

There are a large number of protocols currently used in web services of one kind or another, each with its own pros, cons and common use cases. However, the only protocol which can be considered truly ubiquitous (and therefore meeting criteria requirement 5 above) is HTTP (Hypertext Transfer Protocol), the protocol which forms the basis of the World Wide Web. HTTP has well-defined semantics for retrieving and updating data over the internet (see Fielding *et al*, 1999 for details). It has client and server implementations available for almost all programming languages and operating systems, the most familiar examples being web browsers on the client side, and the Apache *httpd* web server on the server side.

#### 3.2 Identification and representation of data

The HTTP protocol also offers a convenient way for us to identify our datasets, in the form of URLs. The various parameters which together identify the dataset are encoded within the URL itself, according to a defined schema. For example, if the parameters are *observatory*, *start date*, *end date* and *cadence*, a suitable URL schema would be:

[http://geomag.bgs.ac.uk/web/service/obsdata/{observatory}/{cadence}/data?start={start\\_date}&end={end\\_date}](http://geomag.bgs.ac.uk/web/service/obsdata/{observatory}/{cadence}/data?start={start_date}&end={end_date})

Using this URL schema, the one-minute-mean data for the first hour of January 31<sup>st</sup>, 2012 from Lerwick observatory would be identified by and retrieved using the URL:

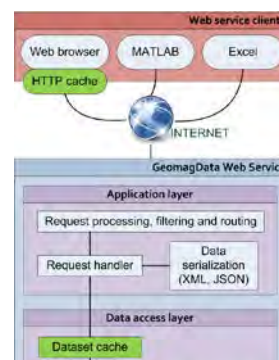
<http://geomag.bgs.ac.uk/web/service/obsdata/ler/minute/data?start=2012-01-31T00:00&end=2012-01-31T01:00>

According to our requirements, we must also provide a way for the user to specify the format in which the data should be returned. The HTTP protocol defines a number of headers which may be sent as part of an HTTP message. One of these is the 'Accept' header, which allows the client to specify the preferred data format (or 'media type', in the language of the HTTP specification). Thus the user can specify the format by setting the 'Accept' header to the appropriate value, such as 'application/xml', 'application/json', or 'application/x-iaga2002'. With some HTTP clients, for example web browsers, it is difficult to modify the HTTP

request headers. To allow users to specify the required data format using such clients, we also allow the format to be specified as an additional URL parameter, for example by appending `&format=xml` to the URL.

### 3.3 Authentication and Authorisation

Some datasets, for example the derived data products used in our quality control procedures, should only be accessible to observatory operations staff. In order to satisfy this requirement, we need a way to authenticate (identify) the user, and authorise each request according to our data access rules. There are a number of authentication protocols in use on the web, but the simplest and most widely supported are the HTTP-Basic and HTTP-Digest authentication protocols that are part of the HTTP specification (see Franks *et al*, 1999). Both protocols require the client to send a username and password along with each request, which the web service can then use to establish the identity of the user. Once authenticated, the web service can then decide whether to allow or to deny access to the requested resource based upon the data access rules of the organisation.



**Figure 1.** Architecture of the web service, showing the relationship between the major components. Some example clients are shown at the top.

## 4 IMPLEMENTATION

At BGS we already have the IT infrastructure and skills required to deploy web applications based on Java Servlet technology. Therefore when evaluating potential web service technologies we restricted ourselves to those that would be compatible with our existing infrastructure. Out of the various Java-based frameworks suitable for developing this kind of web service, we decided to use the *Restlet*<sup>3</sup> framework. *Restlet* is a popular open-source framework designed as an implementation of the “Representational State Transfer” (REST) style of information system architecture (for more details, see chapter 5 of Fielding, 2000). The design of our web service broadly conforms to the architectural constraints of REST<sup>4</sup>, so the *Restlet* framework is a good fit. *Restlet* is mature, well documented, and has an active user community - all important considerations when deciding on the adoption of an open-source technology. The main benefit of using a framework such as this is that many of the components required to build the web service are provided by the framework, significantly reducing development time.

The architecture of the web service comprises two main layers (see Figure 1). The data access layer sits between the SAN data store and the web application layer. At the bottom, the data access layer is responsible for the low level interaction with the file system. It provides an interface to the web application layer, allowing access to the data without other parts of the system having to know anything about how the data are stored. This de-coupling of the application logic and the data access logic promotes resilience to change in the system; if we want to change our data storage system, for example by switching from using plain text files to using a relational database, then we need only provide a new implementation of the data access layer. No changes need be made to any other part of the system, and the change of data store will be transparent to the application logic.

The application layer, which is implemented using components provided by *Restlet*, encapsulates all of the application-level logic, such authentication and authorisation, parsing of request parameters, content-negotiation (determining in which format to return the data) and preparing the response.

The web service is packaged as a standard Java WAR package and deployed in an Apache *Tomcat* web application container. The web service runs on a 64-bit server with an 8-core CPU running at 2.8GHz and 16GB of RAM. This hardware allows us to service a very large number of concurrent requests; the software will also run satisfactorily in a less demanding environment on a standard consumer laptop computer with only 2GB of RAM.

<sup>3</sup> <http://www.restlet.org>

<sup>4</sup> One of the constraints that define the REST architectural style is the ‘hypertext constraint’: that is, all state transitions (including locating and navigating between resources) must be afforded via hypertext links. In our case, since we use a fixed URL schema to allow users to construct their own URLs rather than having them follow links from the web service root, our web service breaks the hypertext constraint.

## 4.1 Improving performance

There are a number of steps involved in a typical request-response cycle between the client and the web service. The client sends a request to the web service, which is then processed by the application layer. The application layer requests the data required to fulfil the request from the data access layer. These data may be spread across a number of files, each of which must be read into memory and parsed by the data access layer. The parsed data are then passed back to the application layer, where they must be re-serialized into the format requested by the client. Finally, the response containing the serialized data is then sent back across the network to the client.

Each of these steps adds to the amount of time taken to process the request. In particular, the reading and parsing of data from the disk, and the transmission of the data across the network are the most time-consuming steps, with the time taken increasing approximately linearly with the amount of data requested. In order to meet our performance requirement, we have implemented a number of strategies to minimize the latency of the typical request-response cycle. These optimisations are discussed briefly in the following sections.

### 4.1.1 Data access layer disk cache

The most time-consuming part of servicing a request for data is in reading and parsing the data from disk. To mitigate this we added caching functionality to the data access layer (see Figure 1). The cache stores frequently-requested data in RAM, so that subsequent requests for the same data may be serviced without having to access the disk at all (assuming the data on disk hasn't been modified in the interim). The caching technology used is Ehcache<sup>5</sup>, a caching solution implemented entirely in Java. The cache is configured to use 8GB of the web server's RAM, which is enough to hold approximately 50,000 observatory daily time-series files. This form of caching decreases the time taken to respond to requests for commonly-accessed data by a factor of more than 100.

### 4.1.2 Compression of response

Even when data can be retrieved from the cache rather than disk, it must still be sent over the network to the client. The amount of data to be transferred can be quite large, especially since data formats such as XML are rather verbose. For example, three hours of one-second resolution data from one of our observatories translates to over 1MB of XML data.

We can reduce the amount of data to be transferred using a compressed encoding, such as GZIP. GZIP encoding is widely supported among HTTP clients, and is often handled completely transparently to the user. HTTP clients that can handle GZIP encoding indicate this capability to the web service by adding the 'accept-encoding: gzip' header to the request. The compressed response is typically 95% smaller, thus decreasing response transmission time by a factor of 20. Response compression incurs its own performance penalty though; it takes time for the server to compress the data, and for the client has to decompress it before it can be used. Although we did not measure the additional time taken for data compression/decompression, estimates based on user feedback indicate that response times decreased around ten-fold after compression was enabled on the server. This improvement was observed on a high-speed LAN; when accessing the web service over the Internet, the improvement observed should be greater still.

### 4.1.2 HTTP caching

The best optimisation is to avoid having to request the data in the first place. The HTTP protocol has a number of features which allow the web service to instruct the client as to which data may be cached (either locally, or by an intermediary server), and how to ask the web service if cached data is still fresh or should be replaced (see Fielding *et al*, 1999). HTTP caching is fully supported by the *Restlet* framework, and is widely supported in HTTP clients such as web browsers. This makes HTTP caching particularly useful in increasing the performance of interactive browser-based tools which make use of the web-service, such as the data visualization tool discussed later in Section 5.2.

<sup>5</sup> <http://ehcache.org>

By implementing disk caching, response compression, and HTTP caching, we have been able to increase the performance of the web service to meet the requirement that common requests are handled in less than a second. In fact, with these enhancements in place the majority of requests to the web service are handled in around a tenth of a second.

## 5 USING THE WEB SERVICE

### 5.1 Accessing the data via various HTTP-compatible clients

Because the web service operates using plain HTTP, any client capable of making an HTTP request can retrieve data using the web service. Since HTTP is one of the fundamental protocols of the internet, many modern software applications and tools are capable of retrieving data over HTTP. However in order to use the data, the application must not only be able to communicate with the web service, it must also understand the format in which the data are returned. Our web service exposes data in three formats: XML, which is one of the most commonly-used data exchange formats; JSON, which is rapidly becoming the de-facto data exchange format for web browser-based applications (see Crockford, 2006); and IAGA-2002, which is a standard data exchange format in the geomagnetism community (see McLean, 2011).

Examples of software packages which can easily consume data from our web-service are MATLAB, R, Mathematica and Microsoft Excel, all of which are capable of retrieving and parsing XML data from a URL.

In addition, almost all programming languages either have an HTTP client and XML parsing software included, or has them readily available as a third-party extension. This means it is easy to develop new software tools which make use of our latest geomagnetic observatory data in real-time.

### 5.2 Developing a browser-based data visualisation tool

In addition to providing external users with access to our data products, the web service also allows tools for internal use to be developed much more quickly than would be the case if the developer had to work directly with the data on the file system. The simple interface to the data provided by the web service frees the developer to concentrate on implementing the functionality of the tool.

We have developed a browser-based data visualization tool (see Figure 2), which uses the data provided by the web service to give a convenient overview of the data being recorded at each observatory and assist observatory operations staff in carrying out daily quality control checks on the data. The interface allows the user to browse plots of observatory data as well as various derived data products used to monitor the data quality. The tool was developed using the JavaScript language, and uses the *jQuery*<sup>6</sup> and *flot*<sup>7</sup> libraries for the user interface and magnetogram plotting, respectively.



**Figure 2.** Screenshot of the data visualisation web application, plotting geomagnetic observatory data obtained in real-time from the web service.

The simple HTTP interface provided by the web service, coupled with the fact that the data can be delivered in the JavaScript-native JSON format makes building this kind of interactive web application relatively straightforward. Furthermore, because the client is not tightly coupled to the server – the client depending only on the semantics of the HTTP protocol and the URL schema we defined – we were able to make extensive modifications to the web service during its development with no impact to the client.

<sup>6</sup> <http://jquery.com>

<sup>7</sup> <http://code.google.com/p/flot/>

## 6 SUMMARY AND FUTURE DEVELOPEMENTS

Our new geomagnetic data web service provides a number of benefits to users, both within and outside the organization, who wish to make use of our observatory data products:

- Ease of access: software clients need only know how to access a URL and parse the response; no knowledge of how the data are stored in the repository is required.
- Reduction of code duplication: low level data access code is isolated in the web-service software – client software need not duplicate this code. This leads to faster and more reliable software development.
- Increased resilience to change: since the low-level data access details are abstracted away by the web service, changes to the way the data are stored (location, storage format, structure) need only be reflected in a single place – the data access layer of the web service – while clients using the data are unaffected.
- Interoperability: the web service can provide data in a variety of standard formats, reducing the need for client-side format translation and making it easier to integrate BGS geomagnetism data with existing software and systems.

Currently, our geomagnetism data web service is only available to users within BGS. However, we will make the web service publicly available in the near future, giving academics and the public unprecedented access to BGS geomagnetic data products.

## 7 ACKNOWLEDGEMENTS

This paper is published with the permission of the Director of the British Geological Survey (Natural Environment Research Council).

## 8 REFERENCES

Crockford, D. (2006) RFC 4627: *The application/json Media Type for JavaScript Object Notation (JSON)*. Retrieved December 15, 2011 from the World Wide Web: <http://www.ietf.org/rfc/rfc4627.txt>

Fielding, R. T., Gettys, J., Mogul, J. C., Nielsen, H. F., Masinter, L., Leach, P. J. & Berners-Lee, T. (1999) *RFC 2616: Hypertext Transfer Protocol -- HTTP/1.1*. Retrieved December 15, 2011 from the World Wide Web: <http://tools.ietf.org/html/rfc261>

Fielding, R. T. (2000) *Architectural Styles and the Design of Network-based Software Architectures*, Doctoral dissertation, University of California, Irvine, USA

Franks, J., Hallam-Baker, P., Hostetler, J., Lawrence, S., Leach, P., Luotonen, A., Sink, E. & Stewart, L. (1999) *RFC 2617: HTTP Authentication: Basic and Digest Access Authentication*. Retrieved December 15, 2011 from the World Wide Web: <http://tools.ietf.org/html/rfc2617>

Macmillan, S. (2007) *Observatories : an overview*. In: Gubbins, D.; Herrero-Bervera, E., (eds.) *Encyclopedia of Geomagnetism and Paleomagnetism*. Netherlands, Springer, 708-711, 1054pp. (Encyclopedia of Earth Sciences).

McLean, S. (2011) *IAGA2002 Data Exchange Format*. Retrieved December 15, 2011 from the World Wide Web: <http://www.ngdc.noaa.gov/IAGA/vdat/iagaformat.html>

Peltier, A. & Chulliat, A. (2010) *On the feasibility of promptly producing quasi-definitive magnetic observatory data*. *Earth Planets Space*, 62(2):e5-e8, doi:10.5047/eps.2010.02.002

# DATA HANDLING WITHIN THE INTERNATIONAL VLBI SERVICE

*Dirk Behrend*

*NVI, Inc./ NASA Goddard Space Flight Center, Code 698.2, Greenbelt, MD 20771, USA  
Email: dirk.behrend@nasa.gov*

## ABSTRACT

*The International VLBI Service for Geodesy and Astrometry (IVS) is a globally operating service that coordinates and performs Very Long Baseline Interferometry (VLBI) activities through its constituent components. The VLBI activities are associated with the creation, provision, dissemination, and archiving of relevant VLBI data and products. The data and products are stored in dedicated IVS components called 'Data Centers.' The three Primary Data Centers provide identical data holdings. We give a brief overview of the organizational structure of the IVS and describe the general data flow among the various IVS components from preparing observational plans to creating the final products.*

**Keywords:** Very Long Baseline Interferometry, VLBI, International VLBI Service, IVS, space geodesy, data center, data, products

## 1 INTRODUCTION

Very Long Baseline Interferometry (VLBI) is one of the most accurate methods used to measure the Earth and its orientation in space. It is one of four space-geodetic techniques, the others being SLR (Satellite Laser Ranging), GNSS (Global Navigation Satellite Systems), and DORIS (Doppler Orbitography and Radiopositioning Integrated by Satellite), which are used to determine the celestial and terrestrial reference frames, the Earth orientation parameters (EOP), atmospheric parameters as well as other ancillary parameters. The EOP parameters are precession/nutation, Earth rotation (UT1), and polar motion. Each space-geodetic technique has its own strengths and unique capabilities. VLBI is unique in its ability to measure precession/nutation and UT1. VLBI employs large radio telescopes to observe compact radio sources, usually quasars, in order to estimate the vector between the telescopes. The VLBI observable is the difference in arrival time of a radio signal at two (or more) telescopes; hence, VLBI requires at least two radio telescopes to furnish useful observations. The VLBI technique dates back to the late 1960s; high-precision data, however, have been collected from the mid-1980s onward.

In this paper we concentrate on the data aspect of the VLBI technique. After a brief overview of the international organizational structure, we describe the general data flow among the various VLBI components and then take a closer look at the VLBI data repositories. The technical and scientific aspects of VLBI are covered elsewhere; the interested reader is referred, for instance, to Sovers, Fanselow, & Jacobs (1998) and Schuh & Behrend (2012) as well as references therein.

## 2 INTERNATIONAL VLBI SERVICE FOR GEODESY AND ASTROMETRY

Geodetic/astrometric VLBI activities on global and regional scales are organized through the International VLBI Service for Geodesy and Astrometry (IVS), which is an international collaboration of institutions that operate or support VLBI components. The IVS was established in 1999 as a service of the International Association of Geodesy (IAG), recognizing the need to move away from the *ad hoc* basis of the VLBI operational activities, which were mostly organized through national or bi-lateral agreements until then (see, e.g., Schlüter & Behrend, 2007). In 2000, IVS became a service of the International Astronomical Union (IAU) and the Federation of Astronomical and Geophysical Data Analysis Services (FAGS). Following the dissolution of the FAGS federation at the end of 2008, IVS applied for membership in the newly established World Data System (WDS) and was approved by the end of 2011. IVS interacts closely with the International Earth Rotation and Reference Systems Service (IERS), which is tasked by IAU and IUGG (International Union of Geodesy and Geophysics)

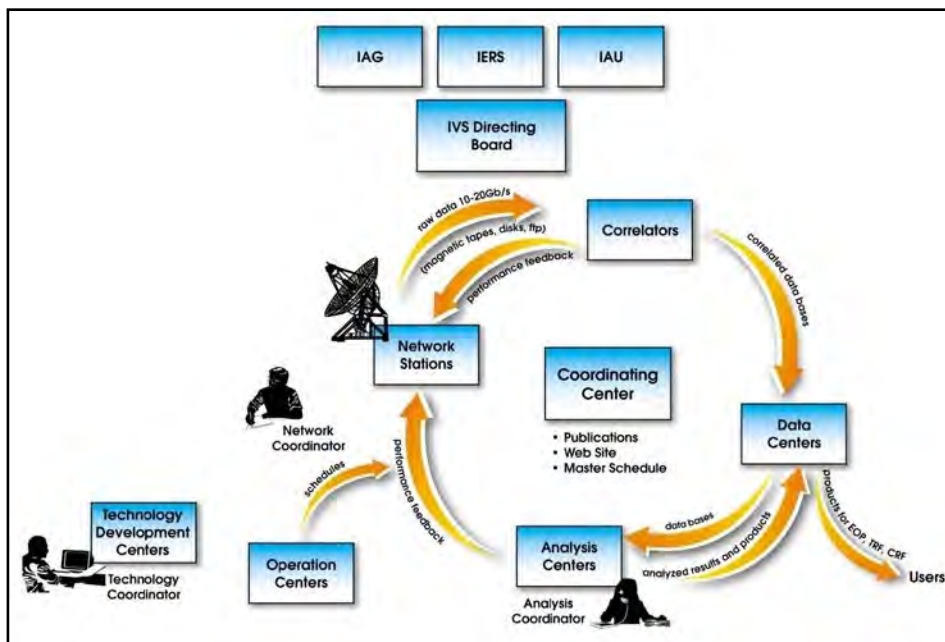


with maintaining the international celestial and terrestrial reference frames (ICRF and ITRF). The IVS currently consists of around 80 permanent components, which are supported by about 40 organizations in 20 countries. The IVS permanent component types are: Network Stations, Operation Centers, Correlators, Data Centers, Analysis Centers, Technology Development Centers, and the Coordinating Center (see Figure 1). Up-to-date information about the service and its activities can be found online under the URL <http://ivscc.gsfc.nasa.gov>.

The mission objectives of the IVS include the provision of support for geodetic, geophysical, and astrometric research and operational activities as well as the integration of VLBI into a global Earth observing system. To meet these objectives, IVS coordinates VLBI observing programs, sets performance standards for VLBI stations, establishes conventions for VLBI data formats and data products, issues recommendations for VLBI data analysis software, sets standards for VLBI analysis documentation, and institutes appropriate VLBI product delivery methods to ensure suitable product quality and timeliness. The VLBI products currently available include the five EOP parameters, the TRF, the CRF, and tropospheric parameters. All VLBI data and products are archived in IVS Data Centers and are publicly available. The IVS data set extends from 1979.

### 3 GENERAL DATA FLOW

The general flow of VLBI data within the IVS is centered around a “data feedback loop” as depicted in Figure 1. Raw VLBI data are recorded by the IVS Network Stations. The IVS observational network currently consists of 40–45 radio telescopes worldwide. Subsets of these telescopes participate in 24-hour observing sessions (8–10 stations) that are run several times per week and in 1-hour intensive sessions (2–3 stations) for UT1 determination every day. The individual observing networks are planned ahead for an entire calendar year by the Coordinating Center in the so-called Master Schedule. About one to two weeks prior to the actual observation date, an Operation Center prepares the individual recording schedule of the session and uploads it to a Data Center. This schedule is subsequently downloaded by the stations and the Correlator that processes the data.



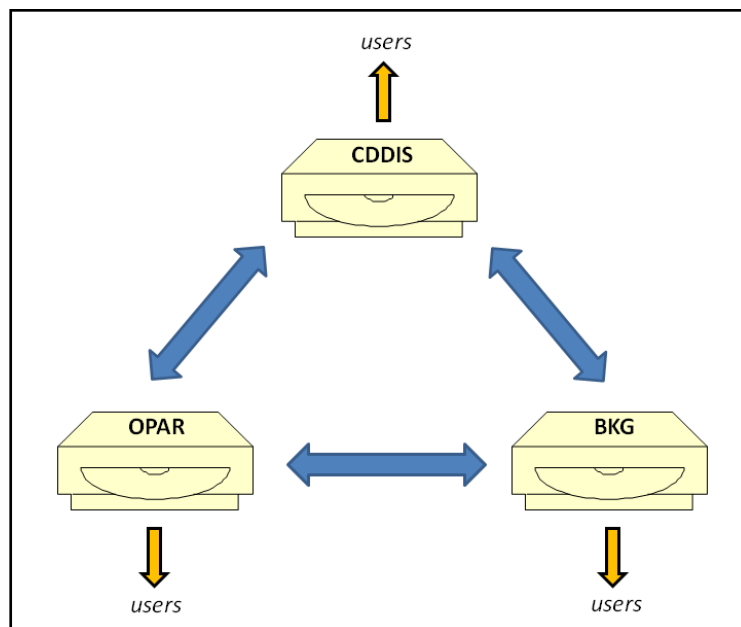
**Figure 1.** Organization of the IVS and general data flow between the various component types of the service.

The raw VLBI data are recorded on storage media and then shipped to the pre-determined Correlator. With about 1–2 TB of raw VLBI data per station per 24-hour session, the entire session occupies up to 15–20 TB of storage space. In the correlation process the Correlator reduces the raw data to the ‘VLBI observables plus metadata’ in the so-called database (*db*) format, a binary format of about 1–2 MB size per 24-hour session. The database is uploaded to the Data Center; the raw VLBI data are erased and the storage media are prepared to record the next VLBI session. The *db* format was created in the mid-1970s for use with a specific VLBI analysis software. It has been the *de facto* standard for archiving and distributing geodetic/astrometric VLBI sessions until now. The IVS is currently working on the design and implementation of a new VLBI data structure, which is based on the NetCDF data storage format and uses modularization and wrapping techniques (Gipson, 2008; Gipson 2010).

Analysis Centers download the databases from the Data Centers, analyze the VLBI data, and upload the analyzed results and products (e.g., EOP, CRF, TRF) to the Data Centers. Special Analysis Centers called Combination Centers use the results of the individual Analysis Centers to prepare the final IVS products. The combined solution is more accurate and robust as compared to the individual solutions of the Analysis Centers. Users can download the final products (either the combination solution or the individual Analysis Center solutions) from the Data Centers. Finally, in order to close the loop, the Correlators and Analysis Centers provide feedback to the Network Stations on their performance and suggest possible improvements.

## 4 DATA CENTERS

The Data Centers play an essential role within the IVS as the repositories of all geodetic/astrometric VLBI data and products. In addition to the data and products mentioned in Section 3, the Data Centers also archive auxiliary information such as station log files, correlator reports, and documentation files. As the radio telescopes constitute a very large financial investment, it is important that none of the collected data gets lost or unusable. Furthermore, users should be able to access the data and products efficiently and reliably. For that, the IVS is supported by three Primary Data Centers: CDDIS (Crustal Dynamics Data Information System), BKG Leipzig (Bundesamt für Kartography und Geodäsie), and OPAR (Observatoire de Paris). The three Primary Data Centers mirror their data holdings several times per day in a predetermined scheme in pairs of data centers (Figure 2).



**Figure 2.** Mirroring scheme between the three IVS Primary Data Centers at CDDIS, OPAR, and BKG.

There are basically two ways the Primary Data Centers update their holdings. First, data and product files can be uploaded to a special incoming area of an individual Primary Data Center by an IVS component (e.g., Correlator or Analysis Center) using authenticated FTP from a registered IP address. Automated scripts check whether the names of the incoming files are either registered with the Coordinating Center in special code files or can be constructed from the information contained in the Master Schedule and, to a limited extent, whether the files comply with the expected data structure. If the tests are successful, the scripts relocate the incoming files to their appropriate archive directories. Secondly, new data and products are added in the mirroring process. Users are only allowed to retrieve files from the Data Centers using anonymous FTP; they are not allowed to upload files. The basic directory structure is identical for all Primary Data Centers. A detailed description of this structure is, for instance, given in Noll (2010); although this reference describes CDDIS only, the VLBI part is directly transferable to BKG and OPAR.

The main geodetic VLBI data and products available at the Data Centers are summarized in Table 1. Auxiliary data files such as schedule files, station log files, or correlator reports are not included in the list, since they mostly support VLBI operations and are of marginal relevance for most scientific applications. They may be

considered metadata. All data and products are freely available at no cost to the user. The earliest VLBI data and products date back to 1979.

**Table 1.** Summary of the major geodetic VLBI data and products available in the three Primary Data Centers (from Noll, 2010).

Data set	Processing level	Granule	Time span
Correlated experiment databases	Data	Daily	1979–date
Baselines	Derived product	Daily	1979–date
EOP (all)	Derived product	Daily	1979–date
Station positions and velocities (TRF)	Derived product	Daily	1979–date
Source positions (CRF)	Derived product	Daily	1979–date
Zenith tropospheric wet delay	Derived product	Weekly	2002–date

Noll (2010) reported the usage statistics of CDDIS for the year 2008. By volume, VLBI accounted for about 2% of archive downloads, similar to SLR (also 2%) and DORIS (1%), whereas GNSS accounted for the major bulk of the downloads (95%). About 400 user organizations accessed CDDIS on a regular basis in 2008 in order to retrieve VLBI-related files. They downloaded over 13.5 TB of data and products in about 640,000 files. The users recruited from government agencies (53%), educational institutions (29%), networks (15%), commercial companies (2%), and some miscellaneous groups (1%). The other two Primary Data Centers are expected to have similar usage numbers.

## 5 CONCLUSIONS

The IVS was formed at the end of the last century to serve the scientific community in the fields of geophysics and astrometry. The service brought under one roof the operational activities and standardized the data flow from capturing the raw VLBI data, to correlating and preparing databases, to creating the final VLBI products. Various component types specialize in given aspects of the VLBI technique. The task of archiving and distributing the VLBI data and products rests with the IVS Primary Data Centers.

The concept of having three Data Centers with identical data holdings by means of a daily mirroring process has proven to be very successful. It ensures continuous and reliable access to the data without possible disruptions through maintenance work or IT security issues. The growing size of the repositories necessitates that the computing facilities be upgraded in terms of data capacity and speed. A challenge for the future will be the integration of near real-time data generation into the overall data flow.

## 6 REFERENCES

- Gipson, J. (2008) IVS Working Group 4: VLBI Data Structures. In: Finkelstein, A., Behrend, D. (Eds.), *Measuring the Future*, Saint Petersburg, Russia, Nauka, ISBN 978-5-02-025332-2, 143-152.
- Gipson, J. (2010) IVS Working Group 4: VLBI Data Structures. In: Behrend, D., Baver, K. (Eds.), *IVS 2010 General Meeting Proceedings*, NASA/CP-2010-215864, Greenbelt, MD, USA, 187-191.
- Noll, C.E. (2010) The crustal dynamics data information system: A resource to support scientific analysis using space geodesy. *Adv. Space Res.* 45(12), 1421-1440.
- Schlüter, W., & Behrend, D. (2007) The International VLBI Service for Geodesy and Astrometry (IVS): current capabilities and future prospects. *J. Geod.* 81(6-8), 379-387.
- Schuh, H., Behrend, D. (2012) VLBI: A Fascinating Technique for Geodesy and Astrometry. *J. Geodyn.*, submitted.
- Sovers, O.J., Fanselow, J.L., Jacobs, C.S. (1998) Astrometry and geodesy with radio interferometry: experiments, models, results. *Rev. Mod. Phys.* 70(4), 1393-1454.

# THE ACTIVITIES AT WORLD DATA CENTER FOR GEOMAGNETISM MUMBAI, INDIA

*M Doiphode\*, R Nimje and S Alex*

*\*Indian Institute of Geomagnetism, New Panvel, Navi Mumbai, 410218, INDIA*

*Email: mahendra\_d@yahoo.com*

*Email: rakesh\_nimje@yahoo.com*

*Email: salex@iigs.iigm.res.in*

## ABSTRACT

*The World Data Centre for Geomagnetism, Mumbai is functioning as a division of Indian Institute of Geomagnetism, Navi Mumbai since its full fledged activities commenced in 1991 in coordination with International Council of Scientific Union (ICSU) Panel on World Data centres. The responsibility of the compilation of final hourly absolute values of nine of the Indian magnetic observatories and deposition to the world data centres is undertaken at the centre. We have utilized the full advantage of technology advancement in upgrading the data preservation and conservation policy at various levels. In the recent years the centre has prioritized its activities related to digital preservation to ensure digital archiving of magnetic data from the traditional media and also the digital conservation of very old hand written/printed data volumes and magnetograms. In view of the scientific importance of data from Colaba-Alibag magnetic observatory, old magnetograms and data volumes are converted to digital images for long term preservation. In the digital preservation process, the creation of metadata is become an important component to store information related to old and current scientific records for future use. The centre is also hosting database driven website to make datasets available online to global scientific community.*

**Keywords:** ICSU, Digital Preservation, Colaba-Alibag Magnetic observatory, Geomagnetic data archival

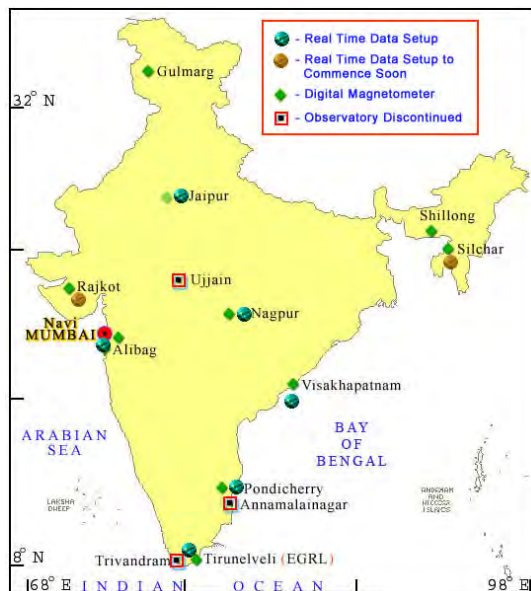
## 1 INTRODUCTION

Main objective of this paper is to introduce the various data related activities at world data centre(WDC) for geomagnetism, Mumbai, which is located at historic site of COLABA-ALIBAG magnetic observatory at Mumbai (Geog. Long. 72°52'E, Lat. 18°53' N). This centre is operated by Indian institute of Geomagnetism (IIG) which is an autonomous research organization under the Department of Science and Technology, Government of India. WDC for geomagnetism Mumbai is a part of International Council of Scientific Union (ICSU) World data centre system.

The Colaba Observatory located at Mumbai was built in 1826, however the geomagnetic measurements were started in 1841 and it was continuously operational till 1906. Later geomagnetic observations are continued at Alibag (Geog. Long. 72°52'E, Lat. 18°38'N) magnetic observatory. Colaba-Alibag combined series makes the geomagnetic data for a period of more than 160 years. This large geomagnetic data set has unique importance, as it provides the opportunity to relook in to old geomagnetic storm events and understand the physical processes associated with it. For example, the most intense 1–2 September 1859, magnetic storm in recorded history is studied by Tsurutani et al., (2003) using geomagnetic field records from Colaba observatory. Also such a large time series observatory data is useful in the study of long term change in geomagnetic activity, which has important implications for secular change in solar activity, global climate change, and the prediction of magnetic storm occurrence likelihood [Love, 2011].

IIG presently operates a network of nine Magnetic Observatories in the Indian longitude, which are shown in Figure 1. Geomagnetic field data is recorded continuously at these observatories, which extend from dip equator to the northernmost latitude of India. WDC for geomagnetism, Mumbai is actively involved in geomagnetic data depositary in India. Geomagnetic data at all the stations operated from IIG is collected at WDC geomagnetism

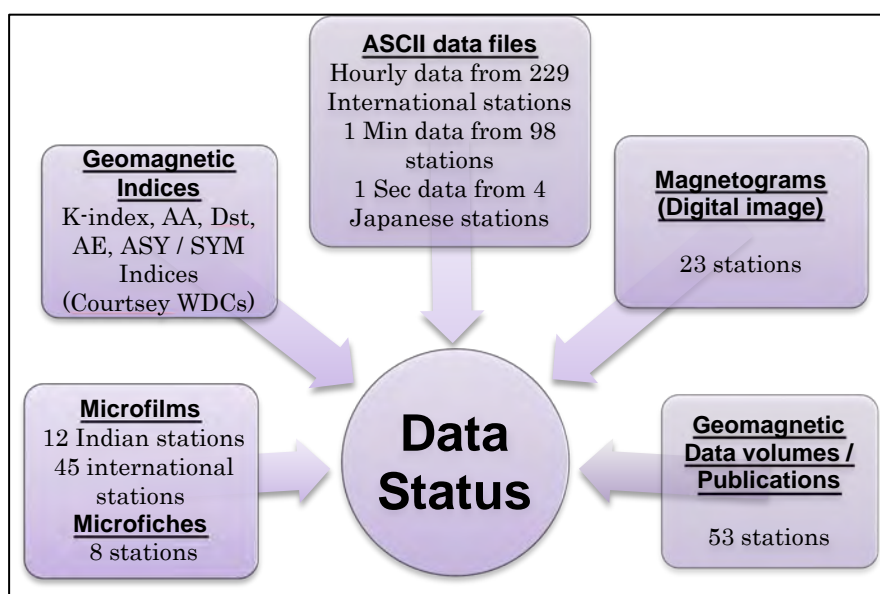
Mumbai along with geomagnetic activity indices supplied from a worldwide network of magnetic observatories. The Data Services at the WDC are available for scientific use at <http://www.wdciig.res.in>. Activities of WDC for geomagnetism Mumbai are elaborated in section 2. Old magnetic field data preservation and digitization process is explained in section 3. Early magnetic data recording at Colaba observatory is briefed in section 4 and the paper is summarized in section 5.



**Figure 1.** Shows network of magnetic observatories operated by Indian Institute of Geomagnetism.

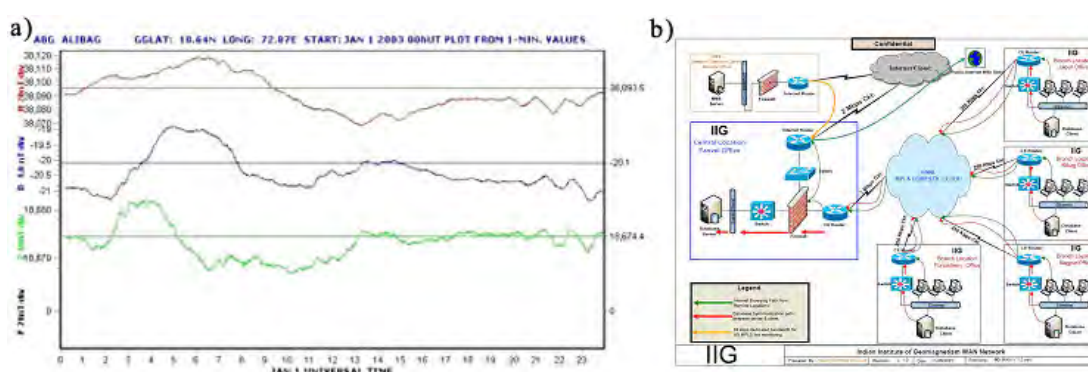
## 2 ACTIVITIES OF THE DATA CENTRE

The centre collects the geomagnetic data from Indian geomagnetic observatories, other international research organizations and the World data centers. We take this opportunity to forward our thanks to all these institutes, centers and organizations for their constant valuable data supply and support. Figure.2 is the current data status of the centre which has vast geomagnetic data set collection from Indian and international geomagnetic observatories in various data types.



**Figure 2.** WDC for Geomagnetism, Mumbai current data status and their types

Centre is having datasets from traditional media like punch cards, magnetic data tapes, microfilms/microfiches, printed data volumes, analog records like magnetograms. Most of these traditional old magnetic records are in current data forms also like ASCII data files on compact disks and DVDs, digital images of magnetograms, online data sets and images, real time digital data and plots. Earlier the centre was preparing html data catalogue to submit it to WDC system. However with the advancement of information technology and common media of internet for scientific data exchange and communication, the centre implemented various IT technologies for its day to day data handling purposes like now centre has hosted its own webportal in 2007 to provide geomagnetic data from Indian observatories online to the scientific community. The website contains hourly magnetic data (H, D, Z component), diurnal variation of plots of 1min resolution and data from few international observatories. Figure 3a shows an example of diurnal variation of H, D and Z component of geomagnetic field recorded at Alibag on 1 January 2003 with 1min sampling interval. The content is upgraded regularly and presently the high resolution magnetogram images are also ready to upload for online data users. Currently more than 500 scientific users are registered and regularly using data exchange facilities on this website. The average website uptime is 67%. This webportal is having online data access by onetime simple registration formalities without any restrictions and free of charge as a part of wdc data exchange policies.



**Figure 3.** (a) Variation of H, D and Z component recorded at Alibag on 1 January 2003 at 1min sampling interval is shown as a function of universal time (UT). (b) Pictorial representation of realtime data transmission network diagram of initial four Indian magnetic observatories.

As a part of technology advancement in science most of the Indian observatories are equipped with digital fluxgate magnetometer (DFM). The Institute has implemented Central REAL TIME DATA ACQUISITION SYSTEM at head quarter Panvel for real time magnetic data collection from 6 remote observatories using MPLS VPN data transmission technology. Figure 3b shows pictorial diagram of realtime data transfer from observatories. Also center is planning to make these realtime magnetic field data plots online through our WDC webportal.

### 3 OLD GEOMAGNETIC DATA PRESERVATION AND DIGITIZATION

The center is equipped with infrastructure for preserving the valuable geomagnetic records in analog formats like hand written/printed data volumes and magnetograms. In 2005 IAGA has funded for archival and retrieval of old Indian magnetic records and technical support was provided by World data centre for geomagnetism, Kyoto, Japan. Under this project the centre has converted old magnetograms of Colaba observatory into high resolutions digital images with the help of high resolution digital camera setup. Some of these records were digitized in 1 hour and 1 minute resolutions. The Center has taken steps to preserve the oldest geomagnetic paper records and data volumes by using preventive and curative conservation technologies. Through this activity centre succeed to curate large set of deteriorated data volumes and these processed data volumes can withstand for another 50 years for future generations. Preventative process is also done on recent good condition volumes to increase their durability, which will help centre in long term storage of geomagnetic records.

### 4 MAGNETIC DATA RECORDING AT COLABA OBSERVATORY

As mentioned earlier Colaba-Alibag observatory is a very old observatory and has geomagnetic data measurements for more than 160years. Geomagnetic data was recorded very systematically during the initial stage using eye observations prior to photographic recording. Figure 5(a) shows image of sample eye observation sheet of 8-10 July 1859 at Colaba and Figure 5(b) shows two days magnetogram image recorded

during 03-04 August 1882 at Colaba with header details. During 1847-1872 the hourly eye observations of the instrument were made on all days in the week except on Sundays and holidays. Whenever disturbance observed in the movement of magnets, eye observations were made at every 15 min and for severe disturbance at 5 min resolution. Figure 5(b) shows the geomagnetic data sheet of August 1882. Also during these period corrections to geomagnetic data is incorporated time to time to get the good quality magnetic data. Allowance for the temperature correction is made by reducing every scale readings to a uniform temperature of 80°F. No allowance is made for any correction for the effect of Moisture on the suspension wire, as the silver suspension wire of the magnet is supposed to be unaffected by moisture.[Moos, 1910]

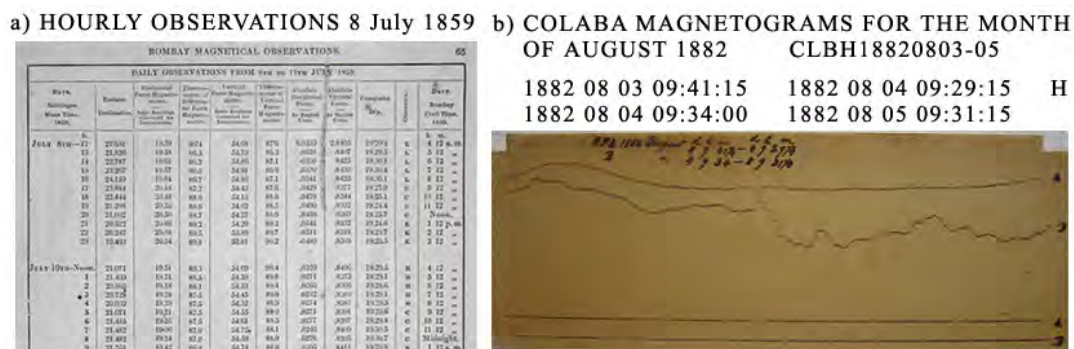


Figure 5. (a) Sample final geomagnetic data sheet . (b) Sample two days magnetogram image with header

## 5 FUTURE PLANS AND SUMMARY

WDC for geomagnetism Mumbai is established in 1991 with coordination of International Council of Scientific Union. This centre provides geomagnetic data from network of magnetic observatories operated by IIG, India together with geomagnetic data from few international observatories. Geomagnetic data and its information can be archived from our WDC web portal <http://www.wdciiig.res.in/>. For any data set the quality of the data is foremost before it is used for scientific purpose. The main aim of our center is to provide best quality geomagnetic field data online from network of observatories to scientific community. Center has following future plans:

1. To enhance Magnetic Data quality and consistency.
2. To extract and organize the metadata information in internationally acceptable metadata format for Magnetic Data.
3. Long term preservation and conservation of the old magnetic records for future reference.
4. The Implementation of IT technologies for Data Handling and global availability.
5. Up gradation of the web portal to add more online data services like real time variation plots of Indian stations.

## 6 ACKNOWLEDGEMENTS

We are thankful to all our observatory staff who is involved in data collection, maintenance and processing of magnetic data.

## 7 REFERENCES

Love, J.J. (2011) Secular trends in storm-level geomagnetic activity, *Ann. Geophys.*, 29, 251–262, doi:10.5194/angeo-29-251-2011.

Moos, N.A.F. (1910) Magnetic Observations made at the Government Observatories, Bombay 1846-1907, Parts I and Part II, Government Central Press Bombay, India.

Nimje, R., Doiphode, M. & Alex, S. (2008) Archival of Geomagnetic Data Management in Digital Form and its Retrieval, *DESIDOC*, 279-284.

Tsurutani, B.T., Gonzalez, W.D., Lakhina, G.S. & Alex, S. (2003) The extreme magnetic storm of 1–2 September 1859, *J. Geophys. Res.*, 108(A7), 1268, doi:10.1029/2002JA009504.

# JAPANESE CONTRIBUTION TO THE WORLD DATA CENTER FOR OCEANOGRAPHY

*A Seta<sup>1\*</sup>, S Wakamatsu<sup>1</sup>, T Miyake<sup>1</sup> and Y Iwabuchi<sup>1</sup>*

*<sup>1</sup> Japan Oceanographic Data Center, Japan Hydrographic and Oceanographic Department, Japan Coast Guard, 2-5-18, Aomi, Koto-ku, Tokyo, 135-0064 Japan  
Email: [jodc@jodc.go.jp](mailto:jodc@jodc.go.jp)*

## ABSTRACT

*The Japan Oceanographic Data Center has been submitting oceanographic data to the World Data Center for oceanography through the framework of the International Oceanographic Data and Information Exchange committee sponsored by the UNESCO/IOC. In the World Ocean Database 2009 which is the compiled database of WDC for oceanography, the Japanese contribution has reached about 16% of the total. Japan is one of the main data suppliers for the WDC for oceanography. JODC would like to contribute to the World Data System as in the past with WDC.*

**Keywords:** World Data Center, World Data System, Oceanography, Data management, International Oceanographic Data and Information Exchange

## 1 INTRODUCTION

The Japan Oceanographic Data Center (JODC) was established in the Hydrographic Department, Maritime Safety Agency (at present Japan Coast Guard), in 1965 in accordance with the resolution adopted by the Intergovernmental Oceanographic Commission (IOC) of UNESCO in 1961 as well as the reports of the Council for Marine Scientific Technology in 1963 and 1964. Since its establishment the JODC has been fulfilling the role of the synthetic marine data bank of Japan in the collection of marine data obtained by various marine research institutes and organizations concerned in Japan and in providing users with these data. (Michida 1997)

The JODC has also been submitting oceanographic data to the World Data Center for Oceanography (WDC-A; Silver Spring, NOAA) as the National Oceanographic Data Center of Japan under the framework of the International Council of Science (ICSU) and the International Oceanographic Data and Information Exchange (IODE) committee of the UNESCO/IOC.

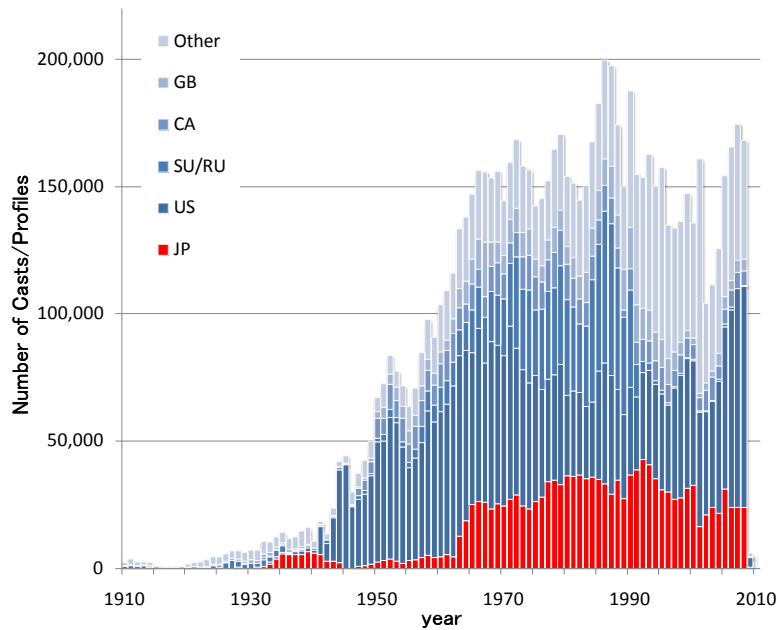
## 2 JAPANESE CONTRIBUTION TO THE WORLD DATA CENTER

The data which is submitted to WDC through the IODE framework is compiled as the “World Ocean Database (WOD)” by NOAA. In WOD09, the most recent version of WOD released in September 2009, the Japanese contribution has reached about 16% of the total (Boyer et al. 2009). That is the second largest contribution by nation next to the United States. With respect to the number of the ocean station data (OSD) which includes low resolution CTDs and XCTDs, the Japanese contribution has reached more than 20 % (Table 1). The number of stations submitted by Japan has been at a constant level of around 30,000-40,000 per year since 1965 when JODC was established (Figure 1).

**Table1.** Comparison of the contribution to the WOD09

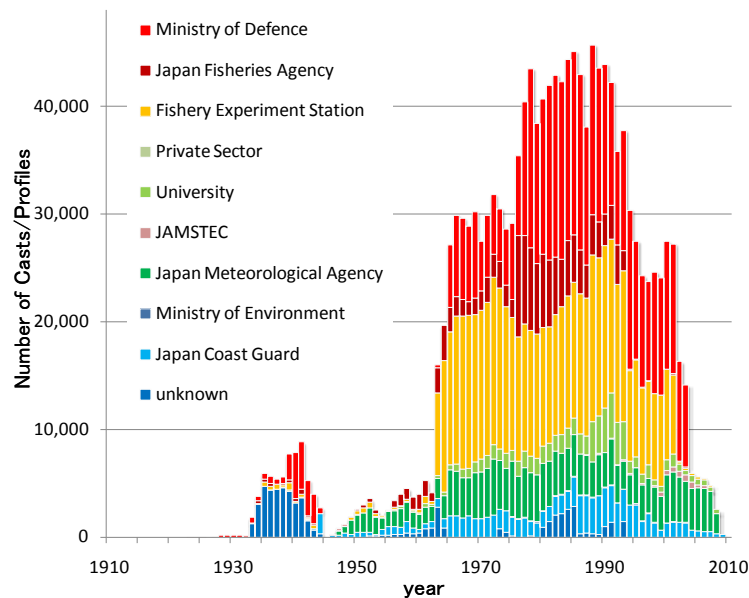
Coutry (ISO code)	United States (US)	Japan (JP)	Russia (Soviet Union) (RU, SU)	Canada (CA)	United Kingdom (GB)	Other	Total	% of total of Japan
Total	3,253,194	1,413,011	1,064,316	534,522	509,298	2,176,398	8,950,739	
(OSD, included)	(374,130)	(541,722)	(577,877)	(119,815)	(130,297)	(797,457)	(2,541,298)	21.3%
% of total	36.3%	15.8%	11.9%	6.0%	5.7%	24.3%		





**Figure 1.** Time series of the number of station data by Nation

Figure 2 shows the breakdown of Japanese organizations that submitted data to the WDC through the JODC. The oceanographic surveys which were carried out by the Fisheries and Defense agencies occupy the majority of the total. The data collection framework in Japan is based on voluntary participation. The JODC has a “Domestic coordinating committee” that meets every year, in order to collect data efficiently and to exchange information about survey plans. This allows for effective cooperation of the exchange of data and information.



**Figure 2.** Time series of the number of Japanese station data submitted by JODC

The JODC also initiated contributions to the IOC/IODE Global Oceanographic Data Archaeology and Rescue (GODAR) project (Levitus et al. 2005) with the goal of locating and rescuing oceanographic data that is at risk of being lost due to the media decay. Results of the project are also reflected in the WOD. A significant amount of historical Japanese data is included in the WOD. Over half of the data collected before 1945 is currently used in the WOD (Figure 1).

Figure 3 shows the historical transition of Japanese OSD data distribution. It indicates that Japan carried out many marine surveys, mainly in the Western Pacific, before 1945. After 20 years of these geographically limited survey activities, the OSD data obtained by Japanese surveys began to be acquired from a much broader area of the world ocean.

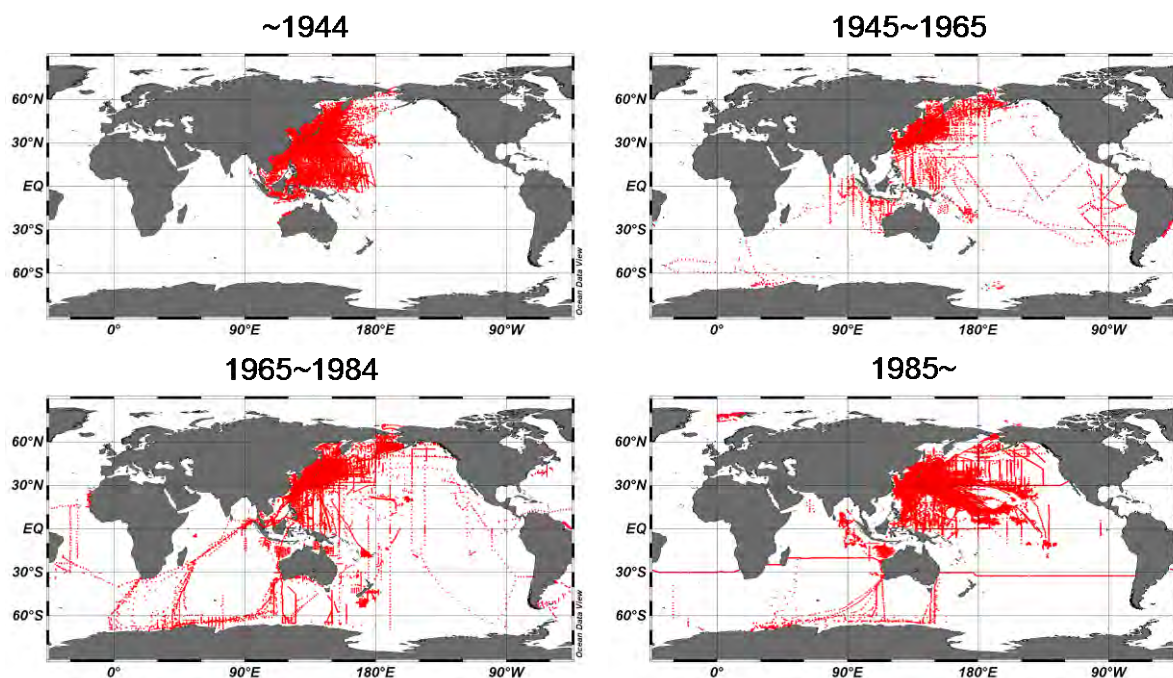


Figure 3. The transition of Japanese OSD data distributions every 20 years

### 3 FUTURE PERSPECTIVES

In March 2011, IODE adopted the statement with respect to the role of IODE in World Data System (ICSU/WDS) in its 21st session. In this statement, IODE expresses its strong interest in sharing data and information with the WDS. The JODC, as one of the national oceanographic data centers of the IODE, intends to contribute to the WDS as in the past with WDC.

### 4 REFERENCES

Boyer, T.P., Antonov, J.I., Baranova, O.K., Garcia, H.E., Johnson, D.R., Locarnini, R.A., Mishonov, A.V., O'Brien, T.D., Seidov, D., Smolyar, I.V. & Zweng, M.M. (2009) World Ocean Database 2009, Chapter 1: Introduction, NOAA Atlas NESDIS 66, Ed. S. Levitus, U.S. Gov. Printing Office, Wash., D.C., 216 pp., DVD.

Levitus, S., Sato, S., Maillard, C., Mikhailov, N., Caldwell, P. & Dooley, H. (2005) Building Ocean Profile-Plankton Databases for Climate and Ecosystem Research, NOAA Technical Report NESDIS 117, U.S. Gov. Printing Office, Wash., D.C., 29 pp.

Michida Y. (1997) Activity of the Japan Oceanographic data Center, *Special Issue of 'Umi no Kenkyu'*, 17-23.

# DATA AND INFORMATION ACTIVITIES OF ICSWSE, KYUSHU UNIVERSITY, JAPAN

*Shuji Abe<sup>1\*</sup>, Kiyohumi Yumoto<sup>1</sup>, Akihiro Ikeda<sup>2</sup>, Teiji Uozumi<sup>1</sup> and George Maeda<sup>1</sup>*

<sup>1</sup>*International Center for Space Weather Science and Education (formerly Space Environment Research Center), Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan*

<sup>2</sup>*Kagoshima National College of Technology, 1460-1, Shinkou, Hayato-cho, Kirishima City, 899-5193, Japan*

\* Email: abeshu@serc.kyushu-u.ac.jp

## ABSTRACT

*In this paper, we introduce data and information activities of the International Center for Space Weather Science and Education (ICSWSE), Kyushu University, Japan. The principal data source is the MAGDAS (MAGnetic Data Acquisition System) project which is a global network of geomagnetic observations operated by collaborations between ICSWSE and institutions in many countries. We operate 66 stations including more than 30 stations distributed along the 210° magnetic meridian and more than 10 stations along the magnetic equator. We have established a semi-automatic data acquisition system via the Internet. Provisional data plots and geomagnetic indices derived from the project are available to the scientific community.*

**Keywords:** Geomagnetism, Ground based observation, Magnetometer, Data citation, Metadata database, MAGDAS, ULTIMA, IUGONET, ICSWSE

## 1 INTRODUCTION

The International Center for Space Weather Science and Education (ICSWSE; formerly Space Environment Research Center, SERC, established in 2002), Kyushu University, was established in 2012 principally for the purpose of conducting research in space weather and related fields. A high priority is placed on collecting ground-based observational data. One of the major data collection efforts of ICSWSE became to be known as the “MAGDAS (MAGnetic Data Acquisition System)/CPMN (Circum-pan Pacific Magnetometer Network) Project”, whose Principal Investigator is the Director of ICSWSE (Prof. K. Yumoto).

The MAGDAS Project is collaboration between ICSWSE and many host institutions in various countries. The observational instruments, such as magnetometers, are owned by, and installed by ICSWSE. The host collaborative institutions maintain these. During the installation, ICSWSE instructs collaborators how to operate and maintain the instruments. These instruments are locally maintained by host collaborators, and ICSWSE remotely monitors the instruments. MAGDAS data are collected by ICSWSE, and distributed to researchers after required processing. Host institutions can use non-processed data maintained by them, but cannot redistribute them without asking ICSWSE. Non-processed data are not recommended for scientific usage (for more details, see Section 4). In addition, for capacity building of host institutions, ICSWSE organizes conferences for MAGDAS and related issues<sup>1</sup>.

As of March 2012, 66 MAGDAS magnetometers have been installed all over the world. Each MAGDAS instrument sends observational data to ICSWSE in near real-time via the Internet. We use these data for space weather research and for other applications, for example nowcasting and forecasting of solar-terrestrial events. Further information about the MAGDAS project is given in the next section.

It is important that any user can easily get detailed information related to MAGDAS. To meet the public demand we need to provide various MAGDAS information via our ICSWSE website. In addition, we plan to provide our MAGDAS information through the optimized metadata database system, IUGONET (Hayashi et al., 2012), with various software for data analysis developed in cooperation with other institutions (see Section 3).

<sup>1</sup> <http://www.serc.kyushu-u.ac.jp/news/MAGSessRes2010/index.html>,  
<http://www.serc.kyushu-u.ac.jp/news/MAGDASSchool2011/>,  
<http://iswimagdas2012.dirgantara-lapan.or.id/>

However, to enable the MAGDAS Project to continue for many years, it is necessary to secure funding for the project. An important aspect of funding is in making sure that credit goes to data providers when their data are used in publications. Therefore, we have established “Data Citation Rules” to ensure that the credit goes to the appropriate data providers. This will help us to secure future funding to continue our observations. For long-term retention, preservation, and open access to scientific data, we need additional funds and human resources. This is a common problem for researchers wishing to ensure long-term preservation and provision of their data.

In this paper, we will introduce the following three topics: First, we describe the general concept of MAGDAS project and its scientific applications. Second, we introduce the metadata database system and the analysis software for scientific usage of our data. Third, we mention the data usage rules and citation for long-term observation and collaboration.

## 2 MAGDAS/CPMN

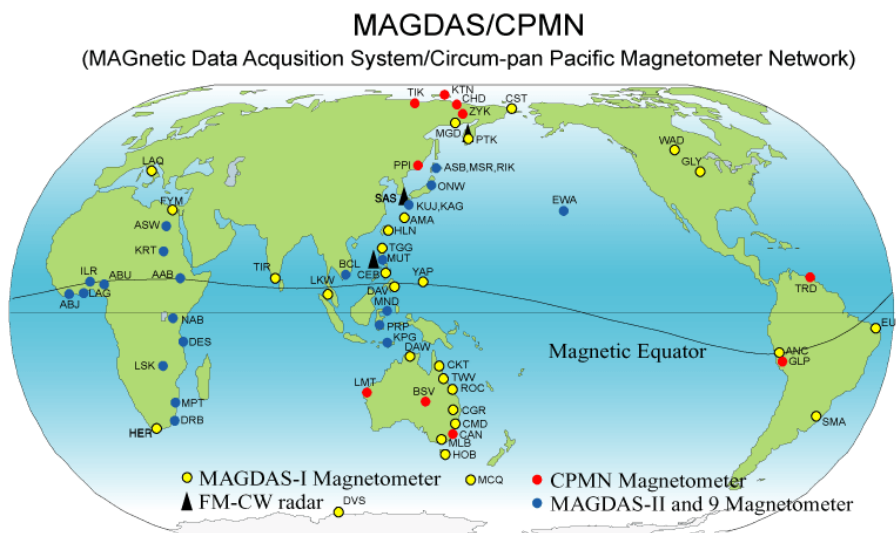
MAGDAS/CPMN is a world-wide ground-based magnetometer network (Yumoto & the 210 MM Magnetic Observation Group, 1996, Yumoto & the CPMN Group, 2001, Yumoto & the MAGDAS Group, 2006, Yumoto & the MAGDAS Group, 2007). The Circum-pan Pacific Magnetometer Network (CPMN) has been conducted since 1996 by Prof. K. Yumoto, Kyushu University, as the principal investigator. The magnetometer system of each CPMN station is installed by a group of 3-axial (horizontal (H), declination (D), and vertical (Z) components, respectively) ring-core type fluxgate magnetometer, a data logging/transfer unit, and a power supply. The maximum sampling interval is 1 second. GPS signals are received to adjust a time-keeping system of the data logger/transfer unit. These data are logged into compact-flash memories (in some cases, magneto-optical disks or cassette tapes are used). Geomagnetic data are recorded as relative values. In 2005, MAGDAS project was launched, and the CPMN was merged into MAGDAS. Since then, MAGDAS/CPMN is the generic name of our magnetometer networks. MAGDAS magnetometer system is based on CPMN magnetometer system (3-axial ring-core type fluxgate magnetometers, data logging/transfer units, and power supply units, GPS time adjustment systems, compact flashes for data storage), but there are some differences between the MAGDAS and CPMN systems. For example, geomagnetic data are recorded as absolute values in MAGDAS, in contrast as relative values in CPMN. Thus, we can estimate the total field strength from the three components of MAGDAS magnetometers. In addition, for more accurate total field observations, we can connect an additional Overhauser magnetometer, one of the magnetometers to measure the total magnetic field values by using nuclear Overhauser effect (e.g., Anderson and Freeman, 1962), to MAGDAS magnetometer logger. A MAGDAS magnetometer sensor has two tilt meters and one thermometer to monitor the environment of the sensor. MAGDAS magnetometer system has a function to send near real-time data to ICSWSE via the Internet. The instrument is entirely self-contained, except for power and network connections. MAGDAS-9 is a second generation magnetometer system for MAGDAS project, and now we are replacing the first generation magnetometers (called MAGDAS-I) by MAGDAS-9 magnetometers. Figure 2 shows a MAGDAS-9 system and its specifications. MAGDAS-II mainly consists of the CPMN magnetometer system, but data transmission units are newly added for near real time geomagnetic monitoring.

The principal components of current MAGDAS/CPMN are three magnetometer chains: 210 Magnetic Meridian (210MM) in Asia region, Magnetic Equator, and 96MM (Africa region). In addition, we have FM-CW (Frequency Modulated Continuous Wave) radar network to monitor global disturbances of electric and magnetic fields along 210MM. Figure 1 shows the distribution of MAGDAS/CPMN stations. Real time data transfer via the Internet is available. We also have long-term ground magnetometer data since 1985. Real time data (e.g., data plots) are available via our portal website (<http://www.serc.kyushu-u.ac.jp/data/>).

We also provide some useful geomagnetic indices estimated from MAGDAS/CPMN magnetometer data. The Pc 5 index shows quantitatively-estimated level of Pc 5 pulsation activity at low latitudes. The reference paper for Pc 5 index is now in preparation. We use hourly mean amplitudes of low-latitude Pc 5 observed at a MAGDAS/CPMN station to produce the hourly-mean local low-latitude Pc 5 index. According to a correlational analysis, we found a good correlation between estimated solar wind velocities from the Pc 5 index and in-situ data taken by the ACE satellite, which observes the solar wind velocity near the Earth since 1998 (Stone et al., 1998). This index enables us to obtain better understanding on the physical connection between Pc 5 pulsations and the temporal variations of the solar wind velocity. The estimated PC 5 index is opened to scientists via website (<http://www.serc.kyushu-u.ac.jp/pc5/>).

The EE-Index (generic name of three indices; *EDst*, *EU*, and *EL*), estimated from MAGDAS/CPMN real time data, has been proposed to be useful to monitor temporal variations of the equatorial electrojet in a long-term

manner (Uozumi et al., 2008). The one-hour mean value of the horizontal components of geomagnetic variations observed at the nightside (between 18 and 06 in local times) MAGDAS/CPMN stations in the magnetic equatorial region are found to show temporal variations similar to those of the *Dst* index (<http://wdc.kugi.kyoto-u.ac.jp/dstdir/index.html>). We call the geomagnetic variations in the magnetic equatorial region as the *EDst* index. We can use this index as a proxy of the *Dst* index for the real-time and long-term geospace monitoring. By subtracting the *EDst* from the horizontal geomagnetic component measured at one of the equatorial stations, we can extract the equatorial electrojet and the counter electrojet components from original geomagnetic observations at the observational point. We defined these values as the *EU* and the *EL*



indices, respectively. Anyone can access these indices at the following URL; <http://www.serc.kyushu-u.ac.jp/ee/>. **Figure 1.** Map of MAGDAS/CPMN activities. Black triangles denote FM-CW radar stations. Circles denote magnetometer stations. The color of each circle shows the current affiliation of the magnetometer (see text).



**Figure 2.** Current MAGDAS magnetometer system, called MAGDAS-9. This system consists of a data logging/transfer unit (right), a GPS antenna, a magnetometer main unit, a magnetic sensor, and a cellphone for reference (left), respectively. Total weight of items shown in this figure is 15.5kg. The type of geomagnetic sensor is that of 3-axial components ring-core fluxgate. It has a  $\pm 70,000\text{nT}/32\text{bits}$  dynamic range, and a  $0.01\text{nT}$  resolution. For more details, see our website, [http://www.serc.kyushu-u.ac.jp/magdas/MAGDAS\\_Project.htm](http://www.serc.kyushu-u.ac.jp/magdas/MAGDAS_Project.htm).

### 3 METADATA DATABASE AND ANALYSIS SOFTWARE

The ICSWSE is a member of the IUGONET (Inter-university Upper atmosphere Global Observation NETWORK, <http://www.iugonet.org/en/>). The IUGONET is a research project operated by five Japanese universities and institutions (National Institute of Polar Research, Tohoku University, Nagoya University, Kyoto University, and ICSWSE) to build a metadata database and analysis software for ground-based observations of the upper atmosphere. Users can search for MAGDAS information via the IUGONET metadata database system. In addition, users can analyze MAGDAS data via the IUGONET data analysis software (note: online data access is

restricted now. See below.). These tools allow researchers to use our database and data more easily.

The IUGONET metadata database (Koyama et al., 2012) has several fundamental functions of registering, retrieving, providing and harvesting the IUGONET common metadata format (Hori et al., 2012). We provide the IUGONET metadata database with MAGDAS Data File/Instrument/Observatory/Data Set/Person/Repository metadata which is written in the IUGONET common metadata format.

The IUGONET Data Analysis Software (UDAS) is a plug-in software which is based on the Themis Data Analysis Software suite (<http://themis.ssl.berkeley.edu/software.shtml>). Details are given by Tanaka et al. (2012). UDAS has a function to erase the differences between actual data formats. We provide some procedures for reading our MAGDAS storage data format to UDAS. Thus, UDAS are in place to allow users to download and view MAGDAS data online.

## 4 DATA USAGE AND CITATION RULES OF ICSWSE

To protect the rights of us and researchers who provide data to ICSWSE we regulate data usage implementing various levels of restriction depending on the type of user. Details of our regulations are shown in our Web pages(<http://www.serc.kyushu-u.ac.jp/data/index.php>). Our data usage rules are based upon the data citation rules employed by ULTIMA (Ultra Large Terrestrial International Magnetic Array, <http://www.serc.kyushu-u.ac.jp/ultima/ultima.html>).

## 5 CONCLUDING REMARKS

The goal of MAGDAS is to become the most comprehensive ground-based monitoring system of the Earth's magnetic field. This ground-based network establishes a mutually complementary relationship with space-based observations. MAGDAS played a significant part in the IHY (International Heliophysical Year) which was held in 2007-2008, and ISWI (International Space Weather Initiative) which was held in 2010-2012.

Since MAGDAS is a research project conducted by scientists, we introduced restrictions in various levels for data usage to assure the priority of them, who are managing day to day observations under the project. This is necessary to assure long-term operation of the project because we depend only on competitive financial resources. Although we wish to open our data for general use in the future, we need additional funds and human resources, both of them are difficult to obtain for a small research group. Since such a situation will be common among scientists who are producing data by their own research activities, we propose that the World Data System (WDS) includes “research data” as important data resources to be preserved and opened in the scope of WDS.

## 6 ACKNOWLEDGMENTS

The authors thank all MAGDAS host institutions and their members for maintaining our instruments and for good collaboration with us. The authors also thank all ULTIMA members for their help in establishing ULTIMA Ground Rules and our data usage/citation rules. We thank all IUGONET institutions and members for providing useful metadata database environment and data analysis software. We thank the anonymous referees and the editor for their detail and helpful reviews, comments and suggestions.

## 7 REFERENCES

- Anderson, W. A. & Freeman, R. (1962) Influence of a Second Radiofrequency Field on High-Resolution Nuclear Magnetic Resonance Spectra, *The Journal of Chemical Physics*, 37 (1), 411-5.
- Hayashi, H., Y. Koyama, T. Hori, Y. Tanaka, A. Shinbori, M. Kagitani, S. Abe, T. Kouno, D. Yoshida, S. UeNo, N. Kaneda, M. Yoneda, H. Tadokoro, & T. Motoba (2012) Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project (in Japanese), *J. Space Sci. Info. Jpn.*, in press.

- Hori, T., M. Kagitani, Y. Tanaka, H. Hayashi, S. UeNo, D. Yoshida, S. Abe, Y. Koyama, T. Kouno, N. Kaneda, A. Shinbori, H. Tadokoro, & M. Yoneda (2012) Development of IUGONET metadata format and metadata management system (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- Koyama, Y., T. Kouno, T. Hori, S. Abe, D. Yoshida, H. Hayashi, T. Tanaka, A. Shinbori, S. UeNo, N. Kaneda, M. Yoneda, T. Motoba, M. Kagitani, & H. Tadokoro (2012) Metadata Database Development for Upper Atmosphere (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- Stone, E. C., A.M. Frandsen, R.A. Mewaldt, E.R. Christian, D. Margolies, J.F. Ormes and F. Snow, The Advanced Composition Explorer, *Space Science Reviews*, Volume 86, Numbers 1-4 (1998) 1-22.
- Tanaka, Y., A. Shinbori, M. Kagitani, T. Hori, S. Abe, Y. Koyama, H. Hayashi, D. Yoshida, T. Kono, S. UeNo, N. Kaneda, M. Yoneda, H. Tadokoro, T. Motoba, Y. Miyoshi, K. Seki, Y. Miyashita, T. Segawa, & Y. Ogawa (2012) Development of IUGONET data analysis software (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- Uozumi, T., K. Yumoto, K. Kitamura, S. Abe, Y. Kakinami, M. Shinohara, A. Yoshikawa, H. Kawano, T. Ueno, T. Tokunaga, D. McNamara, J. K. Ishituka, S. L. G. Dutra, B. Dantie, V. Doumbia, O. Obrou, A. B. Rabiou, I. A. Adimula, M. Othman, M. Fairoos, R. E. S. Otadoy, & MAGDAS Group (2008) A new index to monitor temporal and long-term variations of the equatorial electrojet by MAGDAS/CPMN real-time data: EE-Index, *Earth Planets Space*, 60, 785-790.
- Yumoto, K., & the 210MM Magnetic Observation Group (1996) The STEP 210 magnetic meridian network project, *J. Geomag. Geoelectr.*, 48, 1297-1310.
- Yumoto, K. & the CPMN Group (2001) Characteristics of Pi 2 magnetic pulsations observed at the CPMN stations: A review of the STEP results, *Earth Planets Space*, 53, 981-992.
- Yumoto K. & the MAGDAS Group (2006) MAGDAS project and its application for space weather, Solar Influence on the Heliosphere and Earth's Environment: *Recent Progress and Prospects*, Edited by N. Gopalswamy and A. Bhattacharyya, ISBN-81-87099-40-2, pp. 309-405.
- Yumoto K. & the MAGDAS Group (2007) Space weather activities at SERC for IHY: MAGDAS, *Bull. Astr. Soc. India*, 35, pp. 511-522.

# INFORMATION ABOUT THE WORLD DATA CENTERS FOR SOLAR-TERRESTRIAL PHYSICS AND SOLID EARTH PHYSICS, REGIONAL MULTIDISCIPLINARY INITIATIVES OF RUSSIAN-UKRAINIAN WORLD DATA CENTERS SEGMENT FOR OCCURRENCE IN THE WORLD DATA SYSTEM

*N Sergeyeva\**, *E Kharin*, *L Zabarinskaya*, *A Rodnikov*, *I Shestopalov*, *T Krylova*, *M Nisilevich*

*\*Geophysical Center of the Russian Academy of Sciences, 3, Molodezhnaya str., 119296 Moscow, Russia  
Email: n.sergeyeva@gcras.ru*

## ABSTRACT

*Russian World Data Center for Solar-Terrestrial Physics and World Data Center for Solid Earth Physics collect, analyze, archive and disseminate data and information on a wide range of geophysical disciplines starting from the International Geophysical Year 1957-1958 up to present. The Centers provide free and convenient access for users to the great and permanently increasing volumes of data. Russian WDCs participate in the scientific national and international programs and projects such as InterMAGNET, InterMARGINS, International Polar Year. Since 2008 there is an association of five Russian WDCs and one Ukrainian WDC in a regional segment of the World Data Centers.*

**Keywords:** WDC, Geophysics, Russian-Ukrainian Segment, Data storage, Data access

## 1 INTRODUCTION

The World Data Center for Solar-Terrestrial Physics (WDC for STP) and World Data Center for Solid Earth Physics (WDC for SEP) in Moscow, Russia, perform a permanent job on inclusion of new geophysical data sets to the global distributed network information resources and provide the remote access for users to solar-terrestrial and solid Earth physics data. Digital data, metadata, thematic and problem oriented databases, inventory catalogues for all disciplines are available online at the WDCs web sites. Special user interface provides comfortable means for finding, reviewing, visualization, and retrieval of the online data and assignment them to a user.

The WDC for STP, Moscow, was established in 1957 within the framework of the International Geophysical Year 1957-1958. The WDC for SEP, Moscow, exists since 1971. Both are hosted by the Geophysical Center of the Russian Academy of Sciences (GC RAS) and are incorporated into the Laboratory of geophysical data. GC RAS is a public institution that receives funds from the Russian Federation through the Russian Academy of Sciences.

The main functions of the Centers according to the “Guide to the WDC System” are to collect, manage, and archive geophysical data on the underlying principles of long-term secure preservation, assurance of the quality of scientific data, and provision of free and open access to all data for scientific research.

The WDCs for STP and SEP store and disseminate national and foreign multidisciplinary data. Information resources of WDCs include modern and historical results of global observations related to the wide range of geophysical disciplines, obtained during the International Geophysical Year and subsequent international projects, results of geophysical observations on global observing networks and during special experiments and expeditions.

Providing access to data saved up in their archives, the WDCs besides serve as an information and reference node offering links for the information on other data centers and data providers which possess interesting data sets and databases. The Centers are targeted on the scientific organizations, separate researchers, universities and students in the different fields of sciences both in Russia and abroad.



## 2 RUSSIAN WORLD DATA CENTERS FOR SOLAR-TERRESTRIAL PHYSICS AND SOLID EARTH PHYSICS

### 2.1 WDC for Solar Terrestrial Physics, Moscow

WDC for Solar Terrestrial Physics, Moscow, activity extends to following disciplines:

- *Solar Activity and Interplanetary Medium*: sunspot areas and classifications, solar indices, optical observations, magnetic fields, X rays and UV radiation, energetic protons and electrons, proton bursts, solar wind density and velocity, electric and magnetic fields.
- *Geomagnetic Variations*: magnetic variations, pulsations, magnetosphere boundaries.
- *Ionospheric Phenomena*: ionospheric vertical soundings, radioactive absorption, radio interference, flare associated events.
- *Cosmic Rays*: solar and galactic neutrons, mesons.
- Summaries on separate kinds (individual types) of data and on results of special data analyses or processing (solar proton events, catalogues of geomagnetic storms, etc.).

Solar-terrestrial physics data are available in the form of printed tables, analog records and electronic. Printed tables and analog records are stored on paper and on microfilms and microfiches. WDC for STP realizes converting of printed tables and analog records into electronic form by scanning them or copying by the digital camera. Data in electronic form are stored on CD, DVD and Hard disks and are transformed into international formats whenever possible. The Center ensures persistent free access to them. Solar-terrestrial physics data are distributed either through the online access to the WDC's web site (<http://www.wdcb.ru/stp/index.en.html>) (Figure 1), or through the Space Physics Interactive Data Resource (SPIDR) (<http://clust1.wdcb.ru/spidr/>). User can also receive data by request.



Figure 1. Main page of the WDC for Solar-Terrestrial Physics web site

All standard data of the world geomagnetic observatories network, geomagnetic indexes, ionospheric data, cosmic ray data, solar data etc., stored in the WDC for STP, are accessible on the Moscow SPIDR web site and mirrored worldwide by SPIDR sites in Boulder, Paris, Nagoya, Sydney, Beijing, Kiev, and Capetown. The SPIDR is designed to allow a solar terrestrial physics customer to intelligently access and manage historical and modern space physics data for integration with environment models and space weather forecasts. SPIDR is a distributed network of synchronous databases, web portals and web services, allowing to choose, visualize and model data on the solar-terrestrial physics in the Internet.

Additional solar-terrestrial data in non-standard formats are available on the web site of the WDC for STP. They include data from magnetic observatories in Russia and the Former Soviet Union: hourly-mean values for 38 observatories mainly since the IGY (1957); one-minute values from 41 observatories mainly since 1983; global magnetic activity indices ( $aa$ ,  $Kp$ ,  $Ap$ ,  $AE$ ,  $Dst$ ,  $Pc$ , etc.); digital images of magnetograms beginning from 1957; sudden commencement readings since 1868; and catalogue of geomagnetic Pc1 pulsations at the Borok and Mirny observatories for the period 1957-1992.

### 2.2 The World Data Center for Solid Earth Physics, Moscow

The World Data Center for Solid Earth Physics, Moscow, collects and maintains archives of data on geophysical disciplines:

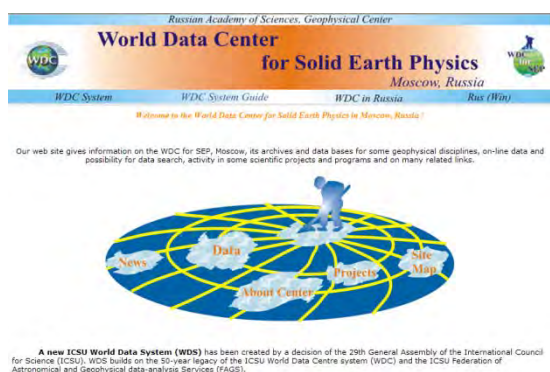
- *Seismology*: wave forms (seismograms), phase data (seismological bulletins), hypocenter data (earthquake catalogues), focal mechanisms, and seismological special data bases.

- *Magnetic Measurements (main magnetic field)*: marine surveys, maps and analytical models of the magnetic field, annual mean values of the magnetic field elements, secular variations.
- *Gravimetry*: marine surveys, measured values of the Earth gravity field, maps of gravity field and its anomalies, satellite data.
- *Heat flow*: catalogues of measured heat flow values, maps of heat flow isolines.
- *Archeo- & Paleo-magnetism*: data catalogues.
- *Recent Movement*: data catalogues.

All data are available in different traditional (paper, microfilms and microfiche) and electronic (separate files and databases) forms on various media. Archive is formed according to the disciplines and types of observations. Each section of the archive contains data represented in the form of tables, descriptions, maps, publications, graphic records (seismograms). All data are registered in the computer database and listed in the data inventory catalogues, which are free accessible on the web site <http://www.wdcb.ru/sep/> (Figure 2). Every data set is accompanied by metadata, detailed documentation and format description. Each new dataset is analyzed and its quality control is provided by means of special computer programs. All received datasets are stored in their initial form and two reserve copies of data are prepared as an indispensable condition.

These data contain the results of observations not only since 1957. Many historical data for earlier time periods are stored in the WDC for SEP. In order to expand the existing electronic geophysical data resources and also for prevention of loss of the old data converting of old data available in the form of publications into digital electronic form and providing the network access to these data is realized.

Users of the WDC for Solid Earth Physics are provided with data in the form of copies from data on paper carriers, CDs with data in electronic form and online on the web site. Any user can contact and get consultation of the WDC specialists by email or phone.



**Figure 2.** Main page of the WDC for Solid Earth Physics web site

### 3 INVOLVEMENT OF RUSSIAN WORLD DATA CENTERS INTO INTERNATIONAL PROJECTS

Russian WDCs participate in the scientific national and international programs and projects. The WDC for STP took part in the ICSU “The Rescue of the Magnetograms” project resulted in digital images of magnetograms from nine observatories of the former Soviet Union covering over 100 observatory-years of valuable data. Now the WDC for Solar-Terrestrial Physics is involved in the modern research project “InterMAGNET” in the part concerning the Earth magnetic field information technologies and data management. WDC for Solid Earth Physics is the participant of the international and interdisciplinary project “InterMARGINS” concerned all aspects of continental margin research.

Both Centers were the active participants of the “International Polar Year 2007-2008”, working in two programs “IPY Data and Information Service for Distributed Data Management – IPY DIS” and “Dataware for Geophysical Research for carrying out of International Polar Year”. Main output of implementing these programs was creation of the special web site containing results of various geophysical observations in Arctic and Antarctic regions carried out in the Former Soviet Union and then in Russia from 1957 up to present and stored in archives of both Centers (<http://www.wdcb.ru/WDCB/IPY/IPY.html>) (Figure 3). Some historical data, for example geomagnetic measurements at drifting stations “North Pole”, has been converted in electronic form specially for this site. The site is permanently supplemented by new data. Besides that the Centers participated in creation of Russian IPY-Info Portal which is an integrate high-quality multidisciplinary information system with

included metadata base, databases, systems of data collection, communication and data storage. Russian IPY-Info Portal serves as the component of the International Portal “IPY Data and Information Service – IPYDIS”. Metadata circulate in a system of data gathering, storage, exchange and processing at international and national levels.



**Figure 3.** Access page to Arctic geomagnetic data of IPY web site

## 4 RUSSIAN-UKRAINIAN WORLD DATA CENTERS SEGMENT

In 2008 five Russian WDCs (for Oceanography, Meteorology, Rockets, Satellites and Rotation of the Earth, Solar-Terrestrial Physics and Solid Earth Physics) and the Ukrainian WDC (for Geoinformatics and Sustainable Development) have united in the regional Russian-Ukrainian Segment of World Data Centers. The Scientific Council for coordination of Segment’s activity was formed.

Since 2009 two joint Russian-Ukrainian projects, aimed at development and strengthening of the Segment and creation of common information space, supported by the Russian Foundation for Basic Research and the Fundamental Researches State Fund of Ukraine, are being implemented.

For efficient storage and process of data and providing users with free and convenient access to them the general distributed multidisciplinary information-analytical system is developed in the framework of the Segment.

The WDCs entering into the Russian-Ukrainian Segment aspire to create a common information space with the uniform multidisciplinary data catalogue, the uniform metadata base and the single access point into the Segment.

The Russian and Ukrainian WDCs are developing an integrated access to common information resources of the Segment. The system will include a complete distributed multidisciplinary base of metadata, a catalog of multidisciplinary information resources, access services – a system of analytical modules, based on different methods of interactive data processing and providing free remote access to data.

## 5 CONCLUSION

WDCs for STP and SEP recently have passed all necessary procedures and became regular members of the World Data System. The Centers hope that their data archives and information resources will serve as the considerable contribution to the development of the World Data System. The further consolidation and solidifying of the Russian-Ukrainian Segment of WDCs and creation of its common information space will serve for further improvement of data management and providing strong connections and more intensive communications among WDS participants for the goal of free and convenient sharing and accessibility of science data and knowledge.

## 6 REFERENCES

- World Data Centers in Russia and Ukraine web site. Retrieved January 12, 2012 from the World Wide Web: <http://www.wdcb.ru/>  
 Space Physics Interactive Data Resource (SPIDR). Retrieved January 12, 2012 from the World Wide Web: <http://clust1.wdcb.ru/spidr/>

# THE APPLICATION OF AN ONLINE DATA VISUALIZATION TOOL, PTPLOT, IN THE WORLD DATA CENTRE (WDC) FOR SOLAR-TERRESTRIAL SCIENCE (STS) IN IPS RADIO AND SPACE SERVICES, AUSTRALIA

*K Wang<sup>1\*</sup> and C Yuile<sup>2</sup>*

*IPS Radio and Space Services, Bureau of Meteorology,  
Level 15, Tower C, 300 Elizabeth Street, Surry Hills, NSW, Australia 2010*

*<sup>1</sup>Email: [mkw@ips.gov.au](mailto:mkw@ips.gov.au)*

*<sup>2</sup>Email: [colin@ips.gov.au](mailto:colin@ips.gov.au)*

## ABSTRACT

*Ptplot is a set of two dimensional signal plotters components written in Java with multiple properties, such as being embeddable in applets or applications, utilizing automatic or manual tick marks, logarithmic axes, infinite zooming and much more. The World Data Centre of IPS applies Ptplot as a multiple function online data plot tool by converting various text format data files into Ptplot recognizable XML files with the AWK language. At present, Ptplot has allowed eight archived solar-terrestrial science data sets to be easily plotted, viewed and downloaded from the IPS web site.*

**Keywords:** Ptplot, Ptolemy Project, Plot, Data Visualization, Infographics, World Data Centre, IPS, Solar-Terrestrial Science, Magnetometer

## 1 INTRODUCTION

The rapid development of sensor, storage and networking technology has resulted in a huge increase in the retrieval and archival of scientific data. Data analysis and data mining is becoming increasingly important and challenging. Along with the rapid increase in data volumes, more powerful and efficient data processing technology and tools are being developed. Data plotting is a key aspect of data visualization and infographics technology and a useful method for a data analyst to locate interesting data.

In data processing, “computer users spend a lot of time doing simple, mechanical data manipulation – changing the format of data, checking its validity, finding items with some property, adding up numbers, printing reports, and the like. AWK is a programming language that makes it possible to handle such tasks with very short programs, often only one or two lines long” (Hao, Kernighan & Weinberger, 1988).

Ptplot (<http://ptolemy.eecs.berkeley.edu/java/ptplot/>) is a 2D data plotter and histogram tool implemented in Java. Ptplot can be used as a standalone applet or application, or it can be embedded in your own applet or application. It has properties: Embeddable in applets or applications; Auto-ranging; Automatic or manual labeling of axes; Automatic or manual tick marks; Logarithmic axes; Live, animated plots; Infinite zooming; Various plot styles: connected lines, scatter plot, bars, etc; Various point styles: none, dots, points, and unique marks; Multiple data sets and a legend; Color or black and white plotting; Error bars; Editable plots; PlotML, and XML language for specifying plots; Compatibility with pxgraph, an older plotting program. There are many interesting demonstrations, download links, and other useful information in its website.

Ptplot is a part of Ptolemy II (<http://ptolemy.eecs.berkeley.edu/index.htm>), but Ptplot is also available as a separate download. The latest version of Ptplot is Ptplot 5.8. The older versions, patches and extensions also can be downloaded from its web site. The Ptolemy II project is developed by the Electrical Engineering and Computer Science, College of Engineering, University of California Berkeley, USA.

IPS applies Ptplot as an online data plot tool to visualize various archived digital text files. Before a text file can be plotted with Ptplot, an AWK program is used to convert the text file into a temporary Ptplot recognizable XML file.

## 2 THE APPLICATION OF PTPLOT

IPS has made available eight different digital data archived in plain text format, as well as several other binary and image datasets. These eight data sets are able to be downloaded, or plotted and viewed online with Ptplot. This paper will demonstrate the process of converting and displaying magnetometer data using AWK and Ptplot.

## 2.1 The first webpage of the magnetometer online display

Figure 1 shows the main part of the first web page of the magnetometer data online display and download ([http://www.ips.gov.au/World\\_Data\\_Centre/1/2](http://www.ips.gov.au/World_Data_Centre/1/2)). IPS has listed 14 magnetometers on its World Data Centre section web page. A magnetometer installed at Casey station in Antarctica will be used as an example to demonstrate the process displaying magnetometer data online.

After selecting a station, year, month, and day, a magnetometer plot consisting of x and y components will be displayed on clicking “Plot Graph” as shown in Figure 1. If the user is interested in a daily magnetometer file, it is able to be downloaded by clicking the “Download Data” button on the same page.

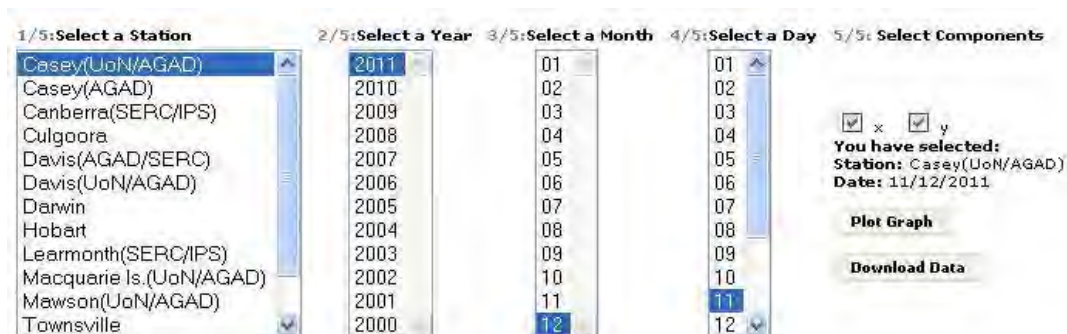


Figure 1. The first web page of the magnetometer online display and data download

## 2.2 The data format of the magnetometer text data file

Before the plot page is introduced, let us to start from the daily original magnetometer data file. The magnetometer data file was originally recorded at Casey in Antarctica and transferred to Sydney in near real-time. This data is automatically processed and archived as a daily text file. Figure 2 is a small part of the daily file recorded on 11<sup>th</sup> December of 2011 at Casey Station of Antarctica. The left three columns are Universal Time of hour, minute, and second respectively and the right two columns are x and y components of the magnetometer records respectively. There are two records every second. So there are about 172800 records in a daily file.

11	59	58.00	-69.00	-129.00
11	59	58.50	-79.00	-124.00
11	59	59.00	-84.00	-125.00
11	59	59.50	-84.00	-124.00
12	0	0.00	-78.00	-128.00
12	0	0.50	-73.00	-123.00
12	0	1.00	-70.00	-135.00
12	0	1.50	-72.00	-137.00

Figure 2. A small part of the original magnetometer text file recorded at Casey Station of Antarctica on 11<sup>th</sup> December 2011

## 2.3 Convert text file to XML file with an AWK program

Behind the front page of Figure 1, on the server side there is a process to convert the original text magnetometer data file to Ptplot recognizable XML file. Figure 3 is the main part of an AWK program that is used to dynamically make the conversion once a date is selected and the “Plot Graph” button is clicked on Figure 1.

```

BEGIN { i = 0
while (getline <"header.plt" > 0)
    print $0
}
{
if ($3 == "0.00") {
    h[i] = $1
    m[i] = $2
    x[1,i] = $4
    x[2,i] = $5
    i++
}
}
END {
print "<title>"b"</title>"
split(a,p,"")
for (n in p) {
    if (p[n] == "x") {
        print "<dataset name=\"\"p[n]\" \">"
        for (j = 0; j < i; j++)
            print "    <p x=\"\"h[j] *60+m[j] ".0\" \" y=\"\"x[1,j] \" \"/>"
        print "</dataset>"
    }
    if (p[n] == "y") {
        print "<dataset name=\"\"p[n]\" \">"
        for (j = 0; j < i; j++)
            print "    <p x=\"\"h[j] *60+m[j] ".0\" \" y=\"\"x[2,j] \" \"/>"
        print "</dataset>"
    }
}
print "</plot>"
}
}

```

**Figure 3.** The main part of an AWK program used to convert text magnetometer files into XML files

Because a daily file has about 172800 records, to read, convert and plot it would take a long time. In order to reduce the server side process time, and deliver a quick online plotting, one record in every minute, that is only 1/120 records will be used to plot. So the condition If (\$3 == "0.00") in Figure 3 is used to selected records. When the third column (second) value of Figure 2 equals 0.00, its fourth (x component) and fifth (y component) column values are read and converted into XML file for further plotting.

Figure 4 is a small part of the XML file, which is saved in server side as a temporary file and is readable by the Ptpplot Java applet. Each XML file consists of a head part and a data part. Each head part consists of lines of Title, xLabel, yLabel, xRange, yRange, default mark type and a series of xTicks, each xTick line consists of an x label value and a position value. The upper part of the Figure 4 shows the last two lines of the xTicks. Each data part consists of two datasets: x and y components. The lower part of Figure 4 shows the first four lines of the x component dataset of the XML file. The middle of the Figure 4 is the Title line, the title "Casey (UoN.AGAD) Magnetogram (nT) on 11/12/2011" will be shown in Figure 5.

```

<tick label="23:58" position="1438.0"/>
<tick label="23:59" position="1439.0"/>
</xTicks>
<title>Casey(UoN/AGAD) Magnetogram (nT) on 11/12/2011</title>
<dataset name="x">
    <p x="0.0" y="-24.00"/>
    <p x="1.0" y="-7.00"/>
    <p x="2.0" y="-140.00"/>
    <p x="3.0" y="-117.00"/>

```

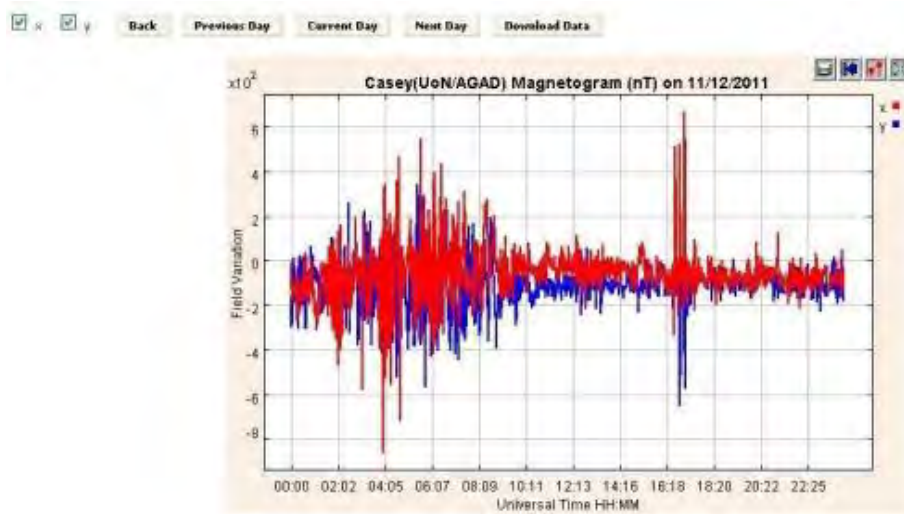
**Figure 4.** A small part of the converted XML file, readable by the Ptpplot Java applet

## 2.4 The properties and functions of Ptpplot

On clicking the "Plot Graph" button (Figure 1), The PHP program will find the data file according to the selected Station, Year, Month and Day in Figure 1 and convert the text data file into a XML file with the AWK program shown in Figure 3 and an XML Head file. Ptpplot reads the XML file and plots the data in a new web page. Figure 5 shows the new web page displaying the plot and other buttons above the plot. The plots of 'previous day' and 'next day' can be easily viewed by clicking the respective buttons (Figure 5). The "Download Data" button in the new webpage also can used to directly download the full data file.

The Ptpplot provides four small buttons on its upper right corner. The left most button is used to print the plot. The plot is infinitely zoomed with the mouse and the second button is used to recover the plot after zooming. A

“Set plot format” window will pop up when the third button is clicked, allowing properties of the plot to be edited. The right most button rescales the plot to fit the data.



**Figure 5.** The plot web page of magnetometer data with other function buttons

### 3 CONCLUSION

The Ptpplot Java applet is a powerful, efficient and versatile data visualization tool. Ptpplot enables a user to make data plots without extensive programming experience and expensive commercial software. The only required programming is to write a script or a program to convert data files into a standard Ptpplot recognizable XML file. IPS has applied Ptpplot on eight data sets for public use. Over twelve thousand online plots are created each year by customers around the world who visit the website.

### 4 ACKNOWLEDGEMENTS

We are thankful to the Electrical Engineering and Computer Science, College of Engineering, University of California Berkeley, for the use of Ptolemy II.

### 5 REFERENCES

Hao, A.V., Kernighan, B.W. & Weinberger, P.J. (1988) The AWK Programming Language, New York: Addison-Wesley.

# Data Intensive Multidisciplinary Science





# LESSONS LEARNED FROM DATA MANAGEMENT ACTIVITIES AFTER GREAT EAST JAPAN EARTHQUAKE IN MARCH 2011

A. Kitamoto<sup>\*1,2</sup>

<sup>\*1</sup>Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Email: kitamoto@nii.ac.jp

<sup>2</sup>PRESTO Project Researcher, Japan Science and Technology Agency

## ABSTRACT

*This paper summarizes our effort towards managing the multi-disciplinary disaster-related data from the Great East Japan Earthquake, which happened on March 11, 2011 off the coast of northeast Japan. This earthquake caused the largest tsunami in the recorded history of Japan, killed many people along the coast, and caused a nuclear disaster in Fukushima, which continues to affect a large area of Japan. Just after the earthquake, we started crisis response data management activities to provide useful information for supporting disaster response and recovery. This paper introduces the various types of datasets we made from the viewpoint of data management processing, and drew lessons from our post-disaster activities.*

**Keywords:** Disaster response, Data management, Meteorological data, Radiation data, Geo-tagging, visualization

## 1 INTRODUCTION

The Great East Japan Earthquake on March 11, 2011 off the coast of northeast Japan caused disasters that extremely impacted on Japan. As a response to the disaster, we decided to start a crisis response to provide various types of data on a website for people in need of emergency response and/or assistance. This was challenging because the variety and amount of data were unprecedented in terms of the number of disciplines involved, and this is especially true for the data from the Fukushima Daiichi nuclear power plant accident. General descriptions of the voluntary projects are already described elsewhere, such as (Utani 2011), so we focus on our own post-disaster activities, and draw lessons from our experiences. As of January 2012, we have produced the following datasets and services on our website (<http://goo.gl/9knll>).

- 1) Mass media online news about the earthquake and related events.
- 2) Meteorological observations and numerical data around Fukushima Daiichi nuclear power plant.
- 3) Radiation monitoring network data around nuclear-related facilities.
- 4) Radiation measurements on all schools in Fukushima prefecture.
- 5) Timelines of major events based on radiation measurements and accidents at the Fukushima plant.
- 6) Geo-tagging of mass media online news for the spatial understanding of events.
- 7) Comprehensive power plant database with aesthetically appealing visualization.

These datasets can be classified by using several facets, such as 1) real-time or archiving, 2) textual or numeric, 3) geographic or temporal, 4) alerting or informational, but we focus on the aspect of the data management process and classify it into several steps: 1) data collection, 2) data grounding, 3) data integration and analysis, 4) data visualization, and 5) data dissemination. The following sections summarize our experiences using the data management process for disaster response after 2011 Great East Japan Earthquake.

## 2 DATA COLLECTION

Datasets were collected from various Internet sources. Some were available before the disaster, and others were released after the disaster. Here, the challenge was to collect and integrate disparate datasets into a single database so that the data could be easily manipulated and visualized for multiple purposes.

The collection of data was a difficult task, not only because of the scattered sources of data, but also because of the non-uniform data formats defined for each data source. Machine-readable formats, such as XML (extensible

markup language) and CSV (comma-separated values) are ideal formats, but those formats were relatively rare. The usage of standard metadata formats, which allows automatic integration of data from many sources, was far from reality. Instead, data were released in PDF (portable document format), where scraping is difficult to automate, scanned PDF, where OCR (optical character recognition) produced poor results, or encrypted PDF, where automatic processing was almost impossible. It seems that some of the organizations wanted to control the usage of their data, and were reluctant to let people use the data in their own manner. People like us, however, are highly motivated for digitizing data against difficulties to obtain a better understanding of the data.

The most reliable way for digitizing data was to manually type the data by reading from PDFs or other types of documents. Thanks to the availability of cloud-based applications such as the Google spreadsheet, even a very time-consuming task could be finished by the collaboration of volunteers. For example, a group of volunteers started the radmonitor311 project (<http://sites.google.com/site/radmonitor311/>) to collect radiation monitoring data from various websites. They used the Google spreadsheet to collect data so that many people can work at the same time and share data easily through the Google spreadsheet API (application programming interface). The usage of cloud-based applications was an important step for aggregating the power of motivated people.

It is important to note that the release of data in a machine readable format is an important step toward realizing an “open government,” which is now regarded as an important agenda in such countries as the United States and United Kingdom. Here, the role of government is to provide data, and the usage of this data can be left as the work for people. The advantage of an open government policy is that better visualization and analysis tools could be produced at a more reasonable cost thanks to the power of volunteers with diverse skills. Disaster response is when this type of policy yields the greatest value, because the need for information and motivation for contribution are at maximum.

### **3 DATA GROUNDING**

Data grounding refers to mapping data on real coordinates such as space and time, which are the most basic facets of disaster-related data. Mapping to space means projecting data on a map, while mapping to time means projecting data on a timeline. When the space and time are represented in a textual representation, we need to convert textual representation into numerical coordinates. An example is the geocoding or geo-tagging of data from place names to latitude and longitude.

When datasets only contain place names without postal addresses or geographic coordinates, geocoding can be performed using the following two steps: 1) place names to addresses and 2) addresses to geographic coordinates. For the first step, we used Internet search engines to look up the postal addresses of places. For the second step, we used an Internet geocoding service to convert a postal address to its given latitude and longitude coordinates. Sometimes the location is only available as a point on a scanned map, or the location is given in a descriptive form (such as 100 m from the park along the road), but in these cases we have to rely on the manual process of geocoding to improve the reliability of points through the interpretation of information from multiple sources.

A more difficult case is the grounding of natural language text, in contrast to the geocoding of a postal address, which is well-structured. This process also consists of two steps: 1) extracting place names from the text, and 2) disambiguating them from the candidates. However, we do not have a simple way to segment place names from other words especially in Japanese where there are no delimiters for a word boundary. Words are also ambiguous, especially when we have two place names with the same spelling. We need additional information to resolve multiple candidates, but text has insufficient information to do this.

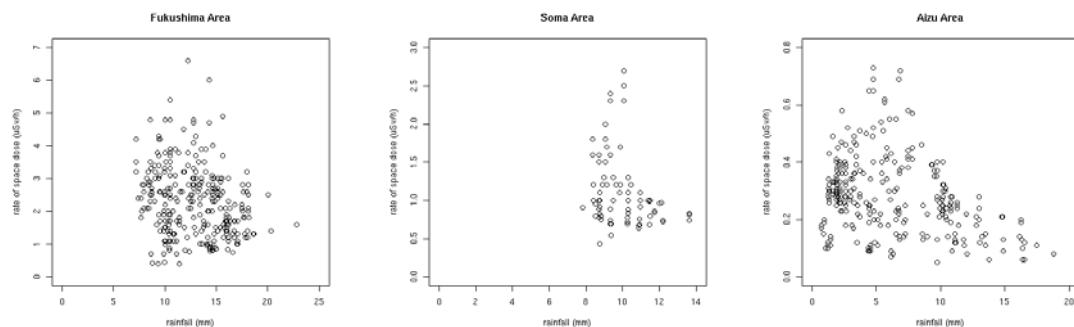
To solve this problem, we are developing a piece of software called “GeoNLP” (<http://goo.gl/5Jq1T>), whose task is roughly described as two steps: 1) toponym recognition and 2) toponym resolution, following the classification made by Leidner (2007). The first step deals with the recognition of a text span that constitutes a toponym, which is a special case of generally named entity recognition. This process depends on a dictionary of place names, or gazetteer. The second step deals with the selection of the correct referent among all the candidate referents (possible locations). We applied several types of heuristics to pick the best referent from among the given candidates, such as the proximity of multiple place names or the similarity of the place name category appearing in the same sentence. GeoNLP was applied to the grounding of mass media online news to categorize the online news articles by their place.

### **4 DATA INTEGRATION AND ANALYSIS**

The data from multiple sources can be compared on a uniform coordinate after data grounding. The task here is to compare different types of data to get a better understanding of the whole picture. The first task was to

integrate multiple timelines to interpret the dispersion of radioactive materials. The following four datasets were integrated: 1) timeline of radiation measurement events from radiation monitoring of Fukushima prefecture since March 15, 2) radiation monitoring for all schools in Fukushima at the beginning of April, 3) timeline of accidental events reported from Nuclear and Industrial Safety Agency since March 11, 4) meteorological simulation data from Japan Meteorological Agency (JMA) for wind and rain every hour since March 11. These four types of data differ in type, scale, and time, but a careful comparison of those datasets helped us understand the mechanism of dispersion for Eastern Japan. We focused on the major release events on March 15 and 21. We found out that the event on March 15 was especially difficult to analyze because of the complex constantly changing wind patterns on that day. However, we identified three phases of dispersion, namely a plume moving to the south, a plume moving to the west, and a plume moving to the northwest. Our interpretation was difficult to validate, however, because there was insufficient observation data in terms of space and time. We suggest that the analysis of the data should be backed up by using a carefully designed simulation study.

The second task was to analyze the relationship between the radiation level and total rainfall on March 15 using the Fukushima school monitoring and weather radar data. This analysis was based on the hypothesis that more rainfall brought more radioactive materials on the ground because rain was an important factor for the fallout of radioactive materials. The Fukushima school monitoring data was obtained at the beginning of April, which is before the decontamination campaign, so we expected the radiation levels to reflect the amount of fallout just after the accident. We divided Fukushima prefecture into nine regions, because the absolute level of radiation differed for each region. Figure 1 shows the analysis results. This indicates that the radiation level and total rainfall does not show a significant correlation. This is probably because the total rainfall is not accurate. We know from JMA observations that the weather in Fukushima on March 15 was weak snow or rain in the afternoon, but weak precipitation is poorly captured especially when the radar site is distant. The calibration of the rainfall by ground observations was not successful because weak rainfall is often measured as 0 mm of rainfall. Since the fallout of radioactive materials is highly affected by small-scale weather conditions and the location of a plume, we concluded that the relationship between the radiation level and total rainfall cannot be analyzed at a reliable level of accuracy using the available data.



**Figure 1.** Relationship between radiation level (obtained from the Fukushima school monitoring data) and total rainfall (on March 15 and 16 from the weather radar data). Panels represent three regions in Fukushima prefecture, Naka-dori, Hama-dori and Aizu, respectively.

## 5 DATA VISUALIZATION

Data visualization for disaster response uses a standard presentation tool such as a map or timeline, because the most basic facets of disaster-related information are space and time. We intensively used Google Maps to visualize the spatial data because of the convenient API they provide. Here, we introduce two examples of data visualization, namely “Wind Map around Fukushima Daiichi Nuclear Power Plant” (<http://goo.gl/gzPmR>) and “Electrical Japan” (<http://goo.gl/HwLVC>).

The first map visualizes the dispersion of radioactive materials from Fukushima Daiichi nuclear power plant. Many people were desperately in need of meteorological data on the wind and rain just after the accident, because they provided information on the transport of radioactive materials; wind contributes to horizontal movement, while rain contributes to vertical movement. Meteorological data, however, were not easily accessible after the accident for three reasons: 1) ground observations became inactive due to the shock of the earthquake, 2) weather forecast simulation data were only available to experts, and 3) the Japanese government decided not to release dispersion simulation results from SPEEDI (System for Prediction of Environmental Emergency Dose Information) to avoid panicking people. Our activity is to solve the second barrier by

improving the access to the visualization of existing data. Atmospheric simulation data for weather forecasting, namely numerical model GPV (Grid Point Value) data from JMA, are the best data for this purpose, because it offers meteorological elements such as pressure and wind on grid points from the surface to the upper atmosphere for selected pressure levels. We developed a system to visualize the GPV data on Google Maps, and released it on March 22. We tried to release the service as quick as possible, but it took 10 days after the accident, which was later than major release events on March 15 and 21. However, the service quickly attracted people's attention, and the page view of the service reached about one million as of January 2012.

The second map, Electrical Japan, focuses on a long-term issue, the energy policy of Japan, because the usage of nuclear power became a hot issue. As the basis of the discussion, we built a comprehensive database of the power plants in Japan with the location and power information to summarize the quantitative data on the current status of electricity generation. This was much more difficult than we first thought because complete information should be reconstructed from the complicated interpretation of fragmented information available from many sources. Moreover, we focused on DMSP (defense meteorological satellite program) nighttime lights of the earth to use nighttime lights as the proxy of electricity consumption. The power plant map for nighttime lights clearly visualizes the relationship between areas where electricity is made and used. The energy consumption of the Tokyo metropolitan area was supported by the electricity generated in rural areas, such as Fukushima. The map was designed for aesthetic beauty to reduce the unnecessary stimulus on this hot issue.

## 6 DATA DISSEMINATION

After the data are visualized on the website, the final step is to promote the services to the general public. Social media, especially Twitter, was used for this purpose, because social media was intensively used for sharing information after the earthquake. We built several Twitter bots to push information, such as @wind\_f1 tweeting wind direction for the next 24 hours. The information in a tweet was limited to 140 characters, but it was enough for the minimal role of a tweet, namely the notification of an update and the "heartbeat" of the system, with a link to a detailed web page. This notification service is only useful for real-time data, not for archival data, but it is always valuable to think about the usage of social media to attract people's attention to valuable data.

## 7 LESSONS LEARNED

We tried our best in the limited time available after the disaster, but the result is far from satisfactory. Lessons learned from our activity can be summarized as follows. First, we need to foresee the evolution of a disaster, and prepare data in advance before they are actually needed. As long as we start a project after something happens, the data become available later than they are actually needed. Solutions are first to improve the speed of software development. Second is to increase the flexibility of software development to make it improvisational and, third is to expand the imagination for potential disasters to prepare data before the crisis actually happens.

Secondly, a mechanism may be necessary to coordinate the voluntary projects started just after the disaster. At the time of a crisis, we cannot afford good coordination among activities due to the limited amount of time and continuous changes to the situation, and we observed many overlapping projects with similar purposes and methods. Of course, independent and rapid development is necessary for extremely rapid response to a disaster, and the situation could be improved if we had a single platform powerful enough to aggregate related efforts. A good example is Google Person Finder, which was quickly recognized as the single central database of safety information and volunteers' power was focused on this platform to achieve a large amount of work.

Finally, we provided most data in Japanese due to limited time, but the data should be provided as multilingual resources, at least in English. This earthquake raised worldwide interests, but Japanese people were accused of providing little English information. A practical solution is to use machine translation, but the accuracy is unsatisfactory. A solution is to limit the type of information using a fixed structure and a small vocabulary.

## 8 REFERENCES

Utani, A., Mizumoto, T. & Okumura, T., (2011) How geeks responded to a catastrophic disaster of a high-tech country – Rapid development of counter-disaster systems for the Great East Japan Earthquake of March 2011. *ACM Special Workshop on the Internet and Disasters 2011*.

Leidner, J.L. (2007) *Toponym Resolution in Text*, PhD thesis, School of Informatics, University of Edinburgh.

# MULTI-DISCIPLINARY APPROACHES TO INTELLIGENTLY SHARING LARGE-VOLUMES OF REAL-TIME SENSOR DATA DURING NATURAL DISASTERS

*Stuart E. Middleton*<sup>\*1</sup>, *Zoheir A. Sabeur*<sup>1</sup>, *Peter Löwe*<sup>2</sup>, *Martin Hammitzsch*<sup>2</sup>, *Siamak Tavakoli*<sup>3</sup>, and *Stefan Poslad*<sup>3</sup>

<sup>\*1</sup>*IT Innovation Centre, University of Southampton, UK, email: sem@it-innovation.soton.ac.uk*

<sup>2</sup>*GFZ German Research Centre for Geosciences, Germany, email: ploewe@gfz-potsdam.de*

<sup>3</sup>*Queen Mary University of London, UK, email: stefan@ecps.qmul.ac.uk*

## ABSTRACT

*We describe our knowledge-based service architecture for multi-risk environmental decision-support, capable of handling geo-distributed heterogeneous real-time data sources. Data sources include tide gauges, buoys, seismic sensors, satellites, earthquake alerts, Web 2.0 feeds to crowd source 'unconventional' measurements and simulations of Tsunami wave propagation. Our system of systems multi-bus architecture provides a scalable and high performance messaging backbone. We are overcoming semantic interoperability between heterogeneous datasets by using a self-describing 'plug-in' data source approach. As crises develop we can agilely steer processing server and adapt data fusion and mining algorithm configurations in real-time.*

**Keywords:** Natural disaster, Tsunami, Semantics, Data fusion, OGC, W3C, TRIDEC

## 1 INTRODUCTION

This paper describes our work in the project Collaborative, Complex and Critical Decision-Support in Evolving Crises (TRIDEC) (Wächter, 2011), building on experience gained from the German Indonesian Tsunami Early Warning System (GITEWS) (Münch, 2011) and Distant Early Warning System (DEWS) (Esbri, 2010). We employ a multi-disciplinary approach to intelligently share and process large-volumes of real-time sensor data, building a knowledge-based service architecture for multi-risk environmental decision-support.

Geo-distributed heterogeneous data management is at the heart of the TRIDEC knowledge-based service architecture (Sabeur, 2011). Real-time sensor proxies publish measurement data from tide gauges, buoys, seismic sensors and satellites, each with their own accuracy and frequency of measurement. Expert reports are published from trusted sources, such as SeisComp3 for earthquake alerts (Hanka, 2010). Real-time Web 2.0 feeds provide rapid crowd sourced multi-format measurements (text, images, video etc.). Lastly on-demand and high-resolution pre-computed simulations of Tsunami wave propagation are available to help (Behrens, 2010).

We describe our system of systems approach to data management, based on a hybrid service oriented architecture (SOA) and event-driven architecture (EDA) (Haener, 2011), using a multi-bus middleware providing performance, scalability and fault tolerance (Tao, 2012). We address semantic interoperability across sensor and information feeds through a self-described data source approach, exploiting metadata and schema from the World-Wide Web Consortium (W3C) and Open Geospatial Consortium (OGC). We provide agility of processing during crises through real-time steerable processing servers and context-aware information filtering.

## 2 PROBLEM STATEMENT

The TRIDEC project is looking at natural crises management for the purpose of tsunami early warning in the North-Eastern Atlantic and Mediterranean (NEAM) region. Following a crisis situation, such as a tsunamigenic earthquake, all the available sensor data will be used for decision support by a National Tsunami Warning Centre (NTWC). Figure 1 shows the effects of a Tsunami in Chile in 2010. Operators on duty need to assess quickly the relevance and the tsunamigenic properties of the earthquake event series, the likelihood of tsunami wave propagation, and the confirmation of tsunami progress. If necessary, the operator on duty will use information logistics and dissemination facilities to disseminate customized, user-tailored warnings to responsible authorities and to people under immediate risk.

It is the nature of natural crises that decision support tasks will change as the crisis evolves, requiring easy integration and re-configuration of data sources and agile processing that can be steered in accordance to current task requirements and appropriate decision support. Multiple NTWCs will jointly share observations from multiple sensor networks in the NEAM region, as well as information bulletins about imminent tsunamigenic events. The timely and reliable dissemination of bulletins and warning messages issued by governmental agencies has legal implications, adding to the challenge when setting up and testing communication channels.

During a natural crisis situation responsible agencies might also elect to forward situation information to other organisations such as Search and Rescue (SAR) or the remote sensing industry.



**Figure 1.** Damages by tsunami: City of Concepcion, Chile imaged on 10/01/2010 (left images) and 27/02/2010 (centre/right images) by the RapidEye satellite constellation. The centre image was taken eight hours after an earthquake of magnitude 8.8 had occurred and the resulting tsunami had affected the shoreline. The right map shows the areas with tsunami-related changes (red layer) on the post-tsunami image. Images courtesy of RapidEye AG, Copyright 2011, all rights reserved.

### 3 APPROACH

#### 3.1 Geo-distributed heterogeneous data sources

Tsunami early warning in the NEAM region involves various warning centres operating on a local, national and international level, linked to each other for centre-to-centre communication. On an international level information is shared across national borders following protocols negotiated between nation states in the Intergovernmental Coordination Group for the Tsunami Early Warning and Mitigation System in the NEAM (ICG/NEAMTWS) as part of the Intergovernmental Oceanographic Commission of the United Nations Educational, Scientific and Cultural Organization (IOC-UNESCO).

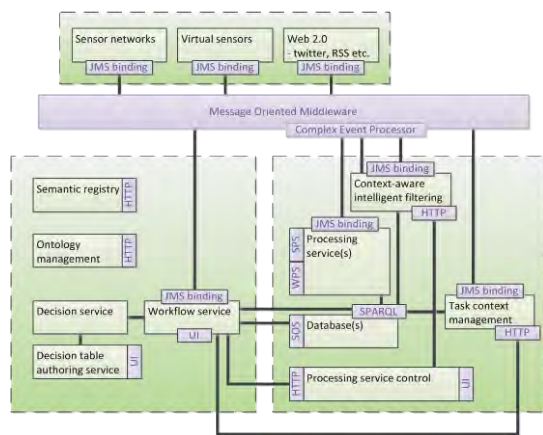
Due to the involvement of many nations and systems data managed is very heterogeneous. In-situ sensor networks provide time series measurements from seismic sensors, tide gauges and deep water buoys. These networks provide high quality configurable measurements but are few in number. Satellite systems provide streams of image data and webcam footage at coastal sites provide video. The scale of measurement message throughput depends on type and number of sensor networks connected, but as an example a seismic sensor network of 171 sensors publishing at 50Hz will lead to about 4 GBytes of data recorded per day.

In addition to 'conventional' sensor measurement systems there are expert reports available from simulations and alerting systems, such as Tsunami simulations and the SeisComP3 earthquake alert system. In recent years vast amounts of Web 2.0 content has become available, such as Twitter messages, YouTube videos and RSS feeds. These Web 2.0 'unconventional sensors' provide rapid in-situ crowd-sourced measurement by people actually experiencing the crisis event, e.g. using mobile devices, albeit with variable quality and have a high noise to signal ratio. In this way proven and reliable sensors are complemented by human sensors.

#### 3.2 Scalable high performance event-driven middleware

A modern warning system following a system of systems approach has to integrate various components and sub-systems such as different information sources, services and simulation systems, taking into account the distributed and collaborative nature of warning systems. A system architecture implementing such a system of systems approach has to combine multiple technologies and architectural styles (Moßgraber, 2012). An important challenge is to promote a communication infrastructure that facilitates environment information services, both sensor-based and human-based, to work together managing disparate information sources providing very large volumes and dimensionality of data. Such a system needs to support: scalable, distributed messaging; asynchronous messaging; open messaging to handling changing clients such as new and retired automated system and human information sources becoming online or offline; flexible data filtering, and heterogeneous access networks. In addition, the system needs to be resilient to handle the ICT system failures, e.g. failure, degradation and overloads, during environment events (Tao, 2012).

For the TRIDEC architecture, see Figure 2, we are looking at open source messaging platforms (Abie, 2009), which adopt a publish-subscribe model with brokers to overcome common problems like overload, interruptions, and the computational overhead of redundant nodes. Using a multi-bus approach supported by a hybrid MOM / SOA design, separating control channels (ESB) from content channels (multi-media) overloading a single communication channel is avoided. Tests (Sachs, 2009) of MOM's implementing a broker pattern show support for message rates of 14k/sec. The message throughput can be higher depending upon network capacity, server configuration, type of message interaction and configuration but that of the order of 100k message rates of 100k/s (Red Hat Inc., 2008)



**Figure 2.** The TRIDEC hybrid service & event oriented multi-bus architecture

### 3.3 Semantic interoperability and the use of metadata driven pre-processing

We employ a self-describing 'plug-in' data source approach to manage semantic interoperability, separately publishing metadata and data. Published metadata describing the phenomena (e.g. water elevation), encoding (e.g. text encoded) and measurement device used (e.g. tide gauge type & id). This metadata is formatted according to the OGC Sensor Web Enablement (SWE) Observation & Measurement (Cox, 2011) model and used to configure 'on-demand' data parsers, which subsequently process all published data. The metadata describing the measurements (URI's, units, data access details) is uploaded to a triple store as an RDF graph. The sensor data (often large with 100,000 measurements+) and complex/binary data (e.g. images or simulations) is stored in a relational databases and/or file storage, allowing high performance queries to large volumes of raw data.

Our data sources use different domain vocabularies to describe their measurements. We maintain a registry of Web Ontology Language (OWL) domain ontologies and inter-domain relationships, allowing automatic semantic mapping between identical and related measurement concepts. These semantic mappings coupled with the data source metadata allow pre-processing and dataset aggregation to be automated. Allowing 'plug and play' data sources without manual integration effort increases the scalability of our system of systems approach.

## 4 RESEARCH CHALLENGES AND IMPACT

One key challenge is now to use MOM technology to support complex event-driven messaging. MOMs can be designed to support message priorities, automatic client failover using configurable connection properties, queued data and metadata that is replicated across all nodes that make up a cluster, retry logic in the client code and persistent published data queues and subscriptions (Wang, 2010). Complex event processing and event cataloguing seem inevitable (Yuan, 2009), but each event-oriented logic operation reduces the message throughput. A resilient model must be implemented that is improved beyond resource hungry methods such as broker mirroring, and geo-resilience.

A few projects have now looked at using the OGC SWE standards for 'plug and play' sensor measurements (Middleton, 2010), exploiting OGC metadata to automate the pre-processing of data. These projects also use catalogue services to overcome scalability issues with geo-distributed data and services. In TRIDEC we are building on these results with a semantic registry and self-described data sources via a MOM.

Coupling OGC XML & W3C RDF has been done at a small scale (Henson, 2009) or via a web portal (Janowicz, 2011), indirectly mapping RDF URL's to OGC SOS's XML queries for data. In TRIDEC we adopt a W3C linked data type approach, making our data directly referenceable via a concrete URL. This allows us to add linked-data style annotations such as uncertainty information and processing provenance records.

## 5 CONCLUSION

Environmental information systems are becoming more and more complex as they increase in scale and scope. Data sources available to such systems of systems are geo-distributed and heterogeneous. Sensor networks such as seismic sensors, tide gauges and satellite systems provide time series measurements, images and video data. Expert reports from semi-automatic simulation and alerting systems is available and recent Web 2.0 provide a way to crowd source 'unconventional' measurements, albeit with variable quality and a very high noise to signal ratio. Scalable middleware solutions supporting system of systems, semantic interoperability between data sources and agility of data processing are key challenges.

In the TRIDEC project we are adopting a multi-Bus system of system model, de-coupling geo-distributed data sources from data processing servers and facilitating a scalable high performance messaging backbone. We are

overcoming semantic interoperability within our heterogeneous multi-domain datasets by using a self-describing 'plug-in' data source approach, exploiting OGC and W3C standards for our information models and using domain ontology mappings to automate pre-processing and aggregation of data from different domains. Lastly are adding agility to our processing servers by orchestrating processing workflows and deploying steerable processing server 'farms' that can adapt the data fusion and mining algorithm configuration as crises develop in real-time.

## 6 ACKNOWLEDGEMENTS

The work presented in this paper is a central part of the research and development in the TRIDEC project (contract no. 258723), supported by the 7th Framework Program of the European Commission.

## 7 REFERENCES

- Abie, H. Savola, R.M. & Dattani, I. (2009) Robust, Secure, Self-Adaptive and Resilient Messaging Middleware for Business Critical Systems, *IEEE Computation World*, pp153-160.
- Behrens, J. Androsov, A. Babeyko, A.Y. Harig, S. Klaschka, F. & Mentrup, L. (2010) A new multi-sensor approach to simulation assisted tsunami early warning, *Nat. Hazards Earth Syst. Sci.*, 10, 1085–1100
- Cox, S. (2011) Observations and Measurements - XML Implementation, OGC ref 10-025r1
- Esbri, M.Á. Esteban, J.F. Hammitzsch, M. Lendholt, M. & Mutafungwa, E. (2010) DEWS: Distant Early Warning System - Innovative system for the early warning of tsunamis and other hazards. In *Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE)*, Lisbon, Portugal, 27-29 October 2010
- Haener, R. Waechter, J. Hammitzsch, M. Lentholt, M. Sabeur, Z. & Poslad, S. (2011) Event-Driven Service Oriented Architecture Foundations for Industrial and Natural Crises Management Scenarios, *Geophysical Research Abstracts* Vol. 13
- Hanka, W. Saul, J. Weber, B. Becker, J. Harjadi, P. Fauzi, and GITEWS Seismology Group (2010) Real-time earthquake monitoring for tsunami warning in the Indian Ocean and beyond, *Nat. Hazards Earth Syst. Sci.*, 10
- Henson, C.A. Pschorr, J.K. Sheth, A.P. & Thirunarayan, K. (2009) SemSOS: Semantic sensor Observation Service, International Symposium on Collaborative Technologies and Systems, CTS '09, Baltimore, MD
- Janowicz, K. Broring, A. Stasch, C. Schade, S. Everding, & T. Llaves, A. (2011) A RESTful proxy and data model for linked sensor data, *International Journal of Digital Earth*
- Middleton, S.E. (2010) SANY fusion and modelling architecture, OGC ref 10-001
- Moßgraber, J. Middleton, S. Hammitzsch, M. & Poslad, S. (2012) A Distributed Architecture for Tsunami Early Warning and Collaborative Decision-support in Crises, *Geophysical Research Abstracts* Vol. 14
- Münch, U. Rudloff, A. & Lauterjung, J. (2011) The GITEWS Project – results, summary and outlook, *Nat. Hazards Earth Syst. Sci.*, 11, 765–769
- Poslad, S. (2009) *Ubiquitous Computing: Smart Devices, Environments and Interactions*, ISBN: 978-0-470-03560-3, Wiley.
- Red Hat Inc. (2008) Red Hat Enterprise MRG – MRG Messaging: Throughput, Red Hat Reference Architecture Series
- Sabeur, Z. Wächter, J. Küppers, A. & Watson, K. (2011) Large Environmental Sensor Observation Data and Knowledge Base for Collaborative Decision-support in Crises, *Geophysical Research Abstracts* Vol. 13
- Sachs, K. Kounev, S. Bacon, J. & Buchmann, A. (2009) Performance evaluation of message-oriented middleware using the SPECjms2007 benchmark. *Performance Evaluation*, Volume 66, Issue 8, pp 410-434.
- Tao, R. Poslad, S. Moßgraber, J. Middleton, S. & Hammitzsch M. (2012) Scalable and Resilient Middleware to Handle Information Exchange during Environment Crisis, *Geophysical Research Abstracts* Vol. 14
- Wächter, J. Fleischer, J. Häner, R. Küppers, A.N. Lendholt, M. & Hammitzsch, M. (2011) Development of Tsunami Early Warning Systems and Future Challenges, *Geophysical Research Abstracts* Vol. 13
- Wang, J. Jiang, P. Bigham, J. Chew, B. VinalMurciano, B. Novkovic, M. & Dattani, I. (2010) Adding resilience to message oriented middleware, 2nd Int. Workshop on software engineering for Resilient Systems.
- Yuan, S. Lu, M. (2009) An value-centric event driven model and architecture: A case study of adaptive complement of SOA for distributed care service delivery, *Expert Systems with Applications* vol. 36, no. 2



# MATHEMATICAL TOOLS FOR GEOMAGNETIC DATA MONITORING AND INTERMAGNET RUSSIAN SEGMENT

*Anatoly Soloviev<sup>1\*</sup>, Shamil Bogoutdinov<sup>1</sup>, Alexei Gvishiani<sup>1</sup>, Ruslan Kulchinskiy<sup>1</sup>, Arnaud Chulliat<sup>2</sup> and Jacques Zlotnicki<sup>3</sup>*

*<sup>1</sup>Institution of the Russian Academy of Sciences Geophysical Center RAS, 3, Molodezhnaya St., 119296 Moscow, Russia (GC RAS)*

*Email: a.soloviev@gcras.ru*

*<sup>2</sup>Institut de physique du globe de Paris (IPGP), 1, rue Jussieu, 75238 Paris cedex 05, France*

*Email: chulliat@ipgp.fr*

*<sup>3</sup>Observatoire de Physique du Globe de Clermont-Ferrand, Campus des Cuzeaux, 24, av. des Landais, 63177 Aubiere Cedex, France*

*Email: jacques.zlotnicki@wanadoo.fr*

## ABSTRACT

*Principally a new approach to detection of anomalies in geophysical records is connected with a fuzzy mathematics application. The theory of discrete mathematical analysis and collection of algorithms for time series processing constructed on its basis represent results of this research direction. These algorithms are the consequence of fuzzy modeling of logic of an interpreter who visually recognizes anomalies in records. They allow analyzing large data sets that are not yielded to manual processing. Efficiency of these algorithms is demonstrated in several important geophysical applications. Plans on extension of Russian INTERMAGNET segment are presented.*

**Keywords:** Magnetic field, Fuzzy sets, Time series, Magnetic observatory

## 1 INTRODUCTION

Detection of anomalies in geomagnetic records is a fundamental task of data analysis. Significance of the process represented by such records is often concentrated in these anomalies. Principally a new approach for solving this task is based on fuzzy logic and fuzzy mathematics application (Zadeh, 1965). Mathematical theory of Discrete Mathematical Analysis (DMA) (Gvishiani, Agayan, Bogoutdinov & Soloviev, 2010) and collection of algorithms for time series processing (e.g., Gvishiani, Agayan & Bogoutdinov, 2008; Bogoutdinov, Gvishiani, Agayan, Solovyev & Kihn, 2010; Soloviev, Chulliat, Agayan, Bogoutdinov & Gvishiani, 2011) constructed on the basis of DMA represent the results of this research direction. These algorithms are a consequence of fuzzy modelling of interpreter's logic who visually recognizes anomalies in records. The goal is its further application to automated analysis of large data sets that are not yielded to manual processing. A sufficient "flexibility" of the algorithms is provided by a wide range of "rectifications" that arise in interpreter operation modelling. Efficiency of the algorithms was demonstrated in several important geological, geophysical and geodynamic applications, including global real-time monitoring of magnetic storms (I. Veselovsky, S. Agayan, R. Kulchinskiy, A. Gvishiani, S. Bogoutdinov, V. Petrov et al., 2011), recognition of artificial disturbances in geomagnetic records (Soloviev, Bogoutdinov, Agayan, Gvishiani & Kihn, 2009), monitoring of volcanoes (Zlotnicki, LeMouel, Gvishiani, Agayan, Mikhailov & Bogoutdinov, 2005). Herein we present an overview of several of these successful applications dealing with magnetic field studies.

The largest global network of on-ground magnetic observations is International Real-time Magnetic Observatory Network (INTERMAGNET) (Love, 2008). The network consists of more than 110 observatories, however only 5 of them are located on the territory of Russia. Geophysical Center of Russian Academy of Sciences (GC RAS) has elaborated a plan on extension of Russian INTERMAGNET segment (Soloviev, 2011). In particular, five new INTERMAGNET observatories in Russia are being deployed by joint efforts of GC RAS and institutions of regional RAS branches. A regional geomagnetic data node of the Russian INTERMAGNET segment is being created on the basis of Russian WDC for Solar-Terrestrial Physics at GC RAS. A particular feature of this node is an automated system for recognition of artificial disturbances in incoming preliminary magnetograms, which is being introduced.

## 2 FUZZY MEASURE OF ACTIVITY AND MAGNETIC STORM MONITORING

A geomagnetic field is subjected to fluctuations of different time scales. In order to describe magnetic activity in the planetary scale the following geomagnetic indices were established: 24-hour Ci-index, 3-hour Kp-index, 1-hour indices Dst, AE and others (<http://www.ngdc.noaa.gov/IAGA/vdat/>). The principal idea of these indices is to give equal estimation of relative strength of disturbances at various observatories. However, more detailed study of the morphology of geomagnetic disturbances and their sources has shown that various indices of geomagnetic activity used nowadays express geomagnetic field activity not on the whole Earth surface but in its separate regions.

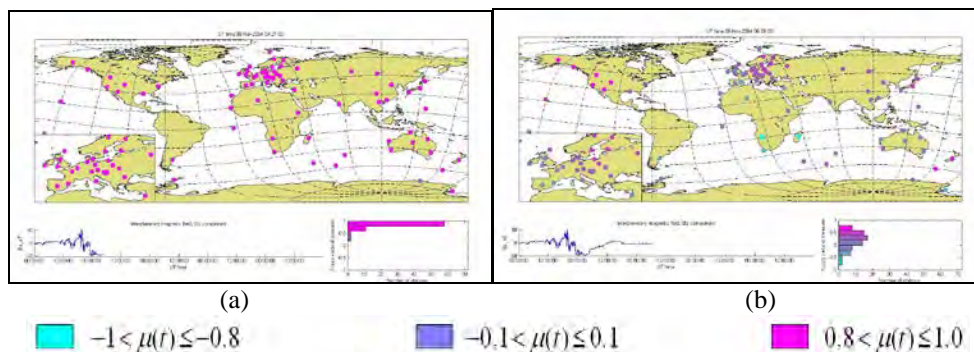
For studying dynamics of geomagnetic disturbances during a storm, it's not enough to use just several standard geomagnetic indices (for example, Cp, AE, Dst, *etc.*). In the process of solar-terrestrial phenomena studies the necessity of simultaneous determination of strength of geomagnetic disturbances at a maximum number of observatories across the Globe has arisen. By now, the largest global network of geomagnetic field observations is INTERMAGNET (Love, 2008). Such a necessity demands an introduction of new parameters independent of geomagnetic latitudes and longitudes.

Performing such kind of analysis by an expert manually is very difficult due to large volume of data involved. To solve this task a new geoinformatics approach based on fuzzy logic methods (Zadeh, 1965) is suggested. In particular we apply algorithm FCARS (Fuzzy Comparison Algorithm for Recognition of Signals) (Gvishiani et al., 2008) constructed on the basis of DMA (Gvishiani et al., 2010). Application of DMA enables processing and studying multidimensional arrays and time series.

FCARS algorithm allows the introduction of a measure of geomagnetic activity  $\mu(t)$ , which estimates geomagnetic activity at each particular time moment at particular geomagnetic record individually. It gives an estimation in a scale [-1, 1], where value -1 corresponds to a calm state and value 1 corresponds to an anomalous state. Testing the measure on magnetograms obtained from several INTERMAGNET observatories showed its high correlation with regional K-index of geomagnetic activity.

By applying measure  $\mu(t)$  to the whole set of magnetograms obtained by all observatories (e.g., INTERMAGNET network) we can have a snapshot of a storm distribution over the Earth's surface at each time moment. In the case of INTERMAGNET data, such picture changes with a time step of 1 minute. To visualize measure  $\mu(t)$  operation we use GIS technology. Thus, the proposed method for geomagnetic activity monitoring provides a new way of studying dynamics of spreading geomagnetic disturbances. It allows the performance of geomagnetic activity monitoring in real time mode.

The proposed toolkit was tested on two strong geomagnetic storms observed during the 23-rd solar cycle. In the first case a complicated storm on 8-11 November 2004 consisting of two parts was considered. In the second case an isolated storm on 15 May 2005 was considered. Before applying the method the selected storms were studied in details in order to investigate their common and specific features. This study involved INTERMAGNET data,  $D_{st}$ -index values (corrected version of  $D_{st}$ ) (Mursula, Holappa & Karinen, 2008), parameters of solar wind and interplanetary magnetic field and data on solar events. Figure 1 illustrates the toolkit application to the first storm monitoring at several time moments.



**Figure 1.** Global monitoring of the first storm in real time basing on INTERMAGNET data ( $H$  component). Two screenshots correspond to the toolkit operation at two different UT time moments: 8/11/2004 04:27 (a) and 9/11/2004 06:09 (b).

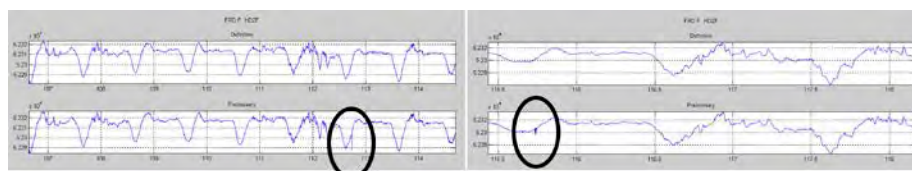
The toolkit visualizes distribution of anomaly parameter  $\mu(t)$  over the whole set of INTERMAGNET observatories in animation mode on a global map (Figure 1). Below on the right a dynamic histogram, which reflects distribution of INTERMAGNET observatories according to different values of  $\mu(t)$  from -1 (no anomaly) to 1 (strong anomaly) for each time moment, is given. Below on the left there is a dynamic plot of  $B_z$  component of interplanetary magnetic field, which evolves correspondingly. A clear conformity between the histogram of observed anomaly and  $B_z$  behavior can be seen.

The result showed that global, regional and local features of the storms have common and individual particularities depending on external stimulating conditions in heliosphere and on the Sun. Based on the analysis of dynamic distribution of geomagnetic disturbances it was shown that the ring current is not always the main contributor to the equatorial geomagnetic perturbations during the development and the main phase of strong geomagnetic storms. This led to a conclusion that the proposed approach gives a more objective and prompt estimation of geomagnetic activity than a number of classical indices.

### 3 DE-SPIKING 1-MINUTE AND 1-SECOND MAGNETIC DATA

Each year data experts at observatories and data centers carry out manual processing and filtering of collected preliminary data sets. The aim of such work is to produce definitive data and make it available to the scientific community worldwide. Despite close cooperation between observatories, approaches to data processing may differ and depend on subjectivity of this or that expert's evaluation. In this connection a mathematical formalization of recognition of artificial disturbances could contribute to significant increase of definitive data quality (Soloviev et al., 2009). In turn, an increase of observed data quality contributes significantly to our knowledge about the Earth's magnetic field.

An important step towards such a mathematical formalization was undertaken by Bogoutdinov et al. (2010). The proposed algorithm SP was applied to recognition of artificial spikes (Figure 2) on magnetograms recorded with 1-minute sampling rate. Since the algorithm operation is adjusted by a set of free parameters, it was first learnt based on 1-year preliminary records of the three components and total field intensity obtained in 2008 by seven INTERMAGNET observatories located in different parts of the Northern hemisphere. Then a learnt algorithm was applied to other 1-year preliminary records of the three components and total field intensity obtained in 2009 by the same observatories. Further, it was applied to 2-year preliminary records from the same observatories obtained during increased solar activity period in 2003 and 2005. In all the cases probability of a target miss was between 0% and 1%, whereas probability of a false alarm varied between 5% and 15%. The probabilities were calculated by comparing the algorithm results with definitive data filtered manually for the same time intervals.



**Figure 2.** Spikes in 1-minute preliminary data (lower plots) removed manually while producing definitive data (upper plots).

Many observatories now modernize their equipment in order to be able to produce 1-second filtered data. While at many observatories the 1-second data cleaning represents a reasonable amount of work, it becomes a daunting task at some observatories, particularly those installed in remote but important locations where no optimal observatory site could be found. In this case the situation is burdened with increase of data volume (86400 values per record per day comparing to 1440 in 1-minute case) and appearance of short-period geomagnetic pulsations similar to artificial spikes and unseen in 1-minute records. Therefore often automated de-spiking tools are much more demanded in the case of 1-second data acquisition.

For that purpose Soloviev et al. (2011) developed SPs algorithm, which is a modified version of SP algorithm, aimed at recognition of artificial spikes on 1-second magnetograms. As in the case of SP algorithm, the algorithm SPs was first learnt basing on 20-day (1-20/07/2009) preliminary records and then examined basing on 10-day (21-31/07/2009) records obtained by magnetic observatory in Easter Island maintained by IGP, France. The algorithm efficiency was estimated by comparing the results of automated preliminary data processing with definitive data for the same time spans. After that it was applied to other 30-day (1-31/08/2009) records with no

definitive data available. The results of the recognition by the algorithm SPs were subsequently evaluated by eye. After a 20-day learning phase in July 2009 the algorithm was able to recognize more than 94% of the spikes on the three components and the intensity recordings in August 2009, while the percentage of false alarms was less than 6%. At all the stages the algorithm showed worse results in processing vertical component Z.

#### 4 INTERMAGNET RUSSIAN SEGMENT

Currently Russian participation in INTERMAGNET program is confined to five observatories, which report preliminary data to geomagnetic information nodes (GINs) in Paris (France) and Edinburgh (UK). A weak development of INTERMAGNET network in Russia and an absence of national GIN induced GC RAS to elaborate a plan on extension of Russian INTERMAGNET segment (Soloviev, 2011). In particular, five sets of geomagnetic equipment compliant with INTERMAGNET standards are ready to be installed in different parts of Russia. Since a network of geomagnetic observations was widely developed in the Soviet Union during the International Geophysical Year in 1957-1958, numerous existing observatories ruled by RAS institutions are considered as possible sites for installing new equipment. These sites include Syktyvkar (Komi Republic), St. Petersburg, Rotkovets (Arkhangelsk region), Novaya Zemlya Islands, Tiksi and others. Apparently, deploying new observatories in auroral zone is of the highest priority in space physics community. However, the major obstacle, which prevents that, is a lack of personnel capable to operate observatories in such locations. In this regard, only those places where the operation of observatories is feasible are considered.

Another goal in the framework of extension of Russian INTERMAGNET segment is creation of a national geomagnetic data node at GC RAS servicing Russian INTERMAGNET observatories. A transmission of magnetograms from functioning and future Russian INTERMAGNET observatories to GC RAS will be performed in a real-time mode. A particular feature of this node is an automated system for data quality control, which involves SP and SPs algorithms, applicable to incoming preliminary magnetograms.

#### 5 CONCLUSION

A new way of study of dynamics of geomagnetic disturbances spreading is presented in the paper. It is based on fuzzy mathematics and GIS technology and involves data of the whole INTERMAGNET network. The measure of geomagnetic activity  $\mu(t)$  is computed according to FCARS algorithm and allows magnetic storms too be followed in real time mode. FCARS algorithm also serves as a basis of an automated system for processing electrotelluric and electromagnetic observations in the framework of the Russian-French project of monitoring volcanic activity on Reunion Island and in Kamchatka.

Application of the measure  $\mu(t)$  to selected INTERMAGNET records has shown its high correlation with regional K-index of geomagnetic activity. The proposed method for geomagnetic activity monitoring was tested on two strong geomagnetic storms. Based on the analysis performed it was shown that the ring current is not always the main contributor to the equatorial geomagnetic perturbations during the development and the main phase of strong geomagnetic storms. This led to a conclusion that the proposed approach gives a more objective and prompt estimation of geomagnetic activity than a number of classical indices. Basing on the analysis it can be concluded that geomagnetic proxies could also serve as an important source of indirect information about solar and heliospheric activity in the past, when direct observations were not available.

The algorithms SP and SPs based on DMA are specifically aimed at recognition of singular artificial spikes with a simple morphology on 1-minute and 1-second magnetograms. The algorithms rely on fuzzy mathematics principles. It was shown that after a learning phase these algorithms are able to recognize artificial disturbances efficiently and distinguish them from natural ones, such as short-period geomagnetic pulsations in the 1s-1min period range. This capability is critical and opens the possibility to use the algorithms in an operational environment. The algorithms were tested on real magnetic data. Small probability values for target miss and false alarm were obtained.

Knowledge of experts who carry out geomagnetic data analysis manually is effectively incorporated into all the developed algorithms at the stage of their learning.

Five new INTERMAGNET observatories in Russia are being deployed by joint efforts of GC RAS and institutions of regional RAS branches. A regional geomagnetic data node of the Russian INTERMAGNET segment is being created on the basis of Russian WDC for Solar-Terrestrial Physics at GC RAS. An automated

quality control system for on-the-fly processing of incoming magnetograms can significantly facilitate and hasten transformation of geomagnetic preliminary data into definitive data.

## 6 ACKNOWLEDGEMENTS

The results presented in this paper rely on data collected at magnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice ([www.intermagnet.org](http://www.intermagnet.org)).

This work is carried out in the framework of collaboration program between GC RAS, IPGP and Schmidt Institute of Physics of the Earth of Russian Academy of Sciences. It is also supported by the Russian Foundation for Basic Research (grant no. 09-07-91051).

## 7 REFERENCES

Bogoutdinov, S.R., Gvishiani, A.D., Agayan, S.M., Solovyev, A.A. & Kihn, E. (2010) Recognition of Disturbances with Specified Morphology in Time Series. Part 1: Spikes on Magnetograms of the Worldwide INTERMAGNET Network, *Izvestiya, Physics of the Solid Earth*, Vol. 46, 11, 1004–1016.

Gvishiani, A.D., Agayan, S.M. & Bogoutdinov Sh.R. (2008) Fuzzy recognition of anomalies in time series, *Dokl. Earth Sci.*, 421(5), 838-842.

Gvishiani, A.D., Agayan, S.M., Bogoutdinov, S.R. & Soloviev, A.A. (2010) Discrete mathematical analysis and geological and geophysical applications, *Vestnik KRAUNZ. Earth Sciences*, 2, Vol. 16, 109-125 (in Russian).

Love, J.J. (2008) Magnetic monitoring of Earth and space, *Physics Today*, 61, 31-37.

Mursula, K., Holappa, L. & Karinen, A. (2008) Correct normalization of the Dst index, *Astrophys. Space Sci. Trans.*, 4, 41–45.

Soloviev, A.A., Bogoutdinov, Sh.R., Agayan, S.M., Gvishiani, A.D. & Kihn, E. (2009) Detection of hardware failures at INTERMAGNET observatories: application of artificial intelligence techniques to geomagnetic records study, *Russ. J. Earth Sci.*, 11, ES2006, doi:10.2205/2009ES000387.

Soloviev, A. (2011), Artificial intelligence in the Earth's magnetic field study and INTERMAGNET Russian Segment, *ICSU CODATA Newsletter*, 100, p. 4.

Soloviev, A., Chulliat, A., Agayan, S., Bogoutdinov, S. & Gvishiani, A. (2011) Automated system for recognition of artificial spikes on 1-minute and 1-second magnetograms (#1170), *XXV IUGG General Assembly "Earth on the Edge: Science for a Sustainable Planet" (28 June - 7 July 2011, Melbourne, Australia), Program Book and Abstracts*.

Veselovsky, I., Agayan, S., Kulchinskiy, R., Gvishiani, A., Bogoutdinov, S., Petrov V. & Yakovchouk, O. (2011) Global, regional and local dynamics of strong geomagnetic storms (#5313), *XXV IUGG General Assembly "Earth on the Edge: Science for a Sustainable Planet" (28 June - 7 July 2011, Melbourne, Australia), Program Book and Abstracts*.

Zadeh, L.A. (1965), Fuzzy sets, *Inf. Control.*, 8, 338-353.

Zlotnicki, J., LeMouel, J.-L., Gvishiani, A., Agayan, S., Mikhailov, V. & Bogoutdinov, Sh. (2005) Automatic fuzzy-logic recognition of anomalous activity on long geophysical records. Application to electric signals associated with the volcanic activity of la Fournaise volcano (Réunion Island), *Earth and Planetary Science Letters*, 234, 261-278.

# A NEW APPROACH TO RESEARCH DATA ARCHIVING FOR WDS SUSTAINABLE DATA INTEGRATION IN CHINA

*WANG Juanle<sup>\*</sup>, SUN Jiulin, Yang Yaping, Song Jia, and Yue Xiafang*

*State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resource Research, Chinese Academy of Sciences, Datun Road, 100101 Beijing, China  
Email: wangjl@igsnr.ac.cn*

## ABSTRACT

*World Data System (WDS) requires that WDS data centers have significant data holdings and sustainable data sources integration and sharing mechanism. Research data is one of the important science data resources, but difficult to be archived and shared. To develop long term data integration and sharing mechanism, a new approach to data archiving of research data derived from science research projects has been developed in China. In 2008, the host agency of World Data Center for Renewable Resources and Environment, authorized by the Ministry of Science and Technology of China, began to implement the first pilot experiment for research data archiving. The data archiving process of the approach includes four phases, i.e., data plan development, data archiving preparation, data submission, and data sharing and management. In order to make data archiving more smoothly, a data archiving environment was established. It includes a uniform core metadata standard, data archiving specifications, a smart metadata register tool, and a web-based data management and sharing platform. Through the last 3 years practice, research data from 49 projects has been collected by the sharing center. The datasets are about 2.26 TB in total size and have attracted over 100 users.*

**Keywords:** World Data System, Data Sharing, Research Data, Data Archiving, China

## 1 INTRODUCTION

Data is one of the most important bases for science research. In general, science data can be divided into two types. One type is operational data derived from operational observation systems, such as meteorology data, seismology data, oceanography data, and so on. These data can be easily collected and shared under the national or departmental data sharing policies. Another type of science data is research data, which is collected and/or produced from scientific research programs or projects, such as International Geosphere Biological Program (IGBP), some national or local research projects, and so on. It is difficult to collect and archive this type of data comparing with the operational data, because the data is collected and hosted by different research teams or scientists separately. With the developments of international, national or regional science research activities, more and more research data will be generated. These data can serve as very important and sustainable data sources for other researches in different and crossing disciplinary fields. How to archive the data and make them to be accessible and reusable by others are a challenge tasks for the science communities including the World Data System (WDS).

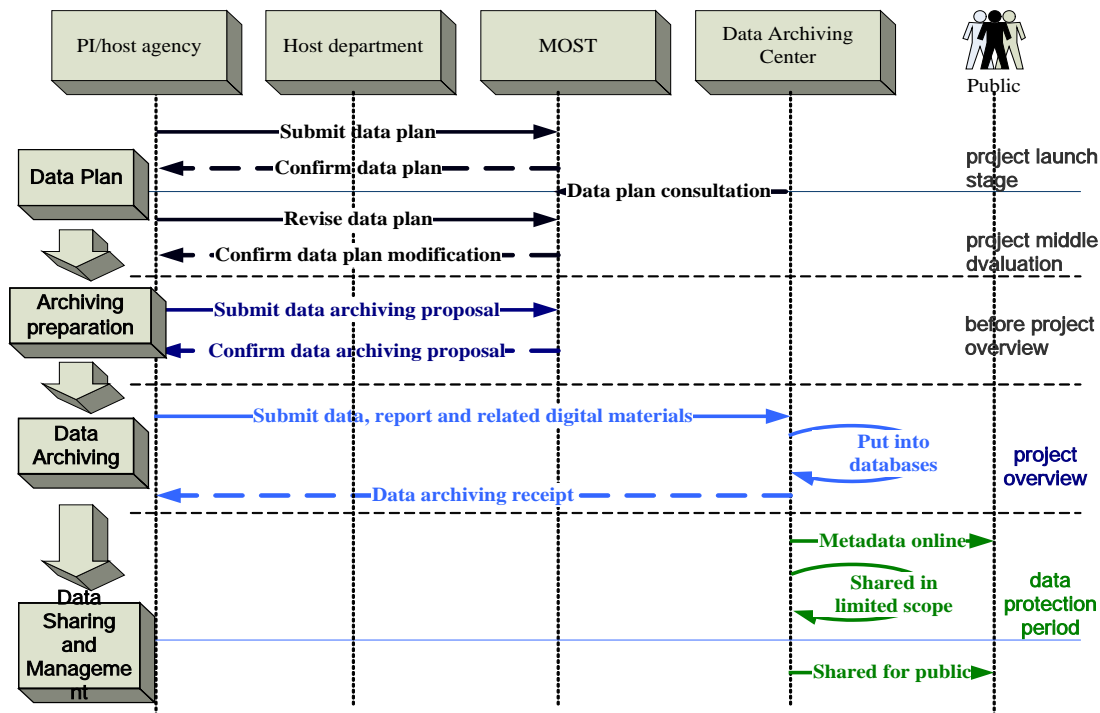
Based on developments of the former World Data Center (WDC) system in China, scientific data sharing has been making sound progresses in the past several years (Wang & Sun, 2007; Xu, 2003 and 2007). Under this background, Ministry of Science and Technology (MOST) of China decided to keep investigations on the data archiving for research projects funded by the government (Lin & Wang, 2008). Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, which is the host agency of WDC for Environment and Renewable Resources in Beijing, is authorized by MOST to design and implement the science research project data archiving experiment. Projects in the resources and environment field of National Key Basic Research Program are specified as the initial participant projects. Through one years' design and preparation, outputting data from these projects has been archived since 2008. This paper will introduce the new research data archiving approach and its progresses in the past 3 years.

## 2 RESEARCH DATA ARCHIVING WORK FLOW

First of all, research data archiving policy for research projects should be in place. After half a year preparation, "National Key Basic Research Program Data Archiving Management Specification on Resource and Environment Field" was published by MOST on 20 March, 2008. This specification not only defines the

responsibilities and duties of data owners, managers and users, but also specifies the data archiving work flow.

There are 4 phases in the data archiving process (shown in Figure 1), i.e., data plan development phase, data archiving preparation phase, data submission phase, and data sharing and management phase. The 4 phases cover the whole project research cycle from the beginning when the project was launched to the end when the project would be overviewed 5 years later.



**Figure 1.** The research data archiving phases work flow

## 2.1 Data plan development phase

Data plan is the guideline of the whole data archiving procedure. Many agencies require their science research projects to manage science research data based on data plans, such as National Institute of Health (NIH, 2003) and National Science Foundation (NSF, 2011).

Data plan for science research data archiving should define the data outputs during the whole project period. Those related output datasets information should be described in the data plan, including data sets' name, main data content description, data types, data formats, data security classification, data protection time period, sharing styles, related software tools, funding sources, etc.

## 2.2 Data archiving preparation phase

Data archiving preparation phase starts once a data plan confirmed and towards the near end of project. In this phase, data archiving center will guide projects to prepare their datasets collected during the research, and provide related technology support for their datasets management. All the projects will collect and manage their data and metadata information in the process of research by using a software tool provided by the data archiving center. At the middle stage of this phase, projects may need to revise their data plans according to status and changes of research projects. All revised data plans should be confirmed by MOST.

## 2.3 Data archiving phase

Data archiving phase will be taken place before the final project overviewed. At this phase, all projects should submit their datasets according to their data plans. It includes 3 steps. (1) Data archiving center provides the related data archiving standards and specifications to each project, including data archiving profile template, metadata standard, data document specification, data quality review specification, data submission format specification, etc. (2) Projects submit their datasets under these standards and specifications to data archiving

center using CD-ROM media. (3) If the dataset and its quality are confirmed by data archiving center, archiving receipts will be given to the related projects. Only those projects with archiving receipts have qualification for final project overview.

## 2.4 Data sharing and management phase

Data sharing and management phase will be conducted under a data management and sharing platform at the data archiving center. The platform provides customized functions for data providers, project managers and the public users respectively. For the data providers, they can submit and edit their research data, review the data services report online; for project managers, they can check the data archiving status online; for public users, they can explore data and access those datasets without sharing restriction online, and may apply those datasets with sharing restriction (e.g., data protect period) offline.

## 3 DATA ARCHIVING ENVIRONMENT CONSTRUCTION

Research data has inherent interdisciplinary features. In order to make these data integration together, a uniform data archiving environment is needed. It includes data archiving standards, data management specifications, related data archiving tools and sharing platform, etc.

### 3.1 Core Metadata Standard

A core metadata standard has been designed for research data archiving. Its metadata elements are listed in Table 1.

**Table 1.** The core Metadata elements for research data archiving

<b>Metadata element</b>	<b>Metadata content definition</b>
Dataset name	Dataset's specified name, which contains information about data thematic attribute, time period and region of data content
Project number	Specified project number allocated by MOST
Abstract	General and brief introduction of data content
Keyword	Significant or descriptive words for datasets
Dataset time	Time period of data content
Dataset format	Description of data storage format
Dataset quality	General evaluation information of dataset quality
Contact information	Contact information of producer(s) or the person(s) who is in charge for data publication or management
Usage restriction	Data copyright or privacy protection
Dataset web link	Website for data accessing

### 3.2 Data archiving specification

According to the requirements of research data archiving, a series of data management standards and specifications were designed and published by data archiving center. These include project data plan specification, data archiving report specification, data archiving document format specification, data archiving CD ROM specification, data quality review report specification.

### 3.3 Metadata collection and management tool

The metadata collection and management tool was designed and developed in Microsoft .Net environment. Its core functions include metadata records collection, review, appending, modification, delete and search. This tool is disseminated to all the projects and used for data preparation and archiving.

### 3.4 Data management and sharing Platform

The science research data management and sharing platform was developed in J2EE framework. All the data management and shared functions will be integrated in the platform, including the functions for data providers, data managers and data users mentioned above.



## 4 APPLICATION AND CONCLUSION

### 4.1 Application

By the end of Oct, 2011, 49 projects in resources and environment field of National Key Basic Research Program have submitted their research data to data archiving center. The size of the data accumulated is about 2.26TB, including more than 1000 datasets. According to the data storage and management types, these data can be divided into attribute data, text data, vector data, remotely sensed data, raster data, picture data and others. These data has their own individual disciplinary classification. A more flexible and integrated data category is under development by the data archiving center and will be published in the data sharing platform in the near future.

The number of registered users in the data sharing platform reached to 103. The website hits are 194704. About 1.5GB data has been downloaded. The top 5 datasets downloaded are listed as follow, “Tibetan plateau GDP change serials datasets (1970-2006)”, “Tibetan plateau ground temperature serials datasets (1951-2006)”, “Tibetan plateau livestock number change serials datasets (1970-2006)”, “Tibetan plateau population change serials datasets (1970-2006)”, and “China palmer drought index datasets”.

### 4.2 Conclusion

This science research data archiving experiment is the pilot initiative for the National Scientific Research Programs in China. It will have far-reaching influence to the scientific research data archiving and sharing projects which are funded by the government. Encouraged by the implementation of data archiving in resource and environment field, MOST will promote the research projects’ data archiving and sharing in broader fields of national funding projects in China. It not only enhances the developments for the data holdings of WDS data centers in China, but also contributes a robust and approved approach to the world community of science data integration and sharing.

## 5 ACKNOWLEDGEMENTS

This work is partially supported by State Key Lab of Resources and Environment Information System, Science & Technology Basic Research Program of China (Grant No. 2011FY110400) and Data Sharing Network of Earth System Science in China. Thanks the contribution of Mr. Shen Jianlei, Dr. Chen Wenjun from MOST of China. Thanks Mr. Wang Jiayi for Metadata tool development. Thanks Dr. Erjiang Fu for English language revision. We also thank the reviewers for their constructive suggestions and the editors for their careful check and revisions, which further improved this article.

## 6 REFERENCES

- Wang Juanle, & Sun Jiulin (2007) Development of China WDC Systems for Data Sharing. *China Basic Science Research*, pp 36-40.
- Guan-Hua Xu (2007) Open Access to Scientific Data: Promoting Science and Innovation. *Data Science Journal*, Volume 6, Open Data Issue, pp OD21-OD25.
- Xu guanhua (2003) Advance for enhance China’s science and technology innovation capacity by data sharing. *China Basic Science Research*, pp 5-9.
- Lin Hai, & Wang Juanle (2008) Data archiving work was launched in national basic research program in resource and environment field. *Advances in Earth Science* 23(8), 895-896.
- National Science Foundation (2011) Chapter II - Proposal Preparation Instructions, Retrieved January 1, 2011 from the World Wide Web: [Http://www.nsf.gov](http://www.nsf.gov)
- National Institutes of Health (2003) NIH Data Sharing Policy and Implementation Guidance. Retrieved March 5, 2003 from the World Wide Web: [Http://grants.nih.gov](http://grants.nih.gov)

# THE STATE OF IPY DATA MANAGEMENT: THE JAPANESE CONTRIBUTION AND LEGACY

*M Kanao*<sup>1\*</sup>, *A Kadokura*<sup>1</sup>, *M Okada*<sup>1</sup>, *T Yamnouchi*<sup>1</sup>, *K Shiraishi*<sup>1</sup>, *N Sato*<sup>1</sup>, and *M A Parsons*<sup>2</sup>

<sup>\*1</sup>National Institute of Polar Research, Research Organization of Information and Systems, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan

Email: \* [kanao@nipr.ac.jp](mailto:kanao@nipr.ac.jp), [kadokura@nipr.ac.jp](mailto:kadokura@nipr.ac.jp), [okada.masaki@nipr.ac.jp](mailto:okada.masaki@nipr.ac.jp), [yamanou@nipr.ac.jp](mailto:yamanou@nipr.ac.jp), [kshiraishi@nipr.ac.jp](mailto:kshiraishi@nipr.ac.jp), [nsato@nipr.ac.jp](mailto:nsato@nipr.ac.jp)

<sup>2</sup> National Snow and Ice Data Center, University of Colorado, 449 UCB, Boulder, Colorado 80309-0449, USA  
Email: [parsonsm@nsidc.org](mailto:parsonsm@nsidc.org)

## ABSTRACT

*Diverse data accumulated by many science projects make up the most significant legacy of the International Polar Year (IPY2007-2008). The Polar Data Center (PDC) of the National Institute of Polar Research (NIPR) has a responsibility to manage these data for Japan as a National Antarctic Data Center (NADC) and as the World Data Center (WDC) for Aurora. During IPY, a significant number of multidisciplinary metadata records have been compiled from IPY- endorsed projects with Japanese activity. A tight collaboration has been established between the Global Change Master Directory (GCMD), the Polar Information Commons (PIC), and the newly established World Data System (WDS).*

**Keywords:** International Polar Year, National Antarctic Data Center, Data Management, Metadata Portals, Polar Information Commons, World Data System

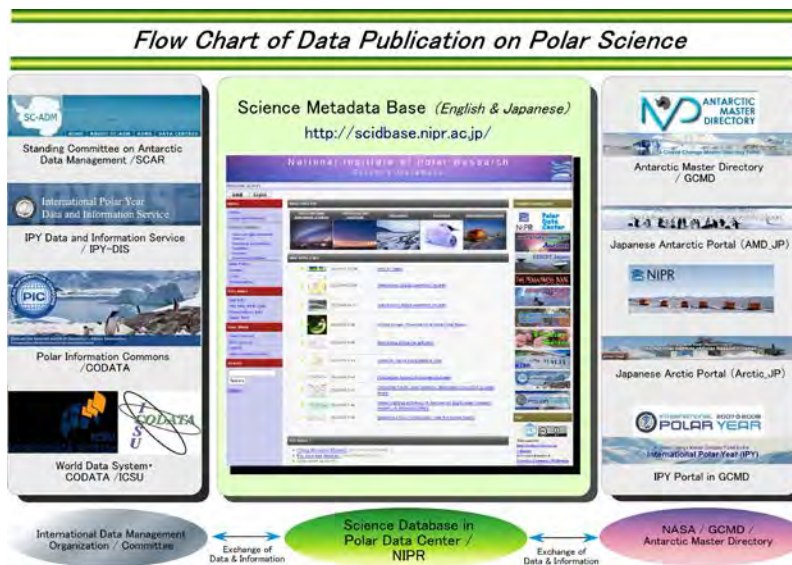
## 1 INTRODUCTION

The International Polar Year (IPY 2007-2008) was the world's most diverse international science program. It was conducted during the 50<sup>th</sup> anniversary of the International Geophysical Year (IGY 1957-1958). The IPY greatly enhanced the exchange of ideas across nations and scientific disciplines to unveil the status and changes of planet Earth as viewed from the polar regions (Rapley, et al., 2004). This sort of interdisciplinary exchange helps us understand and address grand challenges such as rapid environmental change and its impact on society.

The IPY 2007-2008 was jointly led by the International Council for Science (ICSU) and the World Meteorological Organization (WMO). A Joint Committee of WMO and ICSU (IPY-JC) was established in 2004 to arrange the whole IPY program. The same year in Japan, the IPY national committee was initiated under the Science Council of Japan (SCJ). Eventually, Japanese researchers participated in a total of 63 projects endorsed by the IPY-JC (Sato, et al., 2011).

The scientific results from IPY now begin to emerge, but it is clear that deep understanding will require creative use of myriad data from many disciplines. Many of these projects provided well-coordinated observation platforms, and many continue in the post-IPY era. The huge amount of data accumulating during and after IPY should be the most important legacy for IPY if it is well preserved and utilized (Parsons, et al., 2011a; 2011b).

The Polar Data Center (PDC; <http://www.nipr.ac.jp/english/polar-information01.html>) of the National Institute of Polar Research (NIPR) has served as the Japanese National Antarctic Data Center (NADC) with a strong relationship with the Scientific Committee on Antarctic Research (SCAR) under ICSU. During the IPY, we compiled much of the polar data from the endorsed projects involving Japanese activities. In this paper, the state of IPY data management involving Japan, particularly the tasks of the PDC, are demonstrated. A tight linkage is conducted with other science bodies of ICSU, such as the Committee on Data for Science and Technology (CODATA) and the new World Data System (WDS) (Figure. 1).



**Figure 1.** Flow chart of data publication on polar science involving PDC, including IPY data.

## 2 POLAR DATA CENTER

Recent, rapid, technological improvement and development of Earth observation by satellites and ground observation networks both in the Arctic and the Antarctic has led to a large quantity of polar observation data being collected every day. The processing and utilization of these data is an important issue to promote polar science. Our mission is twofold: scientific data management and management of information infrastructure.

At the 22nd Antarctic Treaty Consultative Meeting (ATCM) in 1998, affiliate countries were obliged to ensure that scientific data collected from Antarctic programs could be freely exchanged and used. Following Article No.III.1.c of the 1998 Antarctic Treaty, each country is required to establish a National Antarctic Data Centre (NADC) and to properly provide the data collected from involved scientists. The PDC at NIPR has performed the function of a NADC for Japan. The PDC established a data policy in February 2007, based of the requirements of the Standing Committee on Antarctic Data Management (SCADM) of SCAR. This contributed to the subsequent SCAR Data and Information Management Strategy (SCAR-DIMS; Finney, 2009; de Bruin, T., & Finney, K., 2011).

Regarding aurora data, in particular, we have administered the World Data Centre (WDC) for Aurora since 1981. The WDC for Aurora (<http://polaris.nipr.ac.jp/~aurora/>) is responsible for data archiving and dissemination of all-sky camera observations, visual observations, other optical observations, auroral image and particle observations from satellites, geomagnetic observations, and observations of the upper atmosphere phenomena associated with aurora such as [Ultra Low Frequency \(ULF\)](#), [Very Low Frequency \(VLF\)](#) and [Cosmic Noise Absorption \(CNA\)](#) activities (Kanao, Kadokura, Yamanouchi, & Shiraishi, 2008).

Outside these obligations, the PDC is responsible for archiving and analysis of Earth observing satellite data (Polar Operational Environmental Satellite: POES of NOAA), seismological data (short-period and broadband seismometers) and crustal movement data (Global Positioning System; GPS, Very Long Baseline Interferometry; VLBI) around the Syowa Station (SYO, 69S, 30E), East Antarctica. Finally, the PDC manages various information infrastructures such as: (1) a mainframe and a workstation system, (2) network systems of domestic and related facilities such as Syowa Station, and (3) Earth observing satellite data reception facilities.

## 3 METADATA MANAGEMENT

The PDC has the significant task to archive and deliver the digital data obtained from the polar regions. Summary information of all the archived data (metadata) is available to the polar science community as well as more general interests. The compiled metadata describe all kinds of observed/collected science disciplines (space and upper atmospheric sciences, meteorology and glaciology, geoscience and bioscience) from both long- and short-term projects in the Arctic and Antarctic, particularly data collected by the Japanese Antarctic Research

Expedition (JARE) (Kanao, Kadokura, Yamanouchi, & Shiraishi, 2008). In the science meta-database provided by PDC, a total of 150 metadata records had been accumulated as of October 2011 including metadata from IPY endorsed projects (<http://scidbase.nipr.ac.jp/>). A new content management system for providing the metadata has been in place since April 2011.

The science database provided by PDC has a tight connection with the Antarctic and Arctic Master Directories (AMDs) in the Global Change Master Directory (GCMD) of the National Aeronautics and Space Administration (NASA) (Fig. 1). In addition to the IPY-related data, data from Japanese national and other international projects have been compiled. Moreover, 210 metadata records have been compiled in the Japanese Antarctic portal (URL; [http://gcmd.nasa.gov/KeywordSearch/Home.do?Portal=amd\\_jp&MetadataType=0](http://gcmd.nasa.gov/KeywordSearch/Home.do?Portal=amd_jp&MetadataType=0)) in GCMD.

PDC stores its metadata in our own original format, but this includes the main items listed in the GCMD Directory Interchange Format (DIF). There are tight cross-linkages in corresponding metadata held in the AMD and PDC. Metadata collected by IPY projects for Japan have also been compiled in an IPY portal within the GCMD (<http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=ipy&MetadataType=0>). More than 140 metadata records contributed from Japan are in the IPY portal as of October 2011. This constitutes a significant proportion of all IPY metadata contributed to the GCMD.

It is also noted that there is an Arctic metadata portal in GCMD, describing data about Japanese activities in the Arctic ([http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=arctic\\_jp&MetadataType=0](http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=arctic_jp&MetadataType=0)), but the portal server includes only 15 records at the moment.

## 4 POLAR INFORMATION COMMONS

The Standing Committee on Antarctic Data Management (SCADM) under SCAR has been strongly connected with the activities of the IPY data-management community (IPY Data and Information Service; IPY-DIS). The IPY data policy ([http://classic.ipy.org/Subcommittees/final\\_ipy\\_data\\_policy.pdf](http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf)) emphasizes the need to make data available on the “shortest feasible timescale.” Rapid changes in the polar regions, particularly in the Arctic, make this need to share data more acute because alone, no single investigator or nation can understand these changes. In accordance with the IPY data policy, the data community (IPY-DIS) explicitly recommends that data be formally cited when used, and the IPY Data Committee developed initial guidelines for how data should be cited (Parsons, Duerr, & Minster, 2010). These guidelines harmonize different approaches and they have been adopted by many data centers around the world.

After the end of IPY, a new initiative, the Polar Information Commons (PIC), began as a framework for open and long-term stewardship of polar data and information (Parsons, et al., 2011a). The PIC would serve as an open, virtual repository for vital scientific data and information, and would provide a shared, community-based cyber-infrastructure fostering innovation, improved scientific understanding, and encourage participation in research, education, planning, and management in the polar regions. The PIC builds on the legacy of the IPY and also seeks active participation and ideas from national governments, international organizations, and the scientific and data management communities at large to build this common resource. The PIC was initiated by the International Council of Science (ICSU) Committee on Data for Science and Technology (CODATA) in collaboration with several multidisciplinary science bodies including the World Meteorological Organization (WMO), the International Arctic Science Committee (IASC), the International Union of Geodesy and Geophysics (IUGG), SCAR, Creative Commons, and others. The PIC was officially launched during the IPY Conference in Oslo, June 2010.

The PIC has developed specialized tools that produce a small, machine-readable “badge” that is attached to the metadata or data. This badge asserts that the data are openly available and allows generic search engines or customized portals to automatically identify and locate relevant data, but the badge also requests data users to adhere to basic ethical norms of data use including proper data citation. This service is coupled with a cloud-based data repository for data that may not have a suitable archive elsewhere (<http://www.polarcommons.org/>). NIPR and other Japanese organizations have made significant contributions to the PIC, both by attaching the data/metadata badges and by registration in the cloud-based repository. As of October 2011, Japan was a leading PIC participant and had contributed more than 50 data sets to the PIC.

Polar data can have great relevance for modern, global environmental research well beyond the polar regions. It is critical to explore new approaches like the PIC to develop an effective framework for open, and long-term stewardship of polar data. Data coming from the poles and elsewhere will continue to grow in size and

complexity. The experience of handling IPY data can serve as a valuable case study to examine data management approaches seeking to address issues around complex interdisciplinary science (Parsons, et al., 2011b).

## **5 WORLD DATA SYSTEM**

Through a decision of the 29th General Assembly of ICSU in 2008, a new World Data System (WDS) was established based on the 50-year legacy of two ICSU science bodies — the World Data Centers and the Federation of Astronomical and Geophysical Data Analysis Services. The new WDS aims at a transition from existing standalone WDCs and individual services to a common, globally interoperable, distributed data system that incorporates emerging technologies and new scientific data activities, including polar data as a legacy of the IPY. The new system will build on the potential offered by advanced interconnections between data-management components for disciplinary and multidisciplinary applications.

More than 100 data centers expressed interest to join the new WDS (<http://www.icsu-wds.org/>). The WDC for Aurora, in PDC of NIPR, also expressed interest in joining the new WDS. In October 2010, the ICSU Executive Board accepted the offer from the Japanese National Institute of Information and Communications Technology (NICT) to host and financially support the International Program Office (IPO) for WDS. The office manages and coordinates the establishment and operations of WDS and takes responsibility for outreach and promotional activities.

The first ICSU WDS Conference - Global Data for Global Science – was successfully held at Kyoto University in September 2011, under collaboration with CODATA and the Integrated Risk and Disaster Research (IRDR) of ICSU. It was the first international WDS meeting, and it sought to construct a smooth human network with advanced interconnections between data-management components for disciplinary and multidisciplinary applications across the globe.

The WDS policy of full and open access to data will benefit the international scientific community and ultimately society at large. Many concepts of data publication and data citation should be adopted and promoted by the WDS to facilitate timely release of data. The WDS has agreed to take the necessary steps to archive IPY data and to work with the PIC to preserve, curate, and add value to data in the PIC cloud in order to preserve the legacy of data of the IPY (WDS-SC, 2009).

## **6 CONCLUSION**

The status of IPY data-management in Japan has been summarized in this short paper. Many dedicated data service tasks have been conducted by the staff of PDC in NIPR as a member of NADC under SCAR. Several different aspects of scientific data collected in the polar region have great significance for global environmental research in this century. To construct an effective framework for long-term strategy of the polar data, data must be made available promptly and new Internet technologies such a repository network service like the PIC must be employed.

In addition to the activities in polar science communities of SCAR and the International Arctic Science Committee (IASC), tighter linkages must be established with other cross-cutting science bodies under ICSU, such as CODATA, and WDS. Linkages among these data-management bodies need to be strengthened in the post IPY era.

## **7 ACKNOWLEDGEMENTS**

The authors would like to express their special appreciation to a significant number of collaborators associated with the IPY activities both in national and international projects. They also acknowledge the members of SCADM of SCAR, as well as the IPY Data sub-committee under the IPY-JC for their great efforts to adhere to the data-management issues during the IPY. The authors appreciate the committed individuals of WDS and CODATA for their fruitful discussion and arrangement to initiate the PIC, as well as to create the new data strategy of ICSU. The authors would like to express appreciation to Prof. T. Watanabe and other members of WDS-SC for their arrangement in publishing a special issue on CODATA Data Science Journal, as a proceeding

of the 1<sup>st</sup> ICSU WDS Conference in Kyoto, 2011.

## 8 REFERENCES

de Bruin, T., & Finney, K., (2011) The SCAR Data Policy. *SCAR newsletter*, 27, p3.

Finney, K., (2009) SCAR Data and Information Management Strategy (DIMS) 2009 – 2013. In Summerhayes, C., & Kennicutt, C., (Eds.), *SCAR Ad-hoc Group on Data Management*, 34, Cambridge, Scott Polar Research Institute.

Kanao, M., Kadokura, A., Yamanouchi, T., & Shiraishi, K., (2008) The Japanese National Antarctic Data Centre and the Japanese Science Database. *JCADM newsletter*, 1, p10.

Parsons, M.A., de Bruin, T., Tomlinson, S., Campbell, H., Godoy, Ø., LeClert, J., & the IPY Data Policy and Management Subcommittee, (2011a) The State of Polar Data—the IPY Experience. In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., Rachold, V., Sarukhanian, E., & Summerhayes, C., (Eds.), *Understanding Earth's Polar Challenges: International Polar Year 2007-2008 –Summary by the IPY Joint Committee-*, 3.11, 457-476, Edmonton, Alberta, Art Design Printing Inc..

Parsons, M.A., Godoy, Ø., LeDrew, E., de Bruin, T., Danis, B., Tomlinson, S., & Carlson, D. (2011b) A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science. *J. Information Sci.*, 1-21, DOI: 10.1177/0165551500000000.

Parsons, M.A., Duerr, R., & Minster, J.-B., (2010) Data Citation and Peer Review. *EOS Transactions*, American Geophysical Union, 91(34), 297-298.

Rapley, C., Bell, R., Allison, I., Bindschadler, P., Casassa, G., Chown, S., Duhaime, G., Kotlyakov, V., Kuhn, M., Orheim, O., Pandey, P. C., Petersen, H., Schalke, H., Janoschek, W., Sarukhanian, E., & Zhang, Z., (2004) A Framework for the International Polar Year 2007–2008. *ICSU IPY 2007–2008 Planning Group*. ICSU, Paris, pp 57.

Sato, N., Ito, H., Kanao, M., Kanda, H., Naganuma, T., Ohata, T., Watanabe, K., & Yamanouchi, Y., (2011) Engaging Asian Nations in IPY:Asian Forum for Polar Sciences (AFoPS) (Japanese Section). In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., Rachold, V., Sarukhanian, E., & Summerhayes, C., (Eds.), *Understanding Earth's Polar Challenges: International Polar Year 2007-2008 –Summary by the IPY Joint Committee-*, 5.3, 555-574, Edmonton, Alberta, Art Design Printing Inc..

World Data System - Scientific Committee (WDS-SC) (2009) IPY & WDS -Executive Summary-. 2/7.1 Ver. 1.1

# BEYOND DATA REGULATION: FINDING SOLUTION TO A PERSISTENT PROBLEM OF MARINE DEBRIS AND SEA SURFACE TEMPERATURE MEASUREMENT ALONG COASTLINE OF LAGOS, NIGERIA

*O A Ediang<sup>1\*</sup> and A A Ediang<sup>2</sup>*

*<sup>1</sup>Marine Division, Nigerian Meteorological Agency, PMB1215 OSHODI, Lagos, Nigeria  
Email: ediang2000@yahoo.com*

*<sup>2</sup>The Nigerian Maritime Administration and Safety Agency, 6 Burmal Road, Apapa, Lagos, Nigeria.  
Email: ediang2005@yahoo.com*

## ABSTRACT

*We discuss in this paper environmental changes along the coastal line of Nigeria, especially in the region around Lagos, basing on provisional multi-disciplinary analyses of meteorological and maritime observations. The study has revealed that the environmental change in the Nigerian coastal region has been much more apparent than before (i.e. some few years back 1989-2007). Various kinds of ocean debris, transported mainly by coastal wind, are affecting marine and coastal environment severely. Since the current ocean monitoring system is found to be troubled by ocean debris, it is urgent to establish a new system to obtain reliable observational data to monitor and preserve the environment of the coastal region.*

**Keywords:** Marine Environment, Nigeria, Coastal degradation, Data Analysis.

## 1 INTRODUCTION

Meteorological data have shown that sea surface temperature (SST) plays an important role over the coastline of Nigeria and that SST and the rainfall are linked with each other. Afiesimama (1996) and Indeje (1995) established a relationship between the Pacific Ocean surface temperatures and rainfall over parts of East Africa. An Enso episode is the primarily evidenced through the appearance of sea SST anomalies (WMO, 1996). Adedokun (1978) has noted that upwelling process that takes place, for instances off the Accra Coast, can be weakened or strengthened by increase or decrease in SST respectively, which can result from a weakening or strengthening of the south westerly winds. Edafienene et al. (1997) observed that the temperature of Nigerian coastal water is warmest in April and coldest in August, using SST data for the period of 1989-1997.

In this paper, we discuss sharing data and information about marine debris including a wide variety of man-made items that persist in the marine environment. While ship wrecks and other artifacts indicate that man-made items are already present in the marine environment, the social and technical changes in modern times have added a new dimension to the marine debris problem. We attempt to highlight multi-disciplinary data analysis in finding solutions to a persistent problem of marine debris and SST measurements along the coastline of Lagos, Nigeria.

## 2 STUDY AREA

The Nigeria coastline runs about 860 km along the Atlantic Ocean. It is bounded in the west by the Republic of Benin and in the east by the Republic of Cameroon. It lies generally between latitudes 4<sup>0</sup>10'N and 6<sup>0</sup>20'N, and longitude 2<sup>0</sup>45'E and 8<sup>0</sup>35'E, adjacent to the Gulf of Guinea. A map of the coastal region of Lagos, Nigeria, and an example of aerial photographs of the coastline (from Google Map) are shown in Figure. 1. A common feature of the coastline is its low-lying nature. The coastline has been classified into four broad regions according to differences in general morphology, vegetation and beach types. The regions from west to east are; the Barrier Lagoon Coast, the Transgress Mud Coast, the Niger Delta and the Strand Coast. The Victoria Island beach is known to be a part of the Barrier Lagoon Coast. This region is located to the east of the Eastern Breakwater (East Mole) of the down drift side of the inlet into the Lagos Harbor. To the east of this island, the Kuramo Waters and the Igbosere Creek are located. The Nigerian Coastline is bounded to the north by Five Cowries Creek and to the south by the Atlantic Ocean, where Tin Can and Apapa Port are located. The island beach has suffered degradation resulting from a number of natural and anthropogenic causes in the past two and half decades.



**Figure 1.** A map showing the geomorphology of the Lagos coastal area, Nigeria (left), and an aerial photo of a lagoon (Kuramo Waters) in Lekki Peninsula (right). The aerial photo is taken from Google Map.

### 3 TRANSPORTATION OF MARINE DEBRIS ALONG THE COASTLINE OF LAGOS

The frequency of anomalous transportation of marine debris along the coastline of Nigeria, especially the coastline of Lagos, has increased in the last decade. Marine debris can enter into the marine environment through a variety of vectors. Land-based debris can be transported to the marine environment as a result of:

- Urban runoff,
- Sewer overflow,
- Inadequate garbage management
- Industrial activities
- Terrestrial dumping and littering activities.

Sea based vectors include cruise ships, cargo ships, recreational boats, fishing vessels, and plant forms. The problem, which marine debris prevention along the coastline of Lagos, is that the ocean's ability to move and circulate the debris. The combination of ocean currents and atmospheric winds can transport debris across great distances. It can also retain and concentrate items for later deposition.

## 4 DATA AND METHODOLOGY

### 4.1 Data sources

In this analysis, we use data of the SST and the annual maximum temperatures in the coastal region of Lagos, Nigeria, in the period from 1952 to 2007. We also use daily wind data (speed and direction) at 0900z which were obtained for a period of 11years (1997 -2007) at a maritime meteorological station at Victoria Island, near Lagos harbor. This dataset was provided by the Nigerian Meteorological Agency. Outcomes from independent statistical analyses conducted by the Federal Office of statistics, Lagos and the Institute of Oceanography in Victoria Island are used to compare with our results.

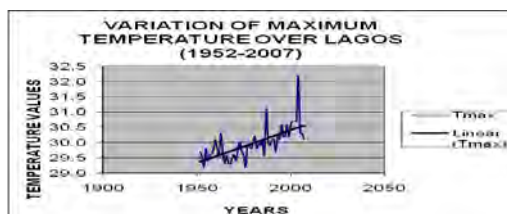
### 4.2 Data analysis

Analyses of the trend and the pattern of sea SST variations were carried out. The monthly averaged data for the period 1989-2007 were statistically treated to obtain the mean yearly SST values which gives adequate and necessary information on the changes in the Nigerian coastal areas. When the SST increased above the mean level in the analyzed interval, the amount of marine debris decreased. On the other hand, the movement of marine debris was significantly enhanced when a sharp decrease in the SST occurred.

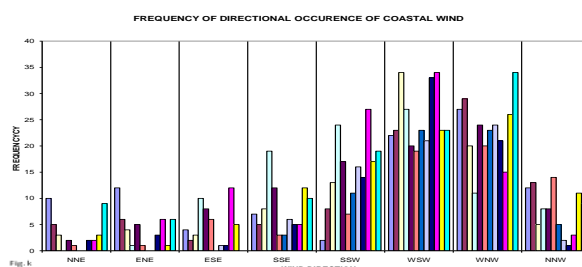
The trend of the annual maximum temperature in the coastal region of Nigeria in 1952-2007 is shown in Figure 2. This analysis shows a clear increasing tendency of the annual maximum temperature as indicated by the linear fit to data, but the increasing tendency is suggested to have been accelerated in 1990s. The increase in temperature results a decrease in the atmospheric pressure, and the movement of marine debris increases, while decrease in



temperature results an increase in pressure, and marine debris movement decreases. Figure 3 shows the monthly frequency of the wind direction in the coastal region of Nigeria, in 1997-2007. It is apparent that the south-westerly winds were dominant in this region over the year. Winds play a huge role in the occurrence and movement of marine debris. Ocean debris is transported mainly by the westerly coastal wind in this region. As shown in Figure 3, winds blew predominantly from WSW and WNW directions. However, there were no cases of wind speeds greater than 19 kts. Within the period of this analysis, the weakest wind speeds were observed in the months between October and January, while the strongest winds were mostly observed in August.



**Figure 2.** Variations of annual maximum temperatures in the coastal region of Lagos, Nigeria, in 1952-2007.



**Figure 3.** Monthly frequencies of wind directions in the coastal region of Nigeria in 1997-2007. The horizontal axis is divided into 8 bins of wind directions. Bars in each bin show the monthly frequencies from January to December, respectively.

## 5 DISCUSSION

Our investigation revealed that the island beach of Nigeria has suffered degradation resulting from ocean surges, influenced by strong winds of a number of natural and anthropogenic causes for the past two and half a decades. Along the coastline of Lagos, persistent debris has played a significant role in the degradation of the marine environment. Such persistent materials include product packages and single-use beverage containers such as aluminum cans and bottles. These items injure and kill marine species. The negative effect is suffered not only by animals, but also by humans who interact with these coastal resources. SST measurement sare frequently influenced by marine debris, and the effect results missing data and incorrect measurements because marine meteorological stations are using rubber sea-temperature bucket thermometers on an hourly basis. The Investigation which was carried out on the coastline of Nigeria revealed that ocean surges were strong in the months of April to October, by the influence of prevailing South-Westerly wind not less than 18 knots. Over the area of  $10^{\circ}5N - 2^{\circ}5N$  in the latitude and  $0^{\circ}E - 10^{\circ}E$  in the longitude, high spring tides of about 1.8 m or more are predicted using a parametric wave model (Afiesimamaet al., 2000). Also, the amount of marine debris along the coastline of Lagos shows seasonal variation and usually highest in the summer months, April-October, due to strong westerly wind (Figure 3).

## 6 COLLABORATIVE DECISION MAKING TO PREVERT MARINE DEBRIS IN COASTLINE OF LAGOS

As technology advances, the concept of Collaborative Decision Making (CDM) has been proved to be useful. In anutshell, CDM will be effective when users of services have a chance to add their expertise to the decision making process. For the process to work effectively, it is important to have tools to view the information seamlessly and data are readily available. Some of Nigerian organizations have been involved; e.g. the Nigerian Maritime Administration and Safety Agency, NIMASA, Lagos State Environmental Protection Agency (LASEPA) and the Nigerian Meteorological Agency. The Nigerian Maritime Administration and Safety Agency,

NIMASA, are committed to the enthrone of global best practices in the provision of maritime services in Nigeria. The core functions include marine pollution protection, marine pollution control, waste management facilities, and marine environment management. Also the Lagos State Environmental Protection Agency (LASEPA) inherits the responsibilities of the Pollution Control Unit, in addition to the functions that were indicated in the edict that later established the agency. The Nigerian Meteorological Agency (NIMET) is a federal government parastatal established by the Act No.9 of 2003. The agency is responsible for production of weather, water and climate information for socio-economic development of Nigeria. The Marine Division in the Applied Meteorological Services of the agency plays its role in information and data to prevent marine debris in the Lagos coastline.

## 7 CONCLUSIONS

Effects of marine debris are seen in the mortality among marine species along the coastline of Nigeria especially the coastline of Lagos, and also influence and affect the measurement of SST. The study has revealed that the change in the Nigerian coastal climate is much more apparent than before (i.e. some few years back 1989-2007). Taking some past years into consideration, it can be seen that the rate at which the coastal change in Nigeria is higher than that before 1989, but in the Nigerian coastal area over Lagos, the rate was even below the average before 1989. However, it shows in the recent years that it is even more above the average. The wind plays an important role in the transport of marine debris along the coastline of Lagos, Nigeria, i.e. the swells (waves generated at a distance) became very active while the sea state becomes slight to moderate over the South Atlantic Ocean extending to North Atlantic. However, the provision of adequate equipment by the Nigerian Meteorological Agency will be a spring board for improvement of monitoring system in the Nigerian Coastal areas. This will enable data to be readily available for research work. Our recommendation to enforce the monitoring system in the Nigerian coastal area is given in Appendix.

## 8 REFERENCES

- Adedokun, J.A. (1978) West African Precipitation and Dominant Atmospheric Mechanism: *Arch. Met Geoph. Biometeorol*, Ser; A 27, 289 – 310
- Afesimama, E.A. (1996) On the variability of the mean sea surface temperature and its influence on the rainfall pattern over coastal station, *West Africa pioc of the 4<sup>th</sup> inter conference on school and popular met and Oceanographic*, Royal-met Soc. Publication, p 321.
- Gbuyero, E. & Afesimama, E.A. (1997) Sea surface temperature analysis from *institu* data at East Mole, *Lagos Nigerian Meteorological Society proceedings of the 12<sup>th</sup> and 13<sup>th</sup> Department Symposia*, p 35.
- Indeje M. (1995) Pacific Ocean sea surface temperatures and east African seasonal rainfall. *Second International Symposium on Assimilation of Obs. Met. and Oceanography*, Vol. II WMO /TD No. 651, p 655.
- WMO (1996) WMO Statement on the status of the global climate in 1995, WMO – No 838.

## 9 APPENDIX

### Recommendation

The Nigerian meteorological Agency and the Institute of Oceanography and Marine research should do more in research works because they understand the impact of marine debris, and can proffer solution based on scientific knowledge and oceanographic conditions. Advanced models developed in Europe and UK will be able to predict the movement of marine debris in coastal areas of Nigeria. The immediate challenge for Nigeria Meteorological Agency and other stake holders in marine industry is to set up and maintain systems which collect data, process, store, retrieve and disseminate them as necessary, especially data related to sea surface temperature and marine debris. Drawing strong inference from this, there is no doubt that the effort toward ensuring a sustainable environment quality and healthy economy is highly desirable and this can be achieved only if there is a co-operation among the government, general public, and the science world. New research on sea surface temperature and its relation to marine debris is needed and the government should continue to support existing private sector.

# VISUALIZATION OF FLUX ROPE GENERATION PROCESS USING LARGE QUANTITIES OF MHD SIMULATION DATA

*Y Kubota<sup>1\*</sup>, K Yamamoto<sup>1</sup>, K Fukazawa<sup>2</sup>, and K T Murata<sup>1</sup>*

*<sup>1</sup>National Institute of Information and Communications Technology, Tokyo, Japan*

*Email: ykubota@nict.go.jp, kaz-y@nict.go.jp, ken.murata@nict.go.jp*

*<sup>2</sup>Research Institute for Information Technology, Kyushu University, Fukuoka, Japan*

*Email: fukazawa@cc.kyushu-u.ac.jp*

## ABSTRACT

*We will present a new concept of analysis by using visualization of large quantities of simulation data. The time development of 3D object with high temporal resolution provides the opportunity of scientific discovery. We visualize large quantities of simulation data using visualization application 'Virtual Aurora' based on AVS and the parallel distributed processing at "Space Weather Cloud" in NICT based on the Gfarm technology. We introduce two results of high temporal resolution visualization which are magnetic flux rope generation process and dayside reconnection using a system of magnetic field line tracing.*

**Keywords:** visualization, cloud computing system, parallel distributed processing, space weather, reconnection

## 1 INTRODUCTION

According as super computer's ability increases we can treat high precision simulation data while we need to visualize the big simulation data. We progress the technique of 3D visualization. Matsuoka et al. (2008) analyzed 3D structures of magnetic flux rope in terrestrial magnetosphere by using 3D visualization techniques. However they visualized simulation data in the specific time and space region in which we need to see physical interest because the computational resource has limit. We succeeded to visualize all time and space by using big storage and parallel distributed processing (Murata et al., 2011). The visualization technique provides the opportunity of scientific discovery because we can see small and large scale structures at the same time. We performed two type visualizations in difference of how to draw magnetic field lines. One is that start points of magnetic field lines are fixed in time. The other is that start points of magnetic field lines move with stream elements following the velocity. 'Frozen-in' concept is that magnetic field line moves on stream elements except magnetic diffusion region. The later visualization means magnetic field line tracing. In chapter 2, we describe how to visualize all step data. In chapter 3, we introduce the visualization result of magnetic flux rope generation in fixed start points and dayside magnetic reconnection by using magnetic field line tracing.

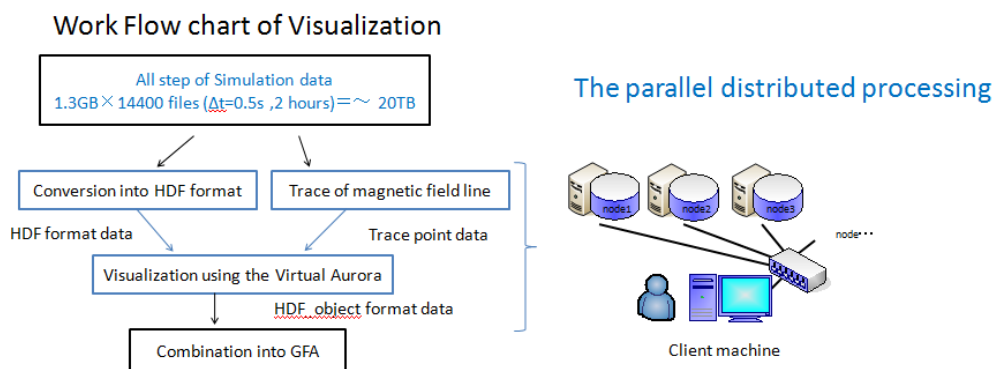
## 2 3D OBJECT WITH HIGH TEMPORAL RESOLUTION

### 2.1 Global MHD simulation data

We use the data generated by an MHD (magnetohydrodynamic) simulation of the interaction between the solar wind and the terrestrial magnetosphere. The number of data points is  $450 \times 300 \times 300$  in the Cartesian coordinates. The simulation covers the time interval of 2 hours with a time resolution of 0.5 sec. Namely the number of simulation steps is 14400. The space grid interval is  $0.2R_E$  where  $R_E$  is earth radius length. Total data size is about 18TB. The data of ACE satellite with 5 minute interval at Galaxy15 (2010/4/5) event is used as the solar wind input. The outer boundary is that the variables are fixed for upstream and the variables are free for downstream. The inner boundary is that the variable is the same of initial condition in radius of  $4R_E$ . In the range from  $4R_E$  to  $5R_E$ , the simulation values are smoothly combined with the initial values.

### 2.2 Visualization techniques using parallel distributed processing

In order to treat the big data, we use the parallel distributed processing at “Space Weather Cloud” in NICT based on the Gfarm technology. The number of processor is above 100 cores. We visualized the simulation data of Galaxy event as drawing the number of above 1000 field lines using visualization application 'Virtual Aurora' developed based on AVS. The work flow of visualization is shown in Figure 1. First we convert the simulation data to HDF format data for Virtual Aurora. Second we visualize the HDF format data using 'Virtual Aurora'. Third we output the result as each time step 3D object file. Last we combine each time step 3D object into the high temporal resolution 3D object which has a large amount of information volume. In case of magnetic field line tracing, we need to read all files in order to trace the streak line. These work flows are complex for the parallel distributed processing. In order to describe the work flow in program, we use Parallel Workflow extension for Rake (Pwrake), which is a tool for the parallel distributed processing. We describe the flexible work flow by using Ruby. In order to look the time development of 3D object, e.g. a file size of 2000 step GFA is 3GB, we developed the 64bit 3D object player in NICT which can zoom the objects and change the view-direction easily.



**Figure 1.** Work Flow chart of Visualization using the parallel distributed processing

### 3 Visualization of simulation results

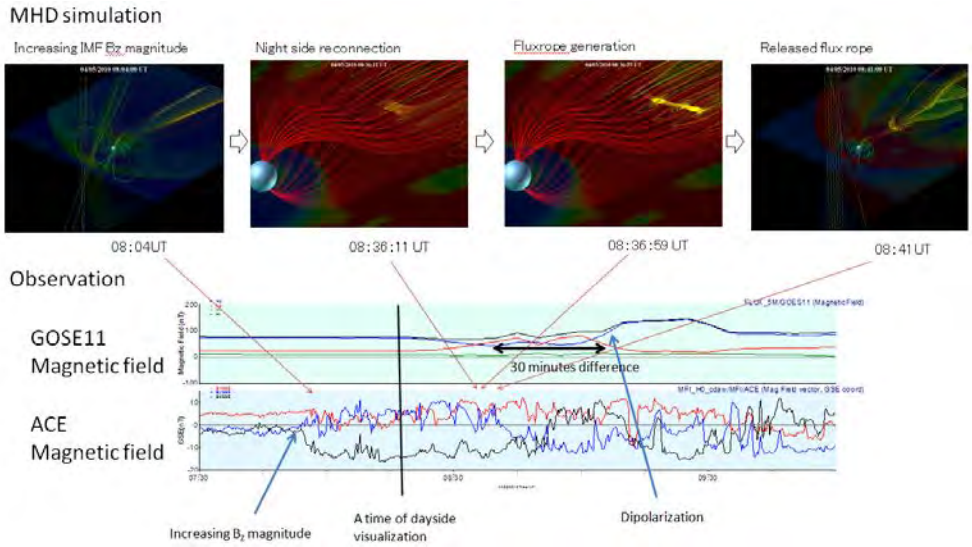
First visualization using the high temporal resolution visualization system with fixed start points is about the generation process of a flux rope, which is a small structure and begins from a spiral field line by magnetic reconnection, related with a release of the solar wind energy at the magnetotail. This knowledge can be acquired by the looking at small and large scale structures with high temporal resolution. Second visualization using the system of magnetic field tracing is about the dayside reconnection. In order to visualize reconnection region, we need to trace the magnetic field lines. We visualized the reconnection region.

#### 3.1 Visualization of magnetic flux rope

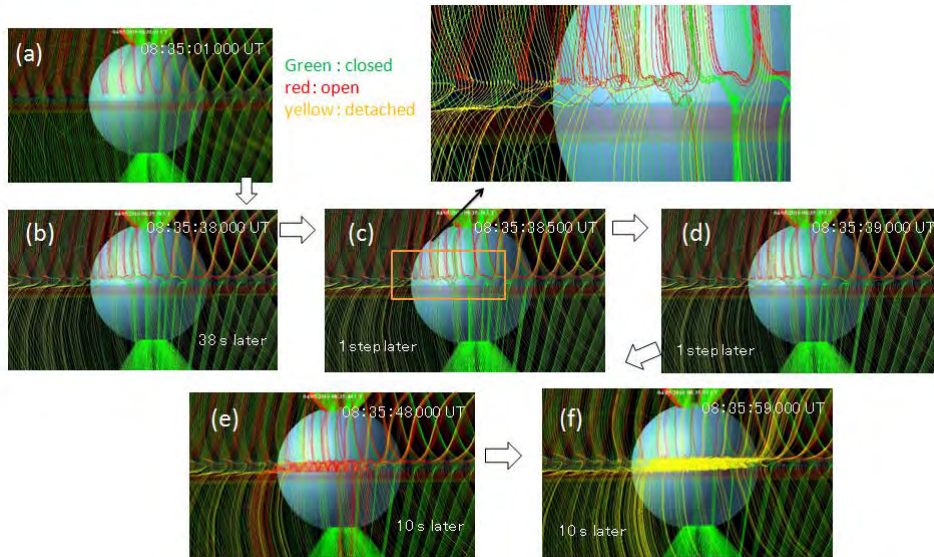
We show a result of galaxy event simulation and observation for two hours. Figure 2 shows that four snapshots of simulation, the solar wind magnetic data observed by ACE satellite and tail magnetic data observed by GOSE11 satellite. The snapshots show pressure color contours of equatorial and meridian cross section and magnetic field line divided into open line (red), closed line (green) and detached line (yellow) according to topology of field line. Intensity of solar wind negative  $B_z$  increases at 8:00 UT as shown in ACE magnetic data. Intensity of magnetic field observed by GOSE11 increases at 9:00 UT. This increase indicates dipolarization. In the MHD simulation, topology of magnetic field lines changes open into detached line because field lines reconnect at about 8:36 UT and a flux rope generates. The flux rope releases to interplanetary space at 8:41 UT. There is a difference of timing about 30 minutes between observational dipolarization and flux rope release in simulation. This difference arises because this simulation does not include ionosphere effect as boundary condition and a magnetic diffusion coefficient is uniform anywhere in the simulation, that is, reconnection is easy to occur in the simulation. The simulation is consistent with the observation except the difference of timing.

We are interested in flux rope generation process. We investigate this process in high time resolution because we visualized this simulation using all step data. Figure 3 shows that topology of magnetic field lines at flux rope generation. We visualized first reconnection, that is, closed field line reconnects and changes into open field lines at 8:35:38.500 UT as shown in Figure (c) and enlargement figure. The open field line is spiral structure and generates flux rope after 10 seconds as shown in Figure (e). The open field lines reconnect and change into

detached field line after 10 seconds and the flux rope is released as shown in Figure (f). Topology of magnetic field lines change during 1 minute for flux rope generation.



**Figure 2.** Comparison between MHD simulation and GOSE11 observation at galaxy event.

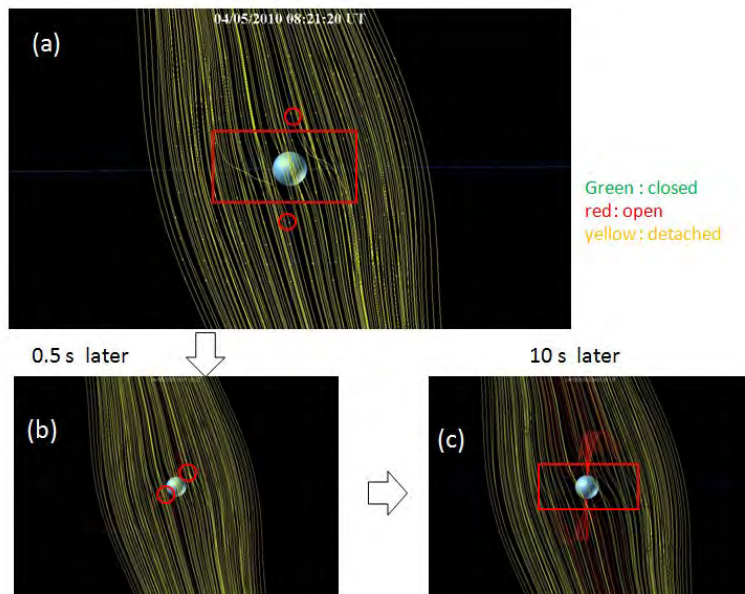


**Figure 3.** Topology of magnetic field lines during 1 minute when the flux rope generates.

### 3.2 Visualization of dayside reconnection using a system of magnetic field line tracing

Dayside magnetic reconnection is a main mechanism that transports solar wind energy into the magnetosphere. We visualize the reconnection region between interplanetary magnetic field (IMF) and terrestrial magnetic field by tracing the IMF. As a first result, we trace the IMF when the IMF is southward ( $B_z = -12\text{nT}$ ,  $B_y = 5\text{nT}$ ) at 8:20 UT. Figure 4 shows dayside reconnection region. We view in the direction from solar to earth. The blue sphere is earth. The yellow lines and the red lines are magnetic field lines. The topology of field lines is indicated by the same color as section 3.1. The yellow points are start points to draw the magnetic field line. Start points are 11point (down to dusk)  $\times$  11point (north to south) and an interval is  $2R_E$ . Figure (a) shows that the magnetic field lines starting from two red circles are bended in the dawn-dusk direction in order to diffuse between the solar wind and terrestrial field. The red square indicates a diffusion region. Diffusion region is from subsolar point to  $\pm 5R_E$  in the dawn-dusk direction. After 0.5 seconds, the field lines first reconnect with terrestrial field line as shown in Figure (b). Solar wind field lines reconnect with terrestrial field through this diffusion region. After 10 seconds, the magnetic field lines in the region from the subsolar point to  $\pm 5R_E$  in the dawn-dusk directions

reconnect with the terrestrial field lines at the dayside magnetopause as shown in Figure (c). In region of flank sides over  $\pm 5R_E$ , magnetic field lines do not reconnect and are transported through the sheath region to the downstream.



**Figure 4.** Visualization of the dayside reconnection region

## 4 CONCLUSION

We have visualized MHD simulation result in two type methods, which have a difference in start points treatment, in high temporal resolution by using parallel distributed processing. First we visualized generation of flux rope for galaxy event. We find that topology of flux rope's magnetic field lines changes during 1 minute. Second we succeeded in visualizing magnetic field line tracing to MHD simulation data. We find that the magnetic field lines in the region from the subsolar point to  $\pm 5R_E$  in the dawn-dusk direction reconnect at the dayside magnetopause and transport the nightside magnetosphere. In region of flank sides over  $\pm 5R_E$ , magnetic field lines do not reconnect and are transported through the sheath region to the downstream.

## 5 ACKNOWLEDGEMENTS

The present work was done by using resources of the OneSpaceNet (the NICT science cloud).

## 6 REFERENCES

Daisuke Matsuoka, Ken T. Murata, Shigeru Fujita, Takashi Tanaka, Kazunori Yamamoto, and Eizen Kimura (2008), Analyses of 3D Structure of Magnetic Flux Ropes via Global MHD Simulations, *Journal of Visualization*, 28(6), 38-46

Ken T. Murata, Shinichi Watari, Tsutomu Nagatsuma, Manabu Kunitake, Hidenobu Watanabe, Kazunori Yamamoto, Yasubumi Kubota, Hisao Kato, Takuya Tsugawa, Kentaro Ukawa, Kazuya Muranaga, Eizen Kimura, Osamu Tatebe, Keiichiro Fukazawa and Yasuhiro Murayama (2011), A Science Cloud for Data Intensive Sciences, *Proceedings of the 1st ICSU World Data System Conference*.

# Application of Information Technologies to Data Systems



# A SCIENCE CLOUD FOR DATA INTENSIVE SCIENCES

Ken T. Murata<sup>1\*</sup>, Shinichi Watari<sup>1</sup>, Tsutomu Nagatsuma<sup>1</sup>, Manabu Kunitake<sup>1</sup>, Hidenobu Watanabe<sup>1</sup>, Kazunori Yamamoto<sup>1</sup>, Yasufumi Kubota<sup>1</sup>, Hisao Kato<sup>1</sup>, Takuya Tsugawa<sup>1</sup>, Kentaro Ukawa<sup>1,2</sup>, Kazuya Muranaga<sup>1,2</sup>, Eizen Kimura<sup>3</sup>, Osamu Tatebe<sup>4</sup>, Keiichiro Fukazawa<sup>5</sup> and Yasuhiro Murayama<sup>6</sup>

<sup>\*1</sup>Space Weather and Environment Informatics Lab., National Institute of Information and Communications Technology (NICT), 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan

Email: ken.murata@nict.go.jp

<sup>2</sup>Systems Engineering Consultants Co. Ltd. 4-10-1, Yoga, Setagaya, Tokyo 158-0097, Japan

Email: kentaro.ukawa@nict.go.jp

<sup>3</sup>Dept. Medical Informatics of Medical School of Ehime Univ., Situkawa, Toon City, Ehime, 791-0295, Japan

Email: ekimura@m.ehime-u.ac.jp

<sup>4</sup>Dept. Computer Science, Univ. of Tsukuba, Tenoudai 1-1-1, Tsukuba science city, Ibaraki 305-8573, Japan

Email: tatebe@cs.tsukuba.ac.jp

<sup>5</sup>Research Institute for Information Technology, Kyushu University, 6-10-1 Hakozaiki, Higashi-Ku, Fukuoka, 812-8581, Japan

Email: fukazawa@cc.kyushu-u.ac.jp

<sup>6</sup>Integrated Science Data System Research Lab., NICT, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan

Email: murayama@nict.go.jp

## ABSTRACT

It is often discussed that the fourth methodology for science researches is “informatics”; the first methodology is theoretic approach, the second one is observation and/or experiment, and the third one is computer simulation. The informatics is expected as a new methodology for data intensive science, which is a new concept based on the fact that most of the scientific data are digitalized and the amount of the data are huge. The facilities to support informatics are cloud systems. Herein we propose a cloud system especially designed for science. The basic concepts, design, resource, implementation and applications of NICT science cloud are discussed.

**Keywords:** Science Cloud, informatics, Observation, Experiment, Computer simulation, Data intensive science, Data-oriented science, Large-scale storage, Super computer

## 1 INTRODUCTION

During these 50 years, along with appearance and development of high-speed computers (and super-computers), numerical simulation is considered to be a “third methodology” for science, following theoretical (first) and experimental and/or observational (second) approaches (Figure 1). The variety of data yielded by the second approaches has been getting more and more. It is due to the progress of technologies of experiments and observations. The amount of the data generated by the third methodologies has been getting larger and larger. It is because of tremendous development and programming techniques of super computers.

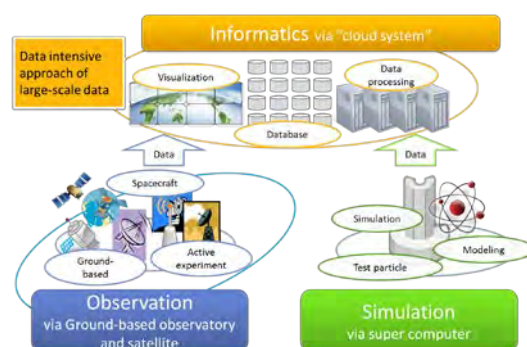


Figure 1. Informatics and Cloud System

Most of the data files created by both experiments/observations and numerical simulations are saved in digital formats and analyzed on computers. The researchers (domain experts) are interested in not only how to make experiments and/or observations or perform numerical simulations, but what information (new findings) to extract from the data. However, data does not usually tell anything about the science; sciences are implicitly hidden in the data. Researchers have to extract information to find new sciences from the data files. This is a basic concept of data intensive (data oriented) science.



As the scales of experiments and/or observations and numerical simulations get larger, new techniques and facilities are required to extract information from a large amount of data files. The technique is called as “informatics” as a fourth methodology for new sciences.

Any methodologies must work on their facilities: for example, space environment are observed via spacecraft and numerical simulations are performed on super-computers, respectively. The facility of the informatics, which deals with large-scale data, is a computational cloud system for science.

This paper is to propose a cloud system for informatics, which has been developed at NICT (National Institute of Information and Communications Technology), Japan. The NICT science cloud, we named as “OneSpaceNet (OSN)”, is the first open cloud system for scientist who is going to carry out their informatics for their own science.

## 2 NICT SCIENCE CLOUD (ONESPACENET)

### 2.1 Overview of Science Cloud

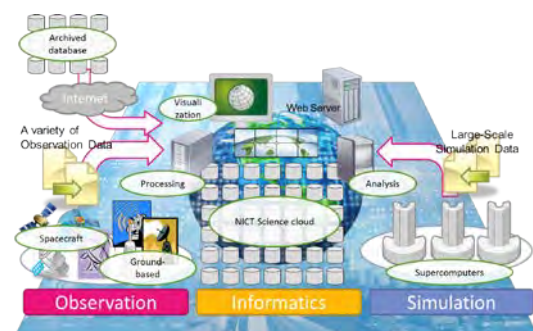
As is discussed above, as the size of data files in any types of science gets larger, we need a new paradigm to analyze the data: that is informatics, a fourth methodology for science as shown in Table 1. The facility to support the fourth methodology is the science cloud (Microsoft, 2009).

The science cloud is not for simple uses. Many functions are expected to the science cloud; such as data standardization, data collection and crawling, large and distributed data storage system, security and reliability, database and meta-database, data stewardship, long-term data preservation, data rescue and preservation, data mining, parallel processing, data publication and provision, semantic web, 3D and 4D visualization, out-reach and in-reach, and capacity building.

**Table 1.** Four methodologies and their facilities

	methodology	facility
First	theory	human being
Second	experiment/observation	e.g., spacecraft
Third	Numerical simulation	super computer
Fourth	informatics	cloud system

Figure 2 is a schematic picture of the NICT science cloud. Both types of data from observation and simulation are stored in the storage system in the science cloud. It should be noted that there are two types of data in observation. One is from archive site out of the cloud: this is a data to be downloaded through the Internet to the cloud. The other one is data from the equipment directly connected to the science cloud. They are often called as sensor clouds. One of the great advantages of the scientific cloud to other legacy systems is its integrated function. A large-scale disk area is provided with the users, but not necessarily for the data file storage. For instance, cluster systems with parallel data processing are also mounted in the NICT science cloud. Since each node is responsible for both data file node and data processing node, users don't have to copy (or move) large size data files to their data processing system sites. Note that it usually takes more than one week to copy data files with 10TB over the Internet.



**Figure 2.** A basic concept of NICT science cloud

## 2.2 Implementation of NICT Science Cloud

One of the definitions of cloud system is its multi-functionality; it has to satisfy a variety of requests from users. Science clouds must be, in general, more functional than commercial clouds; the providers of commercial cloud provide simple services which mainly work on virtual machines. Science clouds have to provide two types of different services: test-beds for developers of the cloud computing (science/technology for cloud) and facilities to perform high-level science (cloud for science/technology). It suggests that the science cloud has to be equipped with many resources to satisfy both science/technology of cloud and cloud of science/technology. The science cloud must be designed and implemented for such a variety of intensive data processing studies.

Figure 3 is a schematic picture of implementation of the NICT science cloud so far. The resources are installed by the end of 2011, thus will be developed step by step in the following years. The cloud system is composed of several clusters of computational resources deployed over the JGN-x network. The JGN (Japan Gigabit Network) is a wide-area network with 10Gbps or more, covering all Japan, from Hokkaido to Okinawa. Most of the access points (APs) are located at research institutions or universities.

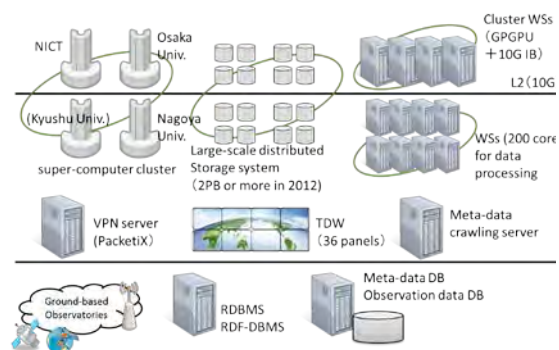
What should be paid attention to the NICT science cloud is that the network over JGN-x is an L2 (layer-2) network. Wide-area network systems are often constructed with L2, because of its easy maintenance and security. The L2 network also has an advantage to L3 in terms of the routing-less data transfer.

Since the NICT science cloud is a wide-area (domestically distributed) cloud, data transfer speed inside cloud is important to performances of the cloud. To avoid long-distance routing inside cloud through the routers widely distributed over JGN-X (over Japan), the L2 network is preferable to L3 network. Another reason is a security; an L2 network is a closed network, thus the traffics between nodes in the cloud stay inside the cloud.

The most important resource in Figure 3 is a set of large-scale storages. In the NICT science cloud, we deploy a set of distributed file storages over the JGN-x. These computing resources are discussed in Section 2.3.

In Figure 3 four super-computers are connected to the science cloud. Virtual super-computing environment also plays important roles in the science cloud, since they usually yield a large size of data to be processed and visualized. For such post-processing, parallel computing environment must be in use. Here, note that these parallel computers are either cluster type or hetero-type.

Large-size displays are also necessary for the use of the science cloud. Since the spatial sizes of numerical simulations get larger (Murata et al., 2007), high-resolution display are required to preview visualized data without data compressions (with full resolutions). This will be discussed in Section 3.4.



**Figure 3:** Implementation of NICT science cloud

## 2.3 Distributed Storage System

As discussed above, one of the most important resources in science clouds is data storage system. Since we need to store all of the digital data from both observations/experiments and simulations, the size and high performance of the data storage are crucial.

In the NICT science cloud, we construct a distributed data storage system named “OSN cloud storage”, equipped over the JGN-x. Since we have many access points over the JGN-x, we are able to deploy data storages at the data centers (DCs) of the JGN-x. To construct the distributed data storage system, we adopt the Gfarm (Grid Datafarm), which is a middleware for such wide-area distributed data storage system (Mikami et al, 2011; Kobayashi et al., 2011). The newest version of the Gfarm is 2.5, and we use 2.4.2 since it is most stable version of the Gfarm.

The deployment of the Gfarm DCs so far is displayed in Figure 4. There are several advantages of these widely distributed storage system using the Gfarm. Most important function of the Gfarm for the OSN cloud storage is data file redundancy. Once a user drops (saves) a data file on the OSN cloud storage, the system automatically

make replications of the file saved at different nodes of the system. When another user accesses the data file, the system provides him/her with most accessible data file (closest to the user). The latency between Hokkaido and Okinawa, the distance between them is more than 3000km, is negligible. Using this system, a user at Hokkaido accesses the file instance located in Hokkaido DC and the user at Okinawa accesses that in Okinawa DC.

What is another important function based on the redundancy of the OSN cloud storage is its backup free service. In Figure 4, there are two replications of each data file automatically created (totally 3 files on the Gfarm system) as discussed above. Assume that one of the data file instance is broken and lost. In this case, the system detects this lost, and makes another replication of the file as soon as possible (see Figure 4). Eventually, the number of the copies of the file remains three on the system. This suggests that neither backup nor restoring is necessary. It usually takes few days or one week to restore a large size, for example 10TB, of backup data files. This often stops researchers' works. The concept of BCP (Business Continuity Plan) or BCM (Business Continuity Management) should be applied not only for business, but for scientific works. To obtain good research results, such continuity is expected as a research environment.

The availability of the distributed storage system is also important from a viewpoint of cost performance. Since the price of data storage (Hard Disk Drive: HDD) is getting lower. This suggests that scalable storage addition is more reasonable than equipping large-size storage at once. Figure 5 is a time-dependent graph of the OSN cloud storage size since October, 2009. Note that the total storage size changes frequently. This means that we have often added or removed small size file system nodes (with about 50TB) on the OSN cloud storage system without terminating service. In the OSN cloud storage system, one file node with 50TB cost about \$5,000 (US dollar); this means the cost for 1TB is about \$100. We don't have to get large budget to develop the storage system.

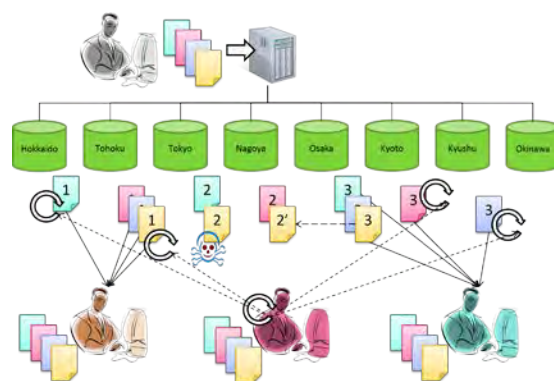
### 3 APPLICATIONS OF THE NICT SCIENCE CLOUD

#### 3.1 Overview

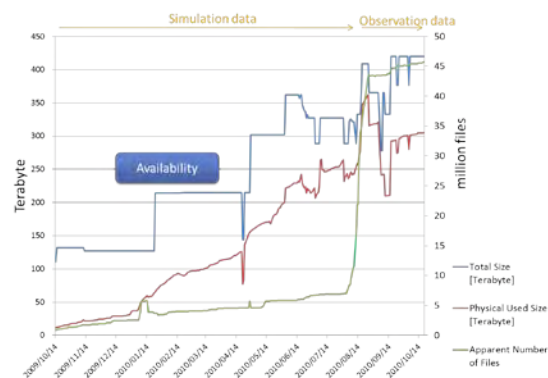
As discussed above, science cloud must be applied for a variety of data intensive research works. In this section, we discuss few examples of good use of the NICT science cloud. Since it is almost one year after opening of the science cloud, we have not archived any outstanding results on the cloud. Most of them are, thus, initial reports. However, these initial results are valuable since they are not derived without supports or uses of the science cloud.

#### 3.2 Large-scale Visualization

One of the effective targets of use of the science cloud is visualization of numerical simulations. As the



**Figure 4:** Distributed Storage in NICT science cloud (OSN cloud storage) at data centers (DCs). The DCs at Hokkaido, Tohoku, Kyoto and Kyushu are under contemplation.



**Figure 5:** Availability of OSN cloud storage

development of super-computers, the size of simulation data is getting larger. The recent trend of the spatial size of numerical simulation is  $1000^3$  (Giga) as the main memory size increases. Herein we consider a computer simulation via Global MHD simulation code (Fukazawa et al., 2006). The  $1000^3$  grid number corresponds to about 10 to 100GB data size since 10 to 100 components are allocated on each grid.

In most of the time-dependent computer simulations like Global MHD simulations, we have discarded most of the numerical data with sampling in time. For example, the time resolution of the present simulation is 0.5 sec., but usual sampling time of visualization is 1 min. It suggests that 99.2% of total simulation data are not used.

The major reason to discard most of the data is post-processing; the data size for all of the data is almost 15TB for 2 hours simulation with 0.5 sec. time resolution. Figure 6 shows comparison of post-processing times between legacy method and parallel visualization via NICT science cloud. Even if one has a large-scale storage for this 15TB data, it is unreasonable to analyze (visualize) the data with legacy method. The left-hand side of the panel in Figure 6 shows that it takes 18 days to simply read this 15 TB data. Visualization with 1 core (1 CPU) takes 16 days, which is not enough, too.

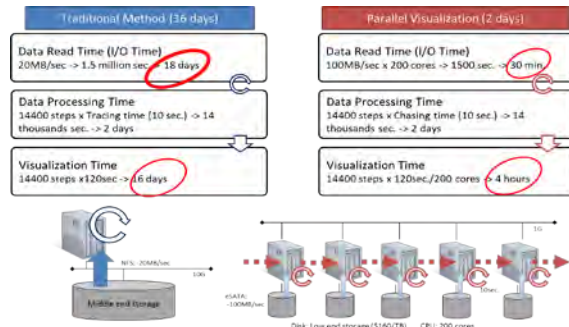
We are now developing a new parallelization method to visualize such a large-scale simulation data using 14 hetero-machines in the NICT science cloud, as shown in the right-hand side panel in Figure 6. Theoretical examination estimates that the data read (I/O) time will be as short as 30 min. and visualization will be parallelized so that it will take only 4 hours.

We have developed a proto-type of the parallel visualization for Global MHD simulations. The performance will be reported in other papers, but it took within one day, and obtained some new visualization results as shown in Figure 7.

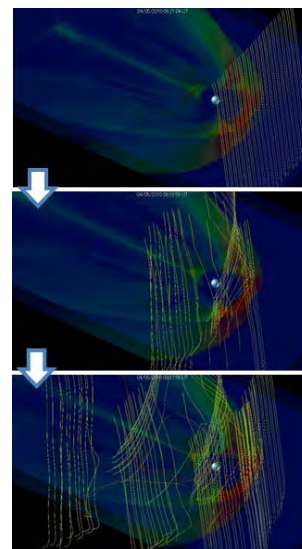
### 3.3 Data Collection (Crawling)

In Figure 2, we discussed that observation and/or experiment data are transferred to the cloud storage. However, apart from the simulation data which comes from super-computers directly connected to the cloud as shown in Figure 2, observation data are usually stored and managed at institutions out of the cloud. We need to independently collect such public data from the institutions through the Internet. For data processing, especially long-term data processing, on-demand data collection system often does not work since the data file download time takes longer time than processing time. To avoid these issues, automatic data collection (crawling) system is crucial.

Figure 8 shows an example of the number of daily data of



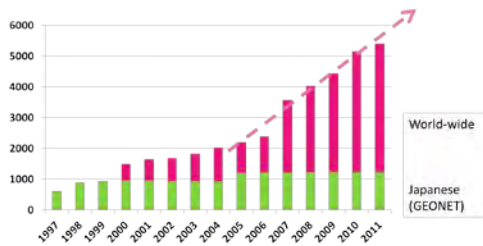
**Figure 6:** Comparison of post-processing times between legacy method and parallel visualization via NICT science cloud (14 cores).



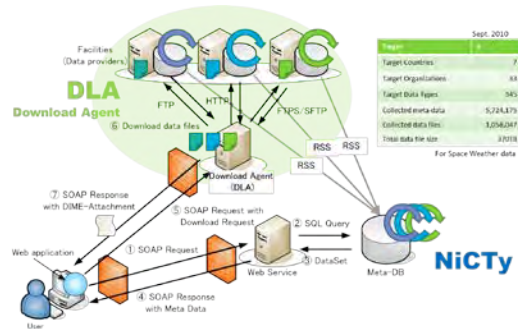
**Figure 7:** High time resolution visualization of Global MHD simulation

permanent GPS receivers collected for research studies in NICT. The number of data files to be collected is increasing, and we need to collect more than 5000 data files a day via FTP/HTTP servers of more than 20 domestic and world-wide institutions now. The policy or way to provide these data from each institution often change without any information to users. It means that manual data file collection is now very difficult.

We have developed an automatic data collection (crawling) system which works on the NICT science cloud. The system has two functions: collection of data file information (meta-data) named as NICTY, and data file crawling based on the meta-data named as DLA (DownLoad Agent) (Ishikura et al., 2006). Figure 9 shows a procedure to collect meta-data and data files using NICTY and DLA. These systems are already in use on the NICT science cloud as shown in Figure 10. The record of meta-data is more than 9 million, and crawled data files are more than 5 million so far.



**Figure 8:** The annual trend of the number of GSP receivers (world-wide since 2000 and Japanese (GEONET))



**Figure 9:** An automatic data crawling system on NICT science cloud: NICTY and DLA (DownLoad Agent).

### 3.4 High Resolution Display

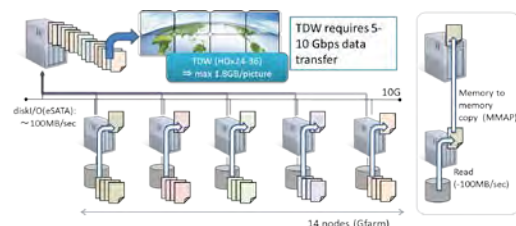
As the size of computer simulation becomes bigger, the spatial resolution of visualized data gets larger. To preview or analyze such large-scale visualized data, we need a high-resolution display which directly refers to the OSN cloud storage. (Otherwise, large scale data must be transferred to the display before previewing; it takes long time and leads to poor usability.) Tiled Display Wall (TDW) is one of the possible solutions in the near future for the issue of the high-resolution visualization. Figure 11 is a picture of a TDW at NICT main hall. The TDW is composed of 25 panels, and each of them has SXGA resolutions. Here we should note that for the best performance of the TDW we need to make use of science cloud. The typical resolution of TDWs is more than 10 HD (HD is high-vision with 1024 x 2048 resolution). It requires more than 1Gbps data transfer. This suggests that we need to prepare 10Gbps network to connect both master server and client servers of TDW. Figure 12 shows a plan to transfer movie data to a TDW using 10Gbps network and distributed storage system discussed in Section 2.3. The I/O time will be a bottle-neck of the data transfer. Note that nominal I/O speed of eSATA disk is as low as 100MB, which is slower than 10Gbps. The authors have a plan to make a parallel data transfer from Gfarm storage nodes to a TDW with as high speed as 10Gbps using memory mapping technique.



**Figure 10:** Automatic data crawling system on NICT science cloud: NICTY and DLA (DownLoad Agent).



**Figure 11:** A Tiled Display Wall at NICT main hall.



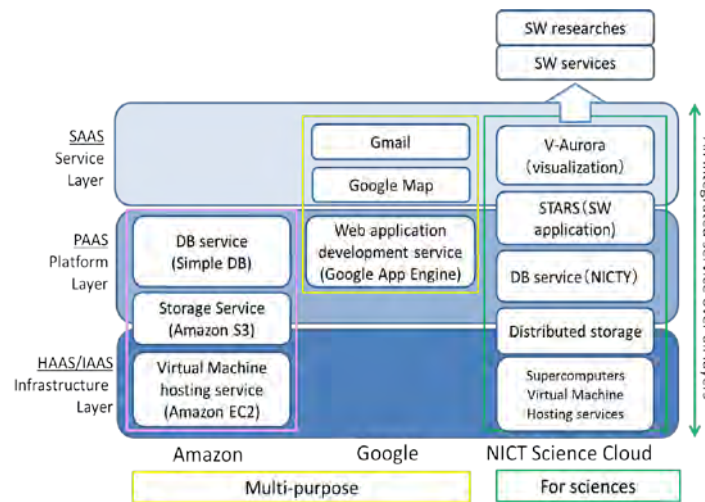
**Figure 12:** High-speed data transfer to TDW.

## 4 CONCLUSION

The first workshop of science cloud was held in February 2010 at Chicago, USA (Grossman et al, 2010). This implies that the concept of the application of cloud system to science is rather new, and nobody has ever succeeded in construction of real science cloud. We need more try-and-error to understand for what our science cloud works, and for what it does not work.

Microsoft is one of the companies who is interested in science cloud systems: they seem to apply their own cloud service, Microsoft Azure (Microsoft, 2010), to science researches as presented at the workshop above. However, one success on one science project does not necessary mean the success of the cloud. The science cloud is an environment of a variety of sciences, and an environment on which any researchers make their own customizations. We often describe a cloud system with three layers: SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service).

In Figure 13 we compare the NICT science cloud named as “OneSpaceNet” (for space weather works) with other famous cloud services. So far, Google cloud services are as SaaS and PaaS. Users mainly make use of the Google cloud service through Web. Another cloud service via Amazon is PaaS and IaaS based. They provide storage and computational resources, but no application services.



**Figure 13.** Comparison of the NICT science cloud for space weather research works with other famous cloud systems.

Murata et al. (2005) has ever attempted to construct a software system to merge different types of data from both spacecraft observations and numerical simulations. Their trial was so challenging that both data are simultaneously previewed on 4D (3D in space and time) space. However, the system was not complete because the inside system was too complicated and data size was too large even though all of the software design was based on object-oriented methodology (Murata et al., 2001). To develop such a large-scale and multi-functional system, we need a computational environment on which we construct the system. The concept of the science cloud is suitable for that; the cloud system generally provides a variety of functions required for science works.

In the present study, we propose a multi-functional science cloud and presented several studies based on the cloud system. The system is still on development, but several research works have started. Many issues are left for more practical uses of the system for research works. New research findings, which are not able to be obtained without science cloud, are expected in the near future.

## 5 ACKNOWLEDGEMENTS

The present work was done by using resources of the OneSpaceNet (the NICT science cloud) and JGN-X (Japan Gigabit Network).

## 6 REFERENCES

- Microsoft (2009), The Fourth Paradigm: Data-Intensive Scientific Discovery, <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>
- Murata, K. T., Yahara, H. & Toyota, K. (2001) Software design via object-oriented methodology & network database for solar-terrestrial observation data, *Database system*, 123-5, database, pp. 31-36.
- Murata, K.T. Yamamoto, K., Matsuoka, D., Matsumoto, H., Okada, M., Mukai, T., Sigwarth, J.B., Fujita, S., Tanaka, T., Yumoto, K., Ogino, T., Shiokawa, K., Tsyganenko, N.A., Green, J.L., & Nagai, T. (2005) Development of the Virtual Earth's Magnetosphere System (VEMS), *Advances in Polar Upper Atmosphere Research*, No.19, pp.135-151.
- Ishikura, S., Kimura, E., Murata, K.T. Kubo, T. & Shinohara, I. (2006) Automatic meta-data collection of STP observation data, *AGU Fall Meeting*, San Francisco, CA, USA.
- Fukazawa, K., Ogino, T. & Walker, R.J. (2006) The Configuration and Dynamics of the Jovian Magnetosphere, *J. Geophys. Res.*, 111, A10207, doi:10.1029/2006JA011874.
- Murata, K.T. Fujita, S., & Tanaka, T. (2007) 3D Visualizations and Analyses of Magnetic Flux Rope, *AOGS 2007 4th Annual meeting*, Bangkok, Thailand.
- Grossman, R., Gu, Y., Sabala, M., Slides, P., Mambretti, J., Szalay, A. & White, K. (2010) An Overview of the Open Science Data Cloud, *The ScienceCloud 2010*, Illinois, USA.  
<http://datasys.cs.iit.edu/events/ScienceCloud2010/p01.pdf>
- Microsoft (2010) Windows Azure Platform, <http://www.microsoft.com/azure/default.aspx>
- Mikami, S., Ohta, O. & Tatebe, O. (2011) Using the Gfarm File System as a POSIX compatible storage platform for Hadoop MapReduce applications, *Proceedings of 12th IEEE/ACM International Conference on Grid Computing (Grid 2011)*.
- Kobayashi, K., Mikami, S., Kimura, H., & Tatebe, A. (2011) The Gfarm File System on Compute Clouds, *Proceedings of 1st International Workshop on Data Intensive Computing in the Clouds (DataCloud 2011)*

# SPASE: THE CONNECTION AMONG SOLAR AND SPACE PHYSICS DATA CENTERS

*J. R. Thieman<sup>1</sup>, D. A. Roberts<sup>2</sup>, and T. A. King<sup>3</sup>*

*\*<sup>1</sup> Code 690.1, NASA/GSFC, Greenbelt, MD, 20771 United States.*

*Email: james.r.thieman@nasa.gov*

*<sup>2</sup> Code 672, NASA/GSFC, Greenbelt, MD, 20771 United States.*

*Email: aaron.roberts@nasa.gov*

*<sup>3</sup> IGPP, 5881 Slichter Hall, UCLA, Los Angeles, CA, United States.*

*Email: tking@igpp.ucla.edu*

## ABSTRACT

*The Space Physics Archive Search and Extract (SPASE) project is an international collaboration among Heliophysics (solar and space physics) groups concerned with data acquisition and archiving. The SPASE group has simplified the search for data through the development of the SPASE Data Model as a common method to describe data sets in the archives. The data model is an XML-based schema and is now in operational use. The use is expanding, but there are still other groups who could benefit from adopting SPASE. We discuss the present state of SPASE usage and how we foresee development in the future.*

**Keywords:** SPASE, Heliophysics, Archive, Data Model, XML Schema, Space Physics, Solar Physics

## 1 INTRODUCTION

The science of Heliophysics, otherwise known as solar and space physics, has been pursued using a variety of instruments both space-based and ground-based to gather data. Figure 1 indicates the many satellites that are now operational in space gathering these data. The picture does not show the many ground-based instruments such as magnetometers, radar facilities, ionosondes, etc. or many of the non-US satellites that also contribute to the accumulation of data. Within NASA the constellation of Heliophysics-related satellites is sometimes called the Heliophysics Great Observatory. With so many different instruments adding mountains of data to the data centers and archives around the world it is difficult for the researcher who may need data from multiple sources for a scientific study to find, retrieve, and analyze the data of interest. In some discipline areas the data repositories have been unified in data formats and methods have been put in place for finding data among all these repositories. Heliophysics has the problem of data being stored in many formats and in many repositories that are not part of a uniform discipline structure to facilitate data search and access. The Space Physics Archive Search and Extract (SPASE) project and the establishment of Heliophysics Virtual Observatories within NASA is an approach to solving this problem. Progress has been made and further progress depends on the acceptance and support by the non-NASA solar and space physics community in general.

Figure 2 gives some idea of the complexity of the Heliophysics Data Environment. Most of what is shown is the NASA-funded aspects of this Data Environment but there are some indications of the non-NASA parts which would complicate the diagram still more were they shown in full. There are many acronyms in Figure 2 and these are spelled out in Table 1.



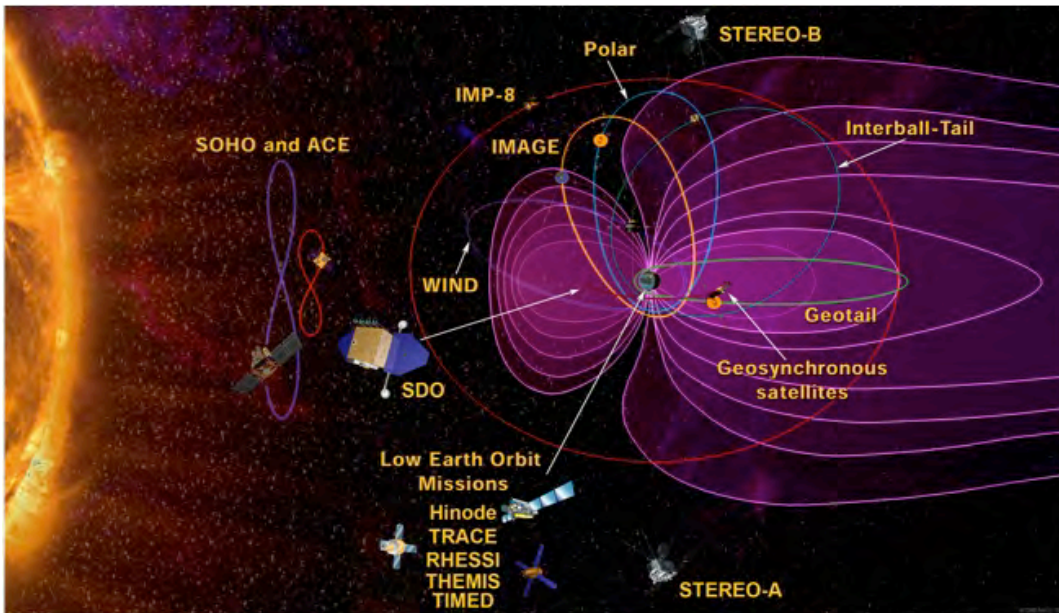


Figure 1. Heliophysics (Solar and Space Physics) Great Observatory

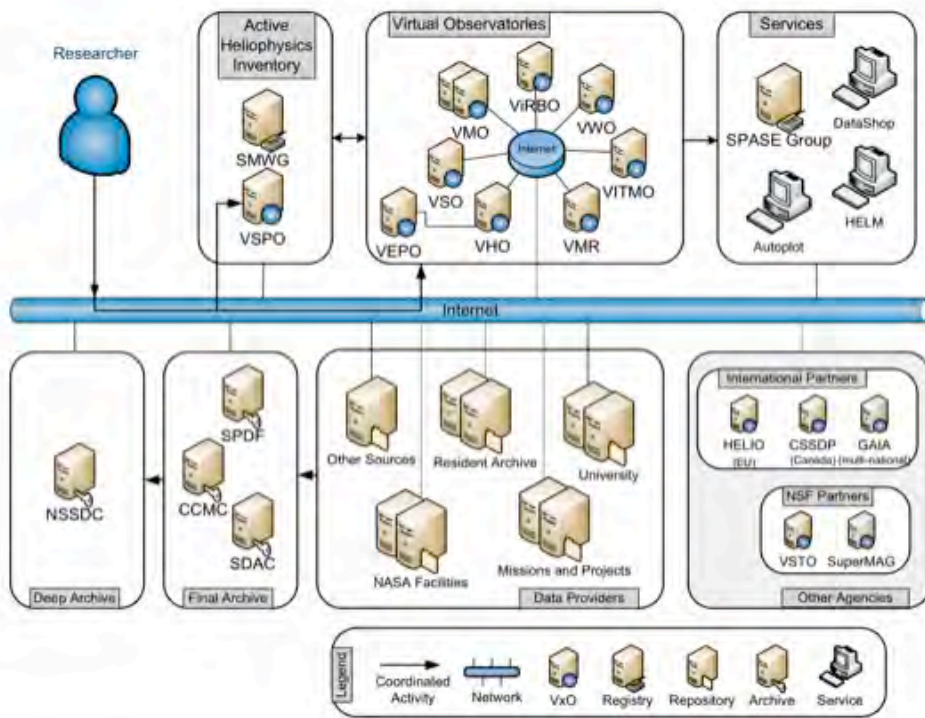


Figure 2. Heliophysics Data Environment

**Table 1.** Acronym list for Figure 2.

CCMC	Community Coordinated Modeling Center
CSSDP	Canadian Space Science Data Portal
GAIA	Global Auroral Imaging Access
HELIO	Heliophysics Integrated Observatory
HELM	Heliophysics Event List Manager
NSSDC	National Space Science Data Center
SDAC	Solar Data Analysis Center
SMWG	Science Metadata Working Group
SPASE	Space Physics Archive Search and Extract
SPDF	Space Physics Data Facility
SuperMAG	The Global Ground-Based Magnetometer Initiative
VEPO	Virtual Energetic Particle Observatory
VHO	Virtual Heliophysics Observatory
VIRBO	Virtual Radiation Belt Observatory
VITMO	Virtual Ionosphere, Thermosphere, Mesosphere Observatory
VMO	Virtual Magnetospheric Observatory
VMR	Virtual Model Repository
VSO	Virtual Solar Observatory
VSPO	Virtual Space Physics Observatory
VSTO	Virtual Solar Terrestrial Observatory
VWO	Virtual Wave Observatory

Figure 2 shows a researcher or user searching for data through the internet. Data are available through the NASA-funded data providers, other federal agency sources, university and other types of repositories as well as a variety of international partners as indicated toward the bottom of the diagram. Usually, each of these providers has different approaches to finding the data and a different user interface that has to be understood in order to locate what is needed. As the funding runs out for particular instruments the data are often transferred to a more comprehensive archive such as SPDF or SDAC and preservation copies are made and put into the Deep Archive at NSSDC. It is a daunting task to locate data of interest among all of these potential sources.

The NASA-funded Virtual Observatories (VxO's) were established to make this task easier within particular subdisciplines of Heliophysics. Using interfaces created by the Virtual Observatories the researcher can search for data within the subdiscipline and may be able to use special searches tuned to the needs of the subdiscipline users. The VxO's have the responsibility of knowing the data sources within their subdiscipline domain and providing a uniform approach to finding and acquiring the subdiscipline data of interest.

Problems arise for the researcher who wishes to compare or combine data of interest from several subdisciplines. Unfortunately, the VxO's do not have a common approach for access to data and the user is again faced with learning a variety of interfaces to do cross-disciplinary data access and retrieval. The lack of uniformity in finding Heliophysics data is partially associated with the variety of formats that are used to encode the data. There is not a single dominant data format in Heliophysics such as there is within Astrophysics in their use of the Flexible Image Transport System (FITS) format. A number of members of the international Heliophysics community agreed that it would be good to work toward a common metadata format and established the Space Physics Archive Search and Extract (SPASE) project to develop a uniform metadata approach across the discipline. With the Heliophysics data described using the common SPASE metadata format, metadata

inventories such as the SMWG and VSPO can be used to do searches for useful data across the entire discipline. This has been the main goal of the SPASE project

## 2 INTEROPERABILITY

Within the complex Heliophysics data environment what we ultimately would like to achieve is interoperability, making it easy for the user to search for and retrieve data and information. This has been the objective of many information systems through the years. The question is, how interoperable should we strive to make the various elements of the data environment? Levels of interoperability were proposed many years ago by the lead author and still may be of use in the discussion today. Figure 3 is an abstract diagram to facilitate the understanding of the levels of interoperability. The usual beginning situation is a group of very disparate elements of the overall data environment, each system element very different from the rest. The system elements are represented by the varying polygonal shapes in Figure 3.

Basic or Level 1 interoperability is achieved when these systems recognize that the user needs information or data from one of the other systems and provides a link for the user to follow to get to that system. Once the users reach the other system they are on their own in terms of learning the new system and how to find what they want.

Level 2 interoperability adds information to be passed to the new system so that the user's needs are at least partially known and the information can be used to help the user find what is wanted. Thus, the connections among the elements of the data environment become pipelines of information rather than simple electronic transfers. It is still necessary within level 2 to learn the nuances of the system to which one is transferred and this may not be straightforward even if the system has some information about the user needs to guide a search.

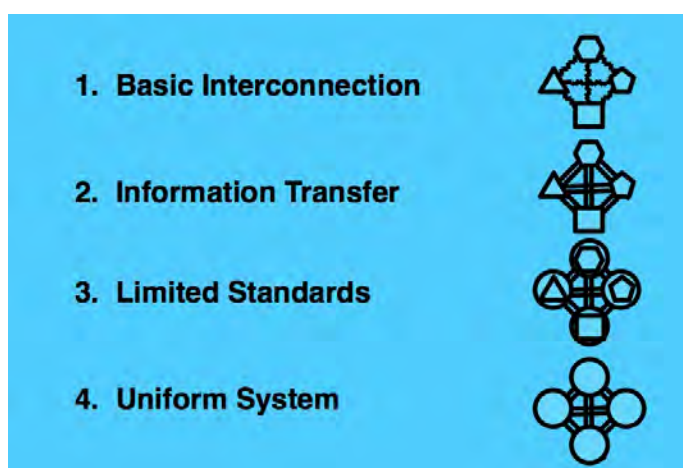


Figure 3. Levels of Interoperability

For Level 3 interoperability the systems are still very different from each other in essence, but they all agree to some standards of information exchange and/or the look and feel of the user interface. Thus the users encounter some familiar aspects among all the systems and are able to use previous knowledge to assist in an efficient search. There is a common "shell" that covers the disparate aspects of the systems. Thus, there is an added layer to each of the systems that provides some commonality among them.

Level 4 interoperability assumes that all of the systems agree to common standards and all of the systems are either created or modified so that to the user any differences among the systems are unnoticeable and it is as though the user were in one common system that has all of the needed information and data. The various elements of this environment may be physically located in very different places but that is irrelevant to the users since they are usually not aware that they have been utilizing different systems.

Levels 1 and 2 interoperability are relatively easy to achieve with internet capabilities as they are now and have been for many years. Level 3 interoperability is the more difficult step since it involves the process of getting many parts of the community to agree to common standards. Needless to say Level 4 interoperability is rarely achieved because of the need to tightly control all elements of the system. Level 4 may only be achievable if the

overall environment is built from the beginning according to exact standards. It is very expensive to modify existing systems to achieve this level of uniformity.

So, in the Heliophysics Data Environment the approach has been to try to achieve Level 3 interoperability through the adoption of the SPASE data model across all systems in the environment. The SPASE project is an international group of representatives of the Heliophysics Data Environment elements that has been developing the SPASE Data Model for many years. The SPASE Data Model is now available in Version 2.2.1 and the descriptive document as well as a variety of other information can be obtained from the main SPASE website at <http://spase-group.org>. This version of the SPASE Data Model has been stable for a relatively long time with only recent minor changes. Thus, the SPASE Data Model is in operational use especially among the VxO's within the NASA-funded Heliophysics Data Environment but still needs wider adoption among all the groups that are part of the global whole of this discipline.

### 3 THE SPASE DATA MODEL

The SPASE Data Model can be described as a grouping of Resources as indicated in Figure 4. The main Resources are those that describe the Data and the Entities associated with the Data. Most Data will be numerical in nature, but they could also be images or Display Data, Catalogs, Documents describing Data or just simple Annotations concerning the Data. The Entities associated with the Data are usually the Observatory (spacecraft, mission, project, etc.) that is the overall facility or group responsible for getting the Data and assuring the availability to the community and the Instrument that was used to acquire or generate the Data. Other Entities that may be associated with the Data are the Registry or Registries with information associated with the Data as well as the Repositories where the Data or copies of the Data are stored. Finally, Services can be associated with the Data which may be useful in interpretation or analysis. Both the Entities and the Data themselves will usually have Provenance information in order to track the origins of the Data and information attached to the Data. Another Resource of importance is the Person or Persons that were involved with the generation of the Data including the contact information needed to enable queries to the knowledgeable individuals.

The other Resource indicated in Figure 4 is the Granule. This is a subset of the overall set of data and usually represents a useful portion of the data for scientific analysis. Granules can be large or small and the number of Granules are often quite large. It becomes quite complex for the SPASE Data Model to describe the Granules in sufficient detail that the user has all the information necessary to analyze the data without having to ask the knowledgeable Person(s) for guidance. Some data formats, such as the Common Data Format (CDF), are self-describing in that they have the necessary information internal to the format to allow correct analysis. This self-describing property of a format is difficult to incorporate, however, and the question in the development of SPASE was whether this capability should be included when other formats are available that do this. This is a question that still is being discussed within the SPASE project.

For the moment, SPASE has sufficient capability to describe the overall metadata associated with the Data, but does not yet have sufficient detail to fully describe and independently analyze the Granule subsets. This feature may not be included in SPASE since it would make the SPASE Data Model much more complicated and difficult to use for data description. It is argued that SPASE should be used for overall data finding and retrieval, but not for data analysis.

### 4 HARVESTING AND EXTRACTION

Whether the SPASE data descriptions are used for data analysis or not, they can be gathered by any system through a harvesting process as indicated in Figure 5. The Virtual Space Physics Observatory (VSPO – <http://vspo.gsfc.nasa.gov>) in particular has the responsibility of maintaining a holding of all of the SPASE data set descriptions. Just as any system can do, VSPO periodically makes a request to the other VxO's for any new SPASE descriptions. These descriptions are usually stored in either a Relational Data Base Management System (RDBMS) as is done by VITMO or in a GIT repository as is done by VMO, VHO, and VIRBO.

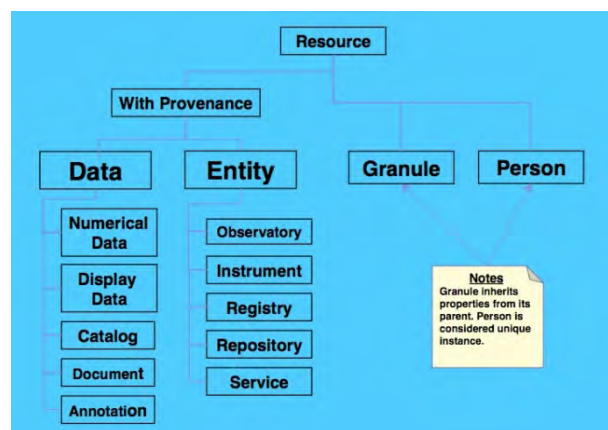
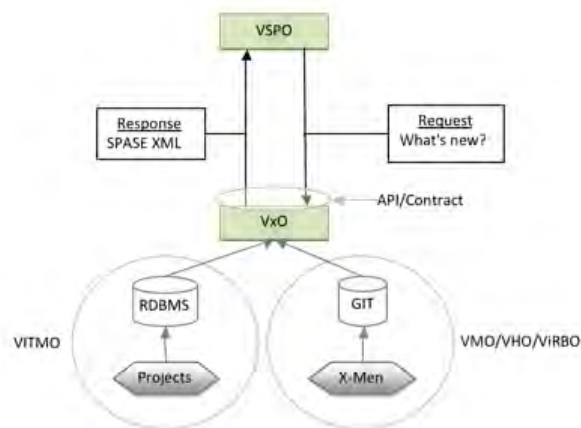


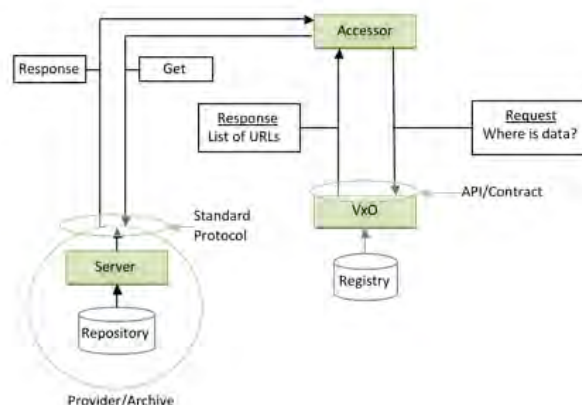
Figure 4. SPASE Information Model



**Figure 5.** Harvesting of SPASE data descriptions from the VxO's and storage in the VSPO.

Generation of the SPASE descriptions is usually done by the projects or missions that have taken the data, but in some cases the VxO's have assigned personnel to do the generation of the descriptions. Since these descriptions are done in XML some of the personnel doing this work have taken on the title of "X-men". It is helpful to have experienced individuals creating the SPASE records, but certainly not necessary. Simple descriptions can be generated with just a basic knowledge of SPASE. The minimal SPASE record should contain enough information that a user can find it when searching for particular observatory or instrument names, personnel names associated with the data, and/or the generic parameters measured by the data.

When a user wishes to extract data of interest they may use any interface that is connected to a registry likely to contain the description for the data. If the query to the registry results in finding the description of the data of interest then there should be an "Accessor" (pointer or URL) that indicates an interest in acquiring the data. In the case of the VSPO this is a "Get Data" button that appears on nearly all data descriptions. When the Accessor is invoked by the user a query is sent to the server connected to the data repository and a response is sent to the user either containing the requested data or indicating how the data may be accessed. Figure 6 is a diagram of this data extraction process.



**Figure 6.** Extraction of data from the VxO's via the SPASE data description.

The success of SPASE depends, of course, on the willingness of the data holders to describe or have their data described using the SPASE Data Model. However, it is usually human nature to not wish to spend time writing descriptions of data holdings, even if it is just a simple high level data description. For this reason a number of tools have been created to help with the data description process. These include: a Generator which can create SPASE descriptions based on Rulesets and external sources of information; several types of Editors which work through the Web or in standalone form, or Editors that use database storage; and a Validator which will check compliance of a SPASE description with the latest version (or earlier versions) of the SPASE data model. Many other tools exist as well and are generally available from the SPASE website (<http://www.spase-group.org>).

If SPASE is brought into widespread usage then it will bring about a relatively uniform approach to the data access within space and solar physics. The data systems information will be accessible in a relatively uniform approach making the data more easily accessed and reaching a level 3 type of interoperability as discussed earlier. Can level 4 be achieved where it seems as though all of the systems are uniform and appear to be part of one common system? Perhaps this might be attained through the widespread use of cloud computing technology which is gaining popularity rapidly. Could a common cloud storage of data and information provide uniformity to the Heliophysics discipline? It is not clear if this is achievable at the present time or even desirable, but it is something to keep in mind. Whether level 4 is reached or not, the general application of SPASE provides a needed connection among the disparate parts of the solar and space physics data community.

## **5 CONCLUSION**

To summarize the contributions of the SPASE project and the Data Model created by the project we review several points. The Heliophysics data environment has historically been very diverse and globally dispersed, but it is now being unified through advancements such as that provided by the SPASE Data Model through a standard metadata approach. The key to this approach is the creation of data descriptions in accordance with the Data Model. It is not easy to get the Heliophysics community to create these data descriptions, but the use of the SPASE tools which have been created for facilitating this process can greatly influence the progress toward creating the descriptions for all data of interest. If this were achieved it would provide a common link within the data environment that would be a type of level 3 interoperability as defined earlier. With the commonality provided by the SPASE Data Model cross-disciplinary research becomes easier and this is a step toward still more uniformity. Will a completely uniform level 4 type of interoperability ultimately be achieved through cloud computing and/or other technologies? This would be an interesting step, but may not be worth the extra effort necessary to achieve it. The main emphasis in the immediate future is the further adoption of the SPASE approach by the global space and solar physics community. Interested readers are invited to contact the authors for additional information.

## **6 ACKNOWLEDGEMENTS**

The lead author would like to acknowledge the contributions of his co-authors in their suggestions for the presentation and especially for the use of some of the graphics they had generated.

# CELL BASED GIS AS CELLULAR AUTOMATA FOR DISASTER SPREADING PREDICTIONS AND REQUIRED DATA SYSTEMS

*Kohei Arai*<sup>1\*</sup>

*1\*: Department of Information Science, Graduate School of Science and Engineering, Saga University, Saga 840-8502 Japan,  
E-mail: aria@is.saga-u.ac.jp*

## ABSTRACT

*A method for prediction and simulation based on Cell Based Geographic Information System: GIS as Cellular Automata: CA is proposed together with required data systems, in particular, metasearch engine usage in an unified way. It is confirmed that the proposed cell based GIS as CA has flexible usage of the attribute information which are attached to the cell in concern with location information and does work for disaster spreading simulation and prediction.*

**Keywords:** Cellular Automata, Geographic Information System, Metasearch

## 1 INTRODUCTION

Satellite data utilized analysis methods need metadata searches before data retrievals (most of countries adopted ISO Metadata standard). Although there are some search engines, CS-W as well as Open search, as standard metadata search engine, there is no unified method for metadata search. In this paper, metadata search is discussed. Meanwhile, Geographical Information System: GIS is totally identical to Cellular Automata: CA which allows predictions, simulations. In this paper, Cell Based GIS as CA (it is referred to CBGISCA hereafter) is proposed.

CA approaches are widely used in disaster spreading simulations such as forest fire (wild fire), flooding, lava flow, landslide, mudflow, etc. CA approaches are simple and easy to develop, and obtain the spectacular displays or visualization. Especially for forest fire prediction, CA approaches were used by many researchers to simulate and show fire spreading for some periods.

Ioannis [1] presented the simple CA model to predict the spreading of fire in both homogeneous and inhomogeneous forests. This model is simple, although it has some addition parameters. Malamud [2] presented a forest-fire CA model by programmed simulating different fire frequencies, spatial and temporal patterns. Ecinas [3] presented a new two dimensional CA approach using hexagonal area of forest, and show the transfer of fractional burned area with different speed. The algorithm seems to be very efficient from the conventional one and it is easily implemented in any computer algebra system, allowing a low computational cost. These approaches [1][2][3] are too simple to show complex forest-fire spreading, such as tree-types, landscape and wind determination.

The other approach, introduced by Song [4], presented an improved model of CA approach with tree species, meteorological conditions and human efforts on fires are looked on as generalized “immunity” of the tree from fire. This model improved the previous model with some addition parameters, but it did not show relation between trees and probability to fire. The probability to fire is important parameter to make the model can adopt some unknown condition with simple approach.

In this paper, a new two dimensional CA approach using fire-control probability, wind characteristics and tree-types is introduced. This approach is simple and easy to develop because it uses phenomenological relationships directly. CA based forest fire prediction requires a plenty of attributed data on conditions, wind speed, wind direction, tree species, humidity, air temperature, topological feature, etc. On the other hands, GIS has databases which includes such attribution data. The proposed CBGISCA consists of 2D cells which represent geographical map, and the aforementioned attribute data. Therefore, CA based simulations can be done with reference to the attribute data effectively. Although this approach is simple, it is an alternate approach to predict forest fire spreading and shows the dangerous area in the future.

Malamud [2] presented the simple CA approach using the programmed model. This model consists of a square

grid, in which at each time step a tree is randomly dropped on a chosen site. Every  $1/f_s$  time steps a match is randomly dropped ( $f_s$  is the sparking frequency). If a tree falls on an unoccupied cell it is planted. If a match drops on a tree, that tree and all non-diagonally adjacent ones are burned in a fire. The keyword of this approach is sparking frequency.

The one of important parameter in forest fire simulation is wind speed. This parameter influences the neighborhood size and sparking probability  $f_s$ . Relation of wind speed, the neighborhood size and sparking probability, introduced by Sullivan [5], is defined. It is the “bubble” convection model of wind speed influences in forest fire simulation, because this model assumes that fire component is a bubble model.

In the following section, the proposed method of metadata search engine is described followed by the CBGISCA together with the example of the proposed forest fire propagation method. Simulation results are also described. Then concluding remarks with some discussions is followed.

## **2 PROPOSED METHODS**

### **2.1 Metadata search engine**

CS-W, Open search is well known metadata search engine. Cadcorp holds metadata in a centralised repository on a geospatial server, GeognoSIS. The repository is accessed through the Open Geospatial Consortium Catalog Server (Web) interface, OGC CS-W. OGC CSW Core (OGC, 2007core): this interface is recommended by the GEO/GEOSS. OGC CSW ISO Application Profile (OGC, 2007iso): this interface is identified by INSPIRE Implementing Rules (IRs) as the reference for ESDI catalog services. Meanwhile, OGC CSW ebRIM/EO is Extension Package (OGC, 2008eo). This extension package of the CSW is recommended by the GMES/ESA-HMA initiative (<http://earth.esa.int/hma/index.html>). On the other hand, OGC CSW ebRIM/CIM (OGC, 2007cim), OGC CSW OpenSearch Extension (OGC, 2008os) and the GENESI-DR (Ground European Network are for Earth Science Interoperations while Digital Repositories) (<http://www.genesi-dr.eu/>) catalog interface which is based on it.

The OGC seeks comment on City Geography Markup Language (CityGML) V1.1. There are OGC Requests Sensor Planning Server (SPS) 2.0 Reference Implementations, OGC seeks comment on candidate Earth observation profile of coverage standard, OGC Seeks Comment on candidate GeoSPARQL standard, The OGC and OpenMI Association to advance computer modeling standards, The OGC Seeks Participants for Hydrologic Forecasting Interoperability Experiment, OGC completes Water Information Concept Development Study, The OGC Announces GEOSS Workshop XLIII: Sharing Climate Information and Knowledge. It would be better that metadata search can be done in a unified way. It is also desired to determine a standard procedure for metadata search.

### **2.2 CBGISCA**

Because geocoded earth observation satellite data can be represented on geographical maps. GIS representation is effective in such case. CA, on the other hand, is effective for estimation and prediction phenomena and is based on cells. Therefore, geocoded data can be treated as cell wise data which results in CBGISCA. Furthermore, GIS does work as neural network then it allows predictions and simulations. In particular, disaster relief and prediction can be done with cellular automata. All the required data for disaster relief and prediction can be represented on cells and can be acquired with Web Map Services: WMS. Meanwhile, Web Geographical Map (Landscape map→Landscape object, Land-use map→Human activity influences), Forest Map (Forest type→Tropical Forest, Homogenous Forest, Hot Spots), and Weather Map (Wind, Season and Temperature) are also retrieved and downloaded from the service servers through the metasearch. WMS provides geographical map data with the information lossy algorithm while WCS (Web Coverage Service) data is represented as a set of cells. Therefore, CBGISCA uses WCS type of geographical maps, cell based representation of maps rather than WMS.

Raster based GIS consist cells (grid) while Geo-coded earth observation data is represented on cells. Meanwhile, data required for simulation and prediction (disaster relief, etc.) is represented on cells. All these data are represented on a cell based GIS and also are used for simulation and prediction as cellular automata. GIS (ISO Standard) representation of cells, GIS based cellular automata is used for disaster relief and prediction. Meta search has to be done with unified way. Therefore, a Standard Clearing House System is highly required. Consistency of the data quality with space and time as well as among the sensors in concern



### 3 APPLICATIONS OF GIS ON CA FOR FOREST FIRE SIMULATIONS

#### 3.1 Proposed cellular automata on GIS method for forest fire spreading simulation

The proposed method is two dimensional CA which uses a square grid of sites. The following four parameters are taken into account, tree-types, wind speed and direction, sparking probability and stopping probability. A number of tree-types with different probability for fire is taken into account.

The proposed forest fire model consists of a square grid of sites of which blank node, tree, and fire are considered as the status. Some trees around fire may be fired depending on the sparking probability  $f_s$  [3], and also fire may be stopped depending on the stopping probability  $f_c$ . Stopping probability  $f_c$  is a constant which depends on the tree material, species. Meanwhile, sparking probability is a variable which depends on material (species), wind speed and wind directions. Tree material (species) parameter shows possibility to be fired. Wind speed parameter defines the size of neighbors, and shows that model uses dynamic neighborhood model. Wild fire propagation direction depends on wind direction parameter.

The algorithm of Cellular Automata for forest fire simulation is represented as follows,

- We begin with a square grid of sites; there are five states:
  - $s=1 \rightarrow$  blank node
  - $s=2,3,\dots,n+1 \rightarrow$  tree (n different tree types). Malamud et.al [1] uses one type of tree.
  - $s=n+2 \rightarrow$  fired
  - $s=n+3 \rightarrow$  stopped or completed fired
- Determine neighborhood (size and shape) depend on wind speed and wind direction. We use the Cardioid concept.
- Trees will be fired by sparking probability  $f_s$ , if there are fire neighbors.
- Fire will be stopped by stopping probability  $f_c$ .

In the proposed method, we define n tree types that have different material. Each material has probability of fire which depends on tree type. Ohgai [6] defines the simple probability of fire depending on material as following:

$$S_{ij} = 1, \text{ if wooden} \\ = 0.6, \text{ if preventive wooden} \\ = 0, \text{ if fireproof.}$$

In CA approach, one of the important parameters is neighborhood rules. The proposed method uses dynamic neighborhood system depends on wind parameters; wind speed and wind direction.

The number of neighbors depends on wind speed. According to Jirou and Kobayashi [7] and Ohgai [6], the relation of the number of neighbors and wind speed is expressed as equation (1):

$$D = 1.15 (5 + 0.5 v) \quad (1)$$

where  $v$  is wind speed (m/s) and  $D$  is limit of distance which fire can spread.

Sullivan [5] uses the bubble concept while we use the “Cardioid” concept for definition of wind direction influence on neighborhood system. It is the proposed model in CA approach for forest fire simulation. This is an original method which differs from the Sullivan and Malamud model. The new limit of distance which fire can spread is represented as equation (2):

$$D^* = D (1.5 + \cos (d)) \quad (2)$$

where  $d$  is wind direction,  $D$  is limit of distance that written in equation 1, and  $D^*$  is new limit of distance of fire can spread.

Figures 1 and 2 show how the wind parameters determine neighborhood system in our approach. In this figure we have two variables,  $r$  and  $\alpha$ . The first parameter is wind speed  $r$  that relates on wind speed. The second parameter is wind direction  $\alpha$ .

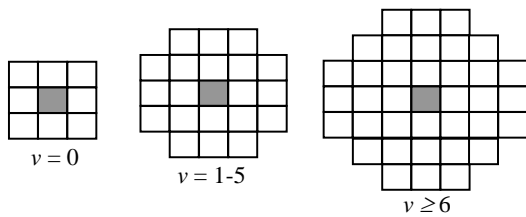


Figure 1. Neighborhood size depends on wind speed

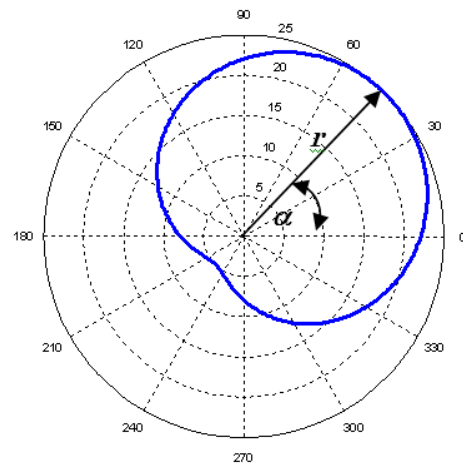


Figure 2. The Cardioid concept for wind influence

### 3.2 Simulation results

In this simulation three tree-types with different probability of fire are set. The probability is randomly selected. The other input parameters are density. It shows the number of trees in the observation area. We select density of around 0.6-1. Figure 3 shows the simulation results in 40 unit time steps with the different density. This simulation uses two probabilities function; sparking probability  $f_s$  and stopping probability  $f_c$ . The number of fired area which depends on sparking probability and stopping probability are shown in Figure 4. Different combination of  $f_s$  and  $f_c$  has the different joint points of the number of tree and the number of fired.

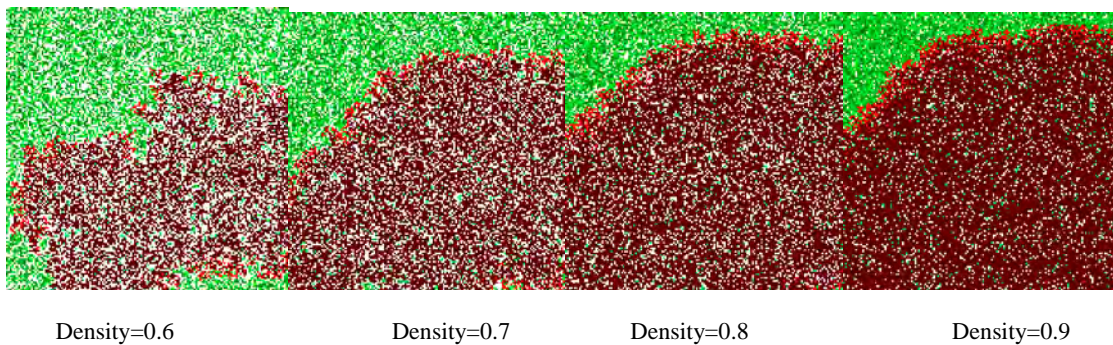


Figure 3. Simulation results on different density.

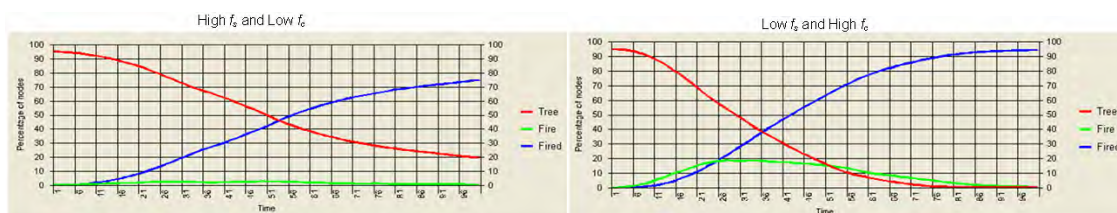


Figure 4. The number of fired area depends on  $f_s$  and  $f_c$ .

## 4 CONCLUSIONS

Conclusion is as follows,

- GIS (ISO Standard) representation of cells
- GIS based cellular automata is used for disaster relief and prediction
- Meta search has to be done with an unified way → Standard Clearing House System is highly required
- Consistency of the data quality with space and time as well as among the sensors in concern

In particular, the proposed CBGISCA allows flexible use of attribute data which are required for disaster prediction with reference to the geographical location information. Also prediction results can be represented in GIS display superimposing with the other attribute data. Therefore, it is easy to check a validity of the prediction results.

## 5 REFERENCES

- [1] Ioannis Karafyllidis, Adonios Thanailakis, A model for predicting forest fire spreading using cellular automata, *Ecological Modeling* 99, 87-97, 1997..
- [2] Malamud, B.D., Turcotte, D.L., 2000, Cellular-Automata models applied to natural hazards, *IEEE Computing in Science & Engineering*, Vol. 2, No. 3, pp. 42-51, 2000.
- [3] L. Hernandez Encinas, S. Hoya White, A. Martin del Rey, G. Rodriguez Sanchez, Modeling forest fire spread using hexagonal cellular automata, *Applied Mathematical Modeling* 31, 1213–1227, 2007.
- [4] SONG Weiguo, FAN Weicheng & WANG Binghong, Self-organized criticality of forest fires in China, *Chinese Science Bulletin* Vol. 46 No. 13 July 2001.
- [5] A.L. Sullivan, I.K. Knight, A hybrid cellular automata/semi-physical model of fire growth, *Complexity International*, Volume 12, 2005.
- [6] A. Ohgai, Y. Gohnai, S. Ikaruga, M. Murakami and K. Watanabe, Cellular Automata Modeling For Fire Spreading As a Tool to Aid Community-Based Planning for Disaster Mitigation, *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning*, 193-209, Kluwer Academic Publishers. Printed in the Netherlands, 2004.
- [7] Jirou K, K Kobayashi, “Large area fire”, in: Fire Institute of Japan (eds.) *Fire Handbook* third edition, Kyoritsu Publication Co., Ltd., Tokyo, p. 508-573 (in Japanese), 1997.

# DATA MINING APPROACHES FOR HABITATS AND STOPOVERS DISCOVERY OF MIGRATORY BIRDS

*XU Qiang<sup>1,2</sup>, LUO Ze<sup>1\*</sup>, WEI Ying<sup>1,2</sup> and YAN Baoping<sup>1</sup>*

<sup>1</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China

<sup>2</sup>Graduate University of the Chinese Academy of Sciences, Beijing, 100190, China

Email: [xuqiang@cnic.cn](mailto:xuqiang@cnic.cn)

<sup>\*1</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China

Email: [luoze@cnic.cn](mailto:luoze@cnic.cn)

<sup>1</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China

<sup>2</sup>Graduate University of the Chinese Academy of Sciences, Beijing, 100190, China

Email: [weiyi@cnic.cn](mailto:weiyi@cnic.cn)

<sup>1</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China

Email: [ybp@cnic.cn](mailto:ybp@cnic.cn)

## ABSTRACT

*This paper mainly focuses on using data mining technology to efficiently and accurately discover habitats and stopovers of migratory birds. The three methods we used are as follows: 1. Density-based clustering method, detecting stopovers of birds during their migration through density-based clustering of location points. 2. Location histories parser method, detecting areas that have been overstayed by migratory birds during a set time period by setting time and distance thresholds. 3. Time-parameterized line segment clustering method, clustering directed line segments to analysis shared segments of migratory pathways of different migratory birds, and discovers the habitats and stopovers of these birds. At last, we analyzed the migration data of bar-headed goose in the Qinghai Lake Area through the three methods above and verified the effectiveness of the three methods, and by comparison, identified the scope and context of use of these three methods respectively.*

**Keywords:** Migratory birds; Flyway; Satellite tracking data; Detection algorithm; Bar-headed goose

## 1 INTRODUCTION

One of the most important tasks to protect migratory birds around the globe is to identify the ecological needs of birds in their breeding and wintering grounds as well as the stopovers during their migration (Berthold, & Terrill, 1991). The information of specific migration routes, net structures of these migration routes and important stopovers during migration is the key to research migratory birds' selection of habitats and stopovers, birds' migration strategy and the influence of global climate change on migratory birds' migration. On the other hand, the role of migratory birds in the spread of avian influenza virus has been a hot topic nowadays. Among the wild birds which have been infected by the H5N1 highly pathogenic avian influenza virus, many are migratory, so migratory bird might be avian influenza virus vectors. As the ecological environment and natural resources of the habitats and stopovers might set the stage for interspecific or intraspecific transmission of avian influenza virus among birds, studying wild birds' migration and detecting these birds' habitats or stopovers efficiently and precisely are of significant value for the research and prevention of the spread of avian influenza virus.

The traditional way of studying bird migration, like bird banding, is simple and easy to carry out, but its result depends on long-time observation and the number and quality of returned birds are under expectation, thus it's impossible to get a whole picture of the track of bird migration in short time (Zhang, & Yang, 1997). In other words, the traditional way is hard to meet the requirements of modern study. The development of satellite tracking technology and its application in biology in recent years provide new opportunities for bird migration study (Cagnacci, Boitani, Powell, & Boyce, 2010). Some of the raw data by using satellite tracking technology is shown in the following **Table 1**.

**Table 1.** Relational representation of raw GPS data.

ID	Animal	Latitude	Longitude	lc94	Date time
930796	BH07_67582	65.448	96.317	LZ	2008-01-30 04:02:00
930948	BH07_67582	65.448	96.317	LZ	2008-01-30 04:02:00

In this chart, **ID** is the recording number, **Animal** is the label of the migratory bird, **Latitude** and **Longitude** showing the specific location, and the **Date time** field signifying time stamp. Obviously, traditional data analysis methods such as drawing-dot or manual statistics method cannot process these high-resolution spatial-temporal data. This paper mainly focuses on using data mining technology to discover habitats and stopovers of migratory birds among the original satellite telemetry data efficiently and accurately, these methods are described as follows:

- **Density-based clustering method.** The habitats and stopovers of migratory bird are the areas where the bird continuously stays for some time, corresponding to the dense regions in space. We use the density-based clustering method to discover these dense regions. Although the location data of the migratory bird may be lost because of some different reasons, these dense regions can characterize the habitats or stopovers of the bird.
- **Location histories parser method.** Given a time and distance threshold, modeling the move status (stay or move) of migratory bird, and then scanning a certain bird's migration route point by point. This method can get the arriving and leaving time of the migratory bird at its every stopover.
- **Time-parameterized line segment clustering method.** We measure the space-time density of moving objects by the spatial distance, the direction of the movement and the time characteristics. We use the time-based plane-sweeping trajectory clustering algorithm to analysis shared segments of migratory pathways of different migratory birds, and discover the habitats and stopovers of these birds.

The following part of this paper is organized as the following: the second section introduces some relevant researches; the third section defines some specific terms; the fourth section elaborates three ways to discover stopovers among GPS data; the fifth section presents the experiments and the result analysis; the last section provides the major conclusions of the paper.

## 2 RELATED WORK

As the improvement in GPS-based radio telemetry and growing international concern about the migratory birds, many international organizations began to trace the birds' migration through satellite positioning technology (Frisch, Vagg, & Hepworth, 2006). There is increasing interest on developing methods to perform data analysis for trajectory datasets (Schiller, & Voisard, 2004) (Stauffer, & Grimson, 2000). A typical data analysis task is to detect the stopovers of the moving objects. We used the same satellite telemetry datasets with (Tang et al., 2009), Tang et al. (2009) proposed a hierarchical spatial clustering method HDBSCAN to find the habitats or stopovers of migratory birds in different spatial scale levels, but HDBSCAN algorithm measures the proximity of birds mainly by Euclidean distance between two points and does not take time information into account. Hariharan, & Toyama (2004), Zheng, Zhang, Ma, Xie, & Ma (2011), Zheng, & Li (2008), Zheng, & Xie (2010) modeled the location histories of human and proposed a method to find the stopovers of human, but their attention focused on personalized recommendation based on location, so they did not study the stopovers in depth. Gaffney, & Smyth (1999), Gaffney, Robertson, Smyth, Camargo, & Ghil (2006) observed that existing trajectory clustering algorithms group similar trajectories as a whole, thus revealing common trajectories. But clustering trajectories as a whole could not detect similar portions of trajectories or could miss common sub-trajectories. The framework and algorithm proposed by Lee, Han, & Whang (2007) did not consider temporal information. Satellite telemetry datasets or GPS-based locations datasets are essentially time series of spatial data. To measure the space-time density of moving objects, this paper defines different distance functions from (Lee et al., 2007) to measure the similarity of different line segments, so that we can find the shared segments of migratory pathways both in time and space. In this paper, we use three data mining methods to discover habitats and stopovers of migratory birds, and analyze in detail the characteristics and the contexts of use of the three algorithms respectively.

## 3 PRELIMINARY

In this section, we clarify some terms used in this paper such as **point**, **line segment**, **trajectory** etc.

**Point:** a point  $P$  is indicated by a tuple  $\langle Lat, Lng \rangle$ , which refers to that one bird once presented in a location at

where the latitude is *Lat* and the longitude is *Lng*.

**Point set:** a point set *PS* consists of a series of points which are generated by one or more birds.

**Trajectory:** a trajectory *TR* is defined as an ordered set of  $\langle position, timestamp \rangle$  pairs ordered by time serials.  $TR = \{\langle P_1, t_1 \rangle, \langle P_2, t_2 \rangle, \langle P_3, t_3 \rangle, \dots, \langle P_n, t_n \rangle\}, \forall(i < j) t_i < t_j$  where  $t_i$  is point  $P_i$ 's timestamp.

**Line segment:** Given a trajectory *TR*, a line segment of *TR* is defined as  $LS_i = \langle \langle P_i, t_i \rangle, \langle P_{i+1}, t_{i+1} \rangle \rangle$ , where  $\langle P_i, t_i \rangle, \langle P_{i+1}, t_{i+1} \rangle \in TR$  represents object moves from position  $P_i$  to position  $P_{i+1}$  during  $[t_i, t_{i+1}]$ . The displacement of moving object is denoted by  $\overline{LS}_i$ , and the duration of  $LS_i$  is denoted by  $LS_i.TD$ .

**Line segment set:** The line segment set of a trajectory *TR* is defined as a collection of two sequential pairs in *TR*,  $LSS = \{\langle P_i, t_i \rangle, \langle P_{i+1}, t_{i+1} \rangle \mid 1 \leq i \leq n - 1\}$

**Stop region:** stop region is the area where the migratory birds stay for some time during their migration. Migratory birds' habitats and stopovers are all stop regions. We use a stop region center's coordinate to indicate the stop region in the following sections.

## 4 THREE METHODS TO DISCOVER THE STOP REGIONS

Migratory routes of migratory birds are long and complex paths (**Figure 1**), and the migratory birds' raw GPS data can't be used conveniently due to its large scale and high complexity. In this section, we will provide three methods to solve the problem, and explain their principle in detail.



**Figure 1.** Migratory pathway of one bar-headed goose captured in the Qinghai Lake Area.

### 4.1 Density-based clustering method

As depicted in **Figure 1**, the dense regions in the picture may be the stop regions from the visual point of view. We can assume that dense regions in spatial-temporal data are equivalent to the stop regions. The GPS position sampling frequency of satellite telemetry device was about once every 2 hours during the day. If a bird stays in a small area more than a certain period of time, the sampling point in this area may be denser than other place. So it is possible to detect the migratory birds' stop regions by finding the dense areas in GPS location history data.

In order to find the dense clusters in spatial data, Ester, Kriegel, Sander, & Xu(1996) proposed the DBSCAN algorithm. The density-based algorithm based on the following notions:  $\epsilon$ -neighborhood is the neighborhood within a radius  $\epsilon$  of a given object; an object is a **Core object** if the  $\epsilon$ -neighborhood of this object contains at least a minimum number (*MinPts*) of objects; an object  $p$  is **directly density-reachable** from object  $q$  if  $p$  is within the  $\epsilon$ -neighborhood of  $q$ , and  $q$  is a core object; an object  $p$  is **density-reachable** from object  $q$  with respect to  $\epsilon$  and *MinPts* in a set of objects,  $D$ , if there is a chain of objects  $p_1, \dots, p_n$ , where  $p_1 = q$  and  $p_n = p$  such that  $p_{i+1}$  is **directly density-reachable** from  $p_i$  with respect to  $\epsilon$  and *MinPts*, for  $1 \leq i \leq n, p_i \in D$ ; an object  $p$  is **density-connected** to object  $q$  with respect to  $\epsilon$  and *MinPts* in a set of objects,  $D$ , if there is an object  $o \in D$  both  $p$  and  $q$  are **density-reachable** from  $o$  with respect to  $\epsilon$  and *MinPts* (Han, & Kamber, 2000). All points within the cluster are mutually density-connected. If a point is density-connected to any point of the cluster, it is part of the cluster as well.

The stop region detection algorithm based on DBSCAN (Ester et al., 1996) is described as follows:

---

**Input:** Point set: *PS*; Radius:  $\epsilon$ ; the minimum number of points to decide the core objects: *MinPts*

**Out Put:** A set of all stop regions *SS*

**DBS\_SR\_DETECTION** (*PS*,  $\epsilon$ , *MinPts*):

$C = 0$ ;

For each unvisited point  $P$  in dataset *PS*

Mark  $P$  as visited;

```

N = P's ε-neighborhood set;
If P is not a core object
  Mark P as NOISE;
Else
  C++;
  Add P to cluster C;
  For each point O in N //find all the objects that density-connected with P
    If O is not visited
      Mark O as visited;
      N' = O's ε-neighborhood set;
      If O is a core object
        N = N joined with N';
      If O is not yet member of any cluster
        Add O to cluster C;
  Return the center coordinates of each cluster;

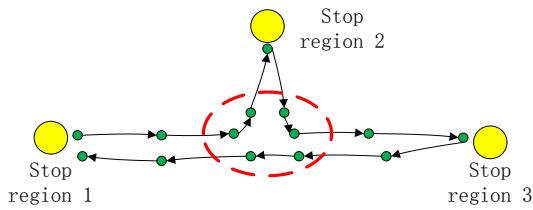
```

---

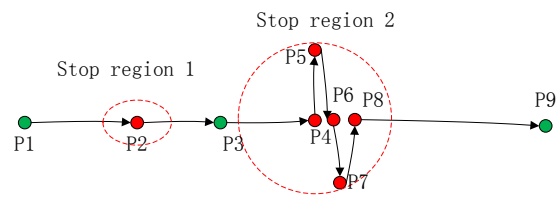
The time complexity of *DBS\_SR\_DETECTION* is  $O(n^2)$ , where  $n$  is the number of points in  $PS$ . If the appropriate spatial index is used, the time complexity of this algorithm will reduce to  $O(n \log n)$ . If  $\epsilon$  and  $MinPts$  are appropriately set, this algorithm can detect arbitrarily shaped clusters, but as for how to choose these two parameters there is no good way. When we use this algorithm,  $PS$  can be either one bird's history location set or multi-birds' history location set. Here,  $PS = \{P_1, P_2, \dots, P_n\}$ , where  $P_i = \langle Lat_i, Lng_i \rangle$ , the points in  $PS$  only contain spatial dimension, and we use great-circle distance as geographical distance formula between two points. Furthermore, the NOISE in *DBS\_SR\_DETECTION* may be significant for the ornithologist, because the object may be flying fast at this location.

## 4.2 Location histories parser method

As stated before, the *DBS\_SR\_DETECTION* only takes the spatial dimension into account, dismissing the time dimension. In fact, birds' migration routes are complex and not regular (**Figure 1**), and the bad climate or other factors in wild environment may cause satellite signal lost. As depicted in **Figure 2**, if we use *DBS\_SR\_DETECTION* algorithm to detect this bird's stop region, we may find out the region surrounded by dotted red line, obviously showing that region is meaningless.



**Figure 2.** A typical migratory route



**Figure 3.** Two kinds of stops of migratory birds

In order to solve the problem above, we need take the time dimension into account. Hariharan et al. (2004), Zheng et al. (2011), Zheng et al. (2008), Zheng et al. (2010) proposed a time and distance threshold based method to discover human's stay point from the historical location data. This method may be useful for detecting the migratory birds' stop regions. The stops of migratory birds may be divided into two kinds:

- As the stop region 1 depicted on **Figure 3**, during the migration, birds may keep stationary for some time because of the bad weather or they need a rest.
- As the stop region 2 depicted on **Figure 3**, the birds may stay in a little area for some time, because they need to find food or for some other reasons.

Both of the stops can be defined as this:

Given a trajectory  $TR = \{\langle P_1, t_1 \rangle, \langle P_2, t_2 \rangle, \langle P_3, t_3 \rangle, \dots, \langle P_n, t_n \rangle\}$ , if there is a subset of  $TR$   $sTR =$

$\{ \langle P_i, t_i \rangle, \langle P_{i+1}, t_{i+1} \rangle, \dots, \langle P_j, t_j \rangle \}$  where  $1 \leq i, j \leq n$  and for  $\forall i \leq k \leq j$ ,  $\text{Dist}(P_i, P_k) \leq Dr$ ,  $\text{Dist}(P_i, P_{j+1}) > Dr$ ,  $\text{Int}(t_i, t_j) \geq Tr$ , the  $\text{Dist}(P_i, P_k)$  denotes the geospatial distance between two points  $P_i$  and  $P_k$ , the  $\text{Int}(t_i, t_j) = |t_i - t_j|$  is the time interval between two points, then the area where the points at  $sTR$  locate is a stop region  $S$  (Zheng, & Xie, 2010). We can also use a quaternion to indicate a stop region  $S = \langle Lat, Lng, ts, te \rangle$ . The  $Lat$  stands for the average latitude of the collection  $sTR$ ; the  $Lng$  stands for the average longitude of the collection  $sTR$ ; the  $ts$  means the bird's arriving time on stop region  $S$ ; the  $te$  means bird's leaving time. We can compute them as:  $S.Lat = \frac{\sum_{k=i}^j P_k.Lat}{|sTR|}$ ,  $S.Lng = \frac{\sum_{k=i}^j P_k.Lng}{|sTR|}$ ,  $S.ts = t_i$ ,  $S.te = t_j$ .

The algorithm that detects all stop regions from a trajectory is described as follows:

---

**Input:** A trajectory:  $TR$ ; Distance threshold:  $Dr$ ; Time threshold:  $Tr$

**Output:** A set of all stop regions  $SS$

**LHP\_SR\_DETECTION** ( $TR, Dr, Tr$ ):

```

i=0, n = |TR|; //the number of GPS points in a GPS logs
While i < n do:
  j=i+1;
  While j < n do:
    Dist = Dist(Pi, Pj)
    If Dist > Dr then
       $\Delta T = \text{Int}(t_i, t_j)$ ;
      If  $\Delta T > Tr$  then
         $S.Lat = \sum_{k=i}^j P_k.Lat / (j - i + 1)$ ;
         $S.Lng = \sum_{k=i}^j P_k.Lng / (j - i + 1)$ ;
         $S.ts = t_i$ ;  $S.te = t_j$ ;
        SS.insert(S);
        i=j+1; break;
  j=j+1;
Return SS

```

---

This algorithm's time complexity in the worst case is  $O(n^2)$ . The data *LHP\_SR\_DETECTION* can process is one bird's trajectory. Before use this algorithm we should sort the bird's location history data by timestamp. This algorithm can't deal with multi-birds' trajectory. A simple method to solve this problem is to combine *DBS\_SR\_DETECTION* with *LHP\_SR\_DETECTION*, which can detect all stop regions of one bird respectively, and then cluster all the stop regions of all birds.

### 4.3 Time-parameterized line segment clustering method

Birds in the same region usually share their habitats or stopovers. As indicated in **Figure 4**, different birds fly from one same place to another, and as a result many similar line segments will be generated between these two places. The sets of starting points and finishing points of each line segments in this cluster may be the stopovers or habitats of migratory birds.



**Figure 4.** A line segment cluster

In order to cluster line segments, the first problem need to be solved is to measure the distance between two objects. The distance function we proposed to measure the distance between two line segments includes both spatial and temporal aspects. We define the distance function between line segments  $LS_i = \langle P_i, t_i \rangle, \langle$



$P_{i+1}, t_{i+1} \gg$  and  $LS_j = \langle \langle P_j, t_j \rangle, \langle P_{j+1}, t_{j+1} \rangle \rangle$  as follow:

$$L\_dist(LS_i, LS_j) = \begin{cases} \frac{\text{dist}(P_i, P_j) + \text{dist}(P_{i+1}, P_{j+1})}{2}, & \text{if } LS_i.TD \cap TW \neq \emptyset \wedge LS_j.TD \cap TW \neq \emptyset \wedge \angle(LS_i, LS_j) \leq \theta \\ \varepsilon + 1, & \text{else} \end{cases} \quad (1)$$

Here  $\varepsilon$  means the spatial threshold;  $\theta$  means the angle threshold;  $\text{dist}(P_i, P_j)$  means the distance between two points  $P_i$  and  $P_j$ , the distance is measured by the great circle distance;  $\angle(LS_i, LS_j)$  means the included angle between line segments  $LS_i$  and  $LS_j$ , which is measured by the spherical angle between two great circles containing line segments;  $TW$  means the time window  $TW = [t_1, t_2]$ .

After defining the distance function between two line segments, we use the DBSCAN (Ester et al., 1996) algorithm to find all the dense clusters. As the object we concerned is line segment, we give some extra description. The set of all the line segments is denoted as  $LSC$ ; the  $\varepsilon$ -neighborhood set of line segment  $LS_i$  ( $LS_i \in LSC \wedge LS_j.TD \cap TW \neq \emptyset$ ) in time window  $TW$  is defined as:

$$N_{(\varepsilon, TW)}(LS_i) = \{LS_k | LS_k \in LSC \wedge LS_k.TD \cap TW \neq \emptyset \wedge L\_dist(LS_i) \leq \varepsilon\} \quad (2)$$

The algorithm can be described as follows:

---

**Input:** The set of all line segments:  $LSC$ ; the time window:  $TW$ ; Distance threshold:  $\varepsilon$ ; Minimum number of line segments:  $MinLSSum$ ; the angle threshold:  $\theta$   
**Output:** A set of stay region  $SS$

**TPLS\_SR\_DETECTION** ( $LSC, TW, \varepsilon, MinLSSum, \theta$ ):

```

LSC_new = {} //get rid of the NOISE in advance
For each line segment LS in LSC
    If LS.TD ∩ TW ≠ ∅
        LSC_new.add(LS);
C = 0;
For each unvisited line segment LS in dataset LSC_new
    Mark LS as visited;
    N = N(ε, TW)(LS);
    If Size of (N) < MinLSSum
        Mark LS as NOISE;
    Else
        C++;
        Add LS to cluster C;
        For each line segment LS' in N
            If LS' is not visited
                Mark LS' as visited;
                N' = N(ε, TW)(LS');
                If Size of (N') ≥ MinLSSum; //if LS' is a core object
                    N = N joined with N';
            If LS' is not yet member of any cluster
                Add LS' to cluster C;
Return SS; //get the set of all stay regions

```

---

The time complexity of the algorithm above is  $O(n^2)$ , where  $n$  is the number of the line segments in  $LSC\_new$ . If spatial index is used, the time complexity will reduce to  $O(n \log n)$ . The algorithm **TPLS\_SR\_DETECTION** can only detect stop regions where the birds leave or arrive at during  $TW$ . In order to find all stop regions, we use the time window size  $TWS$  and time step  $ts$  to replace time window  $TW$  where  $ts \ll TWS$ . Given a set of line segments  $LSC$ ,  $startTime$  means the time of first location in  $LSC$ ,  $endTime$  means the time of last location in  $LSC$ . A set of time window:

$$TW_{set} = \{[startTime, startTime + TWS], [startTime + ts, startTime + ts + TWS], [startTime + 2 * ts, startTime + 2 * ts + TWS], \dots, [startTime + n * ts, endTime]\} \quad (3)$$

We use the time window parameter in  $TW_{set}$  respectively to call the function **TPLS\_SR\_DETECTION**, and then

merge all the results. If the time window size and time step are appropriately set, we can detect all the stop regions. More details are described as follows:

**Input:** The set of all line segments:  $LSC$ ; the time window size:  $TWS$ ; time step:  $ts$ ; Distance threshold:  $\varepsilon$ ; Minimum number of line segments:  $MinLSSum$ ; the angle threshold:  $\theta$

**Output:** A set of all the stay regions  $SS$

**TPLS\_ALL\_SR\_DETECTION** ( $LSC, TWS, ts, \varepsilon, MinLSSum, \theta$ ):

```
Sort  $LSC$  by time;
 $startTime=LSC.getStartTime()$ ; //get the first location's time stamp
 $endTime=LSC.getEndTime()$ ; //get the last location's time stamp
Get the set of the time window  $TW_{set}$ ;
 $SS=\{\}$ ;
For each time window  $TW$  in  $TW_{set}$ 
     $SS_{TW} = TPLS\_SR\_DETECTION(LSC, TW, \varepsilon, MinLSSum, \theta)$ ;
     $SS = SS \cup SS_{TW}$ ;
Return  $SS$ ;
```

## 5 EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

To verify the efficiency of these three methods, we choose the satellite telemetry data obtained from 29 bar-headed geese captured in the Qinghai Lake Area to run a series of tests. Raw data included 471,774 records of position and time information between 25 March 2007 and 5 June 2009. We selected 40,756 records with higher precision estimates to improve the reliability of analysis.

For  $DBS\_SR\_DETECTION$ ,  $PS$  is the location history obtained from a bar-headed goose numbered BH07\_74901, which has 3502 records of time and location information between 31 March 2007 and 23 November 2008. Under the condition of  $\varepsilon = 20\text{Km}$ ,  $MinPts = 10$ , we find 11 stop regions during this bird's migration (**Figure 5**). The distribution of the stop regions we detected are indicated as **Table 2**.

For  $LHP\_SR\_DETECTION$ ,  $TR$  is the trajectory obtained from the same bar-headed goose as above. Under the condition of  $Dr = 20\text{Km}$ ,  $Tr = 48\text{h}$ , we find 31 stop regions (**Figure 6**). These 31 stop regions are distributed as **Table 2**.

For  $TPLS\_ALL\_SR\_DETECTION$ , the GPS position sampling frequency of the satellite telemetry device was about once every 2 hours during the day. We reduced the dimension of data from hours to days by choosing 2 positions that spanned two sampling times closest to a day. These two locations were regarded as starting and ending points of a line segment. The duration between two sampling times was the duration  $TD$  of the line segment. At last we choose 5,959 line segments to make up  $LSC$ . The time interval is from 25 March 2007 to 4 June 2009. Under the condition of  $TWS=60\text{days}$ ,  $\varepsilon = 80\text{Km}$ ,  $MinLSSum = 2$ ,  $\theta = 10$  degrees. Detailed results are as **Figure 7**. The stop regions we detected are: Qinghai Lake Area; The river valleys near Lhasa; Eling Lake and Zaling Lake; Galalacuo Lake Area; Saiyongcuo Lake Area; Zhamucuo, Niri'a cuogai, Ga'e Encuo Nama Area; Cuo'e Lake and Neri puncuo Area; River valleys near Lhasa; and Cuona Lake, Cuo Lake, Nairipingcuo Lake.

**Table 2.** The distribution of stop regions generated by  $DBS\_SR\_DETECTION$  and  $LHP\_SR\_DETECTION$

Area	Stop region ( $DBS\_SR\_DETECTION$ )	Stopregion ( $LHP\_SR\_DETECTION$ )
Qinghai Lake Area	Stop region 9,10	Stop region 1,2,3,4,5,6,7
DonggeiCuona Lake Area	Stop region 11	Stop region 8,9,10
Eling Lake and Zaling Lake Area	Stop region 7	Stop region 11,12
Galalacuo Lake Area	Stop region 8	Stop region 13
Saiyongcuo Lake Area	Stop region 5	Stop region 21,22,23
Zhamucuo, Niri'a cuogai, Ga'e Encuo Nama Area	Stop point 6	Stop point 24,25,26,27,28,29
Cuo'e Lake and Neri puncuo Area	Stop region 2	Stop region 14,15,17,18,19,30
River valleys near Lhasa	Stop region 1	Stop region 16,31



**Figure 5.** The white mark means NOISE, the marks with the same color belonging to one stop region; the number near the mark is the stop region number.



**Figure 6.** A yellow mark is a stop region and the number near the mark is the stop region number.

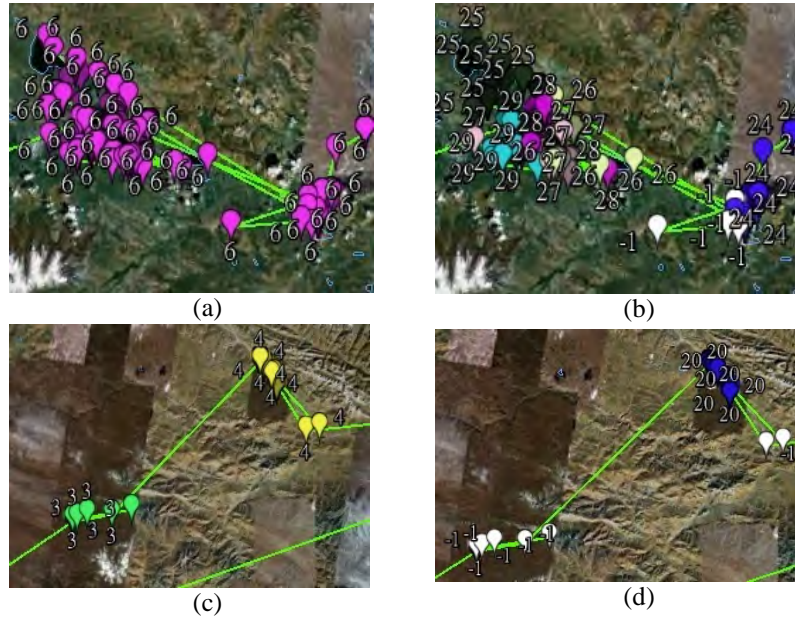


**Figure 7.** Clustering results of long distance segments from 12 June 2007 to 2 April 2009.

From the results above, we can figure out that the stop regions obtained by executing *DBS\_SR\_DETECTION* and *LHP\_SR\_DETECTION* to analyze the same bird's migratory route are similar to each other. Nearly all the stop regions are next to lakes or wet lands (**Figure 5**, stop region 5, stop region 7, stop region 11). While the data handled by *TPLS\_ALL\_SR\_DETECTION* are from all bar-headed Geese, it's not suitable to be compared with the other two algorithms. But we still can find that stop regions obtained by these three algorithms have obvious overlapping areas. Moreover, the result of *TPLS\_ALL\_SR\_DETECTION* is almost the same as the stop regions of bar-headed Geese's migratory routes mentioned in (Tang et al., 2009).

The distance thresholds of *DBS\_SR\_DETECTION* and *LHP\_SR\_DETECTION* are both 20Km, while stop regions obtained from *LHP\_SR\_DETECTION* are much more than those from *DBS\_SR\_DETECTION*. Based on these two algorithms' principles, *DBS\_SR\_DETECTION* only considers the information of spatial dimension, so we can only find out its dense clusters and treat them as stop regions. From microscopic view, this algorithm is unable to analyze data within dense clusters. For instance, **Figure 8(a)** and **Figure 8(b)** indicate the same area.

*DBS\_SR\_DETECTION* treats this area as one stop region, while *LHP\_SR\_DETECTION* obtains several stop regions for it considers both spatial and time dimension. Although these stop regions' spatial positions are next to each other, treating them as different regions still means a lot. What's more, stop region 3(**Figure 8(c)**) discovered by *DBS\_SR\_DETECTION* is treated as noise (**Figure 8(d)**) when executing *LHP\_SR\_DETECTION*. We figure out that birds have 13 position points in the area but never stop there beyond one day and it is probably an exception for this area isn't a perfect stop region. However, *DBS\_SR\_DETECTION* still takes the area as a stop region while *LHP\_SR\_DETECTION* can avoid this incorrect situation. We also notice that stop regions detected by *DBS\_SR\_DETECTION* are without the information of time, while stop regions obtained by *LHP\_SR\_DETECTION* are ordered by time sequence. The three algorithms' further comparison is as **Table 3**.



**Figure 8.** Four special scenarios, (a) and (c) are generated by *DBS\_SR\_DETECTION*, (b) and (d) are generated by *LHP\_SR\_DETECTION*

**Table 3.** Comparison of three methods in detail

	<i>DBS_SR_DETECTION</i>	<i>LHP_SR_DETECTION</i>	<i>TPLS_ALL_SR_DETECTION</i>
<b>Dimension</b>	Spatial	Spatial and time	Spatial and time
<b>Object</b>	Point	Trajectory	Line segment
<b>Raw data</b>	GPS location history	GPS location history	GPS location history
<b>Range</b>	One bird or more	One bird	Multiple birds
<b>Time complexity</b>	$O(n^2)$ or $O(n \log n)$	$O(n^2)$	$O(n^2)$ or $O(n \log n)$

From the experiments above, we find that all these three methods can detect habitats and stopovers on the bar-headed geese's migratory routes. However, their principles lead to their differences in application. *DBS\_SR\_DETECTION* does well in the situation that only cares about stop regions' position. For example, sometimes ornithologists need to know the common stopovers for the whole flock of bar-headed geese during their migratory routes. *DBS\_SR\_DETECTION* may be very suitable for this situation above. The object handled by *LHP\_SR\_DETECTION* is the trajectory, so this algorithm can only deal with one bird's trace once. If we want to analyze more birds' information, we need to perform it multiple times before further processing. This algorithm takes the time factor into account. We can detect stop regions with start and end timestamps, which indicate some bird's arriving and leaving time in some area. This may be useful for studying the relationship between the flyways of migratory birds and climate. The object handled by *TPLS\_ALL\_SR\_DETECTION* is line segment. This algorithm is meaningful only when many birds' trajectories are analyzed. The intermediate products during the process are line segment clusters. According to those clusters, we can easily figure out the fly distance among the stop regions. As indicated in **Figure 7**, observing the lengths of line segment clusters, we find that stop regions around Eling Lake and Zaling Lake are most bar-headed geese's start areas before their long journey. Departing from there, some of the birds make a pit-stop in Niriacuogai Lake, Zamucuo Lake and

Gaencuoname Lake while others fly at one go until Cuona Lake, Cuo Lake and Nairpingcuo Lake. This information may be useful for ornithologists to analyze birds' migration patterns.

## 6 CONCLUSION

Over all, we provide three methods based on data mining for detecting habitats and stopovers on the migratory routes from birds' GPS data. After applying the algorithms on the GPS data of bar-headed Geese captured in the Qinghai Lake region of China, we verify the algorithms' correctness. Having analyzed their principles and distinctions in detail, we give some suggestions about the applying situations of these three algorithms. It'll be helpful for ornithologists to find apposite algorithm for their work. In the future, we'll further study the climate, ecology and other factors in the stop regions on birds' migratory routes.

## 7 ACKNOWLEDGEMENTS

Funding was provided by the Natural Science Foundation of China under Grant No. 90912006; the Natural Science Foundation of China under Grant No. 61003138; The National R&D Infrastructure and Facility Development Program of China under Grant No. BSDN2009-18; Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh 5-Year Plan", e-Science Application of Research on Resources, Disease Monitoring and Risk Assessment of Important Wild Birds in the Qinghai Lake Region under Grant No. INFO-115-D02; Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh 5-Year Plan", Basic Databases of Joint Research Center of Chinese Academy of Sciences and the Qinghai Lake National Nature Reserve under Grant No. INFO-115-C01-SDB2-02; Fund of President of Chinese Academy of Sciences; Found of Director of Computer Network Information Center of Chinese Academy of Sciences; United States Geological Survey (Patuxent Wildlife Research Center, Western Ecological Research Center, Alaska Science Center, and Avian Influenza Program); the United Nations FAO, Animal Production and Health Division, EMPRES Wildlife Unit; National Science Foundation Small Grants for Exploratory Research (No. 0713027); and the Chinese Academy of Sciences (No. 2007FY210700, KSCX2-YW-N-063 and 2005CB523007). The use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government.

## 8 REFERENCES

- Berthold, P., & Terrill, S.B., (1991) Recent advances in studies of bird migration. *Annual Review of Ecology and Systematics* 22, 357-378.
- Cagnacci, F., Boitani, L., Powell, R.A., & Boyce, M.S., (2010) Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Phil. Trans. R. Soc. B* 365(1550), 2157-2162.
- Ester, M., Kriegel, H.-P., Sander J., & Xu, X., (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd international conference on knowledge discovery and data mining*, Portland, OR, USA.
- Fuller, M.R., Seegar, W.S., & Howey, P.W., (1995) The use of satellite systems for the study of bird migration. *Israel Journal of Zoology* 41(3), 243-252.
- Frisch, H., Vagg, R., & Hepworth, H.(Eds.), (2006) Migratory Species and Climate Change: Impacts of a Changing Environment on Wild Animals. CMS Convention on Migratory Species of Wild Animals/UNEP, Bonn, Germany.
- Gaffney, S., & Smyth, P., (1999) Trajectory Clustering with Mixtures of Regression Models. *Proc. 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, San Diego, California, USA.
- Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J., & Ghil, M., (2006) Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models. *Climate Dynamics* 29(4), 423-440.
- Hariharan, R., & Toyama, K., (2004) Project Lachesis: Parsing and Modeling Location Histories. *Geographic Information Science Lecture Notes in Computer Science* 3234, 106-124.
- Han, J., & Kamber, M., (2000) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers
- Lee, J.-G., Han, J. & Whang, K.-Y., (2007) Trajectory clustering: a partition-and-group framework. *Proceedings*

of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W., (2008) Mining User Similarity Based on Location History. ACM GIS '08, New York, NY, USA.

Schiller, J., & Voisard, A.(Eds.), (2004) Location-Based Services, Morgan Kaufmann

Stauffer, C., & Grimson, W. E. L., (2000) Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8), 747–757.

Tang, M., Zhou, Y., Cui, P., Wang, W., Li, J., Zhang, H., Hou, Y., & Yan, B., (2009) Discovery of Migration Habitats and Routes of Wild Bird Species by Clustering and Association Analysis. *Computer Science* 5678, 288-301.

Zhang, Y.F., & Yang, R.L., (1997) Bird Migration Research of China, Beijing published in: China Forestry Publishing House

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W.Y., (2011) Recommending friends and locations based on individual location history. ACM Trans. Web 5(1).

Zheng, Y., & Xie, X., (2010) Learning Location Correlation from GPS Trajectory. 2010.Mobile Data Management (MDM), 2010 Eleventh International Conference, Kansas City, Missouri, USA.

# METADATA MODELLING OF IPv6 WIRELESS SENSOR NETWORK IN HEIHE RIVER WATERSHED

*Wanming Luo, and Baoping Yan*

*Computer Network Information Center, Chinese Academy of Sciences, Beijing, China  
Email: [lwm@cnic.cn](mailto:lwm@cnic.cn), [ybp@cnic.cn](mailto:ybp@cnic.cn)*

## ABSTRACT

*Environmental monitoring in ecological and hydrological watershed-scale research is an important and promising area of application for wireless sensor networks. This paper presents a system design of IPv6 wireless sensor network (IPv6WSN) in Heihe river watershed in Gansu province of China to assist ecological and hydrological scientists collecting field scientific data in an extremely harsh environment. To solve the challenging problems, this paper focus on the key technologies adopted in our project, which is Metadata modeling for IPv6WSN. The system design introduced in this paper provides a solid foundation for effective use of self-developed IPv6 wireless sensor network by ecological and hydrological scientists.*

**Keywords:** Metadata, Ecology, Hydrology, HeiHe River

## 1 INTRODUCTION

Environmental monitoring in ecological and hydrological watershed-scale research is an important and promising area of application for wireless sensor networks (WSN). Its potential to provide dynamic, real-time data about monitored variables of a landscape will enable scientists to measure properties that have not previously been observable.

The adoption of the next generation Internet protocol (IPv6) as the Layer-3 protocol to connect wireless sensors is a promising approach to address current issues of WSN, such as scalability, security, mobility and so on. IPv6 extended address space ( $2^{128}$  instead of  $2^{32}$ ) together with its auto-configuration and mobility capabilities makes IPv6 a suitable protocol for large scale sensor network deployments. Therefore, an IPv6 wireless sensor network is designed and developed by ourselves to assist ecological and hydrological scientists to understand watershed-scale hydrologic cycle and energy balance. The experiment area is the Heihe river watershed, which is a typical continental river basin starting from Qinghai province, through Gansu province and ending Inner Mongolia province in China.

Metadata in IPv6WSN is the descriptive data used to describe the IPv6WSN including the environment, deployment location, data ownership, sensor specifications, sensor status, sensor calibrations and replacements, outlier and error information, etc., which plays a crucial role in processing and properly interpreting raw sensor measurement and management data.

In our project, we define metadata as static, self-describing data for explaining IPv6 wireless sensor networks and node characteristics. We present a metadata model for IPv6 wireless sensor network developed by ourselves for watershed-scale ecological and hydrological research. This model not only involves very rich scientific data but also includes management and control data required by IPv6WSN itself. Obviously, it is necessary to build such a unified metadata model for data transmission and processing. The metadata model introduced in this paper provides a solid foundation for effective use of self-developed IPv6 wireless sensor network by ecological and hydrological scientists.

## 2 METADATA MODELLING

Metadata is structured information that describes, explains, location, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information. The metadata are generally used to describe and structure the principal aspect of data with the aim of sharing, reusing and understanding heterogeneous data sets and allowing the information searching and retrieval [1].

Metadata in IPv6WSN is the descriptive data used to describe the WSN, including the environment, deployment location, data ownership, sensor specifications, sensor status, sensor calibrations and replacements, outlier and error information, etc., which plays a crucial role in processing and properly interpreting raw sensor measurement and management data.

Currently, the metadata need to become an important part of WSN in order to preserve the knowledge of the WSN' status over time. The metadata must describe dynamically the changes of the network status and report them back to other components and systems. For example, if a node changes its location, the system must be able to broadcast a message containing metadata in order to inform other sensor networks and users about these changes. If a node fails, the network must automatically reconfigure new routes to send data. In the same way if a node changes its location, the sensing data (and their metadata) must reflect the new location. [2]

In our project, we define metadata as static, self-describing data for explaining IPv6 wireless sensor networks and node characteristics. The metadata model consists of six categories: *GeneralInfo*, *SensorInfo*, *Processes*, *Position*, *CollectedData* and *ControlInfo*. Some of them also include sub-elements. The IPv6WSN metadata schema is showed in Figure 1.

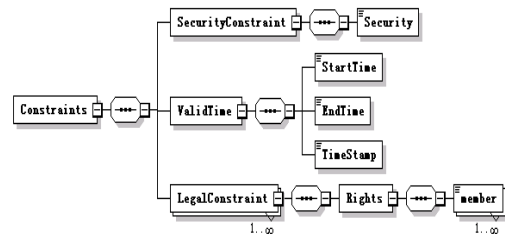


Figure 1. IPv6 wireless sensor network metadata schema.

## 2.1 GeneralInfo Metadata

This metadata includes seven elements, which provide the main information to help users find wireless sensor networks, access to monitoring data attributes, WSN owner's contact information, and the constraints for use of the network and so on. These elements include Identification, Description, Constraints, Properties, References, History and SinkType. The detailed description is showed in Figure 2. Description metadata is used to introduce IPv6WSN supplementary information, and Identification metadata includes the name of IPv6WSN.

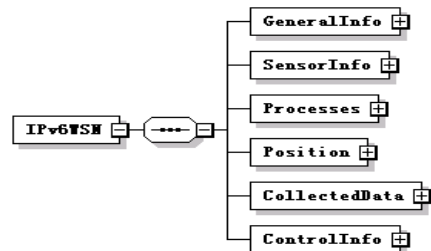


Figure 2. GeneralInfo metadata schema

Constraints metadata consists of three elements: SecurityConstraint metadata, ValidTime metadata and LegalConstraint metadata (see Figure 3). SecurityConstraint describes the security requirements for use of IPv6WSN. ValidTime describes the time interval of IPv6WSN operation. LegalConstraint is used to constraint only registered members can access. This element has multiple members. If LegalConstraint does not exist, IPv6WSN can be used by anyone.

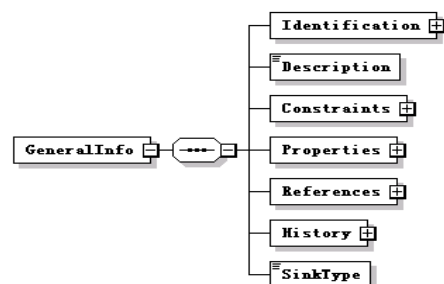


Figure 3. Constraints metadata schema



Properties metadata (see Figure 4) contains one or more attributes. Any wireless sensor networks have their own attributes or characteristics, for example, a wireless sensor network can only collect radiation in a certain frequency range. The application of the network also needs to have some of the data quality requirements, such as acquisition accuracy. Some limitations on WSN in physics or mathematics can also be described in properties metadata.

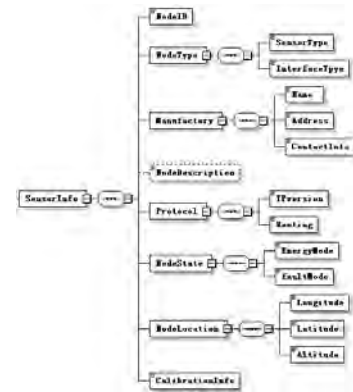


Figure 4. Properties metadata schema

References metadata (see Figure 5) includes at least one Documentations element, which describes the information associated with wireless sensor networks, namely Description, Date, Contact, Format and FileLocation. Each documentation corresponds to a wireless sensor network with the individuals or units, such as wireless sensor network owner. Description describes the general information associated with the object. Date describes the time of the document produced. Format is the format used in the document.

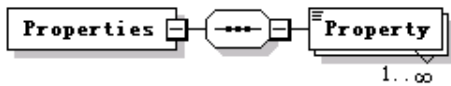


Figure 5. Reference metadata schema

History metadata records the general information and associated changes in wireless sensor networks, which contain one or more Event elements (see Figure 6). Each event has a date element (date), GeneralInfo element and References element.

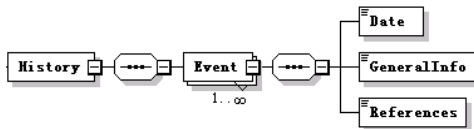


Figure 6. History metadata schema

## 2.2 SensorInfo metadata includes eight elements

The concept of this schema is shown in Figure 7. NodeID is the unique ID of the specified sensor node. The NodeType includes the type of sensors carried by the node and interface conditions. Protocol element reflects the network protocol used by the node, for example IP version (IPv4 or IPv6). NodeState describes the node's energy mode (active or sleep) and operation state (good or fault). NodeLocation includes the node's geographical information. CalibrationInfo records the calibration information, which is very important for improving the sensor accuracy.

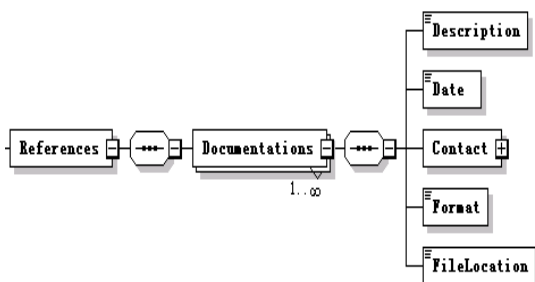


Figure 7. SensorInfo metadata schema

## 2.3 Processes metadata

Processes metadata (see Figure 8) includes at least one process element. Each process represents a process which wireless sensor network can perform. As the wireless sensor node itself has the computing power, a small node with a variety of sensing devices can make some analysis and processing on the physical sensing value. The output is the value processed by the node, for example, the average temperature over time. A process metadata includes input, output and parameters. Input represents the physical phenomena in the natural world the wireless sensor network can sense. Output represents the values after wireless sensor networks process the original sensing data. Parameters describe the requirements or conditions of the process. For example, when the frequency of data collection cannot be less than 2 seconds, otherwise nodes would communicate with each other in the conflict. Parameters are associated with the characteristics of specific wireless sensor network node.

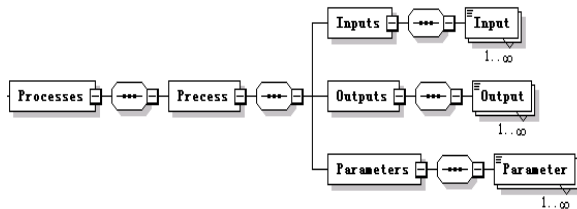
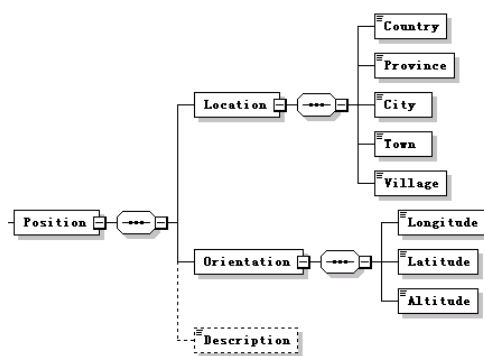


Figure 8. Processes metadata schema

## 2.4 Position metadata



Position metadata is used to describe the geographical location of wireless sensor networks, which includes location element, orientation element and an optional description element (see Figure 9). Location element describes the location information according to administrative division including country, provinces, city, town and village. Orientation element includes longitude, latitude and height. In addition, an optional description can be used to introduce general information of the wireless sensor network deployment area.

Figure 9. Position metadata schema

## 2.5 CollectedData metadata

CollectedData metadata (see Figure 10) describe the available data sets collected by IPv6WSN. In our project, the scientific data need to be collected in the field are mainly multi-disciplinary, multi-scale space-time meteorological, hydrological data, including Lysimeter, Bowen ratio, cosmic rays, soil temperature and moisture profiles, soil heat flux, soil water potential, infrared surface temperature, precipitation, temperature, wind speed, humidity, shortwave radiation, 2cm/5cm/10cm soil moisture and temperature, snow depth and so on. CollectedData metadata includes DataName element, DataType element, DataLength element, Unit element, CollectedTime element, DataDescription element (optional).

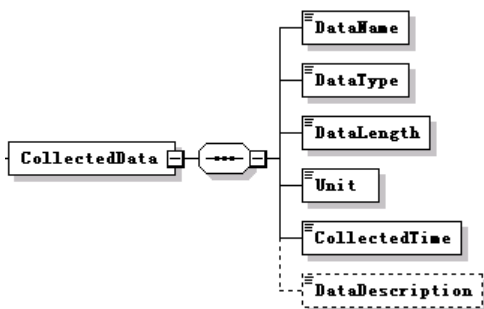


Figure 10. CollectedData metadata schema

## 2.6 ControllInfo metadata

In our project the sensor node not only collects field data but also can be controlled as needed. ControllInfo metadata describes the control instructions or commands sent to sensor node, including NodeID element, SendTime element, ControlType element.(see Figure 11)

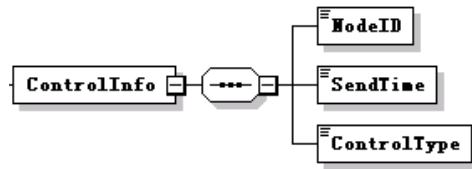


Figure 11. ControllInfo metadata schema

## 2.7 Software system based on the metadata model

Based on the above metadata model, a software system responsible for data management and processing in IPv6WSN is developed (see Figure 12). The programming language is C# and development platform is Visual Studio 2008.



Figure 12. View of the software system

## 3 CONCLUSION

In this paper, we present a metadata model for IPv6 wireless sensor network developed by ourselves in watershed-scale ecological and hydrological research. This model not only involves very rich scientific data but also includes management and control data required by IPv6WSN itself. Obviously, it is necessary to build such a unified data metadata model for data transmission and processing. The metadata model introduced in this paper provides a solid foundation for effective use of self-developed IPv6 wireless sensor network by ecological and hydrological scientists. The conclusion should indicate the significant contribution of the manuscript with its limitations, advantages and applications.

## 4 REFERENCES

1. National Information Standards Organization, understanding metadata, 2004.
2. Daniela Ballari, Monica Wachowicz, Miguel Angel Manso Callejo "Metadata behind the Interoperability of Wireless Sensor Networks" Sensors 2009, 9, pp.3635-3651

# AN INTEGRATED MANAGEMENT SYSTEM OF MULTIPOINT SPACE WEATHER OBSERVATION

*H Watanabe*<sup>\*1</sup>, *K Yamamoto*<sup>1</sup>, *T Tsugawa*<sup>1</sup>, *T Nagatsuma*<sup>1</sup>, *S Watari*<sup>1</sup>, *Y Murayama*<sup>1</sup>, and *K T Murata*<sup>1</sup>

<sup>\*1</sup>*National Institute of Information and Communications Technology, 4-2-1, Nukui-Kitamachi, Koganei, Tokyo, 184-8795, Japan*

*Email: h-watanabe@nict.go.jp*

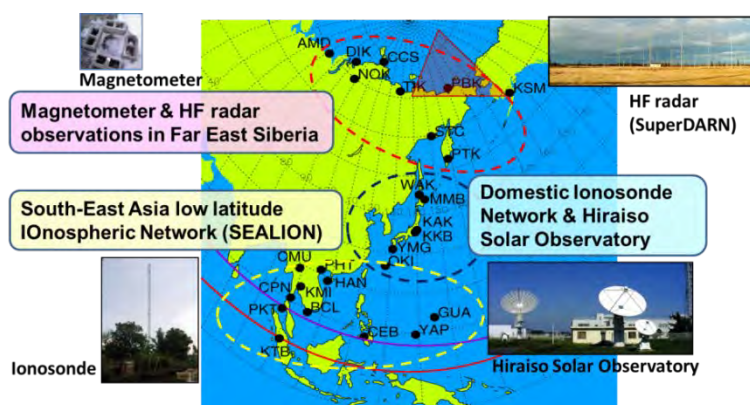
## ABSTRACT

*An outline of a planned system for the global space-weather monitoring network of NICT (National Institute of Information and Communications Technology) is given. This system can manage data collection much easier than our current system, by installations of autonomous recovery, periodical state monitoring, and dynamic warning procedures. According to a provisional experiment using a network simulator, new system will work under limited network conditions, e. g. a 160 msec delay, a 10 % packet loss rate, and a 500 Kbps bandwidth.*

**Keywords:** Space Weather, Data collection, Real time monitoring,

## 1 INTRODUCTION

The NICT (National Institute of Information and Communications Technology) has a project to establish a global observational network of space weather observations (NICT-SWM: Space Weather Monitoring Network) [1-3]. The principal purpose of the project is to improve the reliability of the space weather forecast [4] by introducing real time data obtained by a global network of space-weather related observational facilities, e. g. ionosondes, magnetometers, HF radars and GPS receivers. In this project, NICT will operate about 30 observatories covering a wide area in the northern hemisphere (Figure 1). All observational data will be transferred to NICT (KKB), and stored in a large-scale storage system in a real-time basis. On the other hand, it will become increasingly hard to manage whole the system because of a large number of observational instruments having their own characteristics. The chance of trouble of data transfer networks connecting many observatories will be increased also. Shortage of human resources to maintain the system will be another difficult problem for us. For these reasons, we have developed the integrated management system of global multipoint observations. In this paper, we report an outline of the system.

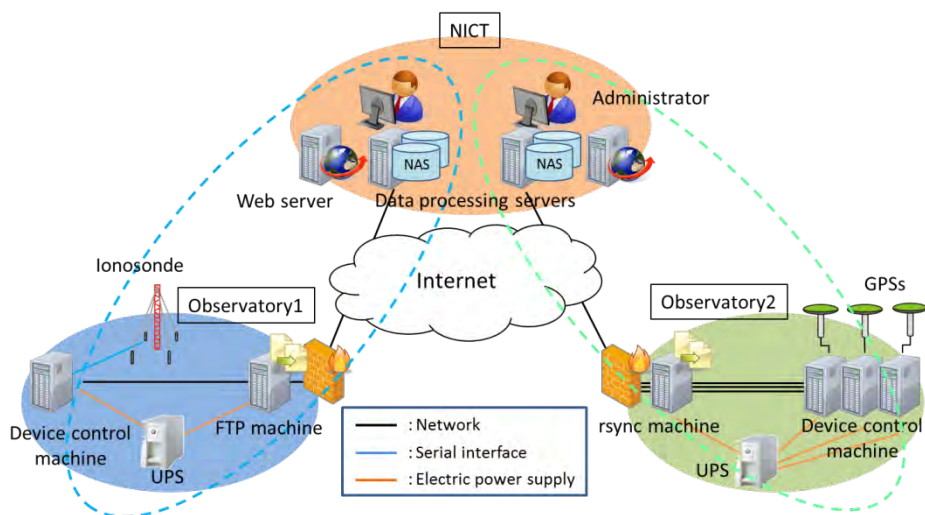


**Figure 1.** Observational network of the NICT-SWM project. The data processing server is located at Kokubunji, Tokyo (KKB).

## 2 CURRENT DATA COLLECTION SYSTEM

Figure 2 shows our current data collection system for space-weather researches and forecasting works. This system is a complicated cluster of about 30 pre-existing observational systems for several independent research

projects, like SEALION [3, 5], having different system architectures, data transfer methods, and secure communication procedures. Each system has been managed by a small number of administrators (researchers), and in an extreme case, only one administrator must manage more than 10 systems in parallel. In addition, the current data collection system has a difficulty in immediate recovery from malfunctions of the system because of complicated situations mentioned above. We will replace the current system with the proposed new system to improve the present situations. Simplification of operational system as a whole and establishment of a rapid failure-recovery mechanism will be the most important points in our new system. This will help us also to reduce workloads of administrators.



**Figure 2.** Current data collection system of the space weather group of NICT.

### 3 PROPOSED SYSTEM

Figure 3 shows a general concept of the proposed system architecture. At a remote station, observational instruments and their data transfer systems are controlled by one agent server with a low power consumption rate. Figure 4 shows the specification of the server. We will apply this concept to other observation systems. The agent server collects observation data and status from observational equipment and stores in a large-scale distributed storage of NICT, which has a 2.2PB disk capacity. We established the storage system already using Gfarm [6], which is a distributed file system developed by Tukuba University for grid computing. Synchronization of data communication between the agent server in each remote station and the storage system in NICT is guaranteed by a combination of VPN (Virtual Private Network) and rsync, which is a free software application for Unix-like systems to synchronize files and directories from one location to another. A data processing cluster takes a role in the status analysis. In addition, we created a real-time monitoring site to manage all systems of NICT-SWM. Figure 5 shows an example of the monitoring page of machine status in the proposed system. The status of a station is indicated by colors of its icon on the Google map. When the color of one of the icons changes from green to red, what an administrator should do is just clicking the icon. The information on machine status of the station will be displayed as shown in the bottom-left of Figure 5, and graphs showing time variations of system parameters are displayed in the right-hand part of the same figure. This new system will allow us to improve the performance of remote management. We confirmed that the system works correctly in a test network environment simulating the network connection between Japan and Seb Island, Philippine, under a limited network condition, e. g. a 160 msec delay, a 10 % packet loss rate, and a 500 Kbps bandwidth. Now, this system is running at some remote stations indicated by the monitoring page.

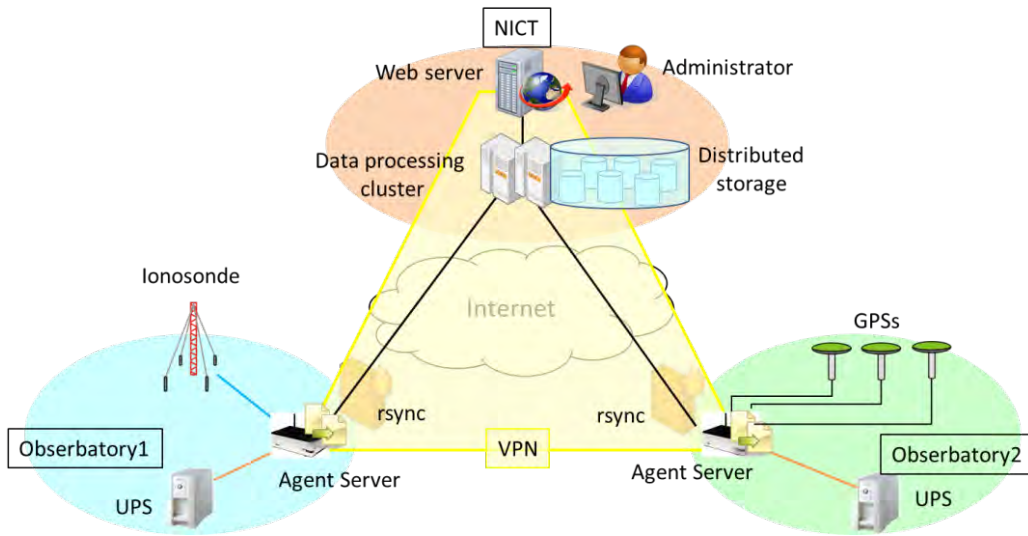


Figure 3. The proposed system architecture of NICT-SWM.



Figure 4. Specification of an agent server.

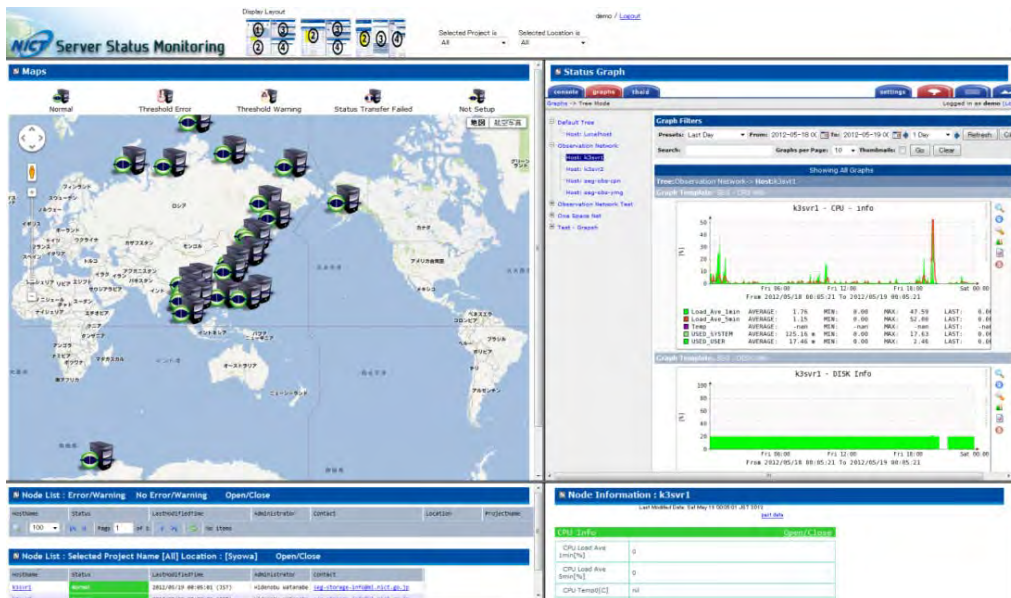


Figure 5. An example of the monitoring page of NICT-SWM.

## 4 CONCLUDING REMARKS

We reported general concept of our new integrated management system of global space-weather observations which has been planned under the space weather project, NICT-SWM. Our system will be pertinent to correct data from many stations distributed in a broad area of the world, in a real-time basis with high reliability. We will evaluate the performance of the proposed system further in detail, and we will set the system in routine operation in near future.

## 5 ACKNOWLEDGEMENTS

This work was conducted by using the machine resource of One Space Net (OSN) in the NICT Science Cloud Service.

## 6 REFERENCES

- [1] Tsutomu Nagatsuma, Monitoring and Forecasting of Geospace Disturbances, and its Importance, Journal of the National Institute of Information and Communications Technology, Vol.56, Nos 1-4, 2009.
- [2] Maki Akioka, Yuki Kubo, Tsutomu Nagatsuma, and Kazuhiro Ohtaka, Monitoring and Warning of Solar Activity and Solar Energetic Particles, Journal of the National Institute of Information and Communications Technology, Vol.56, Nos 1-4, 2009.
- [3] Takashi Maruyama, Susumu Saito, Masabumi Kawamura, Kenro Nozaki, Jyunpei Uemoto, Takuya Tsugawa, Hidekatsu Jin, Mamoru Ishii, and Minoru Kubota, Outline of the SEALION project and Initial Results, Journal of the National Institute of Information and Communications Technology, Vol.56, Nos 1-4, 2009.
- [4]NICT, NICT Space Weather Information Center, [http://swc.nict.go.jp/contents/index\\_e.php](http://swc.nict.go.jp/contents/index_e.php)
- [5]NICT, SouthEast Asia Low-latitude IOnospheric Network (SEALION), <http://wdc.nict.go.jp/IONO2/SEALION/>
- [6]Osamu Tatebe, Kohei Hiraga, and Noriyuki Soda, Gfarm grid file system, New Generation Computing, 28(3):257-275, 2010.

# INTER-UNIVERSITY UPPER ATMOSPHERE GLOBAL OBSERVATION NETWORK (IUGONET)

*H. Hayashi<sup>1\*</sup>, Y. Koyama<sup>2</sup>, T. Hori<sup>3</sup>, Y. Tanaka<sup>4</sup>, S. Abe<sup>5</sup>, A. Shinbori<sup>1</sup>, M. Kagitani<sup>6</sup>, T. Kouno<sup>7</sup>, D. Yoshida<sup>8</sup>, S. UeNo<sup>9</sup>, N. Kaneda<sup>9</sup>, M. Yoneda<sup>6</sup>, N. Umemura<sup>3</sup>, H. Tadokoro<sup>10</sup>, T. Motoba<sup>4</sup>, and IUGONET project team*

<sup>1\*</sup>Research Institute for Sustainable Humanosphere, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan  
Email: hiroo@rish.kyoto-u.ac.jp

<sup>2</sup>Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan

<sup>3</sup>Solar-Terrestrial Environment Laboratory, Nagoya University, Furou-cho, Chikusa-ku, Nagoya 464-8601, Japan

<sup>4</sup>National Institute of Polar Research, 10-3, Midori-cho, Tachikawa, Tokyo 190-8518, Japan

<sup>5</sup>Space Environment Research Center, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

<sup>6</sup>Planetary Plasma and Atmospheric Research Center, Graduate School of Science, Tohoku University, 6-3 Aramaki Aza-Aoba, Aoba-ku, Sendai, Miyagi 980-8578, Japan

<sup>7</sup>Institute for Solid State Physics, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8581, Japan

<sup>8</sup>Weather Information & Communications Service LTD., Fumiei Bldg. 2F, 3-18-36 Minami-Ikebukuro, Toshima-ku, Tokyo 171-0022, Japan

<sup>9</sup>Kwasan and Hida Observatories, Graduate School of Science, Kyoto University, Kurabashira, Kamitakara-cho, Takayama, Gifu 506-1314, Japan

<sup>10</sup>Graduate School of Science, Tohoku University, 6-3 Aramaki Aza-Aoba, Aoba-ku, Sendai, Miyagi 980-8578, Japan

## ABSTRACT

*An overview of the Inter-university Upper atmosphere Global Observation NETWORK (IUGONET) project is presented with brief description of the products to be developed. This is a Japanese inter-university research program to build the metadata database for ground-based observations of the upper atmosphere. The project also develops the software to analyze the observational data provided by various universities/institutes. These products would be of great help to researchers in efficiently finding, obtaining, and utilizing various data dispersed across the universities/institutes. This is expected to contribute significantly to the promotion of interdisciplinary researches, leading to more comprehensive understanding of the upper atmosphere.*

**Keywords** : Metadata, Database, Analysis software, Ground-based observation, Upper atmosphere, Solar terrestrial physics, Earth and planetary sciences

## 1 INTRODUCTION

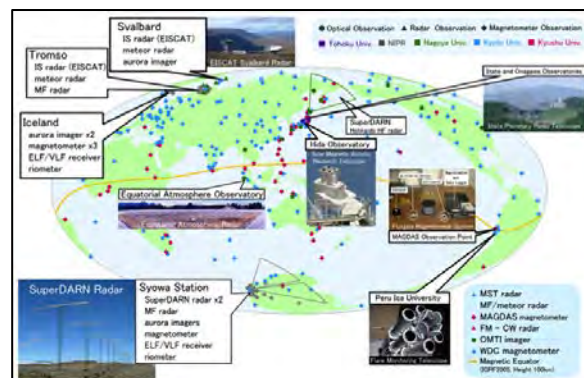
The Earth's upper atmosphere is considered as a compound system consisting of the mesosphere, thermosphere, ionosphere, plasmasphere, and magnetosphere. Although the different atmospheric layers are often referred to as independent regions, they are closely coupled by exchange of material, momentum, and energy through complicated physical processes. While there exist various internal physical processes, the upper atmosphere is strongly influenced by external factors, for example, energy input from the Sun by the ultraviolet radiation, solar wind, etc. and momentum injection from the stratosphere and troposphere by propagating atmospheric waves. What we observe in the upper atmosphere is, therefore, the result of mixing such complicated processes.

To investigate the mechanism of long-term variations of the upper atmosphere, multidisciplinary researches are required with combinations of various types of ground-based observations such as temperature, neutral wind, aurora, geomagnetic field, solar ultraviolet radiation, etc. made at different locations and altitudes. The data or databases of such observations generally have been maintained and made available to the community by each



research organization/group that conducted the observations. Although those data or databases have been well used within certain research communities closely related to the observational activity, they are often difficult to be used by researchers in other research areas due to lack of information on the data. It is also the case that data acquired by some campaign observations have been used by only a very few researchers who were involved in the campaigns and the availability of the data have not been well known for the other people.

A six-year research project, Inter-university Upper atmosphere Global Observation NETWORK (IUGONET, <http://www.iugonet.org/>), has started in 2009 to overcome such problems in data use by National Institute of Polar Research (Space and Upper Atmospheric Science Group), Nagoya University (Solar-Terrestrial Environment Laboratory), Kyoto University (Research Institute for Sustainable Humanosphere, Data Analysis Center for Geomagnetism and Space Magnetism, Kwasan and Hida Observatories), Kyushu University (Space Environment Research Center), and Tohoku University (Planetary Plasma and Atmospheric Research Center). These universities and research institutes (hereinafter, IUGONET institutes) have been leading ground-based observations of the upper atmosphere and the Sun in Japan. Figure 1 shows where and how the IUGONET institutes have been obtaining their data. The data come from observations made at various locations and altitude layers by using various instruments such as magnetometer, airglow imager, radio telescope, solar telescope, atmospheric radar and lidar, etc. They archive a huge amount of and various kinds of observational data, including long-term data obtained over the decades. The IUGONET institutes have formulated a cooperative framework to build a database system for metadata of those observational data. The metadata describes properties of the data, such as observation location and period, type of instrument, data format, and contact information. By sharing such information of the data through the metadata database the IUGONET project intends to facilitate distribution of the observational data among researchers of various disciplines.



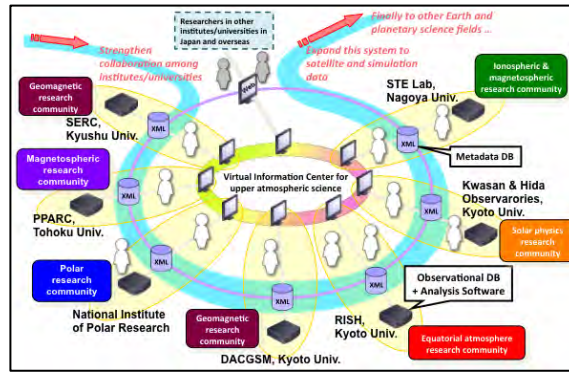
**Figure 1.** World map showing major observation sites from which the IUGONET institutes have been collecting data.

The IUGONET project also develops a data analysis software for those various observational data. It is usually difficult for researchers to use data in an area out of their expertise especially due to differences of data archiving formats. There have been many kinds of data archiving formats used in the long history of the ground-based observations. This might raise another difficulty in promoting the use of data and multidisciplinary researches. To standardize data archiving formats, however, would require too much work to be completed. Instead, the IUGONET project plans to provide researchers with an integrated data analysis software package so that the users could readily handle various data without taking care of difference of archiving formats among them.

An overview of the IUGONET project is described first in this article, followed by the brief introduction of the products to be developed in the project.

## 2 ORGANIZATION AND TIMELINE OF THE IUGONET PROJECT

The project organization chart is shown in Figure 2. The IUGONET project first set up a cooperative framework, named virtual information center for the upper atmospheric science, by introducing remote conference system, electronic mailing list, wiki, etc. to share and exchange any information, opinions, and ideas regarding the project's activities between the IUGONET institutes. Each IUGONET institute newly employed one or two researchers and/or technical assistants dedicated to the project and they organized a core development team. The development team members frequently meet virtually online and discuss many topics of the project in the virtual information center even though they are really far away from each other.



**Figure 2.** Schematic view of the virtual information center of the IUGONET project.

On the basis of the discussions and group works through the virtual information center the IUGONET development team creates metadata of various kinds of observational data archived at their institutes. Then the metadata repository is prepared at each institute, as shown with containers labeled “XML”, and connected to each other on the Internet so that all the metadata can be shared. In addition, an analysis software to handle those observational data is developed. Then the researchers at the IUGONET institutes gather in the virtual information center and discuss new collaborative works using their multidisciplinary data with the developed products. Note that these products mentioned above are made available not only to the researchers within the IUGONET project but also to anyone who is interested in their observational data.

Although the IUGONET project focuses on ground-based observational data of the upper atmosphere, the project plans to realize exchanging of metadata or interoperability between similar e-infrastructure for the satellite-borne and computer simulation data of the upper atmosphere. It also aims at further development of their products for future collaborations with wide variety of disciplines in the Earth and planetary sciences.

Figure 3 shows the project timeline. The main development items of the IUGONET project are designing of the common metadata format to describe their ground-based observational data, building of the metadata database system to archive the metadata, and producing of the analysis software to help users handle those observational data. After the setup of the virtual information center in FY2009, the project started with designing the IUGONET common metadata format and the specification of the metadata database system and data analysis software. According to the specification these products are developed in the second fiscal year. In the project plan, the IUGONET products are made available to the public in the third fiscal year. In the latter half of the project period the generation and archiving of metadata and the development of data analysis software still continue, but more difficult data to be treated, for example, old data poorly documented, undatabased, undigitized, uncomputerized, etc. will be targeted.

ITEMS	FY2009	FY2010	FY2011	FY2012	FY2013	FY2014	REMARKS
Virtual Information Center	Installation & stable operation	Install system			Update system		Construct the integrated research environment (video and/or web conference system, etc.)
	Extension to other disciplines						Wrap up the project and discuss further extension of the system to other disciplines
Metadata DB system	Development	Make prototype	Develop regular system	Release product to public			Design and build the IUGONET metadata DB system on the basis of IDSpace
	Stable operation			Update computers			Conduct regular operation of the metadata DB and customize it as needed
Metadata	Design of metadata format	Revise v1.1 format	Prepare documents	Finalize format, as required			Formulate the IUGONET common metadata format and keep updating it if necessary
	Creation of metadata			Start metadata arrangements (IDSpace)	Target relatively old, undatabased items		Create metadata in the designated format and register them in the metadata DB system
Analysis Software	Survey & Specification of analysis software	Specification	Prepare documents				Design an integrated analysis software to download, visualize, and analyze data provided from the IUGONET institutions
	Programming			Release software to public	Target relatively old, undatabased item		Develop the IUGONET analysis software by using TDAS (a set of IDL subroutines)
Others	Rearrangement of observational DBs		Rearrange DBs or responding to metadata & software development				Rearrange existing observational DBs and newly compile DBs of undatabased items
	Scientific researches			Conduct scientific researches with the IUGONET products			Do interdisciplinary researches using various data from the IUGONET institutions
Management of project website		Build project homepage					Provide project information to the public through the website

**Figure 3.** Development timeline of the IUGONET project.

On the other hand, rearranging of the existing observational databases are necessary in parallel with progress of the development of the main products and this has been continuously addressed at each IUGONET institute from the second fiscal year. The IUGONET development team members will join collaborative researches that use various kinds of observational data so as to examine and improve their products especially in the latter half of

the project. Such research activities by the development team itself will be a strong driver for steady updates of the project products by fixing problems and adding new functions.

### **3 IUGONET COMMON METADATA FORMAT**

There are a variety of metadata formats or data models used to describe data of the Earth and planetary sciences, for example, ISO19115/19139, GCMD DIF, FGDC CSDGM, IPY Metadata Profile, ISTP metadata standard, and so on. The IUGONET development team first investigated these existing formats and finally adopted the Space Physics Archive Search and Extract (SPASE) metadata model (King, Thieman & Roberts, 2010, and Thieman, Roberts, King, Harvey, Perry & Richards, 2010) to describe the upper atmospheric data obtained by the ground-based observations. It is widely used as the common metadata format by Virtual Magnetospheric Observatory (VMO) and other virtual observatories for the solar-terrestrial physics (King, Merka, Walker, Joy & Narock, 2010). Archiving metadata in the SPASE format would promote metadata exchange between such data management organizations all over the world. It is particularly worth noting that the metadata format keeps being maintained and improved by open debates in the SPASE consortium that the researchers in the solar-terrestrial physics from many countries join actively. This is one of the important reasons why the IUGONET has decided to use SPASE as the base of the metadata format.

In addition, to extend the metadata descriptions by SPASE and apply it for the ground-based observational data regarding the upper atmosphere, we have made changes by adding small modifications to the SPASE format, that is, (1) some more words to explain non-digital archival data, (2) words to represent coordinate systems for solar image data, and (3) elements to describe the spatial coverage of each observation. Note that the above modifications (1) and (2) were discussed in the consortium and already have been incorporated into the SPASE metadata model version 2.2.0. The XML schema of the IUGONET common metadata format is available at the project website (<http://www.iugonet.org/data/schema/>).

In the IUGONET common metadata format, not only observational data but also any resources regarding observations, such as instruments, observation sites, researchers, and so on, have their own metadata and all of them are archived in its metadata database. They include metadata referring to each data file (called "Granule"), which enable us to perform a search for data files as well as data sets. See Hori, Kagitani, Tanaka, Hayashi, UeNo & Yoshida et al. (2012) and King et al. (2010) for the details of the metadata format. More than one million of metadata describing various observational data have been archived with the above format and made available for search through the metadata database in late 2011.

### **4 METADATA DATABASE SYSTEM**

The IUGONET development team adopted DSpace as the platform of its metadata database system. DSpace is an open source software to manage digital contents and their metadata in the Dublin Core format and widely used by many academic organizations as their digital repositories. The software contains fundamental functions for registering, retrieving, providing, and harvesting digital data written in various formats. It was confirmed that the system on DSpace could manage metadata written even in the IUGONET common metadata format with some customizations. DSpace was, therefore, expected to fit the project timeline since the IUGONET development team had to establish a stable metadata database system in a short development period.

The IUGONET metadata database system must be continuously maintained even after the termination of the six-year project. This means that its operation and maintenance should go smoothly with anyone who is not one of the original development team members. Since DSpace is in widespread use over the world today, any information concerning operation and maintenance of DSpace-based systems could be easily obtained through various media, especially from the Internet. In fact, most of the IUGONET institutes are managing their academic digital repositories on DSpace. This is another important reason why DSpace was chosen for the base of the IUGONET metadata database system.

The IUGONET metadata database is currently under development and has been opened for beta testing on the Internet at <http://search.iugonet.org/iugonet/>. Users can input there any free word, time period, and/or spatial location to find observational data they are interested in. The web service provides them with information of URL to access the data, if they are available online. Otherwise, at least, information of contact person to ask about details of the data should be given. The description of the metadata database system will be found in Koyama, Kouno, Hori, Abe, Yoshida & Hayashi et al. (2012).

## 5 DATA ANALYSIS SOFTWARE - UDAS

The code of the analysis software for various observational data provided by various universities/institutes, named IUGONET Data Analysis Software (UDAS), is written in the Interactive Data Language (IDL). This is because IDL is a programming language widely used in researches on the upper atmosphere and the solar physics, and therefore a lot of IDL routines produced so far to deal with their observational data can be utilized. UDAS has been developed on the basis of the THEMIS Data Analysis Software suite (TDAS). It is an IDL library developed to analyze data obtained in the Time History of Events and Macroscale Interactions during Substorms (THEMIS) mission (Angelopoulos, 2008). The TDAS library contains a lot of useful functions to download, visualize, and analyze data. It is easy to draw multiple plots of various one- or two-dimensional time series data in a single frame with the TDAS routines. This feature is really suitable for the IUGONET project since it aims at promoting interdisciplinary researches by comparing various kinds of observational data. TDAS is also equipped with the Graphical User Interface (GUI) is so that even users who are not familiar with IDL would be able to readily make quick plots and to perform simple analyses. TDAS was adopted for the data analysis software of the Japanese Energization and Radiation in Geospace (ERG) mission (Miyoshi, Seki, Shiokawa, Ono, Kasaba & Kumamoto et al., 2010), too. Therefore, UDAS has been developed in collaboration with the ERG Science Center.

Tanaka, Shinbori, Kagitani, Hori, Abe & Koyama et al. (2012) mentions further details about UDAS. The software is distributed to the public as a patch for the latest version of TDAS. As of the time of writing, a preliminary version of UDAS (version 1.00.b4) is available for download from the IUGONET website at <http://www.iugonet.org/en/software.html>.

## 6 SUMMARY AND FUTURE SUBJECTS

The IUGONET project has been developing the metadata database system for the ground-based observational data of the upper atmosphere and the integrated analysis software to download, visualize, and analyze the data in order to facilitate the distribution and use of them. The six-year project is currently in its third year and the initial version of the metadata database and data analysis software (UDAS) will be soon released. The metadata already registered will become available to the public through the metadata database and new metadata will be continuously archived even after the release of the products. These IUGONET products would be of great help to researchers in efficiently finding, obtaining, and utilizing various observational data dispersed across various universities/institutes. It is expected that these products contribute significantly to the promotion of interdisciplinary researches, which would lead to more comprehensive understanding of the upper atmosphere, especially the mechanism of its long-term variations.

The project does not confine itself to managing observational data and their metadata of the IUGONET institutes. Instead, it does welcome cooperation with any other universities and research institutes that are interested in the IUGONET activities. It is important to promptly set up a new framework to incorporate these data in order to expand the project.

As mentioned in Section 2 the scientific research using the IUGONET products is one of the major activities in the latter half of the project. While this aims at self-evaluation of the project products to continuously improve them, the project members actively demonstrate to researchers how to use the IUGONET products in the actual scientific studies. Such promotion activities parallel to the upgrade of the developed products would become more important in order to settle them as an essential e-infrastructure in the research communities.

Another future subject of the project is to establish collaborative relations with similar scientific projects that engage in developing e-infrastructure for scientific data. Since the SPASE metadata format is widely accepted in the virtual observatories for the solar-terrestrial physics, as mentioned in Section 3, interoperable access between their metadata services and/or exchange of metadata themselves will be expected. This would provide users of the IUGONET metadata database system with opportunities for using much more science data. It is also important that the IUGONET metadata are utilized by various software and web-based services. There are a lot of such science tools available around the IUGONET, for example, Solar Terrestrial data Analysis and Reference System (STARS) (Murata, Yahara & Toyota, 2001) - a Windows software to search, get, and analyze observational data by using metadata, Conjugation Event Finder (CEF) (Miyashita, Shinohara, Fujimoto, Hasegawa, Hosokawa & Takada et al., 2011) - a web-based service to browse various quick plots available on

the Web over the world, and DATA-showcase system for Geospace In Kml (Dagik) (Saito & Yoshida, 2009) - a visualizing software based on Google Earth that intends to work as a showcase for various data. The IUGONET project would like to collaborate with these projects to build a system to effectively provide the IUGONET metadata to them. This could provide the IUGONET data to more potential users who have never used the IUGONET products. These future challenges would lead to new types of interdisciplinary study on the Earth and planetary sciences.

## 7 ACKNOWLEDGEMENTS

The IUGONET project is supported by the Special Educational Research Budget (Research Promotion) for FY2009 and by the Special Budget (Project) for FY2010 and later years from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. The THEMIS Science Support Team is gratefully acknowledged for allowing us to use TDAS for developing our data analysis software. The authors would like to thank all people who have been involved in setting up and maintaining observational databases at the IUGONET institutes. Special thanks are extended to Ms. Masumi Yokawa for her arrangement and management of observational data at the Space and Upper Atmospheric Science Group, National Institute of Polar Research, and to Dr. Noriko O. Hashiguchi for her maintaining observational databases at the Research Institute for Sustainable Humanosphere, Kyoto University.

## 8 REFERENCES

- Angelopoulos, V. (2008) The THEMIS mission, *Space Sci. Rev.*, **141**, doi: 10.1007/s11214-008-9336-1.
- Hori, T., Kagitani, M., Tanaka, Y., Hayashi, H., UeNo, S., Yoshida, D., Abe, S., Koyama, Y., Kouno, T., Kaneda, N., Shinbori, A., Tadokoro, H., & Yoneda, M. (2012) Development of IUGONET metadata format and metadata management system (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- King, T., Merka, J., Walker, R., Joy, S., & Narock, T. (2008) The architecture of a multi-tiered virtual observatory, *Earth Sci. Inform.*, **1**, pp. 21-28, doi:10.1007/s12145-008-0006-3.
- King, T., Thieman, J., & Roberts, D. A. (2010) SPASE 2.0: a standard data model for space physics, *Earth Sci. Inform.*, **3**, pp. 67-73, DOI: 10.1007/s12145-010-0053-4.
- Koyama, Y., Kouno, T., Hori, T., Abe, S., Yoshida, D., Hayashi, H., Tanaka, Y., Shinbori, A., UeNo, S., Kaneda, N., Yoneda, M., Motoba, T., Kagitani, M., & Tadokoro, H. (2012) Metadata Database Development for Upper Atmosphere (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- Miyashita, Y., Shinohara, I., Fujimoto, M., Hasegawa, H., Hosokawa, K., Takada, T., & Hori, T. (2011) A powerful tool for browsing quick-look data in solar-terrestrial physics: "Conjunction Event Finder", *Earth Planets Space*, **63**, pp. e1-e4, doi:10.5047/eps.2011.01.003.
- Miyoshi, Y., Seki, K., Shiokawa, K., Ono, T., Kasaba, Y., Kumamoto, A., Hirahara, M., Takashima, T., Asanuma, K., Matsuoka, A., Nagatsuma, T., & ERG Working Group (2010) Geospace Exploration Mission: ERG Project, *Trans. JSASS. Aerospace Tech. Japan*, **8**, pp. Tm\_1-Tm\_6.
- Murata, T., Yahara, H., & Toyota, K. (2001) Software design via object-oriented methodology and network database for solar-terrestrial observation data, *IPSJ SIG Technical Reports*, **8**, pp. 31-36.
- Saito, A. & Yoshida, D. (2009) Dagik: A Data-Showcase System for the Geospace, *Data Science Journal*, **8**, pp. S92-S95, doi:10.2481/dsj.8.S92.
- Tanaka, Y., Shinbori, A., Kagitani, M., Hori, T., Abe, S., Koyama, Y., Hayashi, H., Yoshida, D., Kono, T., UeNo, S., Kaneda, N., Yoneda, M., Tadokoro, H., Motoba, T., Miyoshi, Y., Seki, K., Miyashita, Y., Segawa, T., & Ogawa, Y. (2012) Development of IUGONET data analysis software (in Japanese), *J. Space Sci. Info. Jpn.*, in press.
- Thieman, J. R., Roberts, D. A., King, T. A., Harvey, C. C., Perry, C. H., & Richards, P. J. (2010) SPASE and the Heliophysics Virtual Observatories, *Data Science Journal*, **9**, pp. IGY85-IGY93, doi:10.2481/dsj.SS\_IGY-019.

# INNOVATIONS FOR THE CURATION AND SHARING OF AFRICAN SOCIAL SURVEY DATA

*H.L. Woolfrey<sup>1</sup>\**

<sup>\*1</sup>*DataFirst, University of Cape Town, Rondebosch, 7700, Cape Town, South Africa  
Email: [lynn.woolfrey@uct.ac.za](mailto:lynn.woolfrey@uct.ac.za)*

## ABSTRACT

*A substantial amount of data is collected through surveys conducted in Africa, by national statistics offices, international donor organisations, research institutions, and the private sector. However, data management at African national statistics offices is hampered by limited resources. An option for data curation in African countries is the establishment of dedicated institutions for data preservation and dissemination, such as survey data archives and research data centres. DataFirst, at the University of Cape Town, has established an African data centre and is helping to improve African data curation practices through providing data, promoting free curation tools and undertaking data management training in African countries.*

**Keywords:** Survey data, Social Science data, African data, Data curation, Data preservation, Data sharing,

## 1 INTRODUCTION

A substantial amount of data is collected through surveys conducted in Africa, but only a small percentage of this data is preserved in the long-term and an even smaller percentage is disseminated as microdata files to support academic research and policy monitoring. Producers of African data include government statistics offices, international donor organisations, foreign and local universities and other research institutions, and private sector institutions. Their data archiving practices vary, from private survey projects which do not share data, to universities and statistics offices, which have begun to establish infrastructures for the curation and sharing of their data products.

## 2 AFRICAN DATA PRODUCERS AND THEIR DATA CURATION PRACTICES

International donor organisations require country-level data to monitor their regional development projects in Africa. To obtain this data they conduct surveys, independently and in partnership with African governments and regional organisations. Donor organisations often make the microdata files from their projects available for further research. For example, the World Bank provides African data via their online microdata catalogue (World Bank, 2011). African universities and other research institutions are also data producers. However, most of these institutions do not have established data preservation or data dissemination policies or practices. Historically, data sharing among researchers in the region has taken place in an ad hoc manner, and as a result, much valuable quantitative research has been unavailable to the wider research community. Despite the advantages of data sharing espoused by the academic community, many African researchers are still reluctant to share their data. This is because the time and resources required to preserve and disseminate data are not available to them. This is also due to some extent to the dissuading motivations inherent in academic research, which is an environment in which exclusive access to original data can give researchers advantages over rivals in an academic field. Surveys conducted by private sector institutions in African countries also collect valuable data. However, although private sector data producers may provide their data for a fee, generally this data is collected for paying clients and is not available for reuse.

National statistics offices are set up by governments to undertake censuses and surveys to collect statistics to provide evidence for government planning. Most data collected and archived in Africa is official data of this nature. Until recently these statistics were used in-house by government statisticians as evidence for national policymaking. Information based on surveys was provided to the public in the form of reports containing tables of aggregated data. However, with growing international emphasis on the importance of statistical data as a national resource for scientific investigation to foster innovation and to provide feedback for sound national

decision-making, some African leaders have given support for the preservation and distribution of national survey microdata for reuse by researchers. Effective government planning is increasingly seen to depend upon sound policy analysis by researchers utilising survey microdata, which enables them to correct or confirm the findings of government statisticians. The role of the research policy interface, in which sound research by academics enables effective government planning, has come to depend on researchers gaining access to original microdata files (Africa Symposium on Statistical Development 2006:5; African Union, 2009:1).

The management and sharing of data by National Statistics Offices is, however, constrained by several obstacles. These organisations have limited financial and staff resources to curate microdata files and ensure their long-term availability. Despite government lip-service to the value of evidence-based policymaking, and official claims regarding commitment to harnessing empirical data for economic growth, scant government funding is allocated to Statistics offices in African countries. Statistics offices in most countries of the region are chronically underfunded and suffer from shortages of basic equipment such as computers and vehicles (Manning, 2006:1-2; Woolfrey, 2010). Skills shortages and high staff turnover due to low salaries in the public sector also result in a paucity of analytical expertise in these institutions (Lufumpa & Mouyelo-Katoula, 2005:31-32). Government expenditure on statistics in African countries is mainly allocated for data collection, and very little funding is made available to support the long-term preservation and sharing of national data (Kiregyera, 2005: 70-72). The outcome of this is that in many national statistics offices data curation is not practiced in a systematic manner, which has led at times to data losses or the production of unreliable data (Regional Reference Strategic Framework for Statistical Capacity Building in Africa, 2006:131).

### **3 AFRICAN DATA CURATION INSTITUTIONS**

Survey data archives and research data centres are dedicated facilities for the sharing of census and social survey data. In Europe a network of national survey data archives fulfils this purpose cross-nationally. These archives provide the advocacy, institutional links and skilled staff to facilitate data sharing in the region. Data archives acquire, store and disseminate survey microdata for research purposes. (Mochmann, 2005:1). The South African Data Archive, (SADA), based in Pretoria, was established in 1996 to formalise the sharing of South African's official survey data. Researchers can browse SADA's data portal <http://sada.nrf.ac.za/ahlist.asp> and apply online for data to be sent to them on CD or can download the data via an FTP server (South African Data Archive, 2011). SADA is currently the only government funded survey data archive in Africa (Woolfrey, 2010).

Research data centres are university based facilities to give research communities access to census and survey data. These can be set up as national institutional networks for microdata sharing, for example the Canadian Research Data Centre Network (CRDCN, 2011). DataFirst, established at the University of Cape Town in South Africa in 2001, has created a research data centre at the university. This research unit also undertakes projects to support data reuse for research and government planning in African countries. DataFirst's work to assist data usage involves a three prong strategy. Firstly, the unit is working to establish itself as a trusted repository for digital research data to ensure data deposits from African data producers. Secondly, it utilises innovative technology to make this data easily discoverable and obtainable for research purposes. The unit's research data centre provides data access to students and staff at the university. To extend this access to the international research community DataFirst provides online data access via a web portal <http://www.datafirst.uct.ac.za/catalogue3/index.php/catalog>. This site acts as an international portal for African data, and for knowledge exchange around data quality issues pertaining to African microdata. Thirdly, DataFirst undertakes advocacy work concerning data usage for evidence-based policy-making. For example, staff participates in the data advocacy work of regional bodies such as the UN Economic Commission for Africa.

Advancing data analysis skills among African researchers, the unit's original mandate, is still a key part of its work. This is accomplished by hands-on assistance in the data centre and regular workshops in basic and advanced data analysis, which are well attended by researchers from other African countries. Support for data analysis is also provided by a survey analysis webcourse available on DataFirst's website <http://www.saldru.uct.ac.za/courses/>. Research recently undertaken by DataFirst which is in its tenth year of existence has shown that making microdata accessible and providing the tools to manage and analyse this data increases data demand and advances data quality, to the benefit of research and policymaking in the region (Woolfrey, 2010).

Data quality and data curation best practice are further fostered through the unit's Data Quality Project

[http://www.datafirst.uct.ac.za/wiki/index.php?title=Category:DataFirst\\_and\\_Saldru\\_Mellon\\_Data\\_Quality\\_Project](http://www.datafirst.uct.ac.za/wiki/index.php?title=Category:DataFirst_and_Saldru_Mellon_Data_Quality_Project). The project was initiated in 2006 to investigate the comparability and usability of South African government microdata. Project researchers work with the official data producer, Statistics South Africa, to advance the quality of national data products. The project is aimed specifically at South African data, but the work includes innovations to support the quality of microdata produced in or about other African countries. The Data Quality Project focuses on improvements to all the quality dimensions of African data, to ensure their fitness for use. Of concern is the accessibility, relevance, timeliness and accuracy of African data, as well as its comparability and ease of interpretation (US Census Bureau, Methodology and Standards Council, 2006:2-4).

Lessons learned by DataFirst are taken to other African countries through the unit's work with the Accelerated Data Program (ADP). The ADP is funded by the Organisation for Economic and Co-operation and Development (OECD) to advance data curation skills to support evidence-based governance in developing countries. The ADP utilises free and open source data curation software to assist governments to preserve, use and share their national data. The software enables the creation of standardised data descriptions (metadata) to assist data usage and data comparability. The metadata editor is a component of proprietary software developed by NESSTAR, (NESSTAR, 2011) distributed as freeware. Survey metadata created with the editor, and data files, are shared using web-based software, the National Data Archive (NADA 3.1), created by the Development Data Group of the World Bank. The NADA tool allows the creation of online microdata portals to assist best practice in data curation by resource strapped official data producers in developing countries. The NESSTAR metadata editing software, the NADA software, and guides for their usage are distributed by the International Household Survey Network <http://www.surveynetwork.org/home/index.php?q=tools/toolkit> (IHSN, 2011). Since 2008 DataFirst has been working with the ADP to install the data curation software at national statistics offices in African countries, and provide training in their usage. To date the tools have been adopted by national statistics offices in nineteen African countries. Their optimal use is still hampered by a lack of data sharing permissions and policies in some African countries. However, the availability of these resources has overcome many of the technological barriers to effective data curation in these countries, and may set the stage for future cross-country data exchange in the region.

## 4 CONCLUSION

Evidence from the reuse of social survey data produced by national statistics offices, donor organisations and research bodies in Africa can provide feedback to governments to support their development planning. Best practice in data curation is not currently followed by African data producers. Data curation by national statistics offices, which are the main producers of survey data in African countries, are frustrated by technological and human resource constraints (Woolfrey, 2010). Most African researchers also do not curate their data products for reuse, as there are few rewards in academia for devoting resources to data sharing. In South Africa data sharing institutions have been established to assist data reuse. The South African Survey Data Archive is a government archive that disseminates South African national microdata to the research community. At the University of Cape Town in South Africa, DataFirst also curates and shares African social survey data. The work of the unit is aimed at advancing skills for data usage on the continent, and assisting the establishment of data curation infrastructures for future ease of access to African data. Research based on this data is hoped to provide further input to evidence for more effective governance in the region.

## 5 REFERENCES

Africa Symposium on Statistical Development [ASSD]. (2006) Resolutions. In *2006 Africa Symposium on Statistical Development, "The 2010 Round of Population and Housing Censuses"*, Cape Town, South Africa, 30 January – 2 February, 2006. Document ASSD2006/03. Retrieved October 29, 2006 from the World Wide Web: <http://www.statssa.gov.za/asc/WebsiteReports/ASSD2006-03.pdf>

African Union. (2009) African Charter on Statistics. Addis Ababa: Assembly of the African Union, 4 February. Retrieved October 10, 2009 from the World Wide Web: [http://www.africa-union.org/root/AU/Documents/Treaties/text/Charter\\_on\\_statistics%20-%20EN.pdf](http://www.africa-union.org/root/AU/Documents/Treaties/text/Charter_on_statistics%20-%20EN.pdf)

[CRDCN]. The Canadian Research Data Centre Network. (2011) Retrieved September 12, 2011 from the World Wide Web: <http://www.rdc-cdr.ca/>

DataFirst's website information. (2011) Retrieved September 12, 2011 from the World Wide Web:



<http://www.datafirst.uct.ac.za>

International Household Survey Network. (2011) Accelerated Data Program. Retrieved September 12, 2011 from the World Wide Web: <http://www.ihsn.org/adp/>

Kiregyera, B. (2005) A case and some actions for improving statistical advocacy in poor developing countries. *African Statistical Journal* 1: 70-84.

Lufumpa, C.L. & Mouyelo-Katoula, M. (2005) Strengthening statistical capacity in Africa under the framework of the International Comparisons Program for Africa (ICP-Africa). *African Statistical Journal* 1: 30-47.

Manning, R. (2006) Keynote Speech. Forum on African Statistical Development (FASDev II), Addis Ababa, 6-10 February, 2006. Retrieved August 31, 2009 from the World Wide Web: [www.uneca.org/fasdev/speech\\_richard\\_manning.htm](http://www.uneca.org/fasdev/speech_richard_manning.htm)

Mochmann, E. (2005) European cooperation in social science data dissemination. Retrieved August 30, 2009 from the World Wide Web: [http://www.ifdo.org/data/data\\_access\\_conditions.html](http://www.ifdo.org/data/data_access_conditions.html)

NESSTAR website information. (2011) Retrieved September 12, 2011 from the World Wide Web: <http://www.nesstar.com/>

Regional Reference Strategic Framework for Statistical Capacity Building in Africa [RRSF]. 2006. *African Statistical Journal* 2 (May):131-134. Retrieved August 31, 2009 from the World Wide Web: [http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/African.Statistical.Journal\\_Vol2\\_2.Article\\_s\\_6.ReferenceRegionalStrategicFramework.pdf](http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/African.Statistical.Journal_Vol2_2.Article_s_6.ReferenceRegionalStrategicFramework.pdf)

South African Data Archive [SADA]. (2011) Introduction. Retrieved September 12, 2011 from the World Wide Web: <http://www.nrf.ac.za/sada/introduction.html>

US Census Bureau. Methodology and Standards Council. (2006) Definition of data quality: Census Bureau principle. United States Census Bureau. Retrieved August 30, 2009 from the World Wide Web: [http://www.census.gov/quality/P01-0\\_v1.3\\_Definition\\_of\\_Quality.pdf](http://www.census.gov/quality/P01-0_v1.3_Definition_of_Quality.pdf)

Woolfrey, Lynn. (2010) MPhil thesis, University of Cape Town [Unpublished]. [Available from the author].

World Bank central microdata catalog. (2011) Retrieved September 12, 2011 from the World Wide Web: <http://microdata.worldbank.org/index.php/catalog>

# A MATURITY MODEL FOR DIGITAL DATA CENTRES

*W Hugo<sup>1\*</sup>*

*<sup>1</sup>South African Environmental Observation Network, De Havilland Crescent, Pretoria, 0001, South Africa  
Email: wim@saeon.ac.za*

## **ABSTRACT**

*Digital Data and Service Centers, such as is envisaged by the ICSU World Data System (WDS), are subject to a wide-ranging collection of requirements and constraints. Many of these requirements are traditionally difficult to assess and to measure objectively and consistently. As a solution to this problem, an approach based on a maturity model is proposed: this adds significant value not only in respect of objective assessment, but also assists with evaluation of overlapping and competing criteria, planning of continuous improvement, and progress towards formal evaluation by accreditation authorities.*

**Keywords:** Maturity Model, Continuous Improvement, Objective Assessment, Key Performance Areas, Accreditation, Preservation, Data Centers

## **1 INTRODUCTION AND PROBLEM STATEMENT**

Digital Data and Service Centers, such as those envisaged by the ICSU World Data System (WDS), face a variety of key performance areas in respect of their operations, planning, and management – derived from a variety of sources. These may include organizational objectives, user requirements, constraints and requirements imposed by funding agencies, and, of course, the criteria set by the WDS in respect of different categories of membership. In addition, there may be local legal compliance required in respect of preservation and archiving, and technical constraints could include standards for interoperability, cataloguing, processing, and the like.

There are several management problems associated with this wide variety of requirements imposed on a Centre. Examples:

- There is an overlap, though sometimes a subtle difference, in requirements derived from multiple sources.
- Many of the requirements imposed on a Centre cannot be measured objectively, and different observers may come to different conclusions in respect of the current performance of an organization or Centre.
- Knowledge in respect of successful approaches are not easily disseminated or transferred.

## **2 PROPOSED SOLUTION**

A solution to these, and several other smaller management challenges, may be provided by applying the principles of a ‘Maturity Model’ (Humphrey, 1987), analogous to the approach first proposed by Carnegie-Mellon Institute for the assessment and management of organizations involved in software creation and delivery. It provides a framework that addresses many of the management challenges that we have described thus far, and serves as a repeatable and less subjective measuring instrument to assess performance of Digital Data and Service Centers.

## **3 REQUIREMENTS PLACED ON DIGITAL DATA CENTRES**

We will be using a hypothetical data center in the field of earth and environmental sciences to develop our solution. We assume that the data center will be distributed physically (which is increasingly the norm, and adds to the complexity of management), and that it needs to comply with typical interoperability requirements. Such a center might typically expect to

1. **Derive strategic and management objectives** from a business planning process, which, in turn, is subject to financial and other resource constraints, while presumably serving the need of one or more communities. These communities may not all be scientists, and could include the wider public, decision makers, and private enterprise;
2. **Link to a Community of Practice** that imposes constraints and requirements, with the constraints including aspects of mandate and scope of operations, and the requirements often aimed at ensuring interoperability and trouble-free access to the Centre's resources. The latter aspect may include data access policies. The Centre also generally needs to ensure that it meets the requirements of the Communities of Practice that it serves, defining appropriate products and services and service level agreements in the process;
3. **Make provision for physical and software infrastructure** to support its products and services, which may include functions of access, preservation, and processing requirements, as well as measures whereby interruption of service and risk to assets are minimized. This requirement becomes quite complex in the case of a physically distributed system, and may require the separation of archiving/preservation arrangements from those aimed at operational data and services;
4. **Apply due diligence and sound governance** in respect of its operations, covering aspects such as independent oversight, risk management, adequate planning for long-term feasibility, and proper liaison with relevant stakeholders. There may be multiple jurisdictions that impose legal requirements and policy constraints on the Centre.

The large number of requirements and constraints deriving from the above can be arranged into an objective hierarchy (or network, since some of the objectives have multiple links to others), and each of these objectives can theoretically have a goal and current level of performance as a minimum (Brehmer, 2005). This is not new: the process is routinely performed in many private and public organizations as performance management.

The main difficulty lies with the *measurement* of the performance, which, for many of the typical requirements and constraints described above, is often performed arbitrarily and subjectively. The main purpose of this paper is to promote the use of Maturity Models to assist with objective measurement of these.

#### 4 MATURITY MODELS APPLIED TO DATA CENTRE OBJECTIVE HIERARCHIES

The common definition of a Maturity Model is “a [framework] that describes how well the behaviors, practices and processes of an organization can reliably and sustainably produce required outcomes” (SEI, 2012). By creating such a framework, there is several side benefits that can be obtained that will be discussed in detail later on, but the obvious structure in the framework is the descriptions associated with predefined levels of performance. These levels of performance are typically designated as follows:

- **Level 1 (Initial):** Usually associated with ad-hoc approaches, undocumented processes, and little guarantee that a given outcome can be achieved. Knowledge and capacity are centered in individuals. The organization is often ignorant of best practice and of applicable or useful standards and specifications.
- **Level 2 (Repeatable):** Processes are documented in sufficient detail to ensure continuity and allow reliable execution by a number of participants.
- **Level 3: (Defined):** Not only are processes documented, but they are also standardized and aligned where applicable to national or international standards and specifications.
- **Level 4: (Managed and Auditable):** Performance metrics are being collected in respect of achievement of objectives, compliance with standards, and independent audits are performed from time to time to confirm such compliance.
- **Level 5: (Optimized):** Deliberate process optimization is undertaken, and a regime of continuous improvement is possible.

These levels of performance are, of course, generic, and needs to be translated into corresponding descriptions for each of the objective hierarchy elements applicable to a Data or Services Centre. The example in Figure. 1 deals with ‘Meta-Data Interoperability’. Deriving similar descriptions for each performance level across all relevant objectives in the hierarchy leads to a comprehensive ‘Maturity Matrix’.

There are several side benefits and additional uses of this approach, in addition to being able to identify the level that most closely matches current performance (and in the act of doing so, making an objective and repeatable assessment):

- Organizations often do not know where to start. By having access to a maturity matrix, it is possible to evaluate a feasible entry point;
- The matrix can, and should, contain the benefit of prior experience – and each entry may be supported by best practice, standards, guidelines, and specifications.
- It can assist multiple organizations with roughly the same objective hierarchy to align and pursue a shared vision (for example, in the ICSU WDS).
- It assists with planning the next level of performance as a set of explicit, measureable objectives and to priorities such actions that may be needed to achieve it.
- It serves to define a level of performance across a collection of objectives, and as such can be used to envision the requirements imposed by certification or audit authorities, for example, by defining the level of performance required to be certified as a ‘trusted digital repository’.
- It provides and relatively objective way to compare the performance of organizations, should the need arise to do so.

Maturity Matrix or Framework						
		Initial Phase	Repeatable	Defined	Managed/ Auditable	Optimised
		Level 1	Level 2	Level 3	Level 4	Level 5
IN	Meta-Data Inter-operability	No interoperability	Standardised software allow standards-based exchanges on an ad-hoc basis with external applications.	Meta-data interoperability requirements are defined and translated into a set of portal functions and services.	Tasks such as harvesting from participating portals are routine, automated, and can be managed by using portal functionality.	Meta-data duplicates are eliminated, update frequency is managed, optimised harvesting paths, and links are continuously tested.

• **Figure 1.** Example of a Maturity Matrix entry for a Specific Objective

## 5 CONCLUSION

Hence, such an approach can be useful to establish all of the following:

1. Current level of performance;
2. A set of internal objectives and self-assessment against these;
3. A set of future goals and milestones to support a process of continuous improvement;
4. A quality assurance program;
5. Accreditation and external audit mechanisms.

Current work will be extended in the near future to develop specific matrix entries for a wide variety of input requirements, based on the scope discussed in the paper. The intention is to establish this as a community resource that can be edited by any number of collaborators with a view to its refinement, validation, and extension, thereby serving the ICSU WDS specifically, and scientific data systems and services in general.

## 6 ACKNOWLEDGEMENTS

The ideas and concepts described in this paper derive from an informal discussion group on Digital Data Preservation, that in addition to the author, includes Dr Lucia Lötter of the Human Sciences Research Council, South Africa, Dr Heila Pienaar of the University of Pretoria, and Dr Martie v Deventer of the Council for Scientific and Industrial Research, South Africa. The work is supported in part by the National Research Foundation of South Africa and the South African Environmental Observation Network.

Trademark: The ‘Capability Maturity Model’ is a registered service mark of Carnegie-Mellon University.

## 7 REFERENCES

Brehmer, B (2005): “The Dynamic OODA Loop: Amalgamating Boyd’s OODA Loop and the Cybernetic Approach to Command and Control”, *10th International Command and Control Research and Technology Symposium*. Retrieved December 2011 from [http://www.dodccrp.org/events/10th\\_ICCRTS/CD/papers/365.pdf](http://www.dodccrp.org/events/10th_ICCRTS/CD/papers/365.pdf),

GEO (2007): “*Tactical Guidance for Current and Potential Contributors to GEOSS*”, Document 24, Group on Earth Observations, Retrieved December 2011 from [http://earthobservations.org/documents/portal/24\\_Tactical%20Guidance%20for%20current%20and%20potential%20contributors%20to%20GEOSS.pdf](http://earthobservations.org/documents/portal/24_Tactical%20Guidance%20for%20current%20and%20potential%20contributors%20to%20GEOSS.pdf)

Humphrey, W.S (1987): “*Characterizing the Software Process- A Maturity Framework*”, CMU/SEI-87-TR-11, a Technical Report prepared for SEI Joint Program Office. Retrieved December 2011 from <http://www.sei.cmu.edu/reports/87tr011.pdf>

ICSU-WDS (2010), “*Certification of World Data System Facilities and Components*”, Retrieved December 2011 from [http://icsu-wds.org/images/files/Certification\\_summary\\_23Oct2010.pdf](http://icsu-wds.org/images/files/Certification_summary_23Oct2010.pdf)

OAIS (2009): “*Reference Model for an Open Archival Information System*”, Draft Recommendation for Space Data System Standards”, CCSDS Secretariat, Retrieved December 2011 from <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf>

SEI (2011): “*Getting Started*”, Online guidance published by the Software Engineering Institute at Carnegie-Mellon University, Retrieved December 2011 from <http://www.sei.cmu.edu/cmmi/start/>

## 8 APPENDIX: DETAILED OBJECTIVE HIERARCHY

The following sources were applied in deriving an integrated objective hierarchy for a digital data and service center serving the Earth and Environmental data community:

1. Requirements imposed by the **Governance Framework** of the host organization, that includes aspects such as risk management, proper systems engineering and adherence to community requirements, proper technology planning, and the like.
2. Requirements that need to be met for acceptance into the **World Data System** and eventual accreditation (ICSU-WDS, 2010).
3. **GEOSS Architecture** and Interoperability requirements (GEO, 2007).
4. **Open Archival Information System** (OAIS) requirements (OAIS, 2009).

The objectives derived from these sources can be arranged and collated into a hierarchy. The table indicates the source of each objective, and further makes an assessment of its likely contribution to the generic software engineering goals of **Availability**, **Usability**, and **Reliability**. Each of these objectives can be expanded into 5 descriptive levels of performance – part of our future work. A detailed matrix can be obtained as a downloadable spreadsheet: See

<http://data.saeon.ac.za/documentation/it-governance/governance/G328.4.1%20Governance%20Matrix.xlsx>

# GEOPHYSICAL DATA STEWARDSHIP IN THE 21<sup>ST</sup> CENTURY AT THE NATIONAL GEOPHYSICAL DATA CENTER (NGDC)

E A Kihn<sup>1\*</sup> and C G Fox<sup>1</sup>

<sup>\*1</sup> NOAA/ National Geophysical Data Center 325 Broadway E/GC, Boulder CO, USA  
Email: Eric.A.Kihn@noaa.gov

## ABSTRACT

*The World Data Center for Geophysics in Boulder, Colorado is hosted by the National Geophysical Data Center (NGDC). NGDC's vision is to be the world's leading provider of geophysical and environmental data, information, and products. NGDC's mission is to provide long-term scientific data stewardship for geophysical data, ensuring quality, integrity, and accessibility. Faced with ever expanding data volumes and types of data, NGDC is developing more innovative techniques for science data stewardship based in part on data mining and fuzzy logic. Use of these techniques will allow NGDC to more effectively provide data stewardship for its own scientific data archives and perhaps the broader World Data System.*

**Keywords:** Geophysics, Marine Geology, Space Weather

## 1 INTRODUCTION

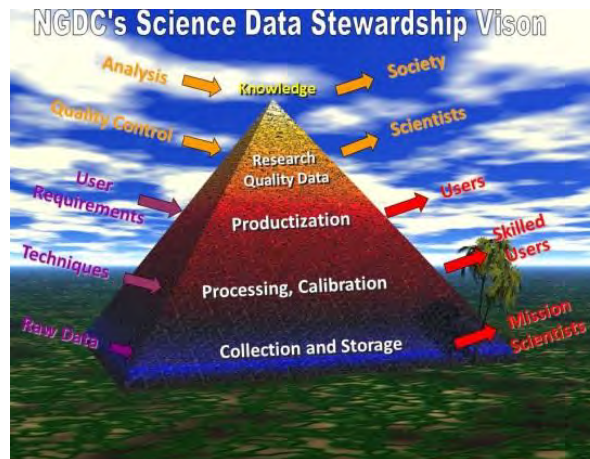
The National Geophysical Data Center (NGDC) [<http://www.ngdc.noaa.gov/ngdcinfo/aboutngdc.html>] is one of three data centers operated by The National Atmospheric and Oceanic Administration (NOAA) to archive and disseminate data collected in executing its environmental mission. NGDC has two primary science divisions each focused on a different domain. The Solar and Terrestrial Physics (STP) division, which focuses on space related and space derived products and information and the Marine Geology and Geophysics (MGG) Division, which focuses primarily on data from the sea floor as well as main field magnetics. A sample listing of the data and applications from each is available in Table 1.

**Table 1.** Sample data products and their application areas

DATA TYPES	APPLICATIONS
Bathymetry	Natural Hazards Assessment & Economic Impact
Digital Elevation Models	Tsunami Inundation Modeling
Gravity & Magnetics	Ocean Mapping
Ocean Drilling	Defense Applications
Seismic Reflection	Cable & Pipeline Routing
Sea Floor Composition	Minerals Exploration
Bottom Pressure Recorder (BPR) Data	Fisheries; Habitats
Natural Hazards Photos	
Significant Earthquakes	Global Change Research
Volcanic Deposits	Climate & Global Change
Solar Imagery	Satellite Operations
NOAA/TIROS Particles	Space Weather Models
GOES Particles and Fields	Electrical Power Networks
Spacecraft Anomalies	Radio Communications
Geomagnetic Variations	Education
Auroral Images	Remote Sensing
Ionospheric Parameters	Global Positioning Satellites
DMSP Particles and Fields	Solar Research
Solar Radiation	Space Research

A primary component of NGDC's mission is to provide scientific stewardship for the data archived at the Center. Here "scientific stewardship" means that in addition to preserving the data for the long term, NGDC focuses on providing calibrated data sets which can reach a broader audience, creating products from raw data thereby exposing the data to a larger audience; providing long term quality control for data sets to create "research quality holdings" and finally propagating the knowledge derived from the data to the community at large. As can be seen in Figure 1, because each of the higher level activities is labor intensive it is performed on a proportionally smaller percentage of the overall data archive thereby reducing the return on investment made in

archiving the data. NGDC is developing tools and techniques that allow the Center to address more of the data at a higher level without increasing overall staff even in the face of increasing data volumes and diversity. The goal is to develop automated “expert systems” that provide stewardship functions without the need for direct staff involvement. The sections below describe the NGDC vision and some early implementations in pursuit of more automated and improved data stewardship.



**Figure 1.** The stewardship pyramid showing decreasing volume for higher order products

## 2 DATA MINING

Data mining is one possible solution in support of stewardship activities. By data mining we mean using mathematical and computational tools to extract previously unknown, and potentially useful, information from the archived data. Data mining uses techniques such as machine learning, and statistical analysis to summarize and present knowledge in a form that is easily comprehensible to humans. By filtering through the vast archives and pointing trained scientists to the more interesting bits of information, data mining enables management of larger and more diverse archives. Some possible applications of these techniques are summarized in Table 2. The first two are addressed with specific examples below.

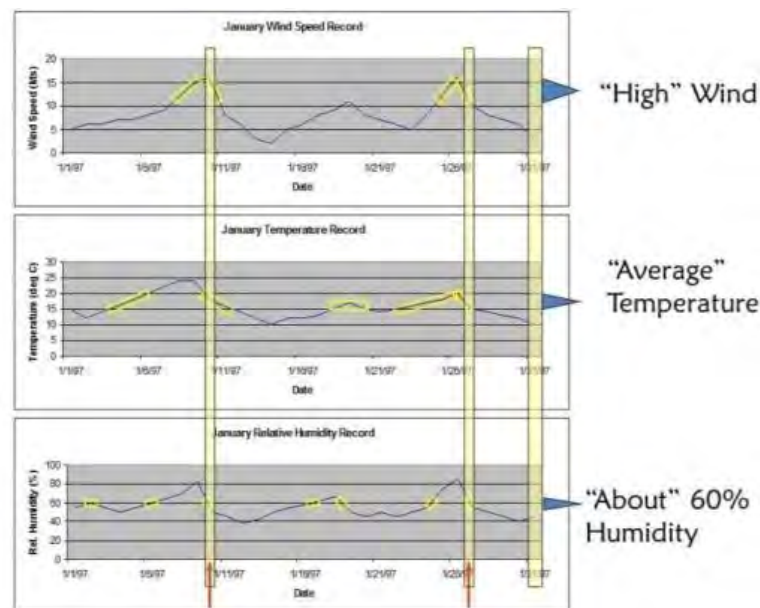
**Table 2.** Applications of data mining.

Applications of Data Mining
• <b>Data quality control</b>
• <b>Human linguistic translation</b>
• Event and trend detection
• Data classification
• Forecast
• Deviation detection

## 3 HUMAN LINGUISTIC TRANSLATION

When attempting to mine data for information natural language is not easily translated into the more computer-friendly terms of simply 0's and 1's. However, natural language is typically how scientist prefer to ask questions when interacting with data: Is the sample “hotter” on average?; Is this observation outside of the “norm”?; Is the sample “changing” with time? Fuzzy logic lets us map human thought and language into computer functions

much closer to the way the brain works. We can aggregate data and form a number of partial truths, which we consider when certain thresholds are exceeded, initiating an action such as flagging the data as suspect or identifying a significant trend. Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth -- truth values between "completely true" and "completely false". It was introduced by Dr. Lotfi Zadeh ( Zadeh, 1965) of UC/Berkeley in the 1960's as a means to model the uncertainty of natural language. The use of "fuzzy" logic allows automated systems to capture some of the natural thought process of a data manager and to apply it to an archive. Applying these techniques, one can search an entire 40-year archive for events described by "High" winds, "Average" temperature, and "about" 60% humidity (perhaps a storm description) and quickly identify when such events are occurring, detect any changes over time, and display the results to a user (Figure 2). Notice that because the language used is natural, the same query would work for data in Alaska or Florida, although what constitutes "average" temperature is obviously quite different between the two. Natural language processing is key to handling large and diverse data volumes and will be expanded at NGDC as ever more automated systems are fielded.



**Figure 2.** A sample search for a typical weather event.

#### 4 DATA QUALITY CONTROL

The Space Weather Reanalysis (SWR) (Kihn, 2007) is a long term reanalysis of space weather data that requires careful quality control of a huge volume of diverse data. The SWR involves taking raw observational data and processing it through linked physical models which produce a higher order product capable of summarizing the state of the space environment. A single instance of bad data can have ripple effects throughout the entire model run. Working with satellite and station data in particular can be tricky, with spikes, baseline shifts, and dropouts all prominent in the data stream (Figure 3). In a typical small scale study it would be possible for a researcher to hand screen the data, but here the data volume requires the application of "intelligent" computer techniques, based on fuzzy-logic, neural computing and other mathematical functions. In particular for this application of data quality control, NGDC developed a system capable of "peer matching"; that is, each station was analyzed to determine a group of peer stations based on location, instrument type and dynamic range. The data mining application was then set to look at the entire 15 year data stream for instances when a given station observed data "unlike" its peers. This much smaller subset of data could then be reviewed directly by an analyst. Notice that in this instance the data mining helps in two ways: by determining a set of peer stations and by allowing a linguistic search for data "unlike" its peers. Application of these techniques allowed for integration of over 15 years of data into the model runs but also left behind a vastly improved data archive, with each station and observation having been screened for quality. NGDC will look to expand usage of such systems as data volumes and diversity increase.



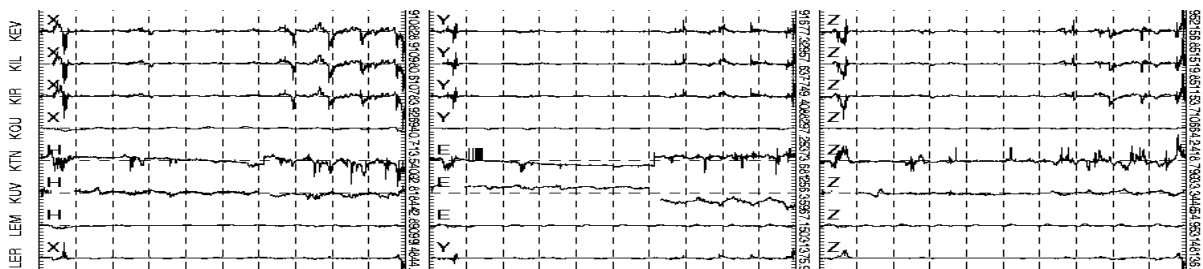


Figure 3. Sample magnetometer data used by the SWR project.

## 5 CONCLUSION

It is clear that increasing data volumes and data diversity demand new tools and methods. While the amount of data and number of data sets tends to increase exponentially (Figure 4) the staff available to manage the data remain level. Mathematical methods exist that provide analysis, classification and forecast methods for large data volumes, specifically the data mining and fuzzy logic systems mentioned above. In particular, fuzzy-based systems hold great promise as knowledge extraction tools, allowing for better information extraction from the vast and diverse archives available. One of the greatest challenges facing science in the coming years is how to effectively utilize the data archived not only at a single center but available across a distributed network such as the World Data System. The techniques described above can play an important role in this effort by better integrating the data and helping scientists to focus on the most relevant bits. Without the development of such tools and systems, the extensive data archives of the World Data System will be vastly underutilized and their scientific potential squandered.

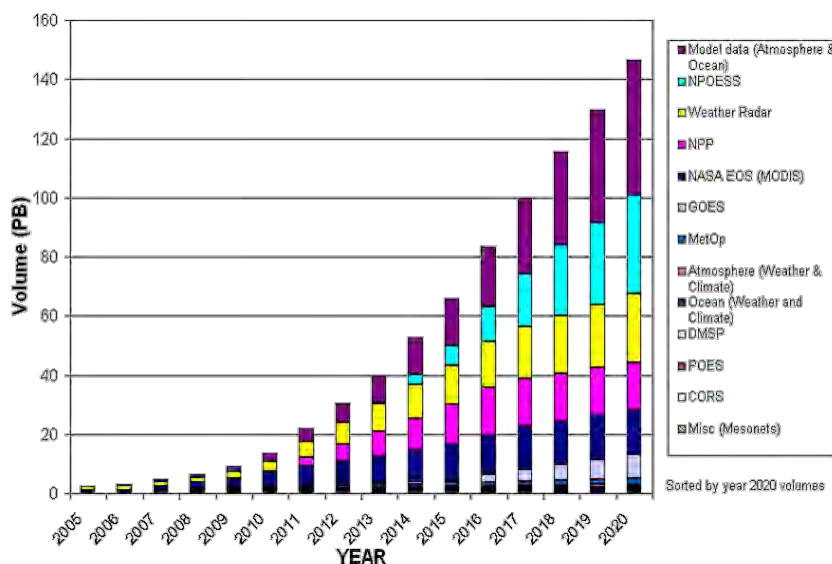


Figure 4. Expected data volume growth at NGDC

## 6 REFERENCES

Kihn, E.A., Ridley A.J. & Zhizhin, M. (2007) The Space Weather Reanalysis, in: *Materials of the International Conference '50th Anniversary of the International Geophysical Year and Electronic Geophysical Year'*, GC RAS, Moscow, doi:10.2205/2007-IGY50conf.

Zadeh, L.A. (1965) *Information and Control* Volume 8, Issue 3, June 1965, Pages 338–353

# APPLICATION AND METADATA FORMAT OF CRYOSPHERE DATA ARCHIVE PARTNERSHIP (CrDAP)

*Hironori Yabuki*<sup>1\*</sup>

<sup>1\*</sup> *Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, Yokosuka, 237-0061, Japan*

*Email: yabuki@jamstec.go.jp*

## ABSTRACT

*An international data archive is critical for understanding the climate system dynamics of the cryosphere. Currently, no such system exists though data collection and integration efforts are ongoing. The Cryosphere Data Archive Partnership (CrDAP) is developing an open system for storing cryospheric observation data and metadata. First stage data handling in CrDAP focused on integrating point observational and photographic data. The metadata structure of CrDAP was extended based on ISO 19115, which is a geographic information metadata standard of the International Organization for Standardization (ISO).*

**Keywords:** Cryosphere, CrDAP, Metadata, ISO 19115

## 1. INTRODUCTION

The Eurasian cryosphere is an integral component of the earth's climate system, comprising frozen surfaces and structures such as glaciers, frozen ground and snow elements. Large fluctuations in the Eurasian cryosphere have been the focus of recent reports, including the IPCC AR-4 report (IPCC, 2007), which expresses great concern regarding the social impact of such fluctuations. Currently, worldwide snow and ice data collection methods are promoted by United States data centers including the National Snow and Ice Data Center (NSIDC) and the National Climate Data Center (NCDC). However, frozen ground and snow data are not stored in an international data organization center such as the World Meteorological Organization (WMO), at which the international and systematic data archive is poor. For a better understanding of the cold regions and wide fluctuations characterizing the Eurasian cryosphere, which encompasses several countries, a wide-reaching and improved coordination of cryospheric data is imperative. The IGOS-Cryosphere (IGOS, 2007) and IPY (Krupnik et al., 2011) have also pointed to the need for such a system and, specifically, one that includes Global Earth Observation System of Systems (GEOSS) data archiving functions.

## 2. PURPOSE

The purpose of this project is to showcase the reality of global environmental changes in the Eurasian cryosphere, by promoting data collection efforts and the cataloging of vital new and legacy cryospheric observations to the public via widespread data digitization. While this project operates in Japan, the data sources will come from all cold regions in the Eurasian countries and will be integrated into our system to promote their release and subsequent publication. Our database, called the Cryosphere Data Archive Partnership (CrDAP; accessible at: <http://www.jamstec.go.jp/acdap/>; Fig. 1) is intended to disseminate observational information to advance scientific understanding of the global climate system.

## 3. THE DEVELOPMENT OF METADATA

### 3.1 Types of data related to cryosphere research

In the framework of cryosphere research, a variety of data types exist in various formats. Designing a metadata structure that incorporates and integrates all data types is a challenging endeavor. Table 1 describes the data

classifications and forms used for planning CrDAP user registration formats. In the first phase of database development, the ground station meteorological data and photographic data were selected, and the structure of the metadata encompassing these data sets was determined. It functioned to archive datasets and metadata sets.



**Figure 1.** The top image of Cryosphere Data Archive Partnership (CrDAP)

**Table 1.** Data classification and format form of planning registration to CrDAP. P is point observation, A is areal observation and G is grid data.

Data category	element	detail	Type
<b>Ground station Observation</b>	(1) Meteorological and climate data	air temperature, humidity, wind, pressure, radiation, precipitation, soil moisture	P
	(2) Snow	depth, density, coverage, etc.	P
	(3) Frozen ground	ground temperature, melting depth, ice volume	P
	(4) Glacier	mass balance, glacier type, velocity, ice depth, ice temperature	P
	(5) Hydrology	river discharge, river frozen condition (icing and melting date), water temperature	P
	(5) Lake	water level, lake area, lake frozen condition (icing and melting date), water temperature	P, A
<b>Remote sensing product</b>	(1) Snow	distribution	G
	(2) Glacier		
	(3) Vegetation		
<b>Map</b>	(1) Frozen ground map	distribution	A
	(2) Vegetation map		
	(3) Glacier map		
<b>Photograph and picture</b>	(1) Glacier photo	aerial photograph, ground-based photograph	P, A
	(2) Frozen ground photo		
	(3) Vegetation photo		

### 3.2 The metadata structure

Before designing the CrDAP metadata structure, we investigated the status of existing atmospheric database projects. Assuming the exchange of metadata to a variety of portals, metadata structure is a necessary condition that conforms to international standards. We found several such projects underway, including the Inter-university Upper atmosphere Global Observation NETwork project (IUGONET; Kouno et al., 2011), which targets data in a polar region archive format. This project serves as an important step toward developing a system for inter-agency metadata cooperation. The Data Integration and Analysis System (DIAS) project (Kinutani et al., 2010), aims to gather all information related to earth observations and others project, and manages unprecedented amounts and levels of data of varied quality. The DIAS project analyzes a data document, collects the metadata and changes the metadata items to correspond to those of ISO 19115 (ISO, 2003). DIAS metadata structure can be viewed as a framework that encompasses the lowest common denominator among common items related to earth observations.

The CrDAP will be compatible with, and potentially extend beyond, the DIAS metadata structure. We will include fine-grained information into the CrDAP metadata structure and into the data sets themselves so that a user can generate a single metadata data set that includes two or more observation points, and retrieve all requested information on an individual observation.

In CrDAP, the metadata item numbers reflect the data types, and the CrDAP metadata structure can be divided into two units, a core unit common among the datasets (e.g., data set name and data provider information) and a variable item unit (e.g., site information) depending on the data item type. The latter unit type enables replacement of metadata as needed. A conceptual diagram of the data and metadata registration flow is shown in Figure 2. A variable item unit can be increased based on international standards, as dictated by the data form type. We patterned our CrDAP metadata structure after these features of the DIAS database, and made it expandable to other metadata objects and data types as needed. We used Microsoft Excel (Microsoft corp.) for entry of CrDAP metadata as an input tool. Because CrDAP assumes that the metadata structure will be populated by users and data providers in the future. Other projects often adopt Web interface registration methods for metadata online; however, the CrDAP project utilized Excel because it is the most common format used to create metadata files and is utilized and stored in an off-line environment, residing securely on the user's desktop. For a metadata author and a site administrator to check the contents of registered work, it is useful to keep the description information associated with the metadata. Finally, the metadata is transformed to XML format using an Excel macro and is used for registration work to the CrDAP site.

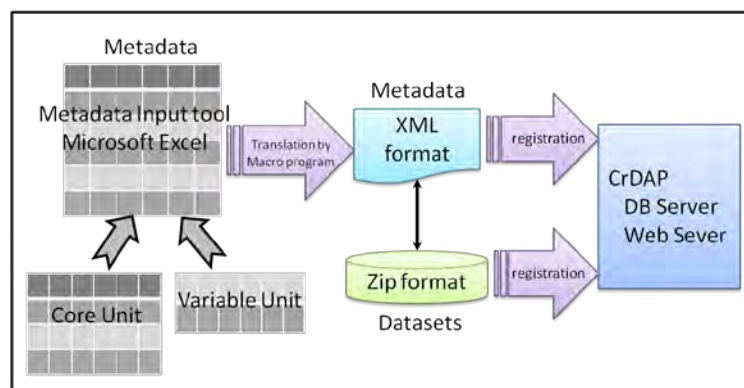


Figure 2. A conceptual diagram of data and metadata registration flow in CrDAP

## 4. APPLICATION OF CrDAP

### 4.1 Simple search and detailed search

In CrDAP, the data register a metadata XML file and pass the contained data set information, including the fine-grained information contained in the metadata itself, to a MySQL relational database management system for data storage and management. A data search can be conducted by querying the MySQL database in CrDAP. The system can distinguish between simple and detailed searches, and indicates whether the data set returned with the specified search conditions is partial or complete. Data search criteria can also request the fine-grained

information embedded within the data sets.

## 4.2 Quick look

CrDAP does not implement a visualization function. However, it is vital to review the data summary before downloading the data set, which we achieved using Quick Look (Apple corp.). To create a Quick Look image, text and binary data are drawn and prepared independently. The resultant Quick Look image can be displayed on a browser presenting the search results. Photographic data is targeted data visualization, have implemented the Web user interface for CrDAP.

## 4.3 Download data

Restricted data set access files are displayed in different colors in the list of search results. A validated user ID and password are required to access the data download function. All accessible data sets are available for download in compressed zip formats.

## 5. CONCLUSION

The CrDAP team has constructed a highly flexible and powerful archive system for storing and managing cryospheric data and metadata. The first phase of development included determining the requirements for registration of the metadata structure for recognizing ground station meteorological and photographic data. The system is downloadable from a Web server, and the CrDAP metadata structure was made to be extended for recognizing and storing other data types, based on the ISO 19115 data standardization criteria. The CrDAP metadata structure performed the design which replaces metadata structure per unit of data type. This new metadata structure is mandatory for future data manipulation requirements, and we can respond to additional requirements by the simple replacement of a unit.

## 6. REFERENCES

IGOS (2007) Integrated Global Observing Strategy Cryosphere Theme Report - For the Monitoring of our Environment from Space and from Earth. Geneva: World Meteorological Organization. WMO/TD-No. 1405. 100 pp.

IPCC (2007) Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.)]. IPCC, Geneva, Switzerland, 104 pp.

ISO (2003) ISO 19115:2003 Geographic information –Metadata. Geneva, International Organization for Standardization (ISO).

Kinutani, H., Shimizu, T., Yoshikawa, M., and Kitsuregawa, M. (2010) A Multidisciplinary Researchers' Collaboration for Disclosure of Earth Science Data in DIAS. The Institute of Electronics, Information and Communication Engineers. Technical Report of IEICE., 110, 328, DE2010–33, 45–50, (in Japanese)

Kouno, T., Koyama, Y., Hori, T., Abe, S., Yoshida, D., Hayashi, T., Shinbori, A., Tanaka, Y., Kagitani, M., Ueno, S., Kaneda, N., and Tadokoro, H. (2011) Development of Metadata Database for Upper Atmosphere Observation by using DSpace. The Third Forum on Data Engineering and Information Management - DEIM 2011 - DEIM Forum 2011 C8-5 (in Japanese).

Krupnik, I., et al., Editors. (2011) Understanding Earth's Polar Challenges: International Polar Year 2007–2008. University of the Arctic, Rovaniemi, Finland /CCI Press (Printed Version), Edmonton, Alberta, Canada and ICSU/WMO Joint Committee for International Polar Year 2007–2008.

# TOWARD A NORMALIZED XML SCHEMA FOR THE GGP DATA ARCHIVES

*Alban Gabillon<sup>1\*</sup>, Jean-Pierre Barriot<sup>1</sup>, Yuri Verschelle<sup>1</sup> and Bernard Ducarme<sup>2</sup>*

<sup>1\*</sup>Université de la Polynésie Française. BP6570, 98702 Faa'a. French Polynesia

Email: {[alban.gabillon](mailto:alban.gabillon@upf.pf),[yuri.verschelle](mailto:yuri.verschelle@upf.pf),[jean-pierre.barriot](mailto:jean-pierre.barriot@upf.pf)}@upf.pf

<sup>2</sup>Royal Observatory of Belgium, Av. Circulaire 3, B-1180 Brussels, Belgium

Email: [bf.ducarme@gmail.com](mailto:bf.ducarme@gmail.com)

## ABSTRACT

Since 1997, the Global Geodynamics Project (GGP) stations use a text-based data format. The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing. As a result, metadata and even data must be checked by human operators. We propose in this paper a new format for representing the GGP data. This new format is based on the eXtensible Markup Language (XML).

**Keywords:** GGP data, XML schema

## 1 INTRODUCTION

Since 1997, GGP stations use a text-based data format known as PRETERNA. The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing. As a result, metadata and even data must be checked by human operators. We propose in this paper a new format for storing and disseminating the data coming from the worldwide GGP network of superconducting gravimeters, in order to streamline the data processing and to enable the scientific community to access these data and their ancillary metadata through distributed, integrated information technology systems and virtual observatories. This new format is based on the eXtensible Markup Language (XML, Bray et al., 2006) that ensures the consistency, reliability and integrity of the data over the Internet and between any data processing platforms. Section 2 of this paper reviews the GGP network of superconducting gravimeters, section 3 outlines the main drawbacks of the current text-based GGP data format. Section 4 presents our new data format based on an XML *schema* (Thompson et al., 2004). Section 5 concludes this paper.

## 2 THE GGP NETWORK OF SUPERCONDUCTING GRAVIMETERS

The Global Geodynamics Project (GGP) is an international network of 25 superconducting gravimeters (Crossley et al., 1999) in operation since July 1997, under the umbrella of the International Association of Geodesy (IAG). The continuous monitoring of timevariable gravity from seconds to years is a tool to investigate many aspects of global Earth dynamics and to contribute to other sciences such as seismology, oceanography, earth rotation, hydrology, volcanology, and tectonics. Another promising application is the use of SG subnetworks in Europe and Asia to validate time-varying satellite gravity observations (GRACE, GOCE) due to continental hydrology and large-scale seismic deformation. GGP plays a small but important role in the Global Geodetic Observing System (GGOS), a primary program of the IAG to coordinate the recording and dissemination of all geodetic data for Earth monitoring, namely the recording of the gravity field and especially its time variations (Crossley & Hinderer, 2009). GGP was incorporated into the IAG as Inter-Commission Project #3.1 in 2003; it is a joint project between Commission 3 (Earth Rotation and Geodynamics) and Commission 2 (The Gravity Field). It is expected to become a full Service of IAG in 2014.

## 3 THE CURRENT GGP DATA FORMAT

All GGP stations use the data format proposed by Wenzel (1996), known as PRETERNA, in which every value (predominantly gravity and pressure), are time tagged in the original units (volt). The only processing is a decimation filter from the original samples to 1-minute values, but no other corrections are done. The full signal is saved with a precision of 7.5+ digits, ensuring that the tides are adequately recorded as well as the smallest

tidal waves. A full discussion of data treatment is given in Hinderer et al. (2007). Users should realize that gaps, spikes and offsets still have to be treated if a clean continuous time series is required, or otherwise avoided if the series is processed as non-contiguous blocks. These 1-minute raw data files are stored at GFZ Potsdam (<http://isdc.gfz-potsdam.de/>). The International Center for Earth Tides, a Service of IAG, provides corrected minute data (i.e. manually cleaned for gaps, spikes and offsets) on their website (<http://www.bim-icet.org/>), but this treatment is designed for tidal analysis and may not be suitable for all purposes, especially long period studies. A GGP 1-minute file is a column-driven file made up of 2 sections, each section being subdivided into 3 parts:

1. The header

1.1 - first ten required lines (ancillary information about the GGP station and instrument)

1.2 – optional text lines inserted by SG group (free comments)

1.3 – two required text lines

2. The data

2.1 – one required introductory line

2.2 – lines of timetagged gravity and pressure data

2.3 – last required termination line

An example of data file is given in Table 1, and the complete data format descriptor (last updated 10 December 2008) is available for download at <http://www.eas.slu.edu/GGP/ggpnews19a.pdf>. This format, in use since 1997, is based on Hollerith punched cards style formats, as FORTRAN character fields (A descriptor), integer fields (I descriptor) and float fields (F descriptor). The main drawbacks of this type of data coding is the lack of data integrity during the data flow processing as described at <http://www.eas.slu.edu/GGP/ggpnews5.pdf>, and the lack of a strict enforcement of data field lengths. As a result, metadata and even data must be checked by human operators. Moreover, this data format includes text-based tags like 77777777 or 99999999 without implicit semantics.

**Table 1.** Current GGP data format

```

Filename           : H2050300.GGP
Station            : Bad Homburg, Germany
Instrument          : GWR CD030_U
Time Delay (sec)   : 45.0      2.0      estimated
N Latitude (deg)   : 50.2285   0.0001 measured
E Longitude (deg)  : 8.6113    0.0001 measured
Elevation MSL (m) : 190.0000   0.1000 measured
Gravity Cal (uGal/V): -67.92    0.02    measured
Pressure Cal (hPa/V): 1.0      0.001   nominal
Author             : P. Wolf (peter.wolf@bkg.bund.de)
yyyymmdd hhmmss gravity(V) pressure(V)
C*****
77777777          0.0      0.0
20050301 000000 -0.504559 993.78749
20050301 000100 -0.502637 993.79867
20050301 000200 -0.500711 993.81193
..
20050320 042800 -1.1410631001.19516
20050320 042900 -1.1415471001.19009
20050320 043000 -1.1420611001.18142
99999999
77777777          0.0      0.0
20050320 161100 -0.151548 998.28556
20050320 161200 -0.146616 998.29147
20050320 161300 -0.141674 998.30143
..
20050331 235700 -0.8851071004.02740
20050331 235800 -0.8876941004.03534
20050331 235900 -0.8902831004.04113
99999999

```

## 4 THE NEW XML DATA FORMAT

Writing GGP files in XML has several advantages:

- XML is a markup language. Data fields are clearly separated by *tags*.
- Since tags are user-defined (XML is not restricted to a predefined limited set of tags like HTML), tags convey semantics specific to the application domain.
- XML files can be automatically analyzed for data treatment/presentation with an XML *parser*.
- An XML file can be checked against an XML schema. An XML schema is a special XML file that specifies a vocabulary identified by a *namespace* (Bray et al., 2006), and some grammatical rules. An XML file that respects the rules dictated by a particular XML schema is said to be *valid*. Checking the validity of an XML file is an *automated* process.

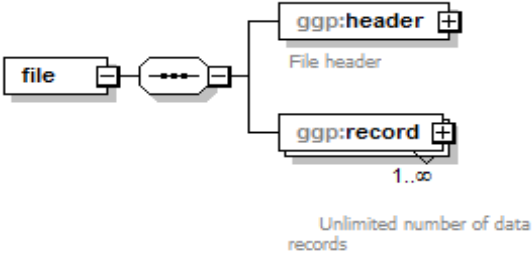
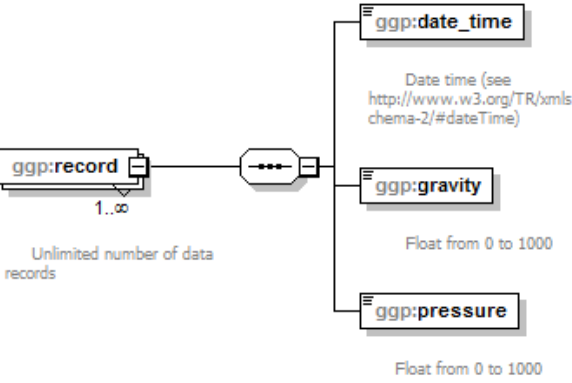
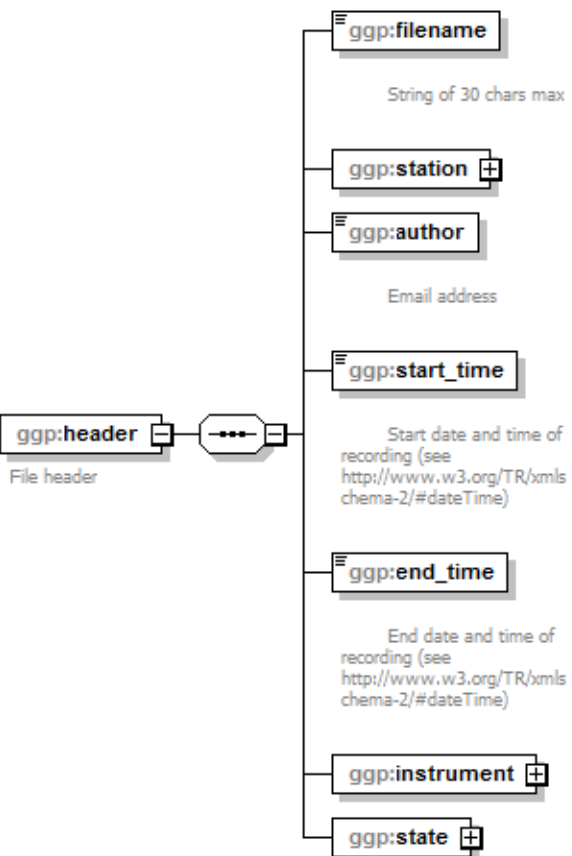
In this section, our objective is to propose an *XML GGP schema*. Our XML GGP schema defines the legal building blocks of the XML GGP files. Our XML GGP schema defines its own namespace identified by the GGP web page URL: <http://www.eas.slu.edu/GGP/ggphome.html>. The schema itself can be accessed at the following URL: <http://pages.upf.pf/Alban.Gabillon/ggp/ggp.html>.

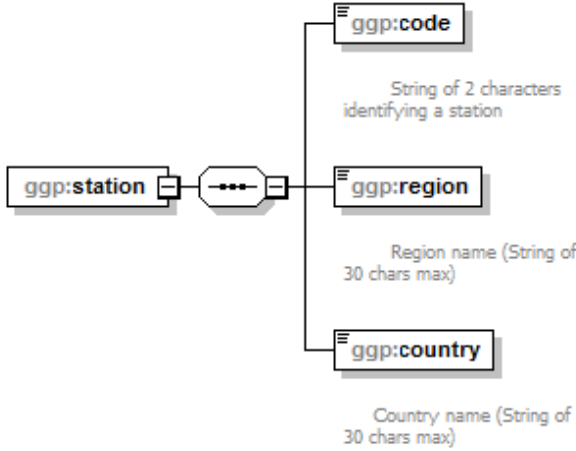
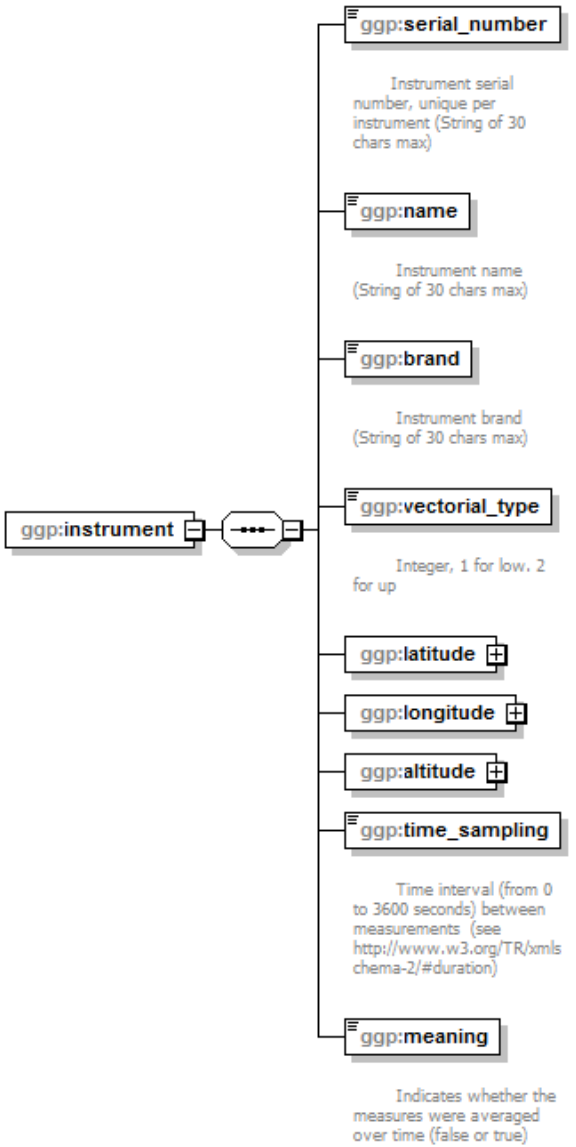
Our schema is described in table 2. Regarding this description we can make the following comments:

- Our schema is a preliminary version of what should become a normalized XML GGP schema officially approved by the IAG.
- Sample GGP files should can be validated online by using the W3C validation service: <http://validator.w3.org/>
- Our schema uses the standardized W3C built-in data types (Biron & Malhotra, 2004).
- We are planning to improve our schema by referring to already official schemas and vocabularies defined by international organizations like the Open Geospatial Consortium (OGC) (<http://www.opengeospatial.org>). Such already existing vocabularies could be used to define some concepts like latitude, longitude etc.
- We are also planning to refer to the Sensor Model Language (SensorML) that provides standard models and an XML encoding for describing the process of measurement by sensors and instructions for deriving higher-level information from observations (Botts & Robin, 2007).
- Checking the validity of a time series and its associated metadata can be done statically from the corresponding XML GGP file. It can also be done dynamically during the data flow processing.

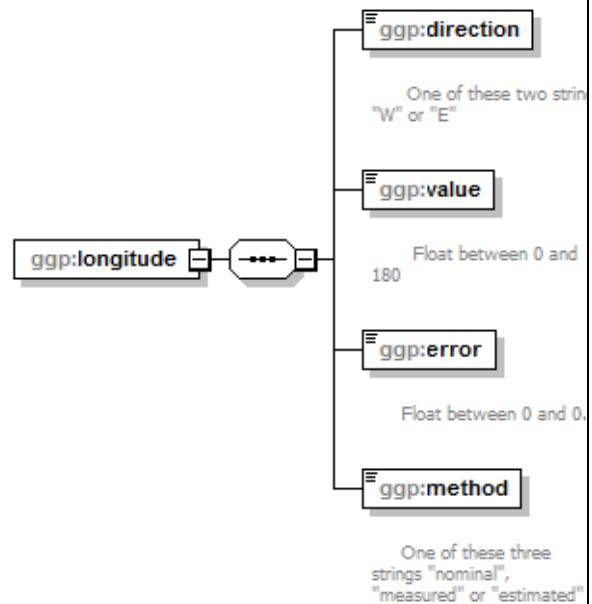
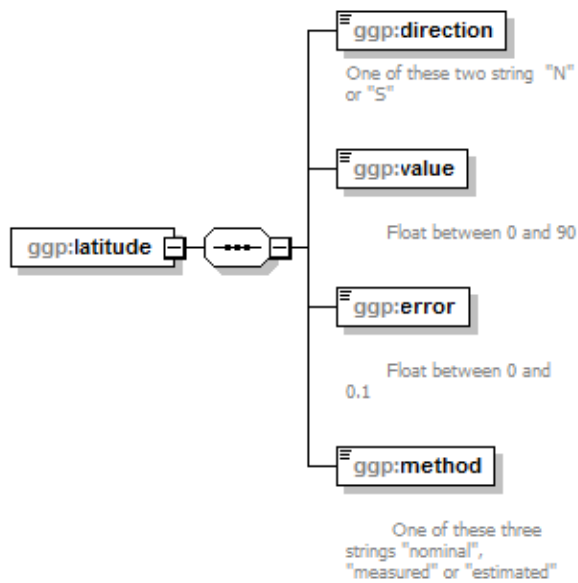


**Table 2.** Schema GGP.xsd Description

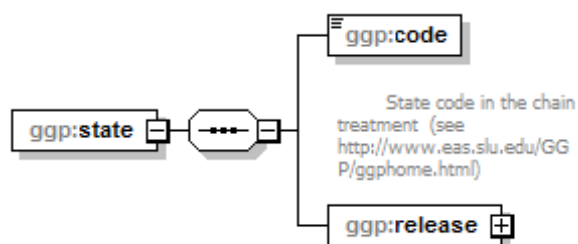
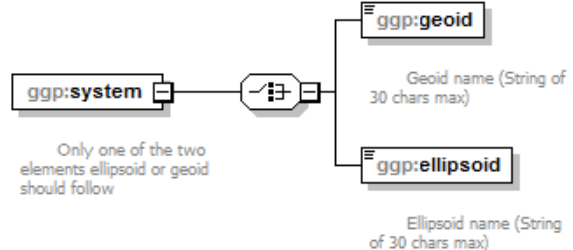
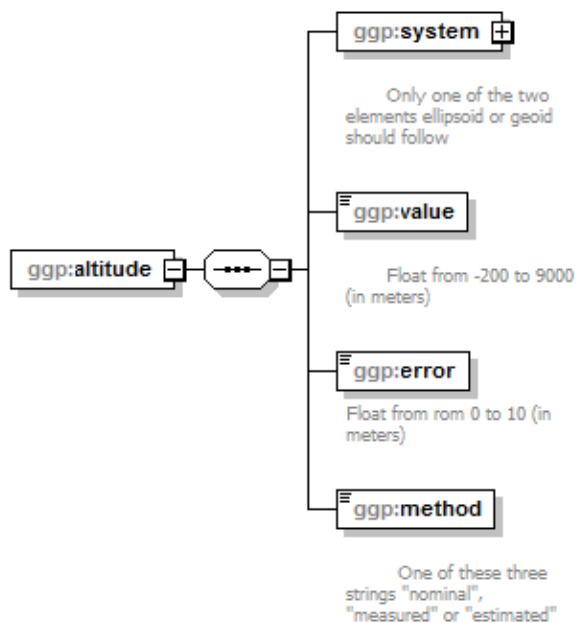
 <p>The diagram shows a 'file' element containing a 'File header' (ggp:header) and an 'Unlimited number of data records' (ggp:record 1..∞).</p>	<p>Our schema divides GGP files into 2 blocks: The <i>header</i> which consists of a set of header fields and the data block which corresponds to the <i>time series</i> and which consists of an unbounded number of data <i>records</i>.</p>
 <p>The diagram shows a 'ggp:record' element (1..∞) containing three fields: 'ggp:date_time' (Date time (see <a href="http://www.w3.org/TR/xmlschema-2/#dateTime">http://www.w3.org/TR/xmlschema-2/#dateTime</a>)), 'ggp:gravity' (Float from 0 to 1000), and 'ggp:pressure' (Float from 0 to 1000).</p>	<p>Each data <i>record</i> consists of 3 fields. The first field records the <i>date</i> and <i>time</i> of the measure in the format specified by the W3C. The second field is the <i>gravity</i> measure. The last field is the <i>pressure</i> measure. Specified bounds (from 0 to 1000) correspond to physical limitations.</p>
 <p>The diagram shows a 'ggp:header' element (File header) containing several fields: 'ggp:filename' (String of 30 chars max), 'ggp:station', 'ggp:author' (Email address), 'ggp:start_time' (Start date and time of recording (see <a href="http://www.w3.org/TR/xmlschema-2/#dateTime">http://www.w3.org/TR/xmlschema-2/#dateTime</a>)), 'ggp:end_time' (End date and time of recording (see <a href="http://www.w3.org/TR/xmlschema-2/#dateTime">http://www.w3.org/TR/xmlschema-2/#dateTime</a>)), 'ggp:instrument', and 'ggp:state'.</p>	<p>The header includes several data fields, <i>filename</i>, <i>author</i>, <i>start_time</i> and <i>end_time</i>. <i>filename</i> and <i>author</i> are self described. <i>start_time</i> corresponds to the date and time of the first data record, <i>end_time</i> corresponds to the date and time of the last data record. Other fields (<i>station</i>, <i>instrument</i> and <i>state</i>) are complex elements containing nested subfields. <i>instrument</i> contains some data regarding the instrument that recorded the time series. <i>station</i> contains some data referring to the station which hosts the instrument. <i>state</i> contains some data which are specific to the version of the time series obtained after a given processing step. Indeed, a given time series can follow a processing chain and each step in the processing chain outputs a new state of the time series.</p>

 <p>The diagram shows a root element <code>ggp:station</code> connected to a container element (a rounded rectangle with three dots). This container element branches into three child elements: <code>ggp:code</code>, <code>ggp:region</code>, and <code>ggp:country</code>.</p> <ul style="list-style-type: none"> <li><code>ggp:code</code>: String of 2 characters identifying a station</li> <li><code>ggp:region</code>: Region name (String of 30 chars max)</li> <li><code>ggp:country</code>: Country name (String of 30 chars max)</li> </ul>	<p>The station element contains 3 fields <i>code</i>, <i>region</i> and <i>country</i> identifying and locating the station.</p>
 <p>The diagram shows a root element <code>ggp:instrument</code> connected to a container element (a rounded rectangle with three dots). This container element branches into several child elements: <code>ggp:serial_number</code>, <code>ggp:name</code>, <code>ggp:brand</code>, <code>ggp:vectorial_type</code>, <code>ggp:latitude</code>, <code>ggp:longitude</code>, <code>ggp:altitude</code>, <code>ggp:time_sampling</code>, and <code>ggp:meaning</code>.</p> <ul style="list-style-type: none"> <li><code>ggp:serial_number</code>: Instrument serial number, unique per instrument (String of 30 chars max)</li> <li><code>ggp:name</code>: Instrument name (String of 30 chars max)</li> <li><code>ggp:brand</code>: Instrument brand (String of 30 chars max)</li> <li><code>ggp:vectorial_type</code>: Integer, 1 for low. 2 for up</li> <li><code>ggp:latitude</code>: (Complex element)</li> <li><code>ggp:longitude</code>: (Complex element)</li> <li><code>ggp:altitude</code>: (Complex element)</li> <li><code>ggp:time_sampling</code>: Time interval (from 0 to 3600 seconds) between measurements (see <a href="http://www.w3.org/TR/xmlschemata-2/#duration">http://www.w3.org/TR/xmlschemata-2/#duration</a>)</li> <li><code>ggp:meaning</code>: Indicates whether the measures were averaged over time (false or true)</li> </ul>	<p>The instrument element contains several fields and complex elements describing the instrument which produced the time series. <i>serial_number</i>, <i>name</i>, <i>vectorial_type</i> and <i>brand</i> are the simple data fields whereas <i>latitude</i>, <i>longitude</i> and <i>altitude</i> are the complex elements containing nested subfields. All these instrument parameters remain unchanged over the various versions (states) of a given time series. <i>serial_number</i>, <i>name</i> and <i>brand</i> are self described. <i>vectorial_type</i> indicates the levitating ball (1 for low, 2 for up). <i>latitude</i>, <i>longitude</i> and <i>altitude</i> contain some data about respectively the latitude, the longitude and the altitude of the instrument at the time it produced the time series. The latitude element (respectively the longitude element) contains some self described fields related to the latitude (respectively the longitude) of the instrument.</p>

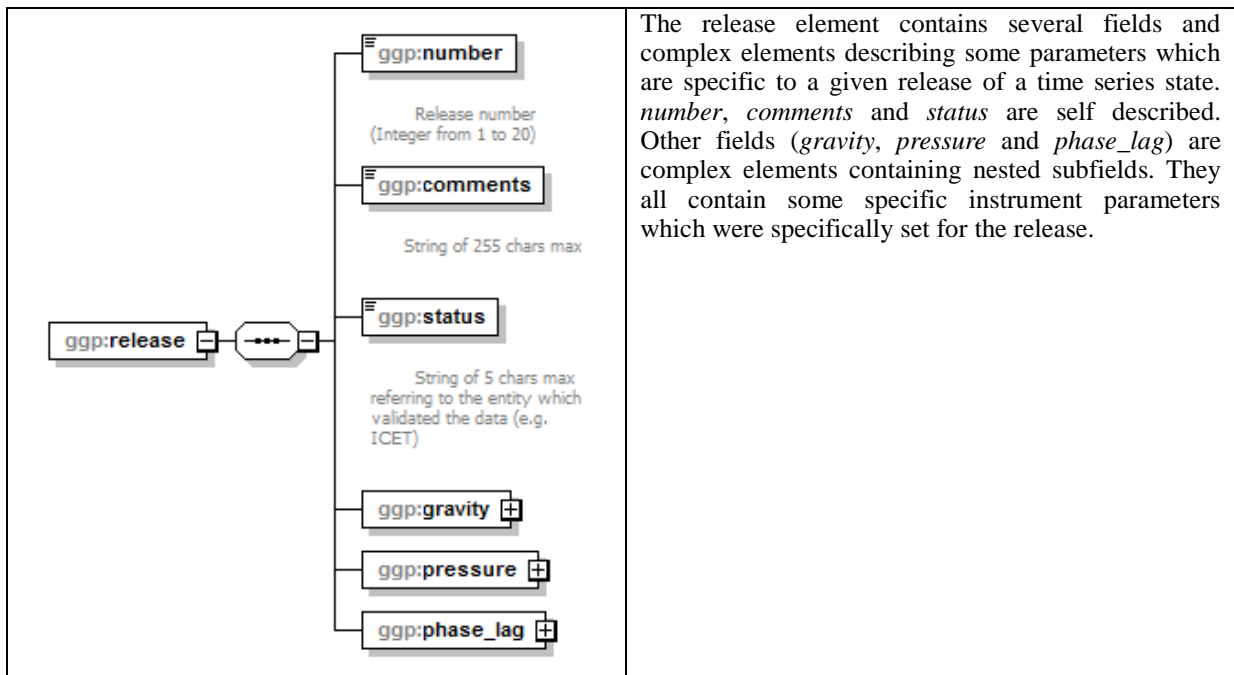
The latitude (respectively longitude) element contains some self described data fields related to the latitude (respectively longitude) of the instrument (see below)



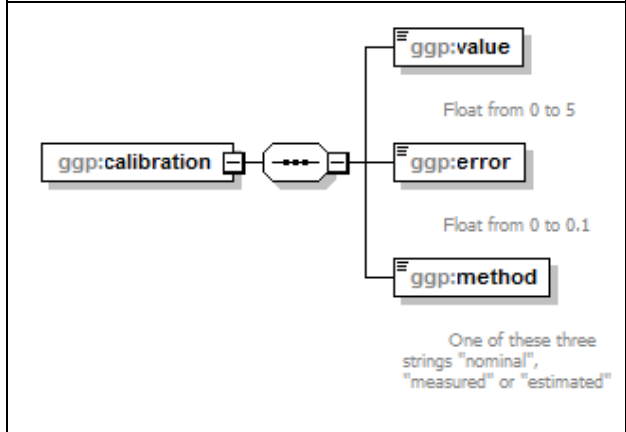
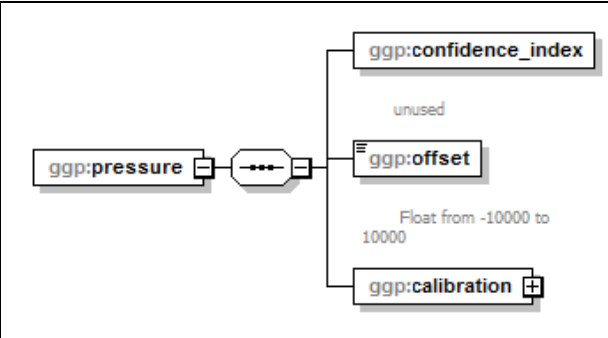
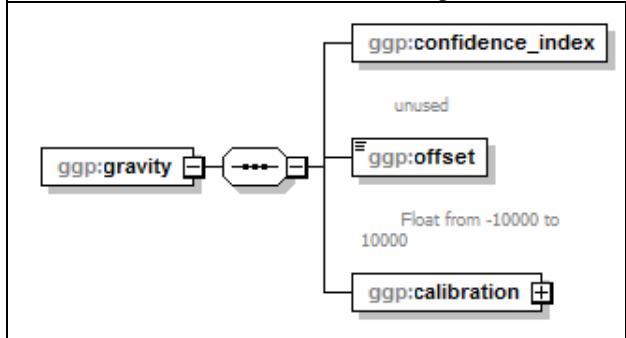
The altitude element contains some self described fields related to the altitude of the instrument. Note that *system* is a complex element which *either* includes a *geoid* data field or an *ellipsoid* data field (see below).



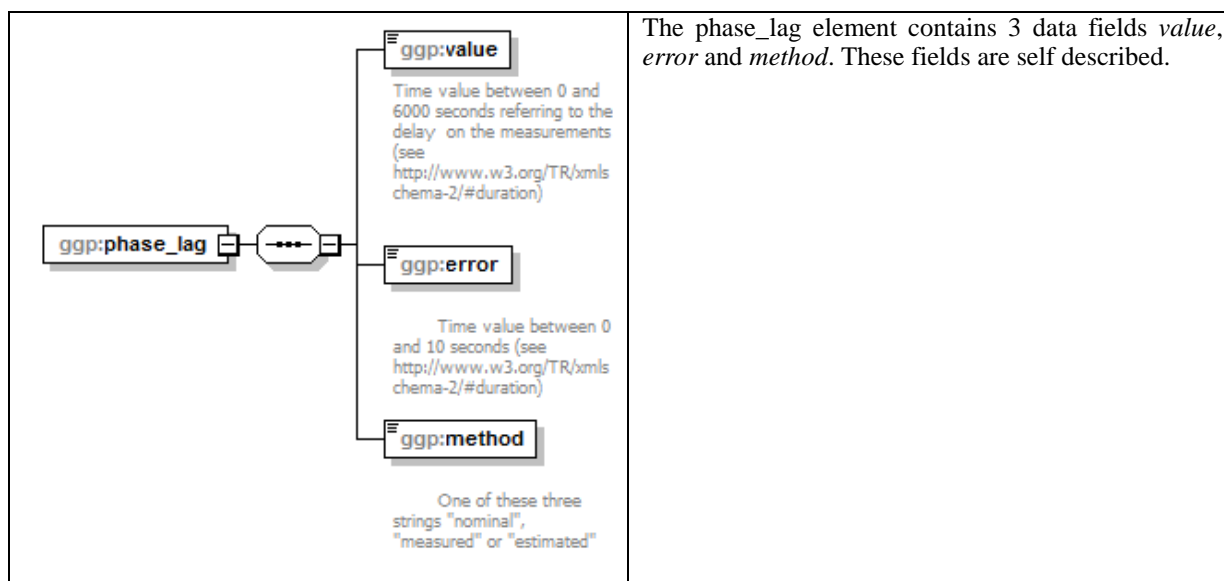
The state element consists of one data field *code* and one complex element *release*. *code* identifies the chain processing step (see <http://www.eas.slu.edu/GGP/ggphome.html>). Each state (i.e. version of the time series) can be subject to several releases (at least one). *release* records some data which are specific to each release.



The gravity (respectively pressure) element (see below) contains 2 data fields and 1 complex element. *confidence\_index* is unused and *offset* (float from -10000 to 10000) indicates a general offset on gravity for the considered data. Note that, contrary to the previous format, there should be only one possible offset value for each time series. *calibration* is the complex element and contains nested subfields (see below).



The calibration element which is nested in both the gravity and the pressure element contains 3 data fields *value*, *error* and *method*. These fields are self described.



## 5 CONCLUSION

We hope that the format we proposed in this paper will serve as a base for the future official GGP data format. We are currently developing a toolbox to allow easy back and forth conversion between the old and our new xml format. We are also writing several XSLT (Kay, 2007) style sheets for visualization of the XML GGP data.

## 6 REFERENCES

- Biron, P. V., & Malhotra, A., (2004). XML Schema Part 2: Datatypes Second Edition. W3C Recommendation 28 October 2004. Retrieved March 3, 2012 from the World Wide Web: <http://www.w3.org/TR/xmlschema-2/>
- Botts M., & Robin, A. (2007). OpenGIS Sensor Model Language (SensorML) Implementation Specification. 2007-07-17. 2007-07-17. OGC 07-000. Retrieved March 3, 2012 from the World Wide Web: <http://www.opengeospatial.org/>
- Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F., & Cowan, J. (2006). Extensible Markup Language (XML) 1.1 (Second Edition). W3C Recommendation 16 August 2006, edited in place 29 September 2006. Retrieved March 3, 2012 from the World Wide Web: <http://www.w3.org/TR/xml11>
- Bray T., Hollander, D., Layman, A., & Tobin, R. (2006). Namespaces in XML 1.1 (Second Edition). W3C Recommendation 16 August 2006. Retrieved March 3, 2012 from the World Wide Web: <http://www.w3.org/TR/xml-names11>
- Crossley, D., & Hinderer, J. (2009). The Contribution of GGP Superconducting Gravimeters to GGOS. In Sideris, M.G., (Ed), *Proceedings of the IUGG, IAG Symposia 133, Perugia 2007, Observing our Changing Earth*: Springer Verlag.
- Crossley, D., Hinderer, J., Casula, G., Francis, O., Hsu, H. T., Imanishi, Y., Jentzsch, G., Kääriäinen, J., Merriam, J., Meurers, B., Neumeyer, J., Richter, B., Shibuya, K., Sato, T. & Van Dam, T. (1999). Network of superconducting gravimeters benefits a number of disciplines, *EOS*, 80, 11, 121/125-126.
- Hinderer, J., Crossley D., & Warburton, W. (2007). Superconducting Gravimetry. In Herring T. and Schubert G., (Eds), *Treatise on Geophysics*, Vol 3, Elsevier.
- Kay M. (2007). XSL Transformations (XSLT) Version 2.0. W3C Recommendation 23 January 2007. Retrieved March 3, 2012 from the World Wide Web: <http://www.w3.org/TR/xslt20/>
- Thompson, H.S., Beech, D., Maloney, M., & Mendelsohn N. (2004). XML Schema Part 1: Structures, Second Edition. W3C Recommendation 28 October 2004. Retrieved March 3, 2012 from the World Wide Web: <http://www.w3.org/TR/xmlschema-1/>
- Wenzel, H.G. (1996). The nanogal software: Earth tide processing package ETERNA 3.30. *Bull. Information des Marées Terrestres*, 124, 9425-9439.

# RESEARCH ENVIRONMENT AND INFORMATION SERVICE OF SPACE WEATHER CLOUD

*S Watari*<sup>\*1</sup>, *H Kato*<sup>2</sup>, *Ken T Murata*<sup>3</sup>, *K Yamamoto*<sup>4</sup>, *H Watanabe*<sup>5</sup>, *Y Kubota*<sup>6</sup>, and *M. Kunitake*<sup>7</sup>

*National Institute of Information and Communications Technology, 4-2-1 Nukuikita, Koganei, Tokyo 184-8795, Japan*

*Email:* <sup>\*1</sup>*watari@nict.go.jp,* <sup>2</sup>*hisa@nict.go.jp,* <sup>3</sup>*ken.murata@nict.go.jp,* <sup>4</sup>*kaz-y@nict.go.jp,* <sup>5</sup>*h-watanabe@nict.go.jp,* <sup>6</sup>*ykubota@nict.go.jp, and* <sup>7</sup>*kunitake@nict.go.jp*

## ABSTRACT

*For researches and information services of space weather, it is important to establish a comprehensive system which enables us to analyze observation and simulation data in an integrated manner. For this, we constructed recently a new computing environment called the “Space Weather Cloud Computing System” of the National Institute of Information and Communications Technology (NICT). Now, the Space Weather Cloud contains a high performance computer, a distributed mass storage system using the Grid Data Farm (Gfarm) technology, servers for analysis and visualization of data, a job service based on the RCM (R&D Chain Management) System, servers for the Solar-Terrestrial data Analysis and Reference System (STARS).*

**Keywords:** space weather, cloud computing system, Gfarm, JGN-X, STARS

## 1 INTRODUCTION

Space weather is the concept of changing environmental conditions in the space from the Sun's atmosphere to the Earth's atmosphere. Space weather variations are affecting human-made infrastructures such as artificial satellite, electric power grids, Global Navigation Satellite System (GNSS), and HF radio communication (Marubashi, 1998; Lanzerotti, 2001). It is difficult to cover whole of this vast space only by existing observational framework. We need a new environment to analyze both observation data and simulation data in an integrated manner (Baker & Barton, 2008; Rankin, 2011). Adding this, the amount of data on space weather has been increasing year by year because of a remarkable increase of new data from ground-based observations, observations using spacecraft, and simulation models. We also need such a computing environment to process these big data. In order to cope with this situation, a new platform called the ‘Space Weather Cloud Computing System (hereafter Space Weather Cloud)’ has been constructed in the National Institute of Information and Communication Technology (NICT). We report details of this system and show examples of its applications.

## 2 OUTLINE OF SPACE WEATHER CLOUD COMPUTING SYSTEM

Figure 1 shows the general concept of the ‘Space Weather Cloud’ of NICT. The system is composed of a super computer, a distributed mass storage system basing on the Grid Data Farm (Gfarm) technology (Tatebe et al., 2001), servers for analysis and visualization of data using IDL (Interactive Data Language) and AVS (Advanced Visual Systems), a job service component based on the R&D Chain Management (RCM) System, a tool for data plot and analysis called Solar-Terrestrial data Analysis and Reference System (STARS), servers to automatically collect metadata (NICTY), streaming servers, Tiled Display Wall (TDW), etc. The Space Weather Cloud can be accessed via networks such as the Internet and the New Generation Network Testbed (JGN-X, <http://www.jgn.nict.go.jp/english/index.html>) which is a new generation network developed by NICT.

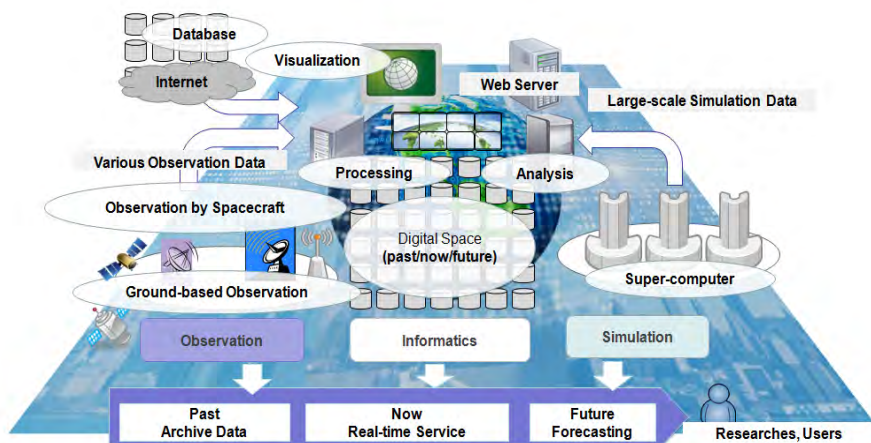


Figure 1. A conceptual diagram of the Space Weather Cloud of NICT.

### 3 INFORMATION SERVICES OF SPACE WEATHER CLOUD COMPUTING SYSTEM

The Space Weather Cloud provides various Web-based services for general users. Figure 2 shows our e-SW Web page (<http://e-sw.nict.go.jp/>), which is the portal of information services of the Space Weather Cloud. Several examples of the services of the Space Weather Cloud are shown in the following sections. Several contents have Japanese and English explanations. We will increase English contents of our services in near future.



Figure 2. The e-SW Web page, which is the entrance of services on the Space Weather Cloud.

#### 3.1 Space Weather Board

The space weather board, shown in Figure 3, is a tool to enable users to customize space weather data. There is a variety of space weather users such as, satellite operators, users of Global Navigation Satellite System (GNSS), and operators of HF communication. This board will be useful for them to customize the display of plots and images on the screen, according to their individual purposes. By using this board, users can select data sets from the component list and arrange them as they want. These users can store their own arrangements in the server for their convenience.

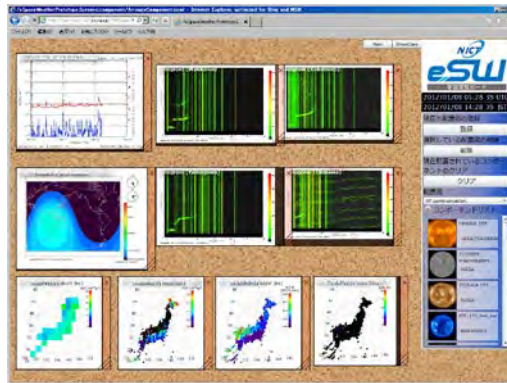


Figure 3. An example of the Space Weather Board Web page (customized to see ionospheric data).

### 3.2 3D View of Real-time Space Weather Simulation

NICT developed magneto-hydrodynamic simulation codes covering the region from the solar corona to the terrestrial ionosphere (Nakamizo et al, 2009; Den et al., 2006; Shinagawa, 2011). We run the simulation in a real-time basis. The results of the simulations are sorted and displayed by the 3D-visualization system of the Space Weather Cloud. Users can access real-time and archived data through the 3D view Web page.

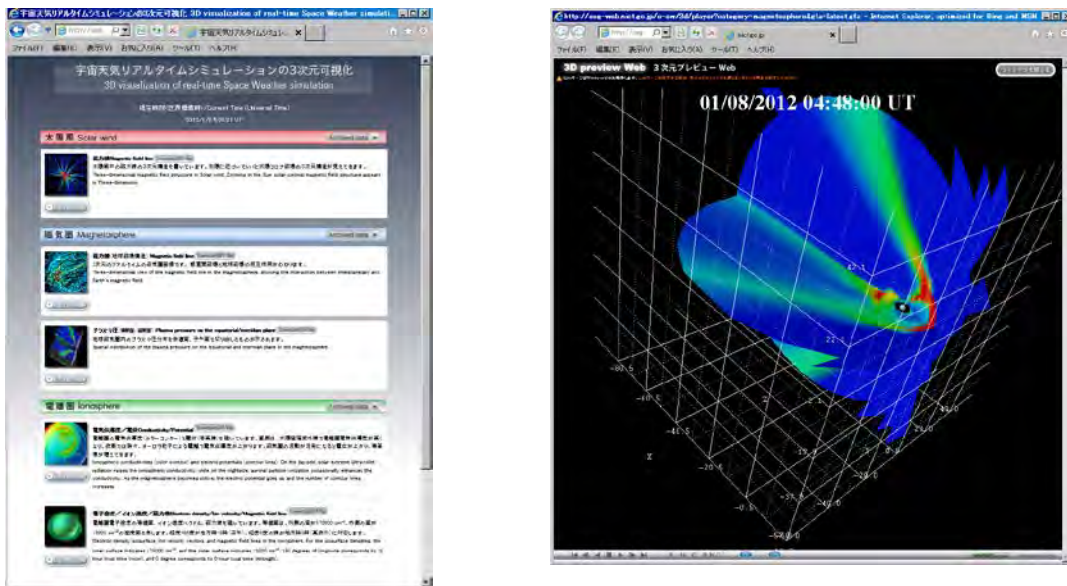


Figure 4. The Web-page of 3D view of real-time space weather simulation (left panel) and an example of the 3D view of the terrestrial magnetosphere (right panel).

### 3.3 Weekly Space Weather News

The concept and terminology of space weather are unfamiliar to general public. To improve this situation, we have started a movie program on the weekly summary of current conditions of space weather, under the name of 'Weekly Space Weather News.' It is delivered by streaming from the server of the Space Weather Cloud. Brief explanations of technical terms of space weather are also provided to help understanding the contents. Examples of scenes of the program are shown in Figure 5.





(a) Opening of 'Weekly Space Weather News'



(b) Outline of space weather of the week



(c) Detail information of space weather of the week



(d) Explanation on terminologies of space weather

**Figure 5.** Scenes of 'Weekly Space Weather News.'

## 4 SUMMARY

The data volume of space weather is increasing year by year by adding new data from many satellites, ground-based observational networks, numerical simulations etc. It is urgent to construct a computing platform to efficiently process both observation and simulation data together (Baker & Barton, 2008; Hey et al., 2009; Rankin, 2011). Our cloud computing system will be an example to meet the demand, and it is expected that new knowledge on the space weather will be extracted from our data intensive studies using the Space Weather Cloud.

## 5 REFERENCES

- Baker, D.N. & Barton, C.E. (2008) Informatics and the 2007-2008 Electronic Geophysical Year, *EOS Transaction, AGU*, 89(40).
- Den, M., Tanaka, T., Fujita, S., Obara, T., Shimazu, H., Amo, H., Hayashi, Y., Nakano, E., Seo, Y., Suehiro, K., Takahara, H., & Takei, T. (2006) Real-time Earth magnetosphere simulator with three-dimensional magnetohydrodynamic code, *Space Weather*, 4, S06004, doi:10.1029/2004SW000100.
- Hey T., Tansley, S. & Tolle, K. (Eds.) (2009) *The fourth paradigm: Data-intensive scientific discovery*, Microsoft Research: <http://resrarch.microsoft.com/en-us/collaboration/fourthparadim>.
- Lanzerotti, L. J. (2001) Space weather effects on technologies. In Song, P., Singer, H.J. & Siscoe, G.L. (Eds.), *Space Weather*, AGU Geophysical Monograph 125, Washington, DC, pp 11-22.
- Marubashi, K. (1989) The space weather forecast program, *Space Sci. Rev.* 51(1-2), 197-214.
- Nakamizo, A., Tanaka, T., Kubo, Y., Kamei, S., Shimazu, H., & Shinagwa, H. (2009) Development of the 3-D MHD model of the solar corona-solar wind combining system A 3-D MHD simulation model of the solar corona-solar wind system, *Journal Geophys. Res.*, 114, A07109, doi:10.1029/2008JA013844.
- Rankin, R. (2011) Space science informatics: A Canadian approach, *EOS Transaction, AGU*, 92(8).
- Shinagawa, H. (2011) Ionosphere Simulation, *Journal of NICT* 56(1-4), 199-207.
- Tatebe, O., Morita, Y., Matsuoka, S., Soda, N., Sato, H., Tanaka, Y., Sekiguchi, S., Watanabe, Y., Iemori, M. & Kobayashi, T. (2001) Grid data farm for petascale data intensive computing, *Technical Report, Electrotechnical Laboratory*, ETL-TR2001-4.

# DIGITAL DATABASE OF LONG-TERM SOLAR CHROMOSPHERIC VARIATION

*R. Kitai<sup>1\*</sup>, S. Ueno<sup>1</sup>, H. Maehara<sup>1</sup>, S. Shirakawa<sup>1</sup>, M. Katoda<sup>1</sup>, Y. Hada<sup>1</sup>, Y. Tomita<sup>2</sup>, H. Hayashi<sup>3</sup>, A. Asai<sup>4</sup>, H. Isobe<sup>4</sup>, H. Goto<sup>5</sup> and S. Yamashita<sup>5</sup>*

<sup>\*1</sup> *Kwasan and Hida Observatories, Graduate School of Science, Kyoto University*

*Email: kitai@kwasan.kyoto-u.ac.jp*

<sup>2</sup> *Department of Astrophysics, Graduate School of Science, Kyoto University*

<sup>3</sup> *Research Institute for Sustainable Humanosphere, Kyoto University*

<sup>4</sup> *Unit for Synergetic Studies of Space, Kyoto University*

<sup>5</sup> *The Kyoto University Museum, Kyoto University*

## ABSTRACT

*From 1926 to 1969, a long term solar full disk observation had been done in Kyoto University. Daily Ca II K (393.4 nm) spectroheliographic images and white light images had been recorded on photographic plates. In this report, we will give the current status of our project to digitize all the images and to construct a database of these images for public use, through the IUGONET system. In addition, we will discuss our perspective on the scientific analysis of the database by taking the solar Ca II K brightness as a proxy measure of the solar UV irradiance onto the terrestrial upper atmosphere.*

**Keywords:** Digital image database, Solar chromosphere, Ca II K full disk image, UV irradiance

## 1 INTRODUCTION

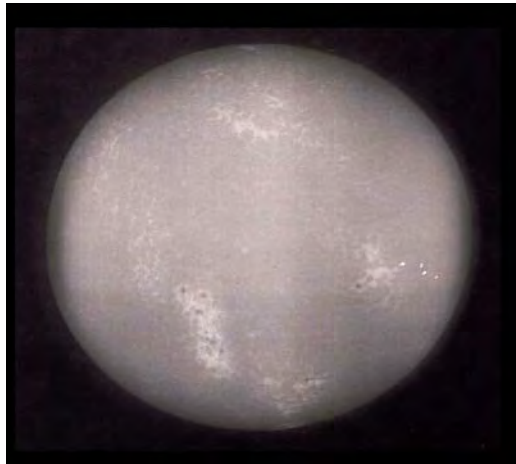
From 1926 to 1969, a long term solar full disk observation had been done in Kyoto University. Spectroheliographic images of Ca II K (393.4 nm) and white light images had been taken on a daily base. All the images were recorded on photographic plates. From the viewpoints of the long-term span of the data coverage and the scarceness of full solar disk images in the first half of 1900s, we think that the data will have a scientific importance. Since we have a risk of aging and degradation of these old photographic plates, we have just started a project to digitize all the plates. We are developing also a digital-image database for public use via IUGONET (Inter-university Upper atmosphere Global Observation NETwork, <http://www.iugonet.org/en/>).

## 2 EQUIPMENT AND HISTORY OF OBSERVATION

In 1926, solar full disk Ca II K observation was started with an Askania spectroheliograph equipped to a 30cm siderostat telescope at the Kyoto University Observatory. In 1929, the spectroheliograph was moved to the newly installed Kwasan Observatory. Then, in 1941, it was moved again to the Ikoma Solar Station in Osaka Prefecture. In spite of two relocations of the spectroheliograph, the observation itself was continued on daily base from 1926 to 1969 without interruption. A photograph of the spectrograph taken at Kwasan Observatory around 1930 is shown in Figure 1. A sample of Ca II K spectrograms taken by this instrument is shown in Figure 2.



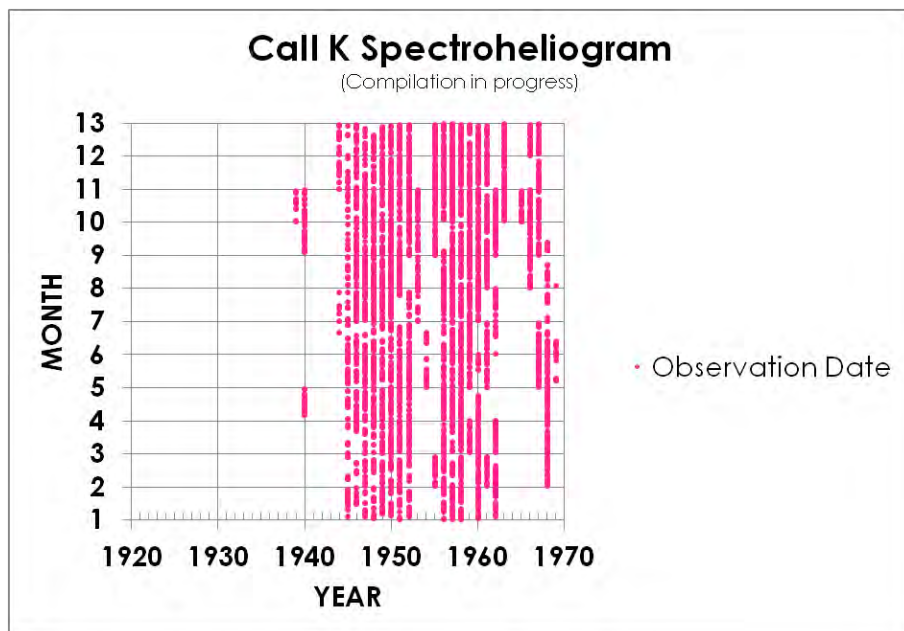
**Figure1.** Spectroheliographic observation at Kwasan Observatory around 1930.



**Figure2.** A sample of Ca II K spectroheliogram (positive) taken on May 24, 1967. We can see dark sunspots and bright plages on the solar disk.

### 3 COMPILATION OF METADATA

Now we have started to compile metadata of the spectroheliograms and finished 50% of them. The distribution of observation dates is shown in Figure 3. A half number of photographic plates taken in the interval from 1926 to 1943 were preserved in the Yamamoto Observatory.



**Figure 3.** Coverage of Ca II K spectroheliographic observations. The data for 1926-1945 are under compilation.

### 4 OUR PROJECT: DATABASE AND ITS SCIENTIFIC APPLICATION

The first target of our database project is to finish the digitization of 44-year solar full-disk chromospheric images and to complete an open database to the public for scientific use. Our database of Ca II K images will be important to complement the existing databases of Ca II K images respectively taken at Mt. Wilson and Kodaikanal Observatories (Foukal et al., 2009) and enable us to perform a cross-check of the trends of long term solar variability estimated from these independent datasets.

One of the scientific targets of our project is to use the database in a study of the heating process of the terrestrial upper atmosphere. A comprehensive review of current researches on long-term variations of the total solar irradiance (TSI) and the spectral irradiance is given by Krivova et al. (2011). These researches were based on solar magnetographic data ( ~40 years span ), sunspot-number data ( ~400 years span ) and  $^{14}\text{C}$  and  $^{10}\text{Be}$  concentration data ( ~  $10^4$  years span), and the solar UV irradiance was estimated with the help of theoretical models of the solar atmosphere. On the other hand, according to a pioneering work by Yokoyama, Masuda and Sato (2006), the total area of Ca II K plages on the solar disk is a good proxy of the solar EUV and UV irradiance on the terrestrial upper atmosphere. Although their analysis is limited only to a two-week span, the presence of a positive correlation between the Ca II K plage area and the UV irradiation measured by satellites is clearly seen. If we can confirm their conclusion by using our Ca II K database for recent 10 years, namely the interval in which satellite data of solar irradiance are available, we will be able to trace the long-term (44 years) variations of the solar UV irradiance in the pre-satellite era, basing on our comprehensive Ca II K database.

## **5 ACKNOWLEDGEMENTS**

This work was supported in part by the research grant for Mission Research on Sustainable Humanosphere from Research Institute for Sustainable Humanosphere (RISH), Kyoto University

## **6 REFERENCES**

Foukal, P., Bertello, L., Livingston, W.C., Pevtsov, A.A., Singh, J., Tlatov, A.G., Ulrich, R.K, (2009), *SP* 255,229.

Krivova, N.A., Solanki, S.K., & Unruh, Y.C., (2011), *J. Atmo. and ST Phys.*, 73, 223

Yokoyama, M., Masuda, S., & Sato, J., (2006), *AGU, Fall Meeting*, abstract #SH43A-1503

# A STATE-SPACE APPROACH TO EXPLORE THE STRAIN BEHAVIOR BEFORE AND AFTER THE 2003 TOKACHI-OKI EARTHQUAKE (M8)

*T Takanami<sup>1\*</sup>, G Kitagawa<sup>2</sup>, H Peng<sup>3</sup>, A T Linde<sup>4</sup> and I S Sacks<sup>5</sup>*

*<sup>1</sup>Earthquake Research Institute, The University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo, 113-0032, Japan*

*Email: takanami@eri.u-tokyo.ac.jp*

*<sup>2</sup>Research Organization of Information and Systems, 4-3-13, Toranomon, Minato-ku, Tokyo, 105-0001, Japan*

*Email: kitagawa@rois.ac.jp*

*<sup>3</sup>School of Information Science & Engineering, Central South University, Changsha, Hunan 410083, China*

*Email: huipeng@mail.csu.edu.cn*

*<sup>4</sup>Department of Terrestrial Magnetism, Carnegie Institution of Washington, DC, 20015-1305, USA*

*Email: linde@dtm.ciw.edu*

*<sup>5</sup>Department of Terrestrial Magnetism, Carnegie Institution of Washington, DC, 20015-1305, USA*

*Email: sacks@dtm.ciw.edu*

## ABSTRACT

*The earth's surface is under continuous influence of a variety of natural forces and human induced sources. A strain data is a good example for such disturbed signals. To determine the geodetic strain behavior before and after the 2003 Tokachi-oki earthquake (M8.0), we decomposed the disturbed strain data into several components of trend, air pressure, earth tide, and precipitations responses. The decomposition of the disturbed strain data and the interpolation of the missing observations are performed very effectively by using the state-space modeling and the Kalman filter/smoothen. The data processing validity is confirmed by the fact that the model derived to fit the strain data matches the GPS data extremely well.*

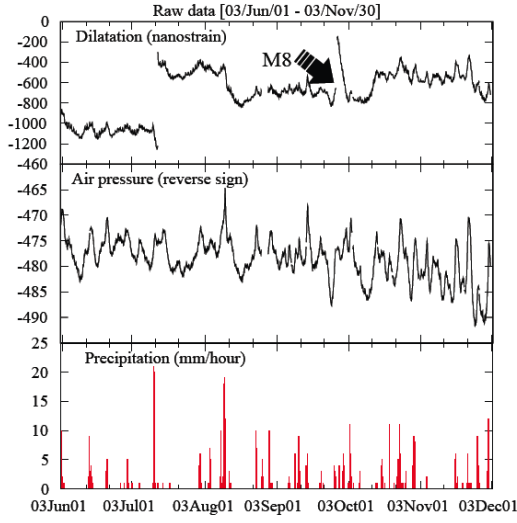
**Keywords:** Signal extraction method, State-space model, Kalman smoother/filtering, Strain, Sacks-Evertson strain meter, 2003 Tokachi-oki earthquake, Slow-slip event.

## 1 INTRODUCTION

On 26 September 2003 a great interplate earthquake (M8.0) struck the Hokkaido corner in the southernmost Kuril trench. The Hokkaido corner is the site of large earthquakes due to the subduction of the Pacific Plate beneath the Hokkaido, Japan at rate of 8.3 cm/yr. The previous great earthquake was the 1952 Tokachi-oki earthquake (M8.2). A Sacks-Evertson borehole strainmeter (Sacks et al. 1971) was installed in November of 1982 to observe the changes before and after such a huge earthquake (Takanami et al., 1998). The observatory is at the located 105 km from the epicenter of the 2003 Tokachi-oki earthquake. Observational data of near surface crustal strain necessarily include changes produced by non-tectonic sources including atmospheric pressure changes, earth tides and precipitation. The continuity of recorded data is also interrupted at times due to power failure and need for instrument maintenance. Thus there is a need to apply processing techniques to remove changes not of interest in seismological studies. We used here the state-space modeling methods for the smoothing and component decomposition tasks developed by Kitagawa and Matsumoto (1996) and Matsumoto and Kitagawa (2003). In principle, it is possible to treat these two tasks simultaneously using state-space modeling and to fit decomposition into components model for the detection of seismic effects to the data with missing and outlying observations. But because of data volume and the need for very high-order models, we adopted a two-stage analysis strategy composing of smoothing by using a simple Gaussian state-space trend model and decomposition into components by assuming the smoothed observation of strain to be characterized by a nonstationary trend and to be influenced by covariate air pressure, tidal, and precipitation effects. In this paper, we show a specific example of time series modeling for signal extraction problem related to geodetic strain change at the 2003 Tokachi-oki earthquake.

## 2 APPLICATION OF STATE-SPACE MODEL

The strain has been measured in the borehole at station KMU of Hokkaido University since November 1982. The time series of observations of strain (Figure 1) includes the irregular offsets (due to instrument reset). Missing data, in both the strain (upper trace) and air pressure records (middle trace), are due to power failures because, at that time, no on site battery powered recording was operational. Such power failures caused by strong torrential rainfalls, shaking effects of large earthquakes, as well as the problems with local power supplies. Two anomalous torrential rainfalls drenched the area around KMU on 10, July (157 mm/day) and on 9,



**Figure 1.** The observations of strain (dilatation), air pressure and precipitation (from top to bottom). The plotted period is from the first of June to the end of November 2003. M8 indicates the occurrence time of the 2003 Tokachi-oki earthquake of magnitude 8. The large jump indicates the reset of observation due to loss of power for 12 hours. Note the jump indicated by M8 is not indicative of the coseismic strain step of the 2003 Tokachi-oki earthquake. The missing data is indicated by a gap in the plotted line.

August (107 mm/day), respectively. In the earlier torrential rainfall, the power supply was also interrupted. Although it is possible to interpolate for the missing data and correct outliers by using a simple non-Gaussian state space model (Kitagawa & Matsumoto, 1996; Matsumoto, 1999), it is almost impossible to restore the missing data at the time of the 2003 Tokachi-oki earthquake. In this paper, we do not deal with such coseismic crustal movement. As to such coseismic behavior, many papers have already been published (e.g. Ozawa et al., 2004; Yagi, 2004; Fukuda et al., 2009; Miwazaki et al., 2008). We address here slow strain changes immediately following the earthquake. Because strain changes induced by non-tectonic changes can mask such slow changes it is necessary to de-convolve those components of the data in order to obtain a reliable estimate of the slow tectonic changes. We used here the state-space modeling method for the smoothing and component decomposition tasks. Successful studies to detect groundwater level have been carried out using the same approach (Kitagawa & Matsumoto, 1996; Matsumoto & Kitagawa, 2003; Matsumoto et al., 2003). As in those papers, the observation data of strain,  $y_n$  can be represented by the following model composed of several components

$$y_n = t_n + P_n + E_n + R_n + \varepsilon_n,$$

$$\varepsilon_n \sim N(0, \sigma^2), \quad n = 1, \dots, N \quad (1)$$

where  $N$  is the number of observations.  $t_n, P_n, E_n, R_n$  and  $\varepsilon_n$  are trend, air pressure effect, earth tide effect, precipitation effect and observation noise components, respectively. The trend components is expressed by the following first-order trend model (Kitagawa & Gersch, 1984),

$$t_n = t_{n-1} + w_n,$$

$$w_n \sim N(0, \tau^2) \quad (2)$$

The other components are as given below,

$$P_n = \sum_{i=0}^m a_i p_{n-i} \quad (3)$$

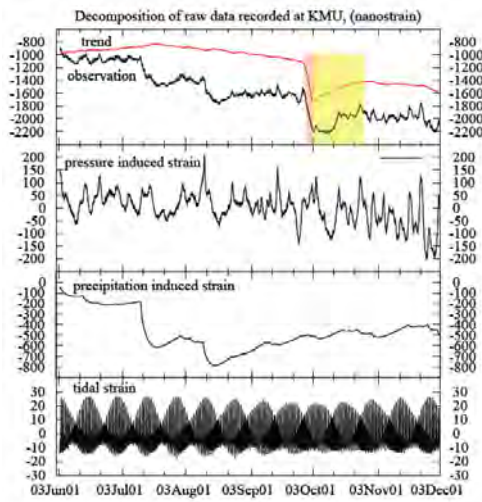
$$E_n = \sum_{i=0}^l b_i e_{n-i} \quad (4)$$

$$R_n = \sum_{i=1}^k c_i R_{n-i} + \sum_{i=1}^k d_i r_{n-i} \quad (5)$$

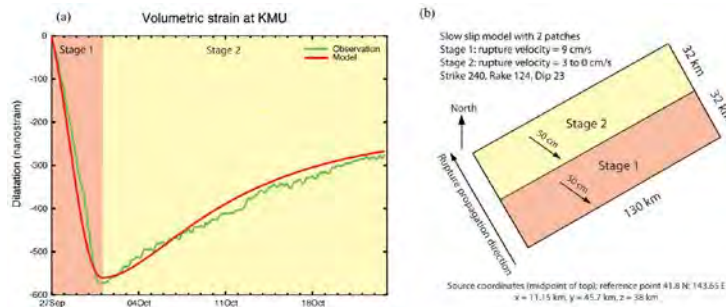
where  $p_n$ ,  $et_n$  and  $r_n$  are the observed air pressure, the theoretical earth tide and the observed precipitation, respectively. For the precipitation effect, we used the ARMAX type model (Box and Jenkins, 1976) as given the model (5) because precipitation effects may continue for a very long time following the precipitation. The regression coefficients  $a_i$  and  $b_i$  can be estimated by the Kalman filter. On the other hand,  $c_i$  and  $d_i$  need to be estimated by numerically maximizing the likelihood function. When the effects of covariates,  $P_n$ ,  $E_n$  and  $R_n$  were removed from  $y_n$ , the trend was expected to be a geodetic strain before and after the Tokachi-oki earthquake, indicated by the label M8. Next we describe the result of decomposition of 6 months of strain observations at KMU into the trend and the several induced strain components.

### 3 RESULTS OF THE IMPLEMENTATION OF STATE-SPACE METHOD

Figure 2 illustrates the decomposition of observations into the trend, the air pressure effect, the precipitation effect, and the earth tide effect. Judging from the smoothed trend excepting the trend just after the 2003 Tokachi-oki earthquake, it is confirmed that the influence of the air pressure, the precipitation and the earth tide were successfully removed by the state-space modeling. Consequently, it turns out that a clear slow change of trend appeared immediately after the 2003 Tokachi-oki earthquake. This indicates that a slow-slip event occurred after the 2003 Tokachi-oki earthquake in the vicinity of KMU. The slow-slip event consists of two stages as shown in Figure 2. The first stage started immediately after the Tokachi-oki earthquake until 30 September with a second stage continuing to 23 October. The strain change after the Tokachi-oki earthquake is characterized by a 4-days contraction followed by a 23-days extension. An interpretational model has been illustrated in Figure 3. According to Linde et al. (1996), we generate a quasi-static time series of deformations as the rupture surface grows with down-dip propagation as shown in Figure 3b. The strain change calculated by the model fits to the observations at KMU extremely well. The GPS data at the various surrounding sites operated by GSI are also suitable to the model well especially considering the simplicity of the model (Takanami et al., 2009). Namely, the model is that a large two-stage slow-slip earthquake (equivalent moment magnitude 7.4) occurred mainly on the ruptured zone of 2003 Tokachi-oki. We might incidentally remark that no pre-seismic strain change was detected by the present work.



**Figure 2.** The decomposition of observations recorded by the borehole strainmeter at KMU. From top to bottom, extracted trend (red line), observations of strain, air pressure effect, precipitation effect and earth tide effect are illustrated. A big variation in trend indicates the slow-slip event occurred immediately after the 2003 Tokachi-oki earthquake. It consists of two stages of higher strain rate for about 4 days (red patch) and lower strain rate for about 23 days (yellow patch).



**Figure 3.** (a) Fitting model curve (red line) inferred from two propagating fault models in Figure 3 (b) to observation data (green line) in stage I (red zone) and II (yellow zone). (b) Schematic propagating fault model for slow slip event consisted of stage I (red) and II (yellow).

### 3 CONCLUSION

We confirmed that the state-space approach was very highly effective in isolating a trend of geodetic strain observations. We handled missing and jumping strain observations and deletion of the air pressure, earth tide, and precipitation effects using the state-space modeling method. A slow slip event was clearly detected immediately following the 2003 Tokachi-oki earthquake (M8.0). It consists of two consecutive stages of a 4-day and a 23-day slow slip occurring largely in the ruptured zone of the 2003 earthquake (M8). We can say that the data processing validity is confirmed by the fact that a model derived to fit the strain data making no use of GPS also fits the GPS data extremely well especially considering the simplicity of the model. If the present data processing did not work well then we would not get such consistency. No pre-seismic strain change was detected.

### 4 REFERENCES

- Box, G.E.P. & Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, (2<sup>nd</sup> ed.). San Francisco: Holden-Day.
- Fukuda, J., Johnson, K.M., Larson, K.M., & Miyazaki, S. (2009) Fault friction parameters inferred from the early stages of afterslip following the 2003 Tokachi-oki earthquake, *Journal of Geophysical Research*, 114, B04412, doi:10.1029/2008JB006166.
- Kitagawa, G., & Gersch, W. (1984) A smoothness prior-state-space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79, 378-389.
- Kitagawa, G., & Matsumoto, N. (1996) Detection of coseismic changes of underground water level. *Journal of the American Statistical Association*, 91(434), 521-528.
- Linde, A.T., Gladwin, M.T., Johnston, M.J.S., Gwyther, R.L., & Bilham, R.G. (1996) A slow earthquake sequence on the San Andreas Fault, *Nature*, 383, 65-68.
- Matsumoto, N. (1999) Detection of groundwater level change related to earthquakes. Akaike, H. & Kitagawa, G., (Eds.), In *The Practice of Time Series Analysis*, New York: Springer-Verlag.
- Matsumoto, N., & Kitagawa, G. (2003) Extraction of hydrological anomalies related to earthquakes. Takanami, T. & Kitagawa, G., (Eds.). In *Methods and Applications of Signal Processing in Seismic Network Operations*, Berlin: Springer-Verlag.
- Matsumoto, N., Roeloffs, E.A., & Kitagawa, G. (2003) Hydrological response to earthquakes in the Haibara well, central Japan-I. Ground level changes revealed using state space decomposition of atmospheric pressure, rainfall and tidal responses. *Geophysical Journal International*, 155, 885-898.
- Miyazaki, S., & Larson, K.M. (2008) Coseismic and early postseismic slip for the 2003 Tokachi-oki earthquake sequence inferred from GPS data, *Geophysical Research Letters*, 35, L4302, doi:10.1029/2007GL032309.
- Ozawa, S., Kaizu, M., Murakami, M., Imakiire, T., & Hatanaka, Y. (2004) Coseismic and postseismic crustal deformation after the Mw 8 Tokachi-oki earthquake in Japan, *Earth, Planets and Space*, 56, 675-680.
- Sacks, I.S., Suyehiro, S., Evertson, D.W., & Yamagishi, Y. (1971) Sacks-Evertson strainmeter, its installation in Japan and some preliminary results concerning strain steps, *Paper in Meteorology and Geophysics*, 22, 195-208.
- Takanami, T., Ogawa, T., Sacks, I.S., Linde, A.T., & Nakanishi, I. (1998) Long-period volume-strain seismogram of the 8 August 1993 Esashi-oki earthquake, off southwest of Hokkaido, Japan and its source mechanism. *Faculty of Science, Hokkaido University, Series 7 (Geophysics)*, 11(2), 523-543.
- Takanami, T., Sacks, I.S., & Linde, T.A. (2009) A strain event related to aftershock activity following the 2003 Tokachi-oki earthquake (8.0). Abstract of 2009 American Geophysical Union Fall Meeting, San Francisco, USA.
- Yagi, Y. (2004) Source rupture process of the 2003 Tokachi-oki earthquake determined by joint inversion of teleseismic body wave and strong motion data, *Earth, Planets and Space*, 56, 311



# METADATA PUBLICATION AND SEARCH SYSTEM IN JAMSTEC

*Y Hanafusa<sup>1\*</sup>, H Saito<sup>2</sup> and Y Abe<sup>3</sup>*

<sup>\*1</sup> *Data Research Center for Marine-Earth Sciences, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan*

*Email: hanafusay@jamstec.go.jp*

<sup>2</sup> *Data Research Center for Marine-Earth Sciences, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan*

*Email: saitoh@jamstec.go.jp*

<sup>3</sup> *Marine Works Japan Ltd., 2-16-32 Kamariyahigashi, Kanazawa-ku, Yokohama, Kanagawa 236-0042, Japan*

*Email: abe@mwj.co.jp*

## ABSTRACT

*Japan Agency for Marine-Earth Science and Technology (JAMSTEC) provides users of its data with comprehensive search services which enable users to find data from JAMSTEC's various data dissemination sites. These are the "JAMSTEC Data Search Portal" which helps users to search for observational data on a map, and the "JAMSTEC Data Catalog" which enables users to find data sites by selecting science keywords. The "Data Search Portal" and the "Data Catalog" have been developed and operated as dedicated metadata publication and search services collaborating with data sites in JAMSTEC.*

**Keywords:** Metadata, Web GIS, DIF, Data Search Service, JAMSTEC

## 1 INTRODUCTION

JAMSTEC has performed very wide variety of observation and research both on the Ocean and the Earth. In 2007 JAMSTEC established the data policy ([http://www.jamstec.go.jp/e/database/data\\_policy.html](http://www.jamstec.go.jp/e/database/data_policy.html)) on handling of data and samples from its research activities, and developed rules and management systems based on the policy. JAMSTEC operates over 100 research cruises a year and disseminates observational data from these cruises in the "Data Site for Research Cruises (<http://www.godac.jamstec.go.jp/cruisedata/e/>)", images in the "E-Library for Deep Sea Images (<http://www.godac.jamstec.go.jp/jedi/e/index.html>)", rock sample information in the "GANSEKI (<http://www.godac.jamstec.go.jp/ganseki/>)", sediment core sample information in the "Core Data Site (<http://www.godac.jamstec.go.jp/coredata/e/>) and biological sample information in the "Marine Biological Sample Database ([http://www.godac.jamstec.go.jp/bio-sample/index\\_e.html](http://www.godac.jamstec.go.jp/bio-sample/index_e.html))". Additionally various research projects operate their own data dissemination sites.

As data management system in JAMSTEC was developed, the numbers of data sites and the amount of opened data increased rapidly. On the other hand users have to search for data in various data dissemination sites designed for the specific data type and there were many contacts to the data management office to inquire whether the specific data is opened or not, and how to find the data. Therefore, JAMSTEC has developed two data search services which enable users to find data by themselves easily.

One is the "Data Search Portal ([http://www.godac.jamstec.go.jp/dataportal/index\\_eng.html](http://www.godac.jamstec.go.jp/dataportal/index_eng.html))" which helps users to search for observational data by specifying the area of interest on a map. The other is the "Data Catalog ([http://www.godac.jamstec.go.jp/catalog/data\\_catalog/index\\_en.html](http://www.godac.jamstec.go.jp/catalog/data_catalog/index_en.html))" which leads users to the appropriate data dissemination sites by selecting science keywords. "Figure 1" shows the schematic flow chart in finding data using these two services. Both systems provide just search service based on the metadata. Users are able to learn the details of the data on the linked data dissemination page and, if needed, to download the data or apply for the use of the off-line data.



**Figure 1.** Schematic flow of data search using “Data Search Portal” and “Data Catalog”.

## 2 DATA SEARCH PORTAL

### 2.1 Objective

The search target of the "Data Search Portal" is an individual observational data from research cruises, moorings and terrestrial observations. It was developed to serve comprehensive data search service through various data dissemination sites by specifying data types and the area of interest. The "Data Search Portal" was set in operation since November 2008.

### 2.2 Method

Using the "Data Search Portal" users are able to confirm the distribution of observations of specific data type and then select the area for search. If needed, additional criteria (research vessel name, cruise period, observation variable, etc.) can be added. The list of search results includes the common metadata (date, location, data id, etc.) and a link to the data dissemination site which enables users to move directly to the relevant data page. This is a simple URL link and the "Data Search Portal" can cooperate with any data system as far as a data page can be specified with URL.

The supposed main users of the “Data Search Portal” are scientists who are searching for the data in the same area at different time or other data type in the same area. So authors developed a map-based spatial retrieval function using a web GIS server (ArcIMS 9.2).

### 2.3 Metadata

Location information in points or lines is extracted from data files and merged into shape files with observation metadata. The metadata includes several common metadata for cruise or dive (cruise number, period, observation variable, etc.) and optional metadata for a specific data type (chief scientist, water depth, area name, etc.). The vocabulary of observation variable is determined originally based on the frequently observed variables in JAMSTEC. Data types in the "Data Search Portal" come mainly from research cruises or dives but some data types come from mooring observations, ocean bottom stations and terrestrial observations. At the end of 2011

over 30,000 observations shown in “Table 1” are opened on the “Data Search Portal”.

**Table 1.** Data types and number of observations on the “Data Search Portal”

Research Field	Data Type	Number of Records
General	Cruise Track, Dive Point	5,764
Oceanography	Temperature and Salinity Profile, Water Chemical Analysis, Current Profile, Primary Production, Sediment Core, etc.	16,635
Meteorology	Marine Meteorology, Terrestrial Meteorology, Atmospheric Composition, Fixed Point Observatory, etc.	196
Solid Earth	Bathymetry, Gravity, Magnetics, Rock Sample, Drilling Hole, etc.	2,981
Biology	Marine Biological Sample, Vegetation	1,175
Images	Still Image, Video	4,824
Total		31,575

### 3 DATA CATALOG

#### 3.1 Objective

While the target of the "Data Search Portal" is individual observation, the target of the "Data Catalog" is a larger data dissemination unit (data sites, databases, datasets, etc.). JAMSTEC aims to provide comprehensive metadata publication and search service for all of the data in JAMSTEC with these two systems complementally. The “Data Catalog” is also designed to be a data publication system for scientists who do not have data dissemination sites by themselves. JAMSTEC has opened the “Data Catalog” on September 2011.

#### 3.2 Method

The user interface of the “Data Catalog” is a classification tree of hierarchical keywords. Supposed users of the “Data Catalog” are scientists searching for data not in their main research field. They are not familiar with the appropriate keyword to search for data in the field. It is easy to understand for those users to select keywords from the tree confirming the structure of the keywords.

The hierarchical keywords of research fields or observations are listed in the category tree. Users are able to narrow down the list of metadata by selecting keywords in the category tree. Going down the tree, related sub categories and metadata of the keyword will appear as shown in “Figure 2”.

The content of each data site is not fully described in the common metadata format. Metadata in the “Data Catalog” is just for finding data sites. Details of the data site and data download depend on the each data site. Users are also able to search for metadata by free text.



Figure 2. Selecting a keyword in the category tree in the “Data Catalog” brings a list of relevant metadata.

### 3.3 Metadata

Authors adopted the “Directory Interchange Format (DIF)” in the “Global Change Master Directory (GCMD)” as the metadata standard of the “Data Catalog”. The structure of DIF metadata standard are shown in the “Directory Interchange Format (DIF) Writer's Guide (<http://gcmd.nasa.gov/User/difguide/>)”. The vocabulary of the keywords in the category tree is also based on the “Earth Science Keywords (Olsen, Major, Shein, Scialdone, Vogel, Leicester et al., 2007)” of GCMD. Because the DIF and its “Earth Science Keywords” cover a wide variety of research fields in the Ocean and the Earth sciences and keywords are controlled by GCMD, it is suitable to describe various research data in JAMSTEC uniformly. The metadata in the “Data Catalog” is also registered to the GCMD.

For domestic users Japanese version of metadata and a set of Japanese keywords of the “Earth Science Keywords” were made. In the Japanese page of the “Data Catalog” users are able to search for and browse metadata in Japanese. A specific metadata page in English is linked to the relevant Japanese page, and vice versa.

## 4 TOWARD THE “ONE STOP DATA SHOP”

JAMSTEC is going to develop easier and more effective data systems for users. One supposed solution is a "One Stop Data Shop", which enables users to search for, browse and get all data in JAMSTEC from a single site. We think "Data Search Portal" and "Data Catalog" are possible prototypes of search system in the "One Stop Data Shop". Integration or cooperation with other data systems in JAMSTEC or other institutions will be expected.

## 5 ACKNOWLEDGEMENTS

The authors thank scientists, ship engineers and onboard technicians who conducted JAMSTEC's research cruises. Colleagues in the data management office contributed in collecting and disseminating data and metadata. Our works in developing metadata publication and search systems greatly depend on their efforts.

## 6 REFERENCES

- Olsen, L.M., G. Major, K. Shein, J. Scialdone, R. Vogel, S. Leicester, H. Weir, S. Ritz, T. Stevens, M. Meaux, C. Solomon, R. Bilodeau, M. Holland, T. Northcutt, R. A. Restrepo, 2007 .NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 6.0.0.0.0

# SOLAR-TERRESTRIAL DATA ANALYSIS AND REFERENCE SYSTEM (STARS) - ITS HIGH POTENTIALITY FOR COLLABORATIVE RESEARCH

*M Kunitake<sup>1\*</sup>, K Yamamoto<sup>1</sup>, S Watari<sup>1</sup>, K Ukawa<sup>1,2</sup>, H Kato<sup>1</sup>, E Kimura<sup>3</sup>, Y Murayama<sup>4</sup>, and K T Murata<sup>1</sup>*

<sup>1</sup>Space Weather and Environment Informatics Lab., National Institute of Information and Communications Technology (NICT), 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan  
Email: kunitake@nict.go.jp

<sup>2</sup>Systems Engineering Consultants Co. Ltd. 4-10-1, Yoga, Setagaya, Tokyo 158-0097, Japan  
Email: kentaro.ukawa@nict.go.jp

<sup>3</sup>Dept. Medical Informatics of Medical School of Ehime Univ., Situkawa, Toon City, Ehime, 791-0295, Japan  
Email: ekimura@m.ehime-u.ac.jp

<sup>4</sup>Integrated Science Data System Research Lab., NICT, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan  
Email: murayama@nict.go.jp

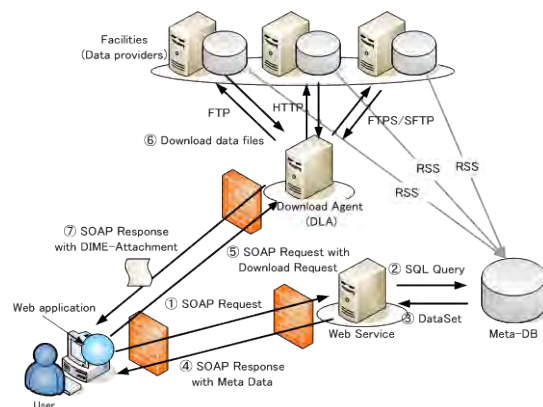
## ABSTRACT

Cross-sectional studies have become important for an improved understanding of various Solar-Terrestrial Physics (STP) fields given the great variety and different types of observations from the sun to the earth. In order to better combine, compare and analyze different types of data together, a system named STARS (Solar-Terrestrial data Analysis and Reference System) has been developed. Cross-sectional study requires cooperative work. The STARS has two functions for cooperative work, “Stars Project List (SPL)” and “Event Listing”. The SPL is used for exchanges of plotting information by cooperating persons. Event list database provides all users of STARS hints for recognizing typical occurrences of STP phenomena.

**Keywords:** Cross-sectional studies, Cooperative work, Combined plot, XML, Collaborative analysis, Common knowledge, Experience sharing, Solar-terrestrial physics, Common use

## 1 INTRODUCTION

A variety of cross-sectional studies have become important for further understanding of Solar-Terrestrial Physics (STP) fields. We need to combine, compare, and analyze different types of data together, for example, both satellite-based and ground-based observation data. To support such cross-over searches and analyses of data, we have developed a system named STARS (Solar-Terrestrial data Analysis and Reference System) (Murata, Yahara, & Toyota, 2005), (Ishikura, Kimura, Murata, Kubo, & Shinohara, 2006). The URL of the brief explanation of STARS is “<http://aoswa.nict.go.jp/application.html>”. The URL of detailed description of STARS is “[http://seg-web.nict.go.jp/e-sw/download/data/STAR5manual\\_e.pdf](http://seg-web.nict.go.jp/e-sw/download/data/STAR5manual_e.pdf).”) Figure 1 shows the overview of the STARS including meta-data and data flow. The STARS has functions to search whether the expected data exist or not, to make a combined plot, and to save the plot or data. Figure 2 shows an example of a combined plot.



**Figure 1.** The structure of the STARS and the flow of meta-data and data

Cross-sectional study often requires cooperative work by researchers whose own specialties are different from each other. Usually, a single researcher cannot cover all of the fields, but one or some of the fields. If findings and experiences of each researcher are exchanged with each other, these exchanges boost to do cooperating analysis. The STARS has two special functions for cooperative work. These are “Stars Project List (SPL)” and “event listing”. In this paper, we focus on these two functions.

## 2 STARS PROJECT LIST (SPL)

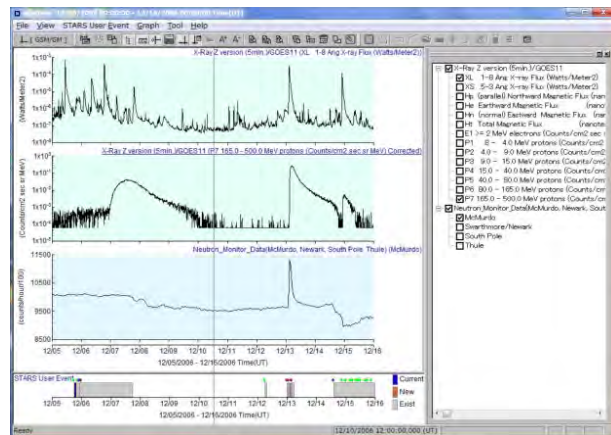
After a user makes a combined plot on the STARS, information about the plot can be stored in an file known as a Stars Project List (SPL). The SPL includes information with which any user can the same combined plot in the STARS. Figure 3 shows an example of the SPL. The information contains start/end date and time, data ID number, plotting status, and details for plotting. Using the any user can easily make the same plot without checking detailed download file options and plotting options. Further, any user can revise the combined plot by adding data file or by changing plotting options. We introduce two use cases of cross-sectional studies by using the SPL.

### Case 1. Plotting information exchange for cooperating analysis

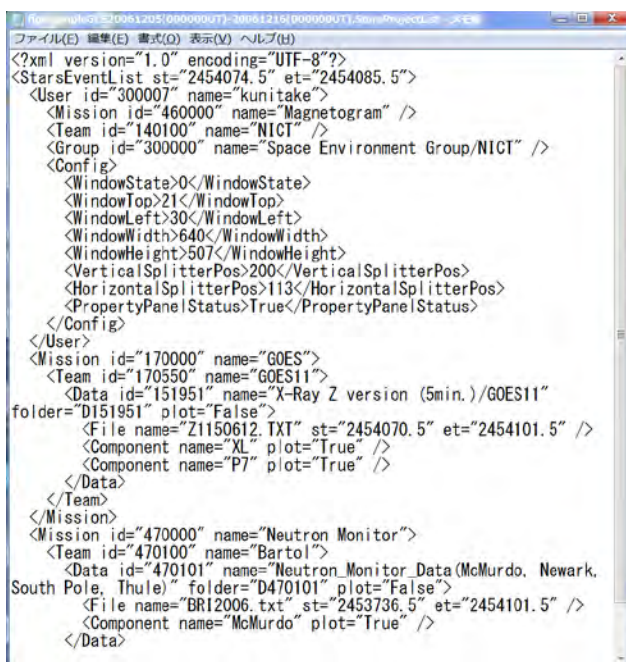
The SPL is used for plotting information exchange between the user “X” and the user “Y”. The detailed example is as following. After the user “X” makes a combined plot, the user “X” stores the plotting information in an SPL file. When the user “X” sends the SPL to another user “Y”, the user “Y” can make the same plot on the STARS based on the information stored in the SPL. So, the user “Y” can easily reach the same standpoint as the user “X” did. Then, the user “Y” modifies the plot based on Y’s own special knowledge after viewing original plot. After the user “Y” makes revised plot, saves an SPL as a new name, and sends the new SPL to the user “X”, the user “X” can know the additional viewpoints by looking through the modified plot. Such an interactive way by exchanging of S PL gives a quick way to do collaborative analysis.

### Case 2. Accumulation of common research knowledge

If many researchers accumulate their SPLs into a common place, accumulated SPLs would be used for SPL database. If a coordinator makes a subset of the database from the SPL database with a clear aim, such an SPL subset is useful not only for plot makers but also for any users of the STARS.



**Figure 2.** An example of combined plot by using the STARS. The time period of the plot includes several phenomena in Solar-Terrestrial physics. The top panel of the plot shows the solar X-ray flux observed at GOES 11 geostationary satellite. The second panel shows the proton flux observed at GOES 11 geostationary satellite. The third panel shows cosmic ray counts observed by neutron monitor on the ground (McMurdo station).



**Figure 3.** Example of SPL. The SPL is an XML file, which includes information with which user worked on the STARS: start/end date and time, the name of downloaded data file, and plot information.

One actual example is our SPL subset website for space weather researchers and users (Figure 4) (URL is [http://seg-web.nict.go.jp/e-sw/spl/index\\_e.html](http://seg-web.nict.go.jp/e-sw/spl/index_e.html)). On the web site, typical outstanding space weather occurrences are collected. Any persons who are interested in one of the occurrences can select the corresponding SPL and download the SPL from this website and they can make a plot.

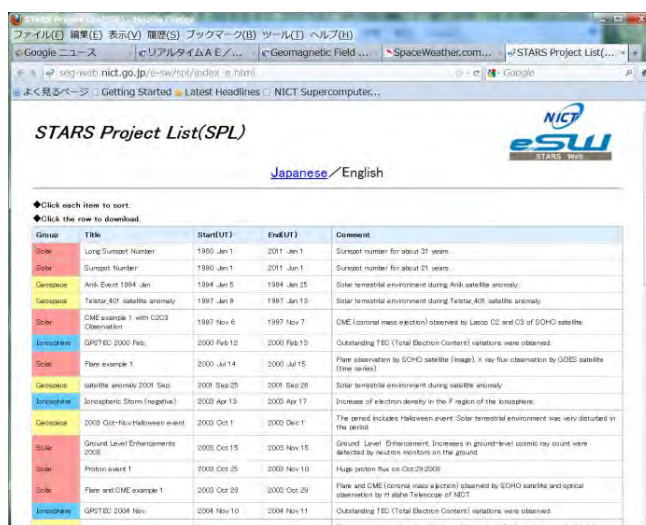


Figure 4. SPL subset website for space weather

### 3 EVENT LISTING

#### 3.1 Making and viewing event list

When a user of the STARS finds an interesting variation of typical phenomena in the plot, then the user recognizes it as an “event”. The user can in turn register the “event” in the event list in the STARS. Each “event” is described in XML and has detailed information (title, start/end time, name who registers the “event” etc.). Registered “events” have been accumulated in the event list database.

Any user can then view “events” which have been already registered by other users as well as by oneself. When many “events” are registered in the event list database by various users and many users share the “event” information by quick viewing, the event list would become common knowledge among users of the STARS.

When a user makes a combined plot on the STARS, the user can know any of the “events” which exist in the analyzing time period. There are two ways. One is to look through the extracted event list. From the whole registered event database, the extracted event list extracts “events” which exist in the analyzing time period. The other is to glance about “event” marks on the combined plot. Each “event” is shown as one pin mark in the combined plot. When a user does double-clicking on a pin mark, detailed information of the selected “event” appears.

#### 3.2 Effectiveness of the event list

The “event” information is shared by users through the event list. As the Solar-Terrestrial Physics (STP) fields have been observed by a wide variety of technique, it is rather hard for one person to become a specialist in all of the observations. If some researchers come together for collaborative research, the total number and kinds of observation which anyone is not familiar with will be minimized. Figure 5 shows schematically the way for an effective usage of the event list to proceed. Suppose that three researchers participate in analyzing data which covers several different fields and that each participant has some special knowledge of one observation. If participant “C” is a specialist in “observation CCC”, but not a specialist in “observation DDD or EEE”, then participant “C” can make a contribution by the registration of “event” #C1. If participant “D” is a specialist in either “observation DDD”, but is not a specialist in “observation CCC or EEE”, then participant “D” can make a contribution by the registration of “event” #D1. After participants C, D, and E have all added their “events” which are related to each participant’s own special observation to the event list, the event list eventually benefits from a richer and more informative set of participants contributing their specific expertise.

## 4 FUTURE WORK

A large number of “events” have been accumulated in the event list in the STARS. Many numbers of SPLs have been collected in another useful list. We are developing a useful website (I-space weather). It is a portal web site. Variety types of services are planned to be customized for space weather researchers in the web site. Information related to space weather forecast is to be shown also. One of the customized services has crossover search functions by key parameters of “event” or by SPL. Another customized service is adding information to SPL. It is possible to add comment description to SPL. It will be good for analysis by cooperating persons and for search by any user.

## 5 CONCLUSION

The STARS is a system which realizes the crossover searches and integrated analyses of ground-based and satellites observations of solar-terrestrial physics. The STARS has several advanced functions (data search, crossover comparison, plotting information exchange by SPL, and common use of event list). Plotting information exchange by Stars Project List (SPL) and common use of event list are useful for collaborating work.

As an SPL contains detailed information of a combined plot, not only the user who made the combined plot but also any other users can easily make the same plot without checking data file download options and plotting options. When any user modifies the plot, the modification can be saved in a new SPL. By information exchange by SPL, cooperating analysis by cross-sectional fields is to be progressed effectively.

If domain experts and specialists in other research fields are expected to register many “events” in the event list database, the database will in turn provide users of the STARS crossover hints for recognizing typical phenomena. In other words, event list database would be used as common research knowledge for all users of the STARS.

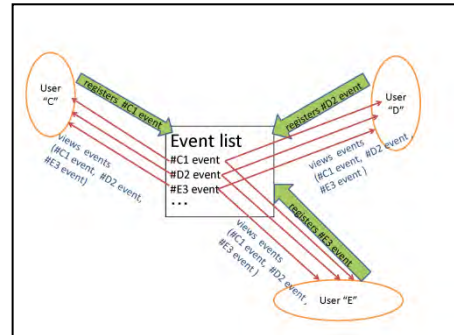
## 6 ACKNOWLEDGEMENTS

We appreciate all of the global institutes for providing data to the STARS. The present work was done by using resources of the OneSpaceNet (the NICT science cloud).

## 7 REFERENCES

Murata K. T., H. Yahara, and K. Toyota (2001), Software design via object-oriented methodology and network database for solar-terrestrial observation data, Database system, 123-5, database, pp. 31-36.

Ishikura S., E. Kimura, K. T. Murata, T. Kubo, I. Shinohara (2006), Automatic meta-data collection of STP observation data, AGU Fall Meeting, San Francisco, CA, USA.



**Figure 5.** An effective usage of the event list. Each person can register each “event”. Any person can view all of the accumulated events from the event list.



# Data Publication



# The World Ocean Database

*S Levitus<sup>1\*</sup>, J I Antonov, O K Baranova, T P Boyer, C L Coleman, H E Garcia, A I Grodsky, D R Johnson, R A Locarnini, A V Mishonov, J R Reagan, C L Sazama, D Seidov, I Smolyar, E S Yarosh, and M M Zweng*

<sup>1\*</sup>National Oceanographic Data Center, 1315 East West Highway, Silver Spring, MD, 20910  
E-mail: Sydney.Levitus@noaa.gov

## ABSTRACT

*The World Ocean Database (WOD) is the most comprehensive global ocean profile-plankton database available internationally without restriction. All data are in one well-documented format and are available both on DVDs for a minimal charge and on-line without charge. The latest DVD version of the WOD is the World Ocean Database 2009 (WOD09). All data in the WOD are associated with as much metadata as possible and every ocean data value has a quality control flag associated with it. The WOD is a product of the U.S. National Oceanographic Data Center and its co-located World Data Center for Oceanography. However the WOD exists because of the international oceanographic data exchange that has occurred under the auspices of the Intergovernmental Oceanographic Commission (IOC) and the International Council of Science (ICSU) World Data Center (WDC) system. World Data Centers are part of the ICSU World Data System.*

**Keywords:** World Ocean Database, ocean profile data, Intergovernmental Oceanographic Commission (IOC), International Council of Science (ICSU), National Oceanographic Data Center, World Data Center.

## 1 INTRODUCTION

The World Ocean Database (WOD) is the largest collection of ocean profile and plankton data available internationally without restriction and with all data made available in a common format. The WOD is available both on DVD and online ([www.nodc.noaa.gov](http://www.nodc.noaa.gov)). The WOD is a product of the U.S. National Oceanographic Data Center (NODC) and its co-located World Data Center for Oceanography (WDC). The WOD is a product built by merging thousands of originators data sets from many different countries and organizations. These data sets, which are sent to NODC/WDC in many different formats, are put into a common database and quality control is performed on the data (Boyer and Levitus, 1994; Conkright et al., 1994). Each data value has quality control (QC) flags associated with it. The data are made available formats including ASCII and netCDF.

The WOD exists because of the international oceanographic data exchange that has occurred under the auspices of the Intergovernmental Oceanographic Commission (IOC) and the International Council of Science (ICSU) World Data Center (WDC) system. We emphasize that it is the originators data sets that represent the archive of oceanographic data maintained by the U.S. National Oceanographic Data Center (NODC). All originators data and accompanying metadata are archived electronically (saved to disk).

It is important to note that the WOD is a product derived from the originators' data stored in the Ocean Archive System (OAS) at NODC. NODC archives the complete originators' data and metadata, exactly as submitted, in the OAS. The original data can be retrieved at any time. The WOD does not completely represent all data in the OAS. WOD contains profile data for 23 oceanographic variables, but the OAS archives many kinds of ocean data that the WOD does not currently store (measurements of metal concentrations, amino acids, etc). In the future, additional variables may be added to the WOD and processed from data in the OAS. Another reason for archiving originators data exactly as they are sent to NODC/WDC is that if a mistake in processing data into WOD is identified, we can access originators' data from the OAS and reprocess these data. WOD is serving a need for a standardized, quality-controlled scientific ocean data product. However, it is still critical to archive originators' data as it was sent to NODC.

In order for the international scientific community to develop climate system forecast capability for seasons, to determine the role of the world ocean as part of the earth's climate system and to improve climate assessments for decadal and longer time-scales, the most complete databases of historical oceanographic data such as the WOD are required. These databases, and scientific products based on these databases, represent the

infrastructure on which much ocean and climate research and assessments are now based. Specifically:

- a) Objective analyses of the data in these databases provide gridded climatologies that are used as initial and boundary conditions for ocean climate simulations and to verify simulations of the climate system;
- b) The data are used to prepare diagnostic studies, particularly for estimating interannual-to-decadal ocean variability of ocean heat content (Levitus et al., 2000);
- c) In recent years the data have been used as the input for ocean data assimilation efforts (Carton and Giese, 2008);
- d) The international scientific community advises national and international bodies on such issues as climate change, e. g. the Intergovernmental Program on Climate Change (IPCC). Hence, the international oceanographic and climate communities should have access to the most complete electronic oceanographic databases possible. Regardless of one's views about the origins of observed changes of the earth's climate system (anthropogenic, internal, or natural), the scientific community needs the best scientific databases possible to perform scientific research on this topic;
- e) Substantial resources have been, and continue to be, allocated for national and international ocean and climate programs such as Tropical Ocean and Global Atmosphere (TOGA), World Ocean Circulation Experiment (WOCE), Global Ocean Ecosystems Dynamics (GLOBEC), Joint Global Ocean Flux Study (JGOFS), Climate and Global Change, Climate Variability and Prediction (CLIVAR), and for the establishment of a Global Ocean Observing System (GOOS). Planners of such programs should have access to all historical oceanographic data in order to optimize measurement strategies for these programs. Scientists analyzing data from such programs need historical data in order to study interannual-to-decadal variability. Operational forecast centers need historical data in order to perform quality control of synoptic data;
- f) The data are a crucial tool to understand fisheries variability, and to manage fisheries and other marine resources;
- g) More specifically the data in the WOD has been used in a variety of products including the production of global climatologies of temperature, salinity, oxygen, and nutrients at ocean depths from the sea surface down to 5,500 m depth (Locarnini et al. (2010), Antonov et al. (2010), Garcia (et al., 2010a,b). Regional atlases have also been prepared (Matishov et al., 2004) and studies of the frequency distribution of ocean variables computed (Levitus and Sychev). These climatologies are used in a variety of ways including initialization of ocean models, and ocean data assimilation studies among others;
- h) The data in WOD have also been incorporated into other atlases including the International Comprehensive Ocean-Atmosphere-Data-Set (Woodruff et al. (2011)).

## 2 WHAT IS INCLUDED IN THE WOD

The following table (Table 1) is a list of the oceanographic variables included in the WOD:

**Table 1.** List of the oceanographic variables included in the WOD

1. temperature;
2. salinity;
3. oxygen;
4. phosphate;
5. nitrate;
6. nitrate + nitrite;
7. silicate;
8. chlorophyll;
9. pH;
10. alkalinity;
11. CO<sub>2</sub> mole concentration;
12. DIC (Dissolved inorganic carbon);
13. Plankton (Biomass, abundance etc.);
14. chlorofluorcarbon-11;
15. chlorofluorcarbon-12;
16. chlorofluorcarbon-113;
17. Tritium [<sup>3</sup>H] (isotope);
18. ΔHe<sup>3</sup> (isotope);

19.  $\Delta C^{13}$  (isotope);
20.  $\Delta C^{14}$  (isotope);
21.  $\Delta O^{18}$  (isotope);
22. Argon (noble gas);
23. Neon (noble gas);
24. Helium (noble gas);
25. Beam Attenuation Coefficient (transmissivity);
26. Meteorological data observed (measured approximately at the sea surface) during the oceanographic measurements may include barometric surface air pressure, wind speed and direction, wave height and direction, and dry and wet bulb temperatures.

The following table is a list of the instrument types used to collect the data included in the WOD:

**Table 2** List of the instrument types used to collect the data included in the WOD

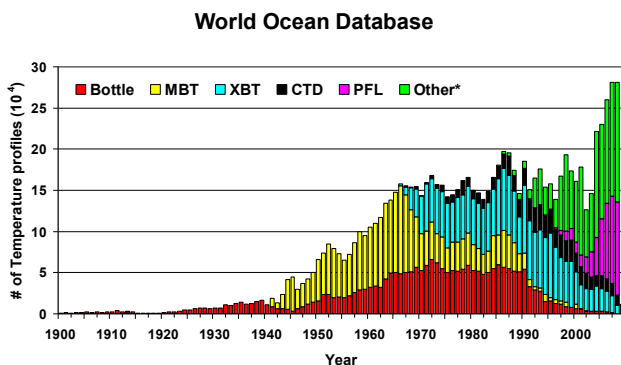
- 1) Ocean Station Data (OSD) (bottles, reversing thermometers);
- 2) Conductivity-Temperature-Depth (CTD);
- 3) Expendable Bathythermograph (XBT);
- 4) Mechanical Bathythermograph (MBT);
- 5) Towed CTD;
- 6) Profiling Floats (PFL);
- 7) Drifting buoy (DRB) (thermistor chains);
- 8) Moored buoy (MRB) (e.g., TAO, PIRATA, TRITON);
- 9) Autonomous Pinniped (APB) (instrumented elephant seals) (Boehlert et al., 2001);
- 10) Gliders (GLD);
- 11) Surface only data (SUR).

### 3 THE WOD IS A COMPLEX, HETEROGENOUS, AND LABOR-INTENSIVE DATABASE TO DEVELOP

Consider the last release of the WOD which is known as the World Ocean Database 2009 (WOD09) (Boyer et al., 2009; Johnson et al. 2009). The WOD09 is a global, comprehensive, integrated, scientifically quality-controlled database with all data in one well-documented format. We characterize WOD09 as a “heterogeneous” database. As an example, the Ocean Station Data (OSD) component of WOD09 contains data from, 65,840 cruises, 3,465 ships and other platforms (buoys, profiling float, etc), 564 institutes, 70 countries, 653 Principal Investigators (P.I.s). Populating the database with these metadata and many other metadata (e.g., instrument codes, scientific methods used, etc) is labor intensive. Populating the database with the actual data is labor-intensive because the data come from so many different sources in so many different formats. Also, both the metadata and data we receive frequently have problems associated with them such as misreported units, locations, etc. which we attempt to fix with the assistance of the data originators.

### 4 TIME SERIES HISTORY OF INSTRUMENTAL DATA WITH TEMPERATURE DATA IN WOD09

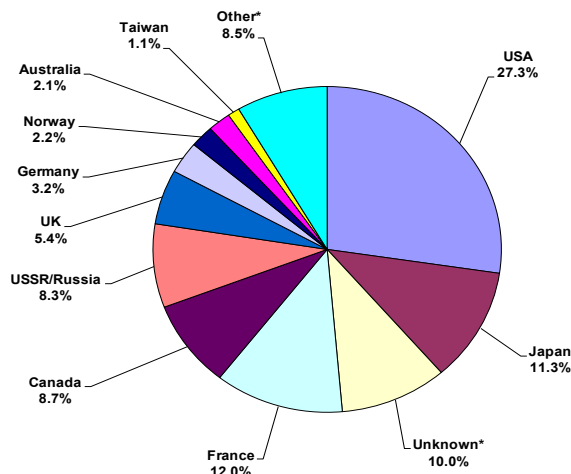
History of ocean temperature profile data set available in electronic form from NODC/WDC contained in WOD09 is shown in Figure 1.



**Figure 1.** History of ocean temperature profile data set available in electronic form from NODC/WDC contained in WOD09. “MBT is the abbreviation for mechanical bathythermographs, “XBT” represents expendable bathythermographs, “CTD” represents Conductivity-Temperature-Depth, and “PFL” represents profiling floats. The category “Other” includes data from gliders, moored buoys, drifting buoys, undulating ocean recorders (e.g., towed CTDs), and instrumented elephant seals.

## 5 NATIONAL CONTRIBUTIONS TO WOD09

National contributions to WOD09 are shown in Figure 2. This figure clearly documents that the WOD is a multinational product. The Intergovernmental Oceanographic Commission (IOC) and the International Council of Science (ICSU) have played major roles in facilitating the international exchange of oceanographic data that have been incorporated into the WOD. The percentages shown are based on the total amount of data in five major datasets (OSD, MBT, XBT, CTD, SUR).



**Figure 2.** Percentage of country contribution in WOD09. The Intergovernmental Oceanographic Commission (IOC) and the International Council of Science (ICSU) have played major roles in facilitating the international exchange of oceanographic data. The percentage shown are based on the total data in five datasets (OSD, MBT, XBT, CTD, SUR).

\*Other – countries contributing < 1%.

\*Unknown – country name not provided with originator’s metadata.

## 6 WOD UPDATING SCHEDULE

The WOD is updated online every three months with all of the data that we have processed for that quarter. Corrections are made for data and metadata reported or found to have been in error and reported to data originators. We compute seasonal temperature anomaly fields for 0-700 m depth and these are placed online ([www.nodc.noaa.gov](http://www.nodc.noaa.gov)) as well as the 0-700 m ocean heat content field and time series of the global integral of this field. We plan to extend these products to include salinity and extend all of the analyses and computations down to 2,000 m depth. Approximately every four years the data in the WOD are subjected to additional quality control and a new database is released on DVD and on-line.

## 7 DATA AVAILABILITY AND ACCESS

As part of its commitment to the scientists, institutions, and countries that have made their oceanographic data available, the GODAR and the WOD projects through NODC/WDC have made all data available on DVD media as well as on-line via the Internet from the NODC/WDC website ([www.nodc.noaa.gov](http://www.nodc.noaa.gov)). Beginning with *World Ocean Database 1998*, all data have been made available on-line.

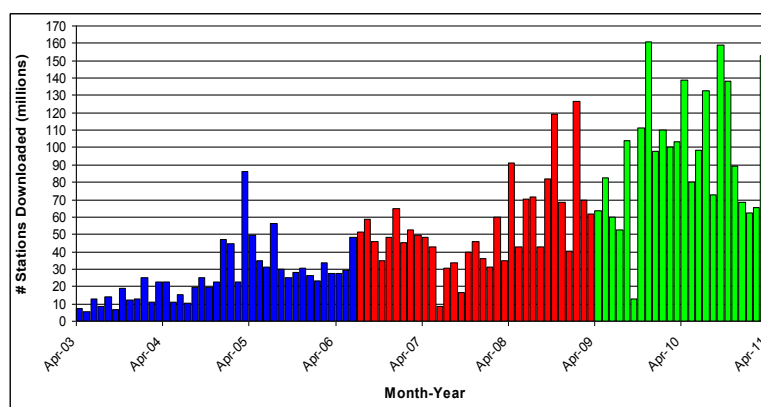
The *World Ocean Database* products come with software conversion routines so that users of software packages, databases, and programming languages such as MATLAB, IDL, PC-Surfer, C, and FORTRAN can access the data. In response to user requests, we have defined the WOD format to be as ‘self-defining’ as possible so as to eliminate, or at least minimize, the need for any structural changes to the format when new data types are added. All code tables, documentation, and software containing metadata are available on-line as well as on the CD-ROMs and DVDs which are used to distribute the WOD series. When a new database is released (every 3-4 years) users can acquire the new database or simply acquire data for those ocean stations that have been added or modified since the previous release. In addition, as corrections are made to the database after a release of WOD, users can acquire any modified data several days after the end of every month. There is a “Help Desk” and “Frequently Asked Questions” for the database available on-line.

Selection software, (WODSelect) (developed by Mr. Tim Boyer and Ms. Olga Baranova), allows users to access

data on-line by specifying geographic area, observation dates, instrument type, measured variables, deepest measurement, country, ship/platform, project name, and institute. Data are made available in a Comma-Separated-Values format. WODselect supports the goals of the IOC, ICSU WDC and United States data exchange systems to promote open access to scientific data. Additionally, it supports the United Nations Framework Convention on Climate Change to “promote and cooperate in the full, open and prompt exchange of relevant scientific, technological, technical, socio-economic and legal information related to the climate system and climate change”.

## 8 WOD DATA DOWNLOADED

All the data in WOD are available online at [www.nodc.noaa.gov](http://www.nodc.noaa.gov). Figure 3 shows the time series of the number of ocean stations downloaded by month since April 2003. Since the implementation of WODselect in April 2003, WODselect has responded to 171 thousand database queries and served over 28 billion stations (1493 GB) via the NODC FTP server. As of April 23, 2011 statistics shows over 114 million (59 GB) stations were downloaded. It is impractical to be limited to downloading one profile at a time. For example the WOD09 contains approximately 9 million temperature profiles and 3.5 million salinity profiles as well profiles of other variables. Therefore a user must be able to easily download multiple profiles.



**Figure 3.** Time series of the number of ocean stations downloaded via the Internet by month from the World Ocean Database. Figure 3. Since the implementation of WODselect in April 2003, WODselect has responded to 171 thousand database queries and served over 28 billion stations (1493 GB) via the NODC FTP server. As of April 23, 2011 statistics shows over 114 million (59 GB) stations were downloaded. Scientists do not want to be limited to downloading one profile at a time. WOD09 contains approximately 9 million temperature profiles, 3.5 million salinity profiles as well as profile for other ocean variables therefore a user must be able to easily download multiple profiles.

## 9 FUTURE WORK

The WOD project continues incorporating as much historical as well as modern oceanographic data as possible. The outlook for continued international cooperation is excellent. The WOD and products and papers based on it are frequently cited in the scientific literature. We plan to continue our work with the cooperation of the international scientific and data management communities to improve the WOD and products based on it.

## 10 ACKNOWLEDGMENTS

We acknowledge the contribution of many individuals, organizations, and countries to the projects described in this document. Scientists and technicians studying the world ocean have undertaken the task of collecting and processing the data. Oceanographic data centers and marine institutes have been particularly helpful through their participation. The NOAA Climate and Global Change Program, the NOAA Earth Science Data and Information System Management Program, the NOAA Environmental Data Rescue Program, and the NOAA Climate Database Modernization Program have supported the work of the GODAR project. NASA contributed to the development of enhanced upper ocean thermal data sets. Much of the international exchange of oceanographic data has taken place under the auspices of the International Oceanographic Data and Information

Exchange (IODE) committee of the IOC and of the ICSU World Data System (formerly known as the World Data Center system).

## 11 REFERENCES

Antonov, J.I., Seidov, D., Boyer, T.P., Locarnini, R.A., Mishonov, A.V., Garcia, H.E., Baranova, O.K., Zweng, M.M. & Johnson, D.R. (2010) *World Ocean Atlas 2009, Volume 2: Salinity*. S. Levitus, Ed. NOAA Atlas NESDIS 68, U.S. Gov. Printing Office, Wash., D.C., 184 pp.

Boehlert, G.W., Costa, D.P., Crocker, D.E. Green, P., O'Brien, T., Levitus, S. & Le Boeuf, B.J. (2001) Autonomous Pinniped Environmental Samplers; Using instrumented animals as oceanographic data collectors. *J. Atm. and Oceanic Tech.*, 18, 1882-1893.

Boyer, T. P. & Levitus, S. (1994) *Quality control of oxygen, temperature and salinity data*. NOAA Technical Report No. 81, National Oceanographic Data Center, Wash., D.C., 65 pp.

Boyer, T.P. Antonov, J.I., Baranova, O.K., Garcia, H.E., Johnson, D.R., Locarnini, R.A., Mishonov, A.V., Seidov, D., Smolyar, I.V., & Zweng, M.M. (2009) *World Ocean Database 2009, Chapter 1: Introduction*, NOAA Atlas NESDIS 66, Ed. S. Levitus, U.S. Gov. Printing Office, Wash., D.C., 216 pp., DVD. (Available online at [www.nodc.noaa.gov](http://www.nodc.noaa.gov)).

Carton, J.A., & Giese, B.S. (2008) A reanalysis of ocean climate using simple ocean data assimilation (SODA). *Mon. Wea. Rev.*, 136, 2999-3017.

Conkright, M.E., Boyer, T.P. & Levitus, S. (1994) Quality control and processing of historical oceanographic nutrient data. *NOAA Technical Report NESDIS 79*, National Oceanographic Data Center, Wash., D.C., 75 pp.

Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Baranova, O.K., Zweng, M.M. & Johnson, D.R. (2010a) *World Ocean Atlas 2009. Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation*. S. Levitus, Ed. NOAA Atlas NESDIS 69, U.S. Gov. Printing Office, Wash., D.C., 342 pp.

Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Zweng, M.M., Baranova, O.K. & Johnson, D.R. (2010b) *World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, and silicate)*. S. Levitus, Ed. NOAA Atlas NESDIS 70, U.S. Gov. Printing Office, Wash., D.C., 397 pp.

Johnson, D.R., Boyer, T.P., Garcia, H.E., Locarnini, R.A., Baranova, O.K. & Zweng, M.M. (2009) *World Ocean Database 2009 Documentation*. Ed. S. Levitus. NODC Internal Report 20, NOAA Printing Office, Silver Spring, MD, 175 pp., (Available online at [www.nodc.noaa.gov](http://www.nodc.noaa.gov))

Levitus, S., Antonov, J., Boyer, T.P. & Stephens, C. (2000) Warming of the World Ocean. *Science*, 287, 2225-2229.

Levitus, S. & Sychev, Y. (2002) *Atlas of temperature-salinity frequency distributions: North Atlantic*. International Ocean Atlas and Information Series, Vol. 4. NOAA Atlas NESDIS 55, U.S. Gov. Printing Office, Wash., D.C., CD-ROMs, 22 pp.

Locarnini, R.A., Mishonov, A.V., Antonov, J.I., Boyer, T.P., Garcia, H.E., Baranova, O.K., Zweng, M.M. & Johnson, D.R. (2010) *World Ocean Atlas 2009, Volume 1: Temperature*. S. Levitus, Ed. NOAA Atlas NESDIS 67, U.S. Gov. Printing Office, Wash., D.C., 184 pp.

Matishov, G., Zuyev, A., Golubev, V., Adrov, N., Timofeev, S., Karamusko, O., Pavlova, L., Fadyakin, O., Buzan, A., Braunstein, A., Moiseev, D., Smolyar, I., Locarnini, R., Tatusko, R., Boyer, T. & Levitus, S. (2004) *Climatic Atlas Of The Arctic Seas 2004: Database of Barents, Kara, Laptev and White Seas- Oceanography and Marine Biology*. NOAA Atlas NESDIS 58, World Data Center for Oceanography-Silver Spring, International Ocean Atlas and Information Series, Volume 9.

Woodruff, S.D., Worley, S.J., Lubker, S.J., Ji, Z.H., Freeman, J.E., Berry, D.I., Brohan, P., Kent, E.C., Reynolds, R.W., Smith, S.R. & Wilkinson, C. (2011) ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *Int. J. Clim.*, 31, 951-967.

# IS DATA PUBLICATION THE RIGHT METAPHOR?

*MA Parsons<sup>1\*</sup> and PA Fox<sup>2</sup>*

<sup>1</sup>*National Snow and Ice Data Center, University of Colorado, UCB449, Boulder, CO 80309*

*Email: parsonsm@nsidc.org*

<sup>2</sup>*Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8<sup>th</sup> St., Troy, NY 12180*

*Email pfox@cs.rpi.edu*

## ABSTRACT

*International attention to scientific data continues to grow. Opportunities emerge to re-visit long-standing approaches to managing data and to critically examine new capabilities. We describe the cognitive importance of metaphor. We describe several metaphors for managing, sharing, and stewarding data and examine their strengths and weaknesses. We particularly question the applicability of a “publication” approach to making data broadly available. Our preliminary conclusions are that no one metaphor satisfies enough key data system attributes and that multiple metaphors need to co-exist in support of a healthy data ecosystem. We close with proposed research questions and a call for continued discussion.*

**Keywords:** data publication, data system design, data citation, semantic Web, data quality, data preservation, cyberinfrastructure.

## 1 INTRODUCTION

Data authors and stewards rightfully seek recognition for the intellectual effort they invest in creating a good data set. At the same time, we assert that good data sets should be respected and handled like first class scientific objects, i.e. the unambiguously identified subject of formal discourse. As a result, people look to scholarly publication—a well-established, scientific process—as a possible analog for sharing and preserving data. Data “publication” is becoming a metaphor of choice to describe the desired, rigorous, data stewardship approach that creates and curates data as first class objects (Costello, 2009; Klump et al., 2006; Lawrence et al., 2011). The emerging International Council for Science World Data System (WDS)<sup>1</sup> and the American Geophysical Union<sup>2</sup> both explicitly advocate data publication as a mechanism to facilitate data release and recognition of providers. Costello (2009) even argues that science needs to adopt the robust principles of “publication” rather than informal “sharing” as a more effective way to ensure data openness and availability. While we strongly support these efforts to recognize data providers and to improve and professionalize data science, we argue in this essay that the data publication metaphor can be misleading and may even countermand aspects of good data stewardship. We suggest it is necessary to consider other metaphors and frames of thinking to adequately address modern issues of data science.

This essay grew out of several conversations between the authors. It began with a “tweet” by Fox at the 2010 CODATA meeting that first questioned the term “publication”.<sup>3</sup> Fox was being deliberately provocative; Parsons is easily provoked; and so the conversations began. About a year later, Parsons was invited to co-convene and speak at a session entitled simply “Data Publication” at the inaugural conference of the WDS. It was a bold move by the WDS to openly question their stated data publication paradigm, and it forced us, the authors, to begin to refine our thoughts beyond casual conversation. The presentation was politely received and generated some interest, enough for us to decide to go ahead and write an essay. We “published” the first draft of our essay on an open blog<sup>4</sup> in December 2011 and asked for community comment. We were overwhelmed by the response. Through comments on the blog, posts on other blogs, and direct e-mail, we received some 70 pages of review comments from more than two-dozen individuals over about six weeks. The reviews ranged from a few casual comments to very thorough and detailed critiques. The conversation was very stimulating, convincing us that it needs to continue more formally.

<sup>1</sup> [http://wds-kyoto-2011.org/WDS\\_Conference\\_Preliminary\\_Report.pdf](http://wds-kyoto-2011.org/WDS_Conference_Preliminary_Report.pdf)

<sup>2</sup> AGU position statement on “The Importance of Long-term Preservation and Accessibility of Geophysical Data” at [http://www.agu.org/sci\\_pol/positions/geodata.shtml](http://www.agu.org/sci_pol/positions/geodata.shtml)

<sup>3</sup> “okay, I’ll say it. The \*term\* data ‘publication’ bothers me more and more. Am leaning toward data release and \*maybe\* review, #CODATA2010” (@taswegian; posted 25 Oct. 2010).

<sup>4</sup> <http://mp-datamatters.blogspot.com/>



It is now almost a year later. The world of data and informatics continues to evolve rapidly. Just in the time since we released the first draft of this essay, Thomson Reuters announced a new data citation index, several new data journals launched, the Research Councils of the UK announced a new policy on open access to research outputs,<sup>5</sup> and US President Obama highlighted data management as a critical new job skill for the 21<sup>st</sup> century in his State of the Union address. In this rapidly changing environment with growing expectations and challenges facing data science, we believe it is critical to be as adaptive as possible. We must do what we can to avoid the negative “path dependence” that can inhibit adaptive evolution of a robust information infrastructure (Edwards et al., 2007). In that light, we present this revised essay. It is much improved by the many cogent comments received, but we are sure we will continue to provoke some disagreement. We remain convinced of our core message that no one metaphor or worldview is sufficient to adequately conceive the entire data stewardship and informatics enterprise. All metaphors have their strengths and weaknesses, their advantages and risks, their clarification and obfuscation. Our position is that this is especially true of Data Publication (Note we deliberately capitalize Data Publication here forward to reflect its status as a recognized metaphor and data management paradigm). As the most established metaphor and narrative, Data Publication may have both the greatest strengths and the greatest weaknesses. If we do not think critically of all our metaphors, we may see only the opportunities and not the risks. Correspondingly, if we do not seek new metaphors, we may miss new opportunities.

With this revised essay, we seek to further stimulate and advance the dialog among data scientists in a way that considers multiple worldviews and helps us conceptualize diverse approaches to science data stewardship and informatics. In Section 2 we discuss briefly the critical importance of metaphor in human communication and cognition. We then explore some existing worldviews and metaphors in Section 3 and examine their strengths and weaknesses in Section 4. We examine some alternative worldviews in Section 5 and conclude in Section 6 with a call to action based on a proposed research agenda.

## 2 THE IMPORTANCE OF METAPHOR AND FRAMING

At a simple level, a metaphor is a figure of speech where a word or phrase is applied to something for which it is not literally applicable. It is something symbolic or representative of something else. But it is much more than that. Metaphor is central to how people communicate and even to how we think and react to the world around us. As Lakoff and Johnson (1980) state in their seminal book *Metaphors We Live By*:

*Metaphor is for most people a device of the poetic imagination and the rhetorical flourish—a matter of extraordinary rather than ordinary language. Moreover, metaphor is typically viewed as characteristic of language alone, a matter of words rather than thought or action. For this reason, most people think they can get along perfectly well without metaphor. We have found, on the contrary, that metaphor is pervasive in everyday life, not just in language but in thought and action. **Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature** (p. 3, our emphasis).*

Understanding this “conceptual system” is central to cognitive science (Lakoff and Johnson, 1980a) and the system is increasingly seen to be fundamentally metaphorical in character (Lakoff, 1993). Lakoff and Johnson (1980) explore some of our most basic metaphors (argument is war, happy is up, sad is down, time is money, love is many things) and show how metaphors help define our modes of thought or worldviews. They show how metaphors help us create the complex narratives we use to understand our physical and conceptual experience. These complex narratives are made up of smaller, very simple narratives called “frames” or “scripts”. Framing and frame analysis are often used in knowledge representation, social theory, media studies, and psychology with much of the work stemming from Erving Goffman (1974).

These frames present a set of roles and relationships between them like characters in a play. They also help us define our terms and make sense of language, because words are defined relative to a conceptual frame. The word “sell” does not make sense without some understanding of a commercial transaction and some of the other roles and terms involved like “buyer,” “money,” and “cost”. Furthermore, by mentioning only one of these concepts like “buy” or “sell”, the whole commercial transaction scenario is evoked or “activated” in the mind (Fillmore, 1976). Similarly, we can see how particular roles and our subtle understanding of them emerge from the publication metaphor with terms like “author,” “editor,” “publisher,” “reviewer,” and “librarian”. We do not define these terms and let readers see what definitions emerge from their own conceptual frame.

5 <http://www.rcuk.ac.uk/research/Pages/outputs.aspx>

Lakoff (2008) further argues that framing is critical to human cognition. The neural circuitry to create a frame is relatively simple and our brain essentially uses framing as a sort of cognitive processing shortcut. If things are understood in the context of a frame, much is already unconsciously understood and need not be consciously processed. We know what to expect. Indeed, the vast majority of human thought is not conscious reflective thought but unconscious reflexive thought. Lakoff (2008) explores the role of this unconscious reflexive thought in politics and morality. While he arguably carries a political bias or agenda into his work, he clearly shows how language, metaphor, and framing play critical roles in any social enterprise. He summarizes the power of language well on page 14:

*Language is at once a surface phenomenon and a source of power. It is a means of expressing, communicating, accessing, and even shaping thought. Words are defined relative to frames and conceptual metaphors. Language ‘fits reality’ to the extent that it fits our body-and-brain based understanding of that reality. [...] Language gets its power because it is defined relative to frames, prototypes, metaphor, narratives, images and emotions. Part of its power comes from unconscious aspects: we are not consciously aware of all that it evokes in us, but it is there, hidden, always at work. If we hear the same language over and over, we will think more and more in terms of the frames and metaphors activated by that language.*

This last point is critical. Thinking in frames is natural and unavoidable. Frames provide a structure for cognition and understanding, but they also, by their nature, present a limited number of possible scenarios. Therefore, metaphors and framing can be extremely useful for describing and conceptualizing new ideas or paradigms, but they can also restrict our thinking and prevent us from seeing necessary alternatives or new possibilities.

We admire and are amused that Lakoff and Johnson turn their own logic back on their own discipline. The concluding sentence of Lakoff and Johnson (1980a) states: “The moral: Cognitive Science needs to be aware of its metaphors, to be concerned with what they hide, and to be open to alternative metaphors-even if they are inconsistent with the current favorites.” We seek to apply that same moral to our discipline of data science. In subsequent sections we examine Data Publication and other metaphors and worldviews around data science and stewardship. We focus on observational and modeled (rather than experimental) sciences, especially interdisciplinary Earth system science, but we believe our ideas, our metaphors, apply broadly.

### **3 CURRENT WORLDVIEWS AND ASSOCIATED METAPHORS**

Currently, we see (at least) five active worldviews on how to most effectively steward and share data in Earth system science. These worldviews vary in their maturity. They, and their corresponding data management approaches, are not mutually exclusive. It is common for data scientists to see themselves as actors in several narratives. Nonetheless, there is usually a dominating perspective that defines particular data management approaches. As Baker and Bowker (2007, p. 129) state “No institution is ever total, nor is any system totally closed. However, it remains true that there are modes of remembering that have very little to do with consciousness on the one hand or formal recording keeping on the other.” This is understandable. As Bruce Barkstrom (2012, personal communication) points out, the data management approaches and their worldviews come from different communities and cultures and are geared toward different users and different data types. There is nothing inherently good or bad about any one approach or worldview unless it is not aligned with community views. Our intent here is not to simply criticize particular systems or methods but rather to unpack our assumptions and understand our frames of thinking and underlying values. Furthermore, we present an admittedly cursory and even stereotypical assessment of the different worldviews. It was clear that our initial draft of this essay offended data scientists from all perspectives with its blithe analysis of the worldviews. As professional data scientists, we do not trivialize the complexity of our discipline, but we do seek to understand how we frame and conceptualize our challenges and strategies. So we must examine some of the stereotypes in which we operate. Broad conceptual understanding can sometimes be at odds with technical precision, but only through understanding our underlying modes of thought, even at a crude level, can we hope to expand and adapt those modes of thought to address the dynamic, complex challenges of data science.

With those considerations in mind, we examine five active worldviews on science data that we name with five metaphors: Data Publication, Big Iron, Science Support, Map Making, and Linked Data. We discuss their attributes in turn below and summarize them in Table 1.

The Data Publication approach seeks to be analogous to scholarly literature publication, and generally emerges from the culture of academic research and scholarly communication. Its focus is often on “research collections” (NSB, 2005) where data are extremely diverse in form and content but tend to be relatively small in size. Data Publication seeks to define discrete, well-described data sets, ideally with a certain level of quality assurance or peer-review. The data sets often provide the basis for figures and tables in research articles and other publications.

Published data are then considered to be first-class, reference-able, scientific artifacts, and they are often closely associated with peer-reviewed journal articles. The Data Publication focus tends to be on curation, archiving, and data quality. Data management systems, like the data, are not well standardized but tend to use relational or hierarchical data structures to organize the data. Further, the standards used across different data systems are fairly high level, e.g. exchange of Dublin Core metadata using protocols such as OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting). Data citation has been an important standards emphasis in Data Publication. Examples of Data Publication can be found in a variety of libraries and university repositories. An especially recognized advocate of Data Publication is the PANGAEA<sup>6</sup> system in Germany. A very explicit form of Data Publication is seen in newly emerging data journals such as *Earth System Science Data*. As mentioned, Data Publication is the most mature of the metaphors in play. Costello (2009) and especially Lawrence et al. (2011) provide much more rigorous descriptions of the paradigm, but it is important to recognize that they describe a desired, not fully realized situation. For example, Lawrence describes a data peer review scheme that is not yet fully or broadly adopted. Furthermore, despite these efforts, there is still incomplete agreement on the definitions and assumptions that arise from the frames of by Data Publication.

The Big Iron approach is akin to industrial production and often comes from more of an engineering culture found with large-scale data producers such as NASA. Big Iron typically deals with massive volumes of data that are relatively homogenous and well defined but highly dynamic and with high throughput. The Big Iron itself is a large, sophisticated, well-controlled, technical infrastructure potentially involving supercomputing centers, dedicated networks, substantial budgets, and specialized interfaces. It may also be a simpler collection of relatively common commodity software and hardware, but the focus is still on large volumes, reducing actual data transfer, computational scaling, etc. Historically less emphasis was placed on archiving, but it is an increasing concern. Big Iron systems rely heavily on data and metadata standards and typically use relational (e.g., MySQL) and hierarchical (e.g. HDF) data structures and organizational schemes. Significant emphasis is placed on consistent, rich, data formats and data production concerns such as careful versioning. Examples of the Big Iron approach include the European Space Agency’s Science Archives<sup>7</sup> or NASA’s Earth Observing System Data and Information System (EOSDIS)<sup>8</sup>. To be fair, nobody usually refers to such data systems as Big Iron. We use the term to be illustrative of a large-scale, production-oriented mode of thinking. “Big data” may be the more common term describing this worldview. It is also worth considering cultural differences across different production paradigms. For example, there are very different concerns around latency, data quality, spatial and temporal resolution, and other issues when addressing operational weather forecasting as opposed to long-range climate analysis, even though the data streams may ostensibly be very similar.

Science Support is viewed as an embedded, operational support structure typically associated with a research station or lab. In environmental sciences, the focus is often on place-based research such as is conducted at long term research stations or sites. Data management is seen as a component or function of the broader “science support” infrastructure of the lab or the project. Science support for a lab is defined differently in different contexts and tends to be very broadly conceived. It may include many things such as facilities management, field logistics, administrative support, systems administration, equipment development, etc. Often, there is no clear line between what is the science and what is the support. For example, data collectors at a field site may be lead investigators on a given research project or lab technicians supporting many projects. In this context, data tend to be the research collections similar to those in the Data Publication metaphor but there is often a focus on creating community collections by characterizing important fundamental processes or particular representative conditions over time. The data are organized in myriad ways, usually geared towards a specific set of intended uses and local reuse in conjunction with other local data. The historical Long Term Ecological Research (LTER) network is a good example of this approach where local science support functions remain constant over time even while a broader, network-level data system is added. Baker and Millerand (2010) describe the process of how the LTER information systems developed both locally and nationally and illustrate the Science Support perspective, where data management is both integrated into the science process, yet also partially outside the process in a support role (Lynn Yarmey, 2012 personal communication).

Map Making is most readily seen in so-called spatial data infrastructures (de Sherbinin and Chen, 2005; FGDC, 1997; NRC, 1993) and their associated geographic information systems (GIS). The perspective emerges naturally from land use and survey agencies that have been creating and working with maps for centuries. Map Making shares attributes of the other paradigms. Maps are certainly used in Science Support and Map Making could be seen as a subset of the Data Publication, but here the analogous publication is a map or an atlas rather than a journal article. On the other hand, national and international spatial data infrastructures often seek to operate the more centrally governed, standardized model of Big Iron. Here, however, the important metaphor is

6 <http://pangea.de>

7 <http://www.sciops.esa.int/index.php?project=SAT&page=index>

8 <http://eosps0.gsfc.nasa.gov/>

it is not the final product or the production process but rather the representation of the data and their associated science questions through a geographical perspective, notably the map<sup>9</sup>. Data in this approach tend to be more fixed in time, i.e. they are more geared toward describing geospatial features rather than dynamic processes. The Map Making focus tends to be on cartographic visualization and intercomparison with uneven attention to preservation. Data are well standardized around a map- (or grid-) based model with an associated (geo)database. Map Making has been especially successful in defining standards around things like coordinate reference systems, map projections, and map transfer protocols. Major examples of map-based systems include the INfrastructure for SPatial InfoRmation in Europe (INSPIRE),<sup>10</sup> OneGeology.org, and Geodata.gov in the US.

Linked Data is based on computer science concepts of the “Web of data” and relies on the underlying design principles behind the Semantic Web,<sup>11</sup> especially as described by Tim Berners-Lee.<sup>12</sup> The paradigm emerges from the culture of World Wide Web development, including non-science and commercial enterprises. The “data” in Linked Data are defined extremely broadly and are envisioned as small, independent bits with specific names (URIs) interconnected through defined semantic relationships (predicates) using model and language standards (e.g. the Resource Description Framework, RDF). The focus to date has been almost entirely on enabling interoperability and capitalizing on the interconnected nature of the web. There is also a major emphasis on *open* data. Scant attention is paid to preservation, curation, or quality. An underlying principle of this approach is that it uses a graph model not a hierarchical or relational model of data organization. This lends itself well to very distributed and interdisciplinary connections but also requires substantial agreement on the formal semantics, i.e. ontologies, to be useful for diverse audiences. Correspondingly, the standards focus, especially in the sciences, has been on the development of formal ontologies. This approach has been applied in a variety of contexts outside science and increasingly in life and medical sciences. There is growing discussion and use in the Earth sciences, such as in the Integrated Ocean Drilling Program (IODP)<sup>13</sup>. In many ways, Linked Data is not as comprehensive a worldview as some of the others. Arguably, it may be seen as a set of techniques or tools used within a broader context such as Data Publication (Bechhofer et al., 2011) that can potentially be accessible by a broad range of data producers, e.g. an individual researcher with programming skills. Again, however, we note the focus of the metaphor. As with Map Making, the metaphorical emphasis is not on the product or the process but the data representation; this time not as a geospatial map but as a network or graph.

#### 4 PROS AND CONS OF THE CURRENT WORLDVIEWS

Each of the worldviews described above have their strengths and weaknesses for understanding and addressing the challenges of data science. Nominally, the data management approaches that emerge from the different worldviews are fully capable of stewarding data according to defined best practice, but the varying perspectives and metaphors focus on different stages of the data life cycle, different audiences, and different challenges. We do not believe that any of the current data management paradigms fully meet all the basic criteria outlined by the ISO standard *Open Archival Information System Reference Model* (ISO, 2003), the broader guidance of the *Association of Research Libraries’ Agenda for Developing E-Science in Research Libraries* (ARL, 2007) or other general community guidance (Arzberger et al., 2004; Doorn and Tjalsma, 2007, Parsons et al. 2011).

We identified seven critical attributes of an effective, comprehensive data stewardship approach, based on the aforementioned guidance and our own worldview and values:

- Established trust (of data, systems, and people).
- Data are discoverable.
- Data are preserved.
- Data are ethically open and readily accessible to humans and machines.
- Data are usable, including some level of understandability.
- Effective, distributed governance of the data system.
- Reasonable credit and accountability for data collection, creation, and curation.

These are by no means all the desirable attributes, but we do not think that any of the current models fully address even these basics. In this section, we provide a cursory, subjective assessment of how the different

9 A broader conception of this metaphor might be “Sense making”. Areas like biological taxonomy and structural chemistry have different constructs for making sense of their information. Maps, however, are especially powerful metaphors and representational tools. Critical geographers have long shown how maps can be tools to assert power and authority and may be viewed as a product of authorial intent rather than objective data presentation (Harley, 1989; Koch, 2004). This is somewhat tangential, but it is another illustration of the power of metaphor in how we conceive of and represent data and their relation to broader conceptions of reality.

10 <http://inspire.jrc.ec.europa.eu/>

11 <http://linkeddata.org>

12 <http://www.w3.org/DesignIssues/LinkedData>

worldviews address these criteria. We examine each worldview briefly and then discuss Data Publication in more detail.

**Table 1.** Summary of attributes (rows) of some major data management related metaphors (columns)

	<i>Data Publication</i>	<i>Big Iron</i>	<i>Science Support</i>	<i>Map Making</i>	<i>Linked Data</i>
<i>Analog</i>	scholarly publication	industrial production	artisanal, task-specific production	cartography	World Wide Web creation
<i>Data characteristics</i>	small volume and diverse form, scale, and topics	high volume and more homogenous in form	small and diverse	geospatial features and attributes	many disparate and named entities
<i>Data organizational models</i>	hierarchical or relational	hierarchical	geospatial, hierarchical, and relational	geospatial and relational	linked graph
<i>Primary Focus</i>	data quality, certification, and preservation	throughput and manageable access	data synthesis and reproducibility,	map-based visualization and intercomparison	interoperability and interconnection
<i>Standards emphasis</i>	data citation	data formats, versioning	local processes	coordinate reference systems, spatial transforms	ontologies
<i>Examples in science</i>	PANGEA, university repositories	EOSDIS,	LTER	INSPIRE, Geodata.gov	IODP, MyGrid, Linked Open Government Data
<i>Metaphorical terminology</i>	data author, publisher, data citation	data producer, processing level, version release	data collector, support staff	data source, feature, layer	data provider, name, link, resource
<i>Cultural context</i>	libraries and university research groups (e.g. NSF science directorates)	system engineering and project management (e.g. NASA, DoD)	place-based research (e.g. focused institutes, NSF)	land use and management (e.g. USGS, local agencies)	computer science and commercial applications (e.g. NSF CISE and W3C)

Data Publication builds from the familiar and conceptually simple model of scholarly literature publication. “Publishers” are distributed and can act autonomously or in concert. Published data are usually well cared for, and often carry assertions that data are of high, or at least well-described, quality. The approach builds from the norms of scientific research and can be well trusted, but there is a corresponding lack of strong governance across systems. There is also little emphasis on data discovery and interoperability across systems. Data are often presented as they were created without explicit considerations of data integration or significant reuse beyond the scientific community. The approach works well for relatively stable data sets, but systems can be difficult to automate and do not always scale well. The attention is on preservation, and formal recognized scholarly contribution with less attention to “big data” issues such as latency, rapid versioning and reprocessing, and computational demands.

Big Iron approaches tend to be highly automated and hence well suited to formal audits and reprocessing. By design, the systems handle large volumes and streaming data well, and can provide very short latency when necessary. The systems usually have defined governance mechanisms with some sort of controlling authority or policy-level certification. On the other hand, Big Iron systems do not handle heterogeneous data well. They tend to be designed around a very consistent data model such as gridded fields. The systems are sometimes overly reliant on automation and tend to assume a certain type of use. Roles are not always well defined and systems are generally not very adaptive. More critically, Big Iron systems tend to underplay the need for preservation (although this is beginning to change). In general, there is more of an engineering focus than science focus, which is both a strength and a weakness in its own right.

Science Support is inherently localized in its focus. Systems may be well established and very useful for the designated community they are supporting. An important strength is the focus on data integration for that community, but data and systems are often not designed for use or access beyond the community. Governance structures across sites are only emerging and completely lacking in some disciplines. Data preservation is variable and largely dependent on the knowledge and interest of local science support staff. In contrast to Big Iron, there is a very strong science focus that makes the data very useful for its intended purpose but systems are more ad hoc and may lack design and preservation rigor.

Map Making is obviously well suited to and correspondingly limited to geospatial representation of data. It can be very useful for integrating data over geographic space, but it typically does not handle temporally dynamic data well. There is an established history of geospatial governance mechanisms with variable success. Of note is the emergent success of the Open Geospatial Consortium in the last decade at establishing widely adopted standards of interoperability. A history of proprietary systems and data formats has hindered data preservation, but that is rapidly improving. Nevertheless, core aspects of data stewardship such as preservation, access, and trust largely depend on the institutional context where the Map Making metaphor is applied.

The Linked Data approach is still fairly new and has not really considered the full data life cycle. Its primary strength is that it is built on a simple, highly scalable model that allows for broad data dissemination and very flexible machine processing. There is no *a priori* assumption of how data are to be used and the model handles extremely diverse data well. In a sense, the approach is data model independent (unlike Map Making for example), but it typically achieves this through a change from the original data model (to RDF). This creates issues for preservation. Indeed preservation is largely ignored in the Linked Data worldview. The approach suffers from poor versioning, auditability, and accountability, and it is generally not very human friendly. It also lacks a controlling central authority; this allows great flexibility but limits preservation and accountability.

We summarize our simplistic analysis in Table 2. While we recognize that most all of our assertions can be countered, we trust the reader can recognize some of the strengths and weaknesses we describe in the systems they are familiar with. More importantly, we have illustrated that by focusing on limited aspects or perspectives of a problem, one can often miss other important issues. None of the metaphors are complete and most data scientists operate in spaces that could be characterized by several of the worldviews. Nonetheless, many might argue that Data Publication is the most mature, well-understood worldview; therefore by better defining and refining Data Publication practice we best serve data science and stewardship. We do not believe that is a complete or wise approach.

Despite well-considered descriptions of formal Data Publication (Lawrence et al. 2011, Costello, 2009), it was clear in the review of this essay that there is no widely understood and accepted definition of what exactly Data Publication means. It was equally clear that “publication” carries many, differing, implicit assumptions that may not be true. A central argument for Data Publication is that the metaphor resonates with researchers. They understand their role in the process, it is said. Yet researchers are not knowledgeable of the refined definition of Data Publication. We argue that this creates false understanding; that the frames and roles of Data Publication create false assumptions that what is true for scholarly literature publication applies to data publication. Furthermore, the metaphor may be too restrictive and not allow researchers *or* data scientists to fully understand and adapt to the modern challenges of data driven science.

To illustrate our concerns we examine three frames that emerge from Data Publication that can create false assumptions and misguided approaches. First, peer review. Data Publication implies some level of imprimatur (Callaghan et al., 2009), and a “published” data set may be assumed to have undergone some sort of peer-review. Yet there are no standards or even agreement on what peer-review of a data set might mean (Parsons et al., 2010). Indeed, de Waard et al. (2006, 2008) demonstrate a rhetorical model of scientific publication that indicates that peer-review of data cannot truly parallel peer review of literature. The model makes the important distinction between the article, which is designed to persuade (Kuhn, 1996; Latour, 1987), and the data, which are intended to be simple fact.

Some communities have made admirable efforts to peer review data, but it is not really the same as traditional peer-review of literature, and the approaches vary. For example, the Planetary Data System has a long established peer-review scheme, but it is actually more like an audit that assures that a data set adheres to best practices of documentation, format, error characterization, etc. (McMahon, 1996). The *Earth System Science Data* journal and other emerging data journals and overlay journals combine the review of the data set with review of a more conventional article that is closely linked to the data (e.g., Callaghan et al., 2009; Pfeifferberger and Carlson, 2011). Lawrence et al. (2011) examine peer-review in depth and provide a useful data review checklist. These are valuable contributions, but we still find the peer-review frame to be limiting. The review of data is fundamentally different than the review of an argument in a paper, and the different

approaches have different meaning and levels of (implied) certification. Traditional human refereeing is appropriate for certain major data sets, but it is too slow and it will not scale to handle the growing deluge of data. We need to consider other models of what is essentially a quality assurance/quality control process and automate where possible. Thinking outside the peer-review frame can help us conceive of these models. For example, tracking how a data set is used over time may be more revealing of its quality and fitness for use than the formal opinion of two or three disciplinary experts. The quality of data depends on the application. Unlike with literature, there may still be value in releasing “poor quality” data because they may be useful for certain applications or because broad exposure of the data may lead to creative solutions to their prior limitations. Too often we have seen purported insufficient data quality used as an excuse to restrict data access. Data quality is a critical and difficult issue fundamentally different from the intellectual merit of a scholarly article. We should not let the Data Publication metaphor limit our thinking of how data quality can be addressed.

**Table 2.** Summary of strengths and weaknesses of the data management worldviews.

	<i>Data Publication</i>	<i>Big Iron</i>	<i>Science Support</i>	<i>Map Making</i>	<i>Linked Data</i>
<i>Trust</i>	good	moderate	good	moderate	Poor
<i>Discovery</i>	poor	moderate	poor	moderate	Good
<i>Preservation</i>	good	poor	variable	poor	Poor
<i>Access</i>	moderate	moderate	poor/moderate	good	Good
<i>Usability</i>	moderate	moderate	good	moderate/good	Moderate
<i>Governance</i>	poor	good	poor	moderate	Poor
<i>Credit and accountability</i>	good	poor/moderate	variable	poor/moderate	Variable

The second frame of concern is the closely related concept of data citation. We strongly support the data citation concept, but we feel that the publication metaphor has created some false expectations around it. Data citation might be better termed data reference. The primary purpose is to aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study (Ball and Duke, 2012; ESIP, 2012). This means that data need to be identified and located, ideally with a persistent identifier, as soon as they are available for use by anyone other than the original creator. Data release often occurs in stages. Data may be initially shared with a small team, later released to a broader group within the discipline sometimes with caveats, and then finally the data are released to the public, i.e. “published”. Technically, data need to be precisely referenced if they are used in a study at any time during those stages, but typically a DOI is not assigned until the final publishing stage. The DOI is meant to assert a sort of imprimatur. This does not seem an appropriate use of the DOI. We understand and appreciate the desire for an imprimatur, but find it odd that it be conveyed with a simple registration of an identifier. Identifiers and locators are necessary before formal, reviewed publication and there is nothing inherent in DOIs that ensure persistence of the data. It still relies on human due diligence (Duerr et al., 2011). We feel that the emphasis on “publication” underplays the often-broad use and evolution of a data set long before it may be formally published and can be making misleading assertions about the meaning and purpose of identifiers.

Another important aspect of data citation is the desire to provide fair credit for the intellectual and technical effort that goes into creating a good data set. Indeed, data citation is often seen as an incentive for researchers to release their data. Unfortunately, in our experience, scientists do not especially welcome data citation. Some like the idea; some see it as diluting citations to their paper. Also, some funding agencies question the idea of recognizing individuals as data authors. We do not necessarily agree with these detractors, but we see a problem in that the Data Publication metaphor has led us to conflate many issues into data citation, including reference, quality assertion, credit, and data discovery. This has only made the precise identification and reference issue more difficult. We need to separate the concerns and come at them from different directions.

Our third concern is with the close association of Data Publication with copyright, and restricted-access literature. While most scholarly publishers agree that data should be openly available regardless of the restrictions on the article, they still assume most data discovery comes through the article and that most data sets have at least one peer-reviewed article associated with them—an arguable assumption at best. If citation in publications is the primary means of identifying data, an unintended side effect may be to actually limit data access and discovery because of the restrictions on the article. We end up reinforcing the hidden “deep web of data” (Wright, 2009). And while data publishers are often strong advocates of open access, some argue that Data Publication necessarily includes licensing of the data set and mandating conditions of use (Klump et al., 2006). We much prefer the norms-based, copyright-free approach of an information commons as adopted by *Earth System Science Data*. Too often, we find that the Data Publication perspective is, as John Wilbanks (2009, personal communication<sup>14</sup>) says, focused on “the container and not the customer”. It requires publishers to spend undue time managing the definition of and access to the container, be it an article or a data set. It also implies a social contract that is not applicable for data. In traditional scholarly publishing authors relinquish certain rights and go through certain processes in exchange for receiving professional credit. The social contract for data could and probably should be much different. For example, semi-blind review may not be appropriate and rights around data are fundamentally different than copyrights on creative works. The focus on the container misses how in a networked world, the proliferation of copies and the customer’s ability to annotate, federate, transform, and integrate the content makes the content more valuable. Openness and flexibility is essential to maximizing value of data, and restricted data discovery and access still remain major inhibitors of data science endeavors (ICSU, 2011). Ironically, while those who advocate data publishing tend to be some of the strongest advocates for open data, we find that the Data Publication container can restrict access, interoperability, and creative use. All the metaphors need to be explored for approaches to *unlocking* the data in the “deep web”.

We have been severely critical of the Data Publication worldview. We do not suggest a wholesale rejection of the metaphor, but rather recognition of how it can sometimes restrict and even misguide our thinking. Scholarly publication is in the process of re-examining its own model (see for example the European Framework LiquidPub project<sup>15</sup>), and we should be open to learning from that process, but we should not assume it will provide a good analog for data. None of the current worldviews described completely address the full needs of robust data stewardship. Data Publication, in particular, has major strengths and is the most evolved, but it may also be the most misleading. Data Publication efforts should certainly continue, but we must remain open to other alternatives. It is critical to avoid the stifling “path dependence” than can inhibit the development of a useful and adaptive sociotechnical infrastructure (Edwards et al., 2007). We must actively challenge our thinking and seek out other worldviews and metaphors.

## 5 ALTERNATIVE WORLDVIEWS AND METAPHORS

While metaphors can limit our thinking, they can also help us conceive alternatives. To say that the Data Publication or any other metaphor is limiting is insufficient. We need to recognize other existing metaphors and actively seek new metaphors that complement each other and help us conceive of all aspects of the e-science data challenge. We believe this needs to be an ongoing conversation in the community, but we offer some initial ideas here.

We see two high-level metaphors that go beyond the data management enterprise and consider the larger whole of science communication: the concepts of infrastructures and ecosystems. The Data Infrastructure metaphor is well established. The geospatial data community has referred to national and global “spatial data infrastructures” since at least the early 1990s (NRC, 1993). More recently, the NSF “Blue Ribbon Advisory Panel on Cyberinfrastructure” has codified the concept, in the US at least, as “cyberinfrastructure” (Atkins et al., 2003). Considering an entire infrastructure helps us recognize the scale of our endeavor—it truly needs to reach across the entire scientific enterprise. But in many ways the concept of a data or information infrastructure is still being defined. More critically, conceptions of infrastructure too often ignore or underplay socio-cultural elements (Bowker et al., 2010). We, therefore find metaphors typically drawn from physical infrastructure concepts like railways and electrical utilities useful but also too simplistic. Indeed, infrastructure can be very difficult to study because it typically exists in the background—invisible and taken for granted (Star and Ruhleder, 1996). We support the developing field of infrastructure studies (Bowker et al., 2010), but despite the rich, sophisticated, and holistic examinations of this literature, data practitioners still tend to view infrastructure as a physical construct rather than the body of relationships defined by Star and Ruhleder (1996).

More recently, we have become intrigued by the metaphor of a “data ecosystem – the people and technologies collecting, handling, and using the data and the *interactions* between them” (Parsons et al., 2011, p. 557). We

14 See brief discussion and slide set at <http://scholarlykitchen.sspnet.org/2009/09/24/john-wilbanks-its-the-customer-not-the-container/>.

15 <http://project.liquidpub.org>



appreciate the extension of the common data life cycle metaphor and the focus on interactions and relationships. As our late friend, Rob Raskin (2012, personal communication), stated, “A characteristic of ecosystems is their interactions with the environment. Often, this role is more than a passive one, in that the ecosystem changes the environment—similar to how data affects the underlying science.” The ecosystem concept emphasizes adaptation, evolution, and diversity rather than a centralized command and control structure. It is similar to what Davenport (1997) and Nardi and O’Day (2000) call an “information ecology,” and it provides a useful perspective. Yet while the obvious metaphors like the seeding and growth of an idea or the evolution of a technology give us a holistic view, they are sometimes lacking in specifics. What is the equivalent of publishing a data set in an ecosystem? Data sprouting, growth, birth, release, culture ...? None of these are completely clear or are likely to truly resonate with researchers and help them understand their role.

Baker and colleagues (Baker and Millerand, 2010; Baker and Bowker, 2007) bring “infrastructuring” and information ecology together. They use the science and technology studies approach of infrastructure studies to examine the infrastructure of an ecology. They may be on to something. Sometimes mixing metaphors is necessary. Perhaps we should not try and find an overarching metaphor for the whole data management process. Perhaps that misses the point. Historically, in literature publication, each publisher filled the multiple roles of archiving, registration, dissemination, and certification of the paper. Van de Sompel et al (2004) and Priem and Hemminger (2012) argue that this model, with thousands of independent publishers each filling all roles, resists innovation and makes it difficult to change any one aspect of the system. Priem and Hemminger argue that we need to “decouple” the journal to create a “Web-like environment of loosely joined pieces—a marketplace of tools that, like the Web, evolves quickly in response to new technologies and users’ needs” (p.1). Van de Sompel et al. make a similar argument for a “scholarly ecology”. We welcome these idea and suggest that similarly, we need to start decoupling or disaggregating the functions of data stewardship to consider each function fully. By disaggregating we can also re-aggregate in new and different ways. For example, people are beginning to consider alternative aggregations of data in ways that connect Data Publication and Linked Data concepts. Alternative forms of information aggregation have been described as “publication packages” (Hunter, 2006) or “research objects” (Bechhofer et al., 2011; Belhajjame et al., 2012). In a modern information ecosystem it is unreasonable to assume one entity would do everything. It is necessary take multiple approaches to manage different types of data. We need to consider all the available paradigms and consider the various functions of data stewardship individually in their own right and as a whole. *We need not one metaphor but many.*

Schopf (2012) argues that we should treat data like how we build production software. That this will make data more readily accessible and available for broad re-use. She states: “We should be treating data as an ongoing process,” which presents a very different perspective than one that views data as a publication or an object. She further argues that this metaphor is readily understood and adopted by scientists. This is an interesting worldview. We like the emphasis on cyclical development and controlled, staged releases (e.g., development, staging, production). The perspective may not fully consider preservation and it creates interesting, perhaps inappropriate, licensing analogs, but it is worthy of further exploration. While recognizing the limitation of Big Iron, other production models could also be worth examining. For example, Morton and Pentico (1993) describe multiple levels of heuristic scheduling systems, and Chase et al. (1998) make careful distinction between manufacturing and service firms. These different classifications of “production” could be examined, much like different classes of “publication”. We also find Van de Sompel’s (2004) description of a value chain useful.

Another metaphor is one of the marketplace or bazaar. We revisit Raymond’s 1999 classic *The Cathedral and the Bazaar*. Metaphorically, considering a bazaar illustrates the need for specialist shopkeepers, mediators, or brokers, who help users understand and make effective use of the data. Indeed, bazaars evolve and thrive on the needs of customers. We note also that a marketplace is a spatial metaphor. People use other spatial metaphors as well. We often hear discussion of an information or knowledge space. Baker and Yarmey (2009) use the concept of a “sphere of influence” to differentiate types of repositories. They introduce the intriguing concept of “sociotechnical distance” created by issues of communication, representation, filtering, and transformation rather than physical distance.

Let us not be afraid to explore, mix, and match these and other metaphors. Let us preserve data in formal, curated *archives*. Let us make data available in rapid, cyclical, carefully versioned and described *releases*. Let us *track* data as it moves through the *marketplace* or *ecosystem*. Let us use many narratives to describe and understand complex processes. Metaphors are prevalent and powerful across the research enterprise. They can help us see new aspects of a problem, but they also create frames of thinking that can limit our perspective and perceived choices. We suggest that, at the present state of evolution toward data as a first class citizen, it is important not to be hidebound by the idea of data publication or any *one* metaphor. We need to disaggregate the roles of data stewardship and reassemble them in new ways. We must be open-minded and consider many metaphors, paradigms, and ways of knowing to fully address the data science challenges of the 21<sup>st</sup> Century. As such, in the next section we put forth our view of a representative (but not comprehensive) research agenda

intended to stimulate further discussion, application, and critical appraisal of current and future worldviews to making data preserved and widely available.

## 6 RESEARCH AGENDA

As mentioned, we seek to foster an ongoing conversation in the data science community. We, the authors, are but dilettantes in cognitive and social science. We are not theorists; we are practicing and teaching data scientists. But we believe data science can learn by examining how other disciplines and theory can inform practical data management approaches. We, therefore, end this essay with a proposed agenda for research and development. We suggest that there are important lessons to be learned from a closer examination of data science by practicing data scientists themselves. Science and Technology Studies (STS) based approaches, rooted in the principles of “Science in Action” (Latour, 1987), have shown to be very useful in understanding actually how science and informatics are actually conducted and how data are handled and perceived (e.g., Baker, and Millerand, 2010; Bowker, and Star, 2000; Harvey and Chrisman, 1998; Parsons et al., 2011; Star and Ruhleder, 1996). We need more STS-based examinations of data science practice that considers sociotechnical, *and cognitive* processes and examines the particular attitudes and perceptions of data stewardship and informatics that emerge from different domain and data science worldviews and ways of knowing. More importantly, we need to use that critical examination to develop creative *solutions* to the challenges of data science and stewardship. Broad, critical, multi-faceted analyses of “data science in action” can reveal potential new sociotechnical solutions to data science challenges. And we can use this analytical framework to examine or test how different solutions are understood, adopted, and adapted by different communities.

As we examine different worldviews, we need a fuller development and understanding of all the roles in the entire data stewardship enterprise. Lawrence et al. (2011) lay out a series of defined roles from a Data Publication perspective. Baker and Bowker (2007) do the same from an ecological infrastructuring perspective. Baker and Yarmey (2009) further examine the specific roles of data curation. Schopf (2012) has yet another examination from a production software perspective. They all emphasize different roles with different terms, and even seem to define the term “role” differently. A deeper comparison of these roles and how data managers and all the players in the enterprise perceive them is warranted. Are the different actors using the same frames and metaphors and in the same way? Is there a difference across disciplinary cultures? How do the worldviews and metaphors of data creators and data users align? Do the metaphors and frames of data scientists help or hinder that alignment? A particularly critical set of roles falls in the category of what Baker and Bowker (2007) call “in between” work. These roles of the intermediary and “middleware” connecting computer science and domain science are central to informatics and data science (Fox, 2011), yet they are also often hidden from view. Similarly, the role of a curator is critical, but as Fleischer and Jannaschk (2011) illustrate, curation can also introduce a bottleneck in data archiving and release processes. They suggest a closer examination of the role of the data manager or curator and automated curation services. In such an examination, we must consider not just the science domain but also the culture from which curators emerge. The culture of an academic library or archive is vastly different from that found in operational weather center, for example. Finally, in the examination of roles, we should use different worldviews to tease out what important roles we have missed. For example, the roles of the financial sponsor or the unintended non-specialist in the overall data ecosystem have not been examined in depth.

In addition to these broader explorations, more specific research ideas emerged from our critique and earlier feedback. Crosscutting issues emerged around data quality, data referencing, and the norms of data sharing. Data quality is an incredibly complex and subjective issue. Given its subjective nature, it seems appropriate to explore flexible means of community annotation and usage tracking as a means to better understand who are using the data for what purpose and how. This and many other aspects of good data stewardship require careful tracking through precise referencing of the data. This need for precise, continuous, dynamic referencing is closely related to but needs to be considered independently from issues of credit, discovery, and quality. For example, Bruce Caron (2011, personal communication) suggested that we consider how we might track “badges” of recognition for the many roles in the data value chain. This might help address tensions around individual vs. institutional credit and accountability; in essence re-evaluating yet another unexplored metaphor of contracts. Going forward, we see many ideas that need further exploration. We close with an initial, incomplete list of short- and longer-term research questions to be explored to identify key components of an enabling data infrastructure that would promote data availability across *many* worldviews and metaphors:

- What informatics and STS approaches can foster new and robust peer norms for science data stewardship and how may they be evaluated?
- What sociotechnical means exist for the precise, continuous, and dynamic referencing and sharing of data from creation all the way to discovery.

- How can we (or should we) evolve peer norms of data sharing to a more “commons” (Bailey and Tierney, 2002, Beagle, 1999) based approach built around ethical rather than proprietary concerns where data are viewed as a networked public good rather than an owned object?
- Can a Contract metaphor (as applied to social networks, for example) be articulated for making data widely available?
- We know that researchers from different disciplines have different attitudes toward data sharing (Key Perspectives Ltd, 2010; Parsons et al., 2011). Do they also have different attitudes when they are presented with different data management metaphors? Do they have different expectations about reuse, credit, user-responsibility, etc. when they “release” data rather than “publish” data?
- How can we identify and track data and related contextual information immediately upon creation? For example, some have suggested the concept of a “Dropbox<sup>16</sup> for the field scientist”, where data and related information are deposited as they are created and are immediately available to curators and potentially other researchers.
- What value-added steps in the data life cycle need to be explicitly credited? How? What approaches (e.g. capability or maturity models) are applicable to determine when computer and information science innovations are ready for data science communities?
- What approaches are needed to bridge the needed domain, data, and computer science disciplines into cohesive collaborations when needed?
- With increasing data intensity, what approaches in the data life cycle need to scale (in numbers of data sets, across disciplines, etc.)?
- What form of improved preservation for large-scale systems are available or need to be developed?
- How can research collections be discovered beyond the context of the scholarly article?
- What are the essential elements of data quality? What standards and technical means are available to capture these elements? How are these balanced or augmented by annotations, recommendations, or qualified citations?
- Should we reexamine our base metaphor of data as a first-class object? Is it sufficient that data simply be accessible, preserved, and usable? Does scientific rigor really require that we give data such formal, independent attention?

It is time for the all stakeholders in the Data Ecosystem (yes, our metaphor) to step outside their comfort zone, examine their worldview, clarify and share it with others, listen to alternate approaches and views, and integrate, assimilate, and evolve. The ideas in this essay must only be a beginning. We hope we have provoked a range of responses and a few ideas from the reader. We look forward to continued discussion, research, *and* action. More metaphors, please.

## 7 ACKNOWLEDGEMENTS

We sincerely thank the dozens of reviewers who have guided us formally and informally over the past year. In particular, the critiques of Bruce Barkstrom, David Carlson, Bryan Lawrence, Chris Rusbridge, and Lynn Yarmey made this a much better essay.

## 8 REFERENCES

ARL Joint Task Force on Library Support for E-Science (2007) *Agenda for Developing E-Science in Research Libraries*. Retrieved February 27, 2011 from the World Wide Web: <http://www.arl.org/bm~doc/ARLESciencefinal.pdf>

Arzberger P., Schroeder P., Beaulieu A., Bowker G., Casey K., Laaksonen L., Moorman D., Uhler P., & Wouters P. (2004) Science and government: An international framework to promote access to data. *Science*. 303(5665):1777-78.

Atkins D.E., Droegemeier K.K., Feldman S.I., Garcia-Molina H., Klein M.L., Messerschmitt D.G., Messina P., Ostriker J.P., & Wright M.H. (2003) *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Retrieved November 26, 2011 from the World Wide Web: <http://www.nsf.gov/od/oci/reports/toc.jsp>

Bailey R. and Tierney B., (2002) Information commons redux: Concept, evolution, and transcending the tragedy of the commons. *The Journal of Academic Librarianship* 28(5): 277–286.

Baker K.S. & Bowker G.C. (2007) Information ecology: open system environment for data, memories, and

<sup>16</sup> <http://www.dropbox.com/>

knowing. *Journal of Intelligent Information Systems*. 29(1):127-144.  
<http://dx.doi.org/10.1007/s10844-006-0035-7>

Baker K.S. & Millerand F. (2010) Infrastructuring ecology: challenges in achieving data sharing. In Parker J., Vermeulen N., & Penders B. (Eds.). *Collaboration in the New Life Sciences*. Surrey, England: Ashgate Publishing.

Baker K.S. & Yarmey L. (2009) Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation* 4(2). <http://dx.doi.org/doi:10.2218/ijdc.v4i2.90>

Ball A. & Duke M. (2012) *How to Cite Datasets and Link to Publications. DCC How-to Guides*. Digital Curation Centre. <http://www.dcc.ac.uk/resources/how-guides/cite-datasets>

Beagle D. (1999) Conceptualizing an information commons. *The Journal of Academic Librarianship* 25(2): 82–89. [http://dx.doi.org/10.1016/S0099-1333\(99\)80003-2](http://dx.doi.org/10.1016/S0099-1333(99)80003-2)

Bechhofer S., Buchan I., De Roure D., Missier P., Ainsworth J., Bhagat J., Couch P., *et al.* (2011) Why linked data is not enough for scientists. *Future Generation Computer Systems*.  
<http://dx.doi.org/10.1016/j.future.2011.08.004>

Belhajjame K., Goble C., & De Roure D. (2012) Research object management: opportunities and challenges. *Data Intensive Collaboration in Science and Engineering (DISCOSE) workshop, collocated with ACM CSCW 2012*.

Bowker G.C. & Star S.L. (2000) *Sorting Things Out: Classification and its Consequences*. Boston, MA: MIT Press.

Bowker G.C., Baker K., Millerand F., & Ribes D. (2010) Toward information infrastructure studies: Ways of knowing in a networked environment. *International Handbook of Internet Research*. Springer Science+Business Media.

Callaghan S., Hewer F., Pepler S., Hardaker P., & Gadian A. (2009) Overlay journals and data publishing in the meteorological sciences. *Ariadne*. (60).

Chase R.B., Aquilano N.J., & Jacobs F.R. (1998) *Production and Operations Management: Manufacturing and Services*. 8th edition. Boston, MA: Irwin/McGraw-Hill.

Costello M.J. (2009) Motivating online publication of data. *Bioscience*. 59(5):418-427.  
<http://dx.doi.org/10.1525/bio.2009.59.5.9>

Davenport T.H. & Prusak L. (1997) *Information Ecology: Mastering the Information and Knowledge Environment*. Oxford, UK: Oxford University Press.

Doorn P. & Tjalsma H. (2007) Introduction: archiving research data. *Archival Science*. 7(1):1-20.  
<http://dx.doi.org/10.1007/s10502-007-9054-6>

Duerr R., Downs R., Tilmes C., Barkstrom B., Lenhardt W., Glassy J., Bermudez L., & Slaughter P. (2011) On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*. 4:139-160. <http://dx.doi.org/10.1007/s12145-011-0083-6>

Edwards P.N., Jackson S. J., Bowker G.C., & Knobel C.P. (2007) *Understanding Infrastructure: Dynamics, Tensions, and Design*. National Science Foundation. Retrieved May 30, 2012 from the World Wide Web: <http://hdl.handle.net/2027.42/49353>

ESIP (Federation of Earth Science Information Partners) (2012) *Data Citation Guidelines for Data Providers and Archives*. Parsons M.A., Barkstrom B., Downs R.R., Duerr R., Tilmes C., & ESIP Preservation and Stewardship Committee (Eds.) ESIP Commons. Retrieved September 1, 2012 from the World Wide Web: <http://commons.esipfed.org/node/308>

FGDC (Federal Geographic Data Committee) (1997) *A strategy for the NSDI*. Retrieved September 3, 2012 from the World Wide Web: [http://www.fgdc.gov/policyandplanning/A%20Strategy%20for%20the%20NSDI%201997.doc/at\\_download/file](http://www.fgdc.gov/policyandplanning/A%20Strategy%20for%20the%20NSDI%201997.doc/at_download/file)

- Fillmore C.J. (1976) Frame semantics and the nature of language\*. *Annals of the New York Academy of Sciences*. 280(1):20-32. <http://dx.doi.org/10.1111/j.1749-6632.1976.tb25467.x>
- Fleischer D. & Jannaschk K. (2011) A path to filled archives. *Nature Geoscience*. 4(9):575-76. <http://dx.doi.org/10.1038/ngeo1248>
- Fox P. (2011) The rise of informatics as a research domain. *Proceedings of the Water Information Research and Development Alliance*. CSIRO e-Publication, online at: <http://www.csiro.au/WIRADA-Science-Symposium-Proceedings> pp. 125-132.
- Goffman E. (1974) *Frame Analysis: An Essay on the Organization of Experience*. New York: Harper & Row.
- Harley J.B. (1989) Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization*. 26(2):1-20. <http://dx.doi.org/10.3138/E635-7827-1757-9T53>
- Harvey F. & Chrisman N. (1998) Boundary objects and the social construction of GIS technology. *Environment and Planning A*. 30(9):1683-694.
- Hunter J. (2006) Scientific publication packages: A selective approach to the communication and archival of scientific output. *The International Journal of Digital Curation*. 1(1):33-52.
- ICSU (2011) *Interim Report of the ICSU ad-hoc Strategic Coordinating Committee on Information and Data*, at [http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report/SCCID\\_Report\\_April\\_2011.pdf](http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report/SCCID_Report_April_2011.pdf)
- ISO (2003) *ISO Standard 14721:2003, Space Data and Information Transfer Systems—A Reference Model for an Open Archival Information System (OAIS)*. International Organization for Standardization.
- Key Perspectives Ltd (2010) *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability*. Digital Curation Center. Retrieved February 5, 2011 from the World Wide Web: [http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS\\_FINAL.pdf](http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf)
- Klump J., Bertelmann R., Brase J., Diepenbroek M., Grobe H., Höck H., Lautenschlager M., Schindler U., Sens I., & Wächter J. (2006) Data publication in the open access initiative. *Data Science Journal*. 5:79-83. <http://dx.doi.org/10.2481/dsj.5.79>
- Koch T. (2004) The map as intent: Variations on the theme of John Snow. *Cartographica*. 39(4):1-13.
- Kuhn T.S. (1996) *The Structure of Scientific Revolutions. 3rd edition*. Chicago, IL :: University of Chicago Press.
- Lakoff G. (2008) *The Political Mind: Why You Can't Understand 21st-Century Politics With An 18th-Century Brain*. New York: Penguin Group.
- Lakoff G. & Johnson M. (1980) *Metaphors We Live By*. Chicago: The University of Chicago Press.
- Lakoff G. (1993) The contemporary theory of metaphor. In Ortony A. (Ed.). *Metaphor and Thought, 2nd edition* Cambridge: Cambridge University Press.
- Lakoff G. & Johnson M. (1980a) The metaphorical structure of the human conceptual system. *Cognitive Science*. 4(2):195-208. [http://dx.doi.org/10.1207/s15516709cog0402\\_4](http://dx.doi.org/10.1207/s15516709cog0402_4)
- Latour B. (1987) *Science in Action: How To Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.
- Lawrence B., Jones C., Matthews B., Pepler S., & Callaghan S. (2011) Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*. 6(2).
- McMahon M. (1996) Overview of the Planetary Data System. *Planetary and Space Science*. 44(1):3-12. [http://dx.doi.org/doi:10.1016/0032-0633\(95\)00101-8](http://dx.doi.org/doi:10.1016/0032-0633(95)00101-8)

- Morton T.E. & Pentico D.W. (1993) *Heuristic Scheduling Systems: With Applications To Production Systems And Project Management*. Wiley-Interscience.
- Nardi B.A. & O'Day V. (2000) *Information Ecologies: Using Technology With Heart*. Boston, MA: MIT press.
- National Research Council (2007) *Environmental Data Management at NOAA: Archiving, Stewardship, and Access*. Washington, DC: National Academies Press.
- NRC (National Research Council) (1993) *Toward a Coordinated Spatial Data Infrastructure for the Nation*. Washington, DC: National Academies Press.
- Parsons M.A., Duerr R., & Minster J.B. (2010) Data citation and peer-review. *Eos, Transactions of the American Geophysical Union*. 91(34):297-98. <http://dx.doi.org/doi:10.1029/2010EO340001>
- Parsons M.A., Godøy Ø., LeDrew E., de Bruin T.F., Danis B., Tomlinson S., & Carlson D. (2011) A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science*. 37(6):555-569. <http://dx.doi.org/10.1177/0165551511412705>
- Pfeiffenberger H. & Carlson D. (2011) "Earth System Science Data" (ESSD) — A peer reviewed journal for publication of data. *D-Lib Magazine*. 17. <http://dx.doi.org/10.1045/january2011-pfeiffenberger>
- Priem J. & Hemminger B.M. (2012) Decoupling the scholarly journal. *Frontiers in Computational Neuroscience*. 6(19). <http://dx.doi.org/10.3389/fncom.2012.00019>
- Raymond E.S. (1999) *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Cambridge, MA: O'Reilly.
- Schopf J.M. (2012) Treating data like software: A case for production quality data. *Proceedings of the Joint Conference on Digital Libraries*. 11-14 June 2012, Washington DC.
- de Sherbinin A. & Chen R.S. (Eds). (2005) *Global Spatial Data and Information User Workshop: Report of a Workshop*. Socioeconomic Data and Applications Center, Center for International Earth Science Information Network, Columbia University. Retrieved February 5, 2011 from the World Wide Web: <http://sedac.ciesin.columbia.edu/GSDworkshop/>
- Star S.L. & Ruhleder K. (1996) Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*. 7(1):111.
- de Waard A. & Kircz J. (2008) Modeling scientific research articles--shifting perspectives and persistent issues. *Proc. ELPUB2008 Conference on Electronic Publishing*.
- de Waard A., Breure L., Kircz J.G., & Van Oostendorp H. (2006) Modeling rhetoric in scientific publications. *International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*.
- Van de Sompel H., Payette S., Erickson J., Lagoze C., & Warner S. (2004) Rethinking scholarly communication. *D-Lib Magazine*. 10(9):1082-9873.
- Wright A (2009) Exploring a 'Deep Web' that Google can't grasp. *The New York Times*. February 22, 2009. <http://www.nytimes.com/2009/02/23/technology/internet/23search.html>

# CONNECTING SCIENTIFIC ARTICLES WITH RESEARCH DATA: NEW DIRECTIONS IN ONLINE SCHOLARLY PUBLISHING

*IJsbrand Jan Aalbersberg\**, Judson Dunham, and Hylke Koers

*Elsevier, Radarweg 29, 1043 NX Amsterdam, The Netherlands  
Email: I.J.Aalbersberg@elsevier.com*

## ABSTRACT

*Researchers across disciplines are increasingly utilizing electronic tools to collect, analyze, and organize data. However, when it comes to publishing their work, there are no common, well-established standards on how to make that data available to other researchers. Consequently data is often not stored in a consistent manner, making it hard or impossible to find data sets associated with an article – even though such data might be essential to reproduce results or to perform further analysis. Data repositories can play an important role in improving this situation, offering increased visibility, domain-specific coordination, and expert knowledge on data management. As a leading STM publisher, Elsevier is actively pursuing opportunities to establish links between the online scholarly article and data repositories. This helps to increase usage and visibility for both articles and data sets, and also adds valuable context to the data. These data-linking efforts tie in with other initiatives at Elsevier to enhance the online article in order to connect with current researchers' workflows and to provide an optimal platform for the communication of science in the digital era.*

**Keywords:** STM publishing, Data repositories, Data-linking, Innovation

## 1 INTRODUCTION

Driven by technological advancements enabling storage and sharing of large volumes of data, experimental data sets (possibly very large) have become an essential part of scientific research. Data in science is ubiquitous, with notable examples spanning a wide range of research disciplines from the Human Genome Project, to Earth Observations, to the Large Hadron Collider. Science, by its nature, produces a lot of data - and this is increasingly true now that many sensory data are born-digital and barriers to storing and processing this data are low.

Making research data available to other scholars provides an impetus to the advancement of science. First of all, it enables others to reproduce a scientific result to assure themselves of its validity – one of the cornerstones of the scientific method, yet often impossible to follow through when parts of the input data or computational methodology remain a black box. Secondly, data may be re-used for other purposes, sometimes unforeseen when the data was gathered. Re-use drives research efficiency by preventing duplication of work, but also opens the door to analyses that were not otherwise possible – in particular when different data sets are combined into a meta-analysis of sorts.

Despite the availability of basic enabling technologies, the potential for sharing research data is far from being fulfilled today. A recent study by PARSE.Insight [1] shows how researchers share their data in many ways: by email, on their university website, as supplementary material to a journal article, in an institutional repository, etc. This makes it hard for others to find data sets and, by the absence of clear and consistent metadata, to interpret them correctly. In addition, researchers are sometimes reluctant to share their data sets. There are several reasons for this: the additional work, worry about incorrect usage, the desire to “monetize” the value in data that was collected by hard work through a series of journal articles, or simply unawareness of existing possibilities.

What appears to be lacking to fully benefit from accessible research data is often organizational in nature: incentives to share data (in the form of academic credits, recognition, or otherwise), and common standards and processes that are widely accepted and consistently followed. Many stakeholders need to align to make that happen, and such a movement appears to be happening at present. An increasing number of funding agencies require their researchers to share data and/or have a data management plan. Domain-specific data repositories are taking up a role as centralized places to deposit and access research data. Organizations like DataCite [2] and the

ICSU World Data System [3] help to create overarching policies, views, or infrastructure to facilitate some of the work for individual data repositories, such as creating persistent identifiers, increasing discoverability, and establishing authority. Last, but not least, publishers are actively establishing connections between the scholarly record and data sets. This helps increase visibility and usage of data sets, integrates data sets into the existing researcher workflow, and provides accurate context to data sets – thereby addressing some of the issues that researchers face with sharing and re-using data. It is worth noting in this context that the international STM publishing community has issued a statement in 2007 to outline their view that “raw research data should be made freely available to all researchers” [4].

Scientific data repositories are numerous and diverse in character. A recent survey identified over a thousand scientific data repositories in Life Sciences alone [5]. The character of these repositories depends on the field and the intended audience - some data repositories just provide researchers a “safe harbor” to store their data sets for perpetuity, others actively curate the literature and organize data into authoritative information resources.

As a globally leading STM publisher, Elsevier has taken a prominent role in establishing (reciprocal) connections between the scholarly article and scientific data repositories. Such connections can take various forms, from clickable hyperlinks in the article text to interactive applications integrated into SciVerse® ScienceDirect® that pull data on-the-fly from a data resource on the web. What they have in common is the goal to present the reader with relevant, trustworthy data and information in the context of a research publication, to provide context to data sets, and to make it easier for researchers to find publications and data sets relevant to their work.

In the remainder of this article we will discuss data-linking initiatives at Elsevier. We will also describe the enabling technologies that Elsevier has invested in to allow for agile and collaborative development of data linking tools, and touch upon the vision underlying these efforts.

## **2 ELSEVIER’S ARTICLE OF THE FUTURE**

The electronic age has brought profound changes to the way scientific research is conducted and captured. Researchers increasingly use electronic tools to perform measurements, and to analyze, organize, and share their material. Consequently research output is increasingly diverse and “rich” in an electronic sense – including data sets, video and other multimedia files, computer code, etc. However, when it comes to publishing, the scientific article – as a vehicle to communicate that research - has shown little adaptation and a scientist often finds herself reducing valuable scientific output to “ink on paper” – which the reader then has to reconstruct to take full advantage of the insights the author wanted to share (many a researcher will recognize the frustration of having to re-key a data table, or use a ruler to determine the location of a peak on a data plot.)

Improvements to the scholarly article over the last few decades have been mostly in terms of delivery (electronically), discoverability (full-text search), as well as a number of smaller-scale, specific enhancements such as the possibility to upload supplementary data. However, in terms of structure and shape, the current article is by and large the same as in the first scholarly journals of the 17<sup>th</sup> century. In order to address this growing mismatch between the frozen “article” concept, and the evolving workflows and needs of researchers, Elsevier has initiated the “Article of the Future” – a project to rethink the scientific article in the electronic age.

The Article of the Future project aims to offer an optimal platform to communicate science in today’s digital world. From this starting point, the concept has been developed in close collaboration with the scientific community, involving feedback from several hundreds of researchers. Very early on in the development, it was recognized that the greatest additional value lies in domain-specific enhancements. Different scientific communities use different tools, are used to different data standards, and might have different attitudes towards communicating research output – so one clearly has to go beyond “one size fits all” to really connect with, and support, the workflow of individual researchers. The Article of the Future improves the online article in essentially three directions:

- Presentation – offering an optimal online browsing and reading experience, which is a basic requirement for online reading and for any further enhancements.
- Content – supporting a richer pallet of author-delivered material, including multimedia files, scientific data and computer code.



- Context – connecting the online article to trustworthy scientific resources to present the reader with relevant information in the context of the article.

A perhaps naïve metaphor for the enhanced contextualization of articles is to think about the article as a roundabout rather than the traditional one-way street. Establishing connections between the article and data repositories perfectly fits into this vision, making the Article of the Future a natural format for seamless integration of data-linking tools.

Initially introduced for Cell Press journals in 2009, the Article of the Future concept was expanded to other disciplines from 2010 onwards. After an initial series of prototype articles on [www.articleofthefuture.com](http://www.articleofthefuture.com), the first phase of this concept has been implemented across Elsevier's publication platform SciVerse ScienceDirect at the time of writing (January 2012). Further enhancements are expected later in 2012.

### 3 LINKING SCIVERSE SCIENCEDIRECT AND DATA REPOSITORIES

SciVerse ScienceDirect supports a number of methods to link with data repositories, including author-based tagging schemes, automatically generated data-linking banners (see also [6]) and data-linking applications. The most appropriate choice depends on the nature of the data and the database, and on how the information is best presented to the reader.

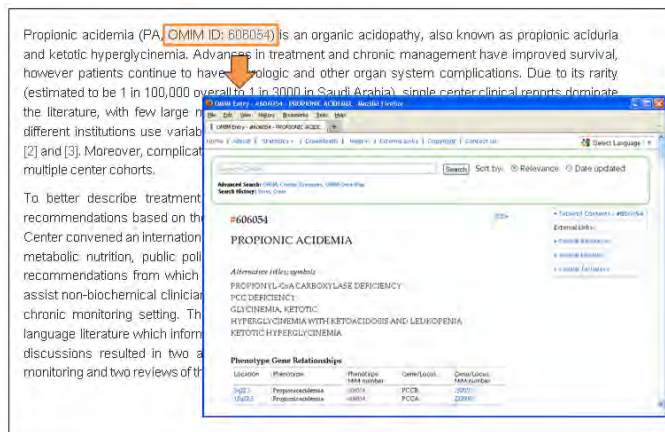
The first method that Elsevier employs to link out to a data repository is by asking authors to explicitly tag entities for which data is available at a repository, for example "OMIM ID: 606054" to refer to a specific data record at the Online Mendelian Inheritance in Man (OMIM) data repository. This tag will be recognized during the publication production process, and show up as a hyperlink in the online article, pointing to the relevant data record.

The second possibility relies on an automatic, on-the-fly linking banner service that has been developed for this purpose. Here a selected database is queried when a reader opens up an online article on SciVerse ScienceDirect. If the data repository recognizes the article DOI, a banner image is returned to indicate that the repository holds relevant data records for this article. Clicking on the banner then directs the reader to those records. A key benefit of this scheme is that it allows for data to be connected to the article after publication (possibly even years thereafter). This is particularly useful to help make past data available using current technology, and also accommodates authors who prefer to make their data available only after an initial embargo or delay.

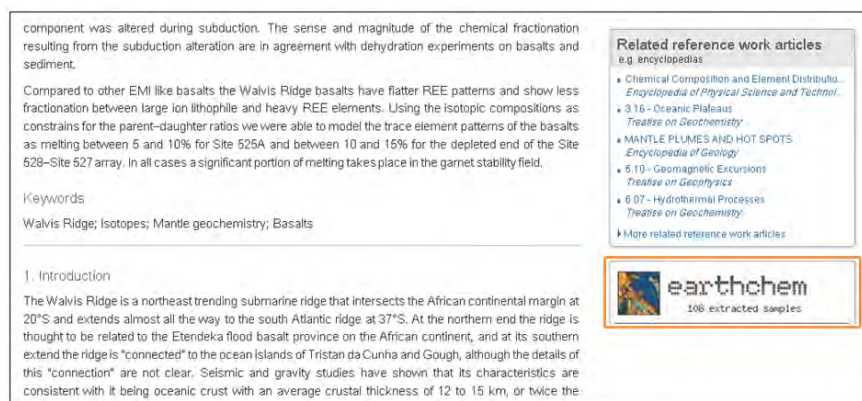
A third, broad category of data-linking methods lies in utilizing Elsevier's SciVerse Application Framework [7] to create dedicated applications that connect with online data repositories and display relevant data and information alongside the online article for interactive exploration. This framework – which will be discussed in detail in the next section - opens the door to the wide range of possibilities for interactive exploration of data in the context of the online article.

At the time of writing, SciVerse ScienceDirect features over 20 linking schemes and data-linking applications to connect articles to scientific data repositories. Covering the wide gamut of subject areas, linked data repositories include the Protein Data Bank [8], Encyclopedia of Life [9], Cambridge Crystallographic Data Center [10], EarthChem [11], PANGAEA [12], the SIMBAD Astronomical Database [13], ClinicalTrials.gov, Data.gov – and many others.<sup>1</sup>

<sup>1</sup> For an up-to-date overview of author-tagged linking schemes and SciVerse (data-linking) applications, please visit <http://www.elsevier.com/databaselinking> and <http://www.applications.sciverse.com/action/gallery>, respectively.



**Figure 1.** Screenshot of an online article page on SciVerse ScienceDirect showing an author-tagged link to the OMIM data repository (see <http://dx.doi.org/10.1016/j.ymgme.2011.08.007>). The inset shows the landing page for the data record at OMIM. Here, as in the other figures, the orange boxes and arrows (pointing from a link to its target) are for illustration purposes only; these are not displayed on SciVerse ScienceDirect.



**Figure 2.** Screenshot of an EarthChem linking banner as displayed next to an online article on SciVerse ScienceDirect (see <http://dx.doi.org/10.1016/j.chemgeo.2010.02.010>).

## 4 THE SCIVERSE APPLICATION FRAMEWORK AND DATA-LINKING APPLICATIONS

### 4.1 The SciVerse Application Framework as an Enabling Technology

The SciVerse Application Framework is a collection of technologies that, amongst others, enable dynamic integration of data and software from third-party sources into Elsevier’s publishing and search platforms SciVerse ScienceDirect, SciVerse Scopus®, and SciVerse Hub. The component systems to support these integrations provide services ranging from platform integration, search and retrieval APIs<sup>2</sup> and content syndication, to advanced services like entity extraction and text mining.

The core component of the SciVerse Application Framework is the platform integration infrastructure, or Framework API, which allows for third-party data and tools to integrate dynamically into the SciVerse user interface. To accomplish this, SciVerse uses a standards-based, open source gadget framework: Apache Shindig, an OpenSocial container which allows for applications, or “gadgets”, to be integrated into container web applications. Elsevier has extended its implementation of this software in a few key ways to offer a range of functionality especially suited for linking with data sets:

<sup>2</sup> APIs, short for Application Programming Interfaces, are machine-readable interfaces that allow for programmatic access to content and services.

- Platform and content integration – applications running within SciVerse ScienceDirect have access to the full-text scientific article, the document metadata, and user input such as search terms.
- User interface extensions – applications have access to a wide array of functions to enable rich user experiences and integration into the surrounding user interface, such as the ability to create links within documents and load applications dynamically within the page.
- Interaction with external resources – applications can access information resources on the web to combine that with Elsevier content into a form most useful to readers.

A cornerstone of the SciVerse Application Framework is the set of SciVerse Content APIs, which allow programmatic access to a broad range of scholarly content and bibliometrics within SciVerse ScienceDirect and SciVerse Scopus. The APIs allow for real time query and retrieval of search results, abstracts, full-text articles, document metadata, author and affiliation data and citation metrics. Created for easy re-use to facilitate new application development, these tools are now frequently used in text analysis and matching operations to identify and link data references within articles.

In cases where on-the-fly analysis tools are too slow, and pre-processing is required, Elsevier also provides a Content Syndication service to deliver large amounts of full-text articles and books to data linking partners and application developers. This service allows for rapid delivery of configurable collections of full-text content in bulk, either downloadable via FTP or delivered on hard disk. New content can be updated via FTP, with configurable delivery schedules that can provide regular content updates as often as hourly, to enable partners to access newly published full-text content with minimal delays.

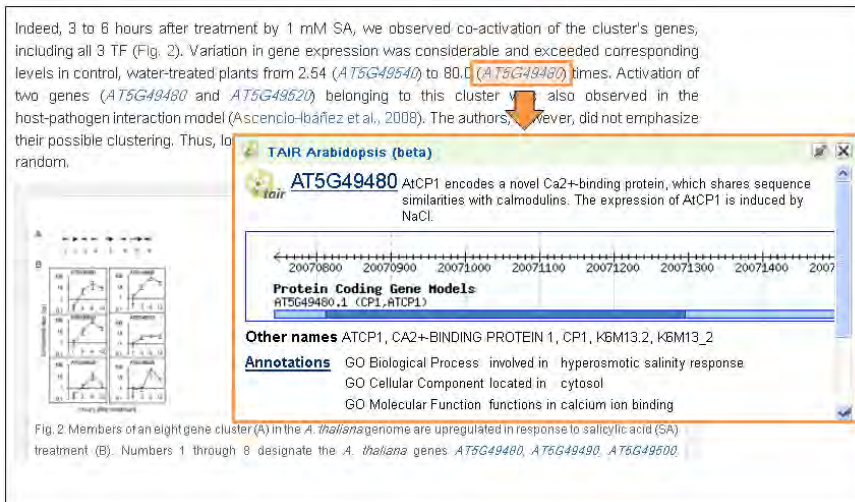
Beyond these basic components, Elsevier is also expanding its services into more advanced areas by creating “Developer Services”. These are meant to enable rapid development and deployment of new data-linking opportunities as they arise. These building blocks also help to let potential partners focus on the scientific implementation of a data application, rather than on the technical details. As an example of this, a prototype entity extraction service was recently made available. This service can be configured with a lexicon containing phrases associated with the entities to be identified. It can then be passed an arbitrary chunk of text in which it identifies entities from the associated lexicon. The response from this service contains the list of entities matched in the text, the number of occurrences, and the location of those occurrences. In this way, data linking applications can be created with very little effort, often requiring little more than the transfer of a lexicon of unique terms or identifiers as a basis.

## 4.2 Some Examples of Data-Linking Applications on SciVerse ScienceDirect

### LIPID Structures & TAIR Arabidopsis

Elsevier recently collaborated with the community-organized LIPID MAPS [14] and TAIR [15] (The Arabidopsis Information Resource) data repositories to develop two applications for researchers in biology. Both applications operate in a similar manner: they identify terms within articles on SciVerse ScienceDirect, pull key information into the article, and link to the associated data record pages – thereby helping researchers to quickly access valuable reference information on Arabidopsis loci (TAIR) or lipids (LIPID Structures).

These applications provide examples of the collaborative development discussed in the previous section, allowing for rapid development and deployment. They use a centralized, configurable text-mining service created by Elsevier. This enables data repositories to focus on their key strengths such as localized collection, aggregation, enhancement and storage of domain-specific data, and selection of the most relevant information for online readers.



**Figure 3.** Screenshot showing a locus recognized by the TAIR application, and the information panel after being opened by the reader (see (<http://dx.doi.org/10.1016/j.gene.2011.09.023>)).

### The Genome Viewer

The Genome Viewer, developed in collaboration with NCBI, combines author-tagged entities with SciVerse's capabilities to interact with external resources and to add a layer of interactivity that lets readers interactively explore data rather than having to digest a long list with information. The application recognizes NCBI Accession Numbers for genetic sequences, collects sequence data from GenBank [16], and collects this in an interactive sequence viewer. Users can easily find locations of specific interest, change the visualization, or download sequence data from within the application.



**Figure 4.** Screenshot showing genetic sequence data visualized by the Genome Viewer (see <http://dx.doi.org/10.1016/j.gene.2011.01.002>).

### PANGAEA

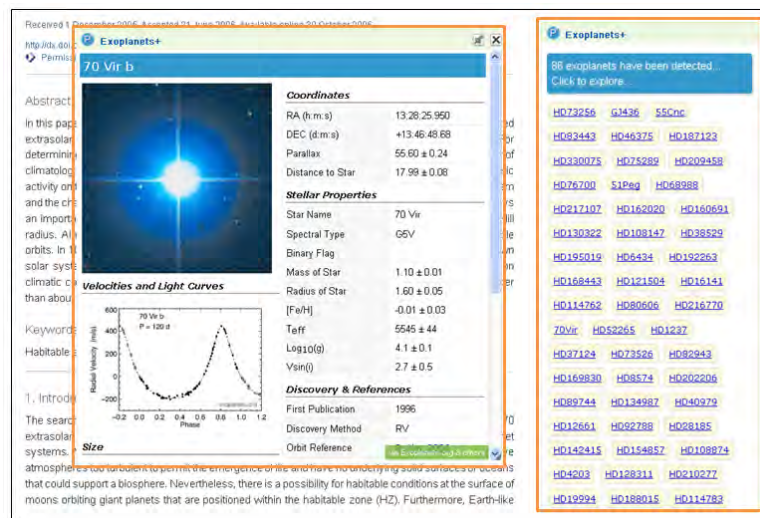
Elsevier and PANGAEA have built an advanced linking service between SciVerse ScienceDirect and the PANGAEA data repository for Earth Sciences research. Authors who submit a paper to a participating journal are encouraged to submit their raw data sets to PANGAEA, where they are archived and assigned a unique, persistent identifier. When the paper is published online, the reader will see an interactive map application that visualizes the geographical locations of the data sets at PANGAEA and offers links to the data records.



**Figure 5.** Screenshot showing the PANGAEA map viewer with the geographical locations of data records at PANGAEA (see [http://dx.doi.org/10.1016/S0031-0182\(01\)00461-8](http://dx.doi.org/10.1016/S0031-0182(01)00461-8)).

## Exoplanets+

The Exoplanets+ Application was one of the contestants for Elsevier’s “Apps for Science” contest. The application searches through full-text articles for exoplanet (extra-solar planet) names that it recognizes from several astronomical data repositories including Exoplanets.org and SIMBAD [13]. If exoplanets are found, the application opens up in the right-hand pane to alert the reader that additional information is available. Clicking on an exoplanet name opens a panel with a compilation of key information, and links to the underlying data repositories.



**Figure 6.** Screenshot showing a list of exoplanet names recognized by the Exoplanet+ application, and the information panel after being opened by the reader (see <http://dx.doi.org/10.1016/j.pss.2006.06.022>).

## 5 CONCLUSION

In a time where scientists are increasingly utilizing electronic tools for their research tasks - think only of the extensive use of automated data collection and processing pipelines, or the ubiquity of software tools to analyze and share research material - new or improved ways to disseminate scientific output are emerging. In particular, several stakeholders are actively encouraging the sharing of research data through scientific data repositories – so as to benefit from increased visibility, domain-specific coordination, and expert knowledge on data management.

Recognizing the need to support changing workflows and researcher needs (in both the author and reader role) Elsevier has taken several initiatives to bring additional value to its online articles on SciVerse ScienceDirect. These include the Article of the Future project, focusing on improved presentation, content and context, and the SciVerse Application Framework, which enables Elsevier and partner organizations to develop specific, interactive tools that the reader can access from within the context of the online article.

Building on these foundations, Elsevier is actively establishing connections between online articles and scientific data repositories. This improves the user experience for readers of SciVerse ScienceDirect by providing simple, one-click access to trustworthy and relevant data. At the same time, this program improves visibility and usage of data repositories, and places data sets in perspective: journal articles often contain essential information about data sets – how were they accumulated, what are their limitations, which conclusions have been drawn from them, etc. – that is essential for correct interpretation and consistent re-use. In this manner, connections between the scientific articles and data repositories add value on both sides, and contribute to a better infrastructure for the dissemination of science in the electronic age.

Elsevier is keen on further expanding its range of data-linking schemes and applications, and welcomes collaboration with interested parties.

## 6 REFERENCES

- [1] <http://www.parse-insight.eu/>
- [2] <http://www.datacite.org>
- [3] <http://www.icsu-wds.org/>
- [4] [http://www.stm-assoc.org/2007\\_11\\_01\\_Brussels\\_Declaration.pdf](http://www.stm-assoc.org/2007_11_01_Brussels_Declaration.pdf)
- [5] Laura Haak Marcial, Bradley M. Hemminger (2010): Scientific Data Repositories on the Web: An Initial Survey. JASIST 61 (10) 2029-2048. doi:10.1002/asi.21339
- [6] IJsbrand Jan Aalbersberg, Ove Kähler (2011): Supporting Science through the Interoperability of Data and Articles. D-Lib Magazine 17 (1/2). doi:10.1045/january2011-aalbersberg
- [7] <http://www.applications.sciverse.com>
- [8] <http://www.rcsb.org>
- [9] <http://www.eol.org>
- [10] <http://www.ccdc.cam.ac.uk>
- [11] <http://www.earthchem.org>
- [12] <http://www.pangaea.de>
- [13] <http://simbad.u-strasbg.fr/simbad>
- [14] <http://www.lipidmaps.org>
- [15] <http://www.arabidopsis.org>
- [16] <http://www.ncbi.nlm.nih.gov/genbank>

# RE-EVALUATION OF GEOMAGNETIC FIELD OBSERVATION DATA AT SYOWA STATION, ANTARCTICA

*K Takahashi<sup>1\*</sup>, Y Minamoto<sup>1</sup>, S Arita<sup>1</sup>, I. Tomofumi<sup>1</sup> and A Kadokura<sup>2</sup>*

<sup>\*1</sup>*Kakioka Magnetic Observatory, Japan Meteorological Agency, Kakioka 595, Ishioka, Ibaraki 315-0116, Japan  
Email: [takahashi\\_kosuke@met.kishou.go.jp](mailto:takahashi_kosuke@met.kishou.go.jp)*

<sup>2</sup>*National Institute of Polar Research, 10-3 Midoricho, Tachikawa, Tokyo 190-8518, Japan  
Email: [kadokura@nipr.ac.jp](mailto:kadokura@nipr.ac.jp)*

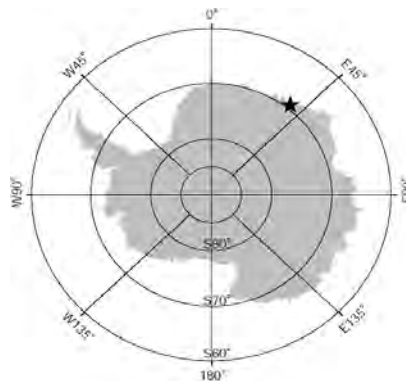
## ABSTRACT

*The Japanese Antarctic Research Expedition has conducted geomagnetic observations at Syowa Station, Antarctica, since 1966. Geomagnetic variation data measured with a fluxgate magnetometer are not absolute but are relative to a baseline and show drift. To enhance the importance of the geomagnetic data at Syowa station, therefore, it is necessary to correct the continuous variation data by using absolute baseline values acquired by a magnetic theodolite and proton magnetometer. However, the database of baseline values contains outliers. We detected outliers in the database and then converted the geomagnetic variation data to absolute values by using the reliable baseline values.*

**Keywords:** Absolute geomagnetic observations, Geomagnetic variation measurement, Baseline value, Syowa Station, Antarctica

## 1 INTRODUCTION

Since 1966, the Japanese Antarctic Research Expedition (JARE) has conducted absolute geomagnetic observations and geomagnetic variation measurements at Syowa Station, Antarctica (N69.006°, E39.590°; Figure 1). The absolute geomagnetic observations have been carried out basically at once per month during geomagnetically quiet period with a magnetic theodolite and proton magnetometer to obtain absolute baseline values. The geomagnetic variation measurements are performed with a three-axis fluxgate magnetometer to obtain continuous readings of variations relative to the baseline. The results are publicly available on the website of the National Institute of Polar Research, Japan (<http://polaris.nipr.ac.jp/~aurora/syowa.magne/magne.main.html>).



**Figure 1.** Location of Syowa Station (star).

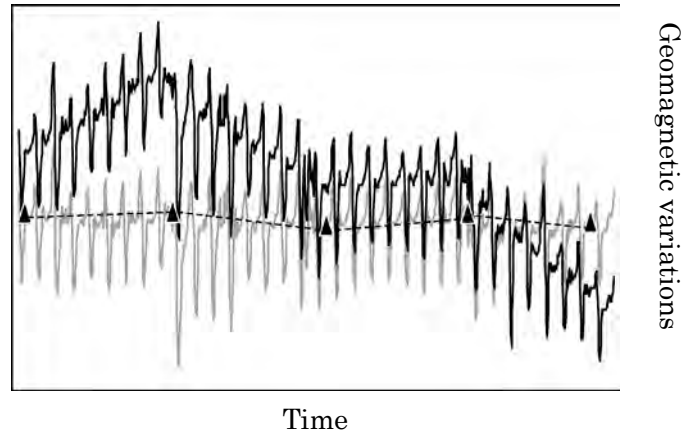
The continuous geomagnetic variation data acquired with the fluxgate magnetometer in general show a drift caused by changes in the sensor temperature and/or tilt.

To convert the continuous geomagnetic variation record to absolute values, it is necessary to correct such drift. Therefore, we use the following correction procedure (Figure 2):

(1) Calculate each baseline value ( $I_B$ ) at each time of the absolute observations ( $t_A$ ) from each absolute baseline value ( $I_A$ ) and the continuous variation data ( $I_C$ ) at  $t_A$  as  $I_B = I_A - I_C$ .

- (2) Calculate the baseline values during other periods by linear interpolation between the successive baseline values at the times of the absolute observations.
- (3) Add the interpolated baseline values to the continuous variation data to obtain absolute variation data.

However, the database of baseline values contains outliers. Therefore, we developed a statistical procedure for objective detection of outliers in the baseline values.



**Figure 2.** Schematic diagram of the correction method. Triangles denote absolute geomagnetic baseline values. Solid and grey lines denote observed and corrected continuous geomagnetic variation data, respectively.

## 2 GEOMAGNETIC OBSERVATIONS AT SYOWA STATION

The three-axis fluxgate magnetometer used by JARE at Syowa Station since 1966 measures three components of geomagnetic variation: the components parallel and perpendicular to the geomagnetic meridian and the vertical component. The variation data are sampled every 1, 2 or 10 s, and the resolving power is 0.1 nT. During the same period, JARE has made absolute geomagnetic observations to investigate the secular variation of declination, inclination, and geomagnetic intensity. These observations are obtained approximately monthly with a magnetic theodolite and proton magnetometer. A search coil magnetometer, which is the G.S.I. (Geographical Survey Institute of Japan) type magnetometer had been used from Mar., 1966 to May, 1997, and after that, a fluxgate declinometer/inclinometer has been used (Ookawa, 1999). The G.S.I. type magnetometer consists of a rotating search coil and Helmholtz coil which are mounted on a steel-free theodolite. Its minimum detectable angle is 0.2 minutes. The fluxgate declinometer/inclinometer consists of a single-axis magnetometer with a fluxgate probe mounted on a steel-free theodolite. Its observation accuracy is 0.1 nT. The observation accuracy of the proton magnetometer is better than 0.1 nT and the accuracy of the absolute observation depends not only on the accuracy of the instruments but also on the individual observer's skill at the operation. In each absolute observation session, four observed baseline values are averaged to determine the baseline value for that session. Baseline values for the horizontal (H) and vertical (Z) components and the declination (D) of the continuous variation data are calculated by using the absolute observations. Figure 3 shows the calculated baseline values from 1997 to 2011. It can be seen in Fig. 3 that all components of the baseline values measured during some absolute observation sessions had a large variance. Then a statistical approach is applied to detect outliers in the observed baseline values.

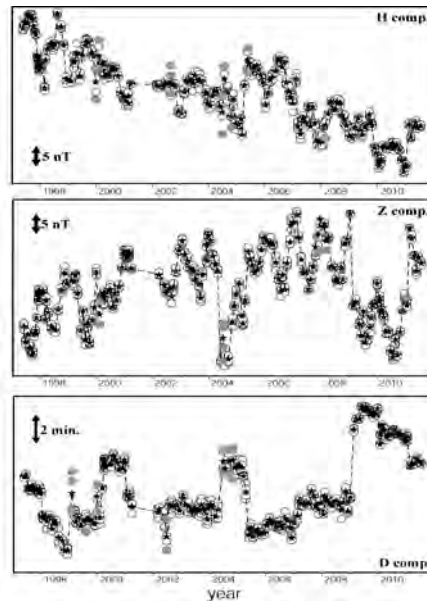
## 3 DETECTING OUTLIERS IN THE DATABASE OF BASELINE VALUES

For objective detection of outliers in a series of observed baseline values, Ito & Fujii (2003) proposed a robust estimation procedure that uses the median and median absolute deviation of the observed data. However, it is difficult to apply this robust estimation procedure to our database because only four baseline values were obtained per observation session.

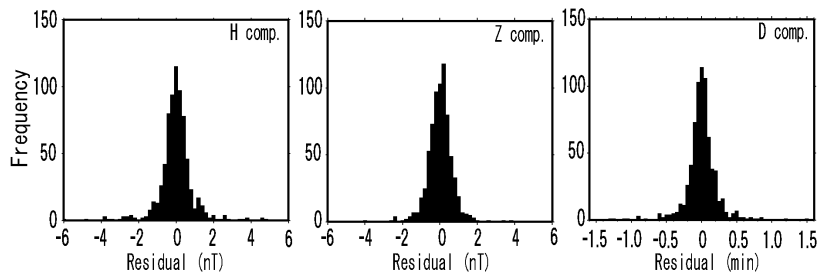
We therefore investigated the distribution of residuals (i.e. differences between the mean and each of the four observed baseline values) for all observations from 1997 to 2011. During this period, 720 observed baseline values were obtained for each component. The frequency distribution of the residuals of each component is bell-shaped, with only a few large-amplitude samples (Figure 4), which suggests that we can assume a



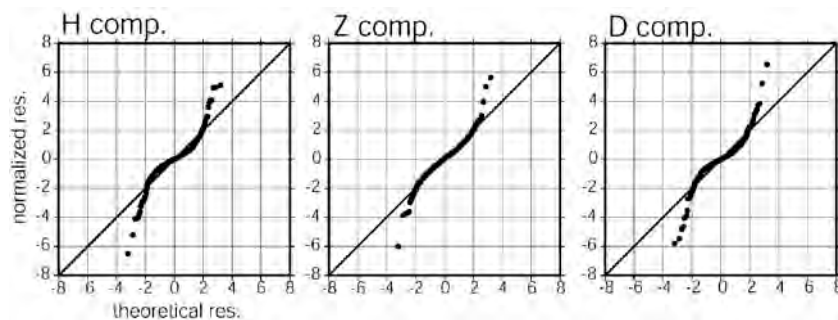
Gaussian-type distribution. In addition, we examined the distribution of the residuals of each component in a normal quantile–quantile plot (Figure 5), in which residuals normalized by the standard deviation of the residuals on the vertical axis are compared to the theoretical residuals predicted given a standard Gaussian distribution. For all components, the normalized residuals with a normalized residual size of about 3 or larger deviate from the 45° reference line. Therefore, we regarded the observed baseline values with normalized residuals in excess of 3 as outliers, because these residuals are larger than would be expected given a Gaussian distribution. This is equivalent to determining as outliers residual amplitudes of the H-, Z-, and D-components in excess of 2.8 nT, 2.0 nT, and 0.7 minutes, respectively. In this way, we identified a total of 42 residuals as outliers: 16 in the H-component, 10 in the Z-component, and 16 in the D-component (Figure 3).



**Figure 3.** Temporal variation of baseline values of the horizontal (top) and vertical (middle) components and the declination (bottom) of geomagnetic observations at Syowa Station from 1997 to 2011. Open circles and stars denote observed and mean values, respectively. Grey circles indicate outliers detected as described in Section 3.



**Figure 4.** Frequency distribution of the residuals of the horizontal (left) and vertical (centre) components in addition to the declination (right) from 1997 to 2011.



**Figure 5.** Normal quantile–quantile plots of the horizontal (left) and vertical (centre) components in addition to the declination (right). The normalized residuals of the observed baseline values are shown on the vertical axes, and the theoretical residuals under the assumption of a normal Gaussian distribution are shown on the horizontal axes.

## **4 CAUSES OF OUTLIERS**

A possible cause of outliers in the absolute geomagnetic observation is magnetic disturbances. A difference in total intensity of geomagnetic field between an absolute observation site and a remote site of proton sensor could be so large during significant magnetic storms (Jankowski & Sucksdorff, 1996). Therefore, it is not suitable for absolute observations under a severe geomagnetic disturbance. To investigate the intensities of geomagnetic disturbance during the observations with outliers, we estimated each variance of the data of geomagnetic intensity obtained during these observations. The variances of the 12 data are less than  $30 \text{ nT}^2$  among the 14 data. This fact suggests that most of the absolute geomagnetic observations with outliers were executed under an inactive geomagnetic field. We also checked the field notes of the absolute observations with outliers. Consequently, we found most of the field notes contain abnormal observation values. Hence, we concluded that the majority of causes for being outliers are artificial magnetic disturbances or mistakes in operations.

## **5 CONCLUSIONS**

We detected outliers in the database of observed baseline values obtained at Syowa Station, Antarctica, by examining the distributions of the residuals of the absolute baseline values observed from 1997 to 2011. The frequency distribution of the residuals supports the assumption that the residuals of all components follow a Gaussian distribution. In accordance with this assumption, we used a normal quantile–quantile plot to determine the outlier thresholds, which were 2.8 nT, 2.0 nT, and 0.7 minutes for the residuals of the H-, Z-, and D-components respectively.

## **6 REFERENCES**

- Ito, N., & Fujii, I. (2003) On automatic determination of criteria to detect outliers for the absolute measurement of the geomagnetic field. Technical Report of Kakioka Magnetic Observatory, Vol. 1, No. 1, pp 1-9.
- Jankowski, J., & Sucksdorff, C. (1996) Guide for Magnetic Measurements and Observatory Practice, International Association of Geomagnetism and Aeronomy (IAGA), Boulder, USA, pp 235.
- Ookawa, T. (1999) Absolute geomagnetic observations at Syowa station, Antarctica –the replacement of magnetic theodolite–. Technical Report of Kakioka Magnetic Observatory, Vol. 38, No. 2, pp 52-56 (in Japanese).

# A DATA-DRIVEN METHOD FOR SELECTING OPTIMAL MODELS BASED ON GRAPHICAL VISUALISATION OF DIFFERENCES IN SEQUENTIALLY FITTED ROC MODEL PARAMETERS

*K S Mwitondi*<sup>\*1</sup>, *R E Moustafa*<sup>2</sup> and *A S Hadi*<sup>3</sup>

<sup>\*1</sup>Sheffield Hallam University, Faculty of Arts, Computing, Engineering and Sciences; Sheffield S1 1WB, United Kingdom (k.mwitondi@shu.ac.uk CC mwitondi@yahoo.com)

<sup>2</sup>George Washington University, Statistics Department, 2140 Pennsylvania Ave., NW, Washington DC, 20052, USA (Shalash@gwu.edu CC moustafa@dmining-technology.com)

<sup>3</sup>The American University in Cairo, Egypt/Cornell University, 291 Ives Hall, Cornell University, Ithaca, NY 14853-3901, USA (ahadi@aucegypt.edu CC ali-hadi@cornell.edu)

## ABSTRACT

*Differences in modelling techniques and model performance assessments typically impinge on the quality of knowledge extraction from data. We propose an algorithm for determining optimal patterns in data by separately training and testing three decision tree models the Pima Indians Diabetes and the Bupa Liver Disorders datasets. Model performance is assessed using ROC curves and the Youden Index; moving differences between sequential fitted parameters are then extracted and their respective probability density estimations are used to track their variability using an iterative graphical data visualisation technique developed for this purpose. Our results show that the proposed strategy separates the groups more robustly than the plain ROC/Youden approach, eliminates obscurity and minimizes over-fitting. Further, the algorithm can easily be understood by non-specialists and demonstrates multi-disciplinary compliance.*

**Keywords:** Bayesian Error, Data Mining, Decision Trees, Domain Partitioning, Data Visualisation, Optimal Bandwidth, ROC curves, Visual Analytics, Youden Index

## 1 INTRODUCTION

Choosing from a range of competing models is a common practice in predictive modelling in which the selection of the optimal model and its performance depends on a combination of factors. In classification, for instance, the consequences of misclassification largely depend on the true class of the object being classified and the way it is ultimately labelled. Generally, the performance of both parametric and non-parametric models depends exclusively on the chosen model, the sampled data and the available knowledge for the underlying problem. For instance, the accuracy and reliability of, say, a medical test will depend not only on the diagnostic tools but also on the definition of the state of the condition being tested. Such variations make model complexity a natural challenge to data modelling (Mwitondi, 2010). Thus, when data sources, repositories and modelling tools are shared it is imperative to work out a unifying environment with the potential to yield consistent results across applications. Achieving this goal requires striking a balance between model accuracy and reliability across applications (Mwitondi and Said, 2011). The paper examines how multiple model performances can be used to devise a generalised strategy for attaining the foregoing balance in predictive modelling. Using a generic two-class scenario it addresses the underlying issues relating to prediction errors and combines model generated numerals and graphics to decipher data patterns for optimality. More specifically, the paper sets off from conventional approaches for group separation – ROC curves (Egan, 1975) and Youden Index (Youden, 1950) to propose an iterative algorithm for detecting separation levels based on estimated data densities. The paper is organised as follows. Section 2 provides an overview of the methods and simulations, Section 3 outlines the modelling strategy, implementation, results and discussions and the concluding remarks and potential future applications are outlined in Section 4.

## 2 METHODS AND SIMULATIONS

The ultimate goals are to illustrate the nature of variation in the performance of various models given the random nature of the data used to train and test them and devise a modelling strategy for optimising the model selection process. The illustrations and the strategy derive from the Bayesian rule as outlined in Berger (1985), the decision trees (DT) domain

partitioning technique as described in Breiman *et al.*, (1984) and the receiver operating characteristics (ROC) analysis as outlined in Egan (1975).

## 2.1 Allocation rule errors due to data randomness

As shown in Table 1, the total empirical error is typically associated with randomness due to the allocation region and randomness due to assessing the rule by random training and validation data (Mwitondi, 2003).

**Table 1:** Error types associated with domain-partitioning modelling (Source: Mwitondi, 2003)

POPULATION	TRAINING	CROSS VALIDATION	TEST
$\Psi_{D,POP}$	$\Psi_{D,TRN}$	$\Psi_{D,CVD}$	$\Psi_{D,TST}$

Thus, given that there are  $X_{i=1,2,\dots,N}$  data points in  $Y_{k=1,2,\dots,K}$  different classes, the overall misclassification error is computed as the sum of the weighted probabilities of observing data belonging to a particular class given that we are not in that class. For instance,

$$\Psi_{D,TST} = \sum_{k=1}^K \sum_{i=1}^N P(C_k)P(X_i \in C_k | Y \notin C_k) \quad (\text{Equation 1})$$

where  $C_k$  and  $P(C_k)$  represent the partition region and the class priors respectively in a typical Bayesian context. Various approaches for minimising this error have been proposed – see, for instance, Reilly and Patino-Leal (1981), Wan (1990), Freund and Schapire (1997) and Mwitondi *et al.* (2002). A commonly acceptable practice is to vary the allocation rule in order to address specific requirements of an application. We illustrate the scenario based on decision trees as in Breiman *et al.*, (1984) and ROC curves (Egan, 1975).

## 2.2 Decision trees modelling

Growing a tree amounts to sequentially splitting the data into, typically, two super sets – A and B – based on a single predictor at a time. The observations in A and B lie on either side of the hyper-plane  $x_j = m$  chosen in such a way that a given measure of impurity is minimised. The splitting continues until an adopted stopping criterion is reached. Selecting an optimal model is one of the major challenges data scientists face. Breiman *et al.*, (1984) propose an automated cost-complexity measure described as follows. Let the complexity of any sub-tree  $f \in F$  be defined by its number of terminal nodes,  $L_t$ . Then if we define the cost-complexity parameter  $0 \leq \alpha < \infty$ , the cost-complexity measure can be defined as

$$R_\alpha(f^\alpha) = R(f) + \alpha L_f \quad (\text{Equation 2})$$

Let  $f_t$  be any branch of the sub-tree  $f^{(1)}$  and define  $R(f_t) = \sum_{t^* \in L_{\alpha, f_t}} R(t^*)$  where  $L_{\alpha, f_t}$  represents the set of all terminal nodes in  $f_t$ . They further show that given  $t$  any non-terminal node in  $f^{(1)}$ , the inequality  $R(t) > R(f_t)$  holds. It can be shown that for any sub-tree  $f_t$  we can define a measure impurity as a function of  $\alpha$  as

$$R_\alpha(f_t) = R(f_t) + \alpha L_{f_t} \quad (\text{Equation 3})$$

Typically, growing a large tree yields high accuracy but risks over-fitting while growing a small tree does the opposite. The measure of impurity will typically return different estimates for different values of  $\alpha$  directly impinging on accuracy and reliability. One way of assessing model performance is to use ROC curves.

## 2.3 ROC curves analysis, optimality and Youden Indexing]

Without loss of generality, consider a binary medical diagnostic test scenario in which patients are tested for a particular disease and there are four possible outcomes – true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this case, the ROC curve is constructed based on the proportions

$$SST = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \text{and} \quad SPT = \frac{N_{TN}}{N_{TN} + N_{FP}} \quad (\text{Equation 4})$$

where  $SST$  and  $SPT$  denote the sensitivity and specificity respectively and  $SST = 1 - SPT$ .  $N_{TP}$  and  $N_{FN}$  denote the number of those with the disease and who are diagnosed with it and those having the disease but cleared by the test

respectively. Similarly,  $N_{TN}$  and  $N_{FP}$  are the number of those without the disease who test negative and those testing positive without having the disease respectively. As with type I and II errors, the usefulness of a test cannot be determined by SST/SPT alone – and so a ROC analysis trade-off is needed. Assuming four possible outcomes – the ROC accuracy (ACCR) and error (ERR) are defined as

$$ACCR = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \leftrightarrow 1 - ACCR = ERR \quad (\text{Equation 5})$$

If we denote the data by  $X$  and the set of class labels as  $C_i = \{Y_1, Y_2\}$  the probability of accuracy can be computed as follows - where the integral is over both classes.

$$P(ACCR) = P(Y_i \in C_i) = \sum_{i=1}^2 P(Y_i) \int P(X|Y_i) dx \leftrightarrow 1 - P(Y_i \in C_i) = P(ERR) \quad (\text{Equation 6})$$

The main goal of predictive modelling is to maximise  $P(ACCR)$  (minimise  $P(ERR)$ ) consistently across applications. By appropriately costing each of the class allocation measures we can make the outcome not only depend on the diagnostic tools and techniques but also on the definition of the state of the tested condition. For instance, one would rather set “low specificity” for a cancer diagnostic test – i.e., let it trigger on low-risk symptoms than miss the symptoms. Our model implementations are focused on striking this balance. One way of determining the optimal cut-off point for the ROC curves is to use the Youden index (Youden, 1950). Its main idea is that for any binary classification model with corresponding cumulative distribution functions  $F(*)$  and  $G(*)$ , say, then for any threshold  $t$ , the relationship  $SST(t) = 1 - F(t) \leftrightarrow SPT(t) = G(t)$  holds. We can then compute the index  $\gamma$  as the maximum difference between the two as

$$\gamma = \max_t \{SST(t) + SPT(t) - 1\} = \max_t \{G(t) - F(t)\} \quad (\text{Equation 7})$$

Within a model, the Youden index is the maximum differences between the true and false positives values and between competing models ordering of the indices highlights performance order.

### 3 MODELLING STRATEGY, IMPLEMENTATION, RESULTS AND DISCUSSIONS

The modelling strategy is based on the methods described in Section 2 and seeks to facilitate the selection of optimal models based on consistency of performance. Two datasets - the Pima Indians diabetes data - 768 observations on 9 variables (NIDDK, 1990) and the Bupa liver disorders data - 345 observations on 7 variables (Forsyth, 1990) – as described in Since variations in  $\alpha$  affect reliability we trained and tested three decision tree models on each of the datasets.

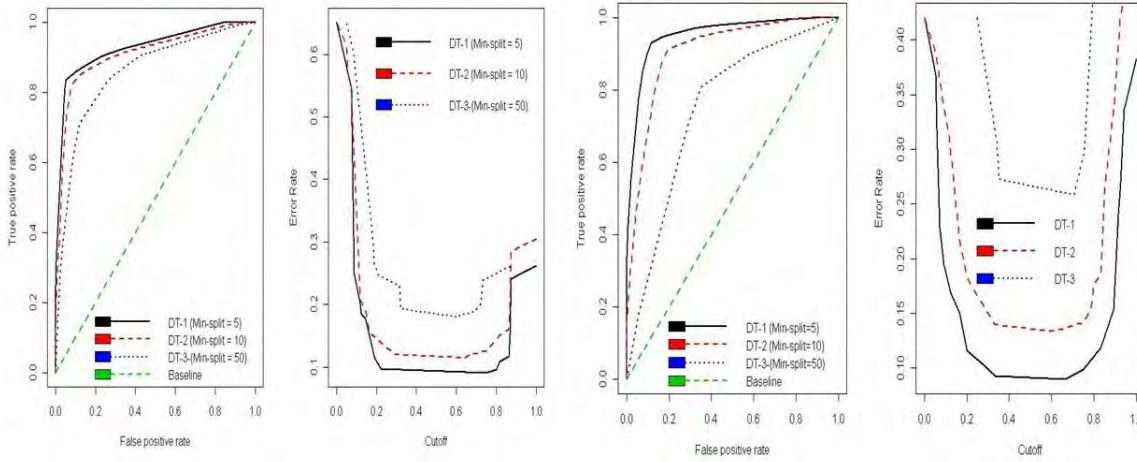
Table 2 are used. The former relate to females of at least 21 years old while the latter relate to blood tests for liver disorders sensitivity to excessive alcohol consumption. Since variations in  $\alpha$  affect reliability we trained and tested three decision tree models on each of the datasets.

Table 2: Data attributes for the Pima Indians diabetes and the Bupa liver disorders datasets

PIMA INDIANS DIBETES DATA		BUPA LIVER DISORDERS DATA	
<b>NTP</b>	Number of times pregnant	<b>MCV</b>	Mean corpuscular volume
<b>PGC</b>	Plasma glucose concentration	<b>ALKPHOS</b>	Alkaline phosphotase
<b>DBP</b>	Diastolic blood pressure	<b>SGPT</b>	Alamine aminotransferase
<b>TSF</b>	Triceps skin fold thickness	<b>SGOT</b>	Aspartate aminotransferase
<b>BMI</b>	Body mass index	<b>GAMMAGT</b>	Gamma-glutamyl transpeptidase
<b>DPF</b>	Diabetes pedigree function	<b>DRINKS</b>	Daily half-pint drink equivalents
<b>AGE</b>	Age (years)	<b>CLASS</b>	Binary target
<b>CLASS</b>	Binary Target		

#### 3.1 Implementation and results

Graphical results for both datasets are presented in Figure 1. Left to right, the first two panels correspond to Pima ROC and predictive patterns while the remaining two correspond to those of the Bupa dataset.



**Figure 1:** Pima (left) and Bupa(right) ROC curves and model over/fitting points

Based on the ROC convention, a classifier is optimal only if it yields results in the top left corner, given the set conditions. Thus, the performance ranking in both cases was DT-1, DT-2 and DT-3. The maximum differences (Youden indices) between the TPR and FPR for each model, the areas under the curve and the minimum prediction errors are highlighted in Table 3, agreeing with this ranking. Note that Bupa’s gaps between DT1 and DT2 and between DT2 and DT3 are much wider than Pima’s implying that DT1 and DT2 performance on the Pima Indians data is almost indistinguishable. Further, repeated simulations are expected to vary depending on factors such as data sources and the settings defined in Section 2.2.

**Table 3:** Performance table for each of the three models on each of the two datasets

	PIMA-1	PIMA-2	PIMA-3	BUPA-1	BUPA-2	BUPA-3
<b>YOUDEN INDEX</b>	0.7838	0.7392	0.5907	0.8128	0.7169	0.4583
<b>AREA UNDER THE CURVE</b>	0.9273	0.9086	0.8556	0.9527	0.9069	0.7487
<b>MINIMUM ERROR</b>	0.0911	0.1146	0.1797	0.0899	0.1333	0.2580

Accuracy of the test is a function of how well it separates the groups and it is assessed by the corresponding area under the curve. Traditionally, a 90%-100% area under the curve is considered excellent while anything about 60% and less is a failure. Since ROC curves may mask or over-fit data estimated parameters, we propose a strategy for enhancing the formulation of a generalised error.

### 3.2 Proposed strategy for optimal model selection

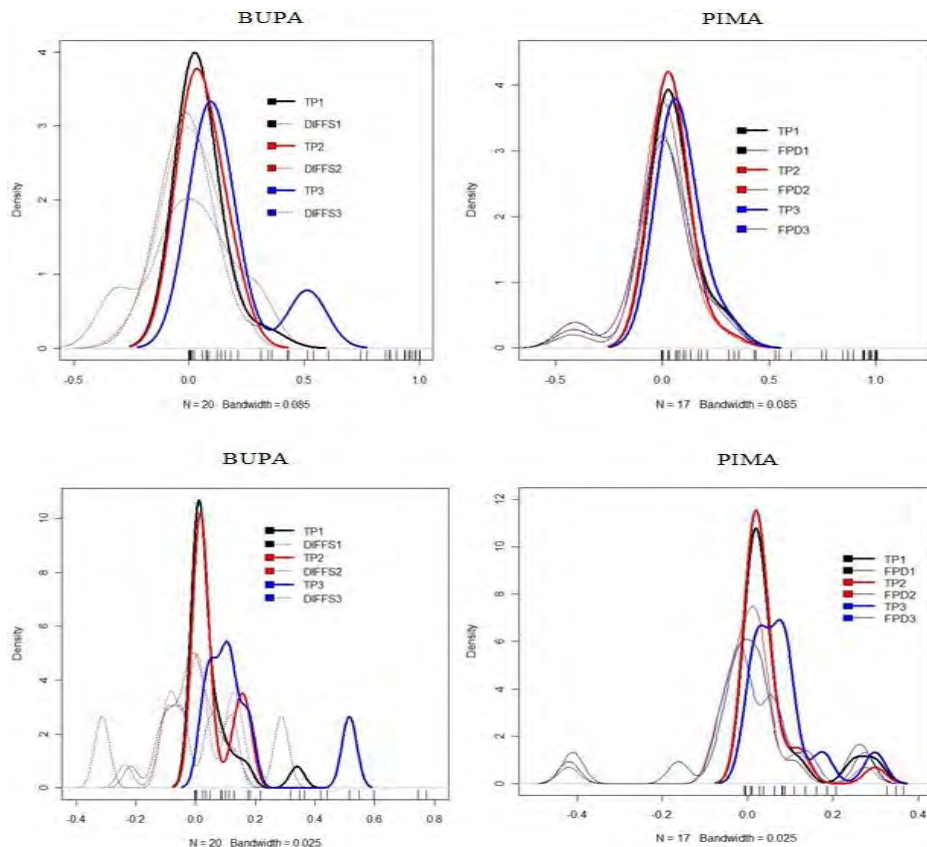
Typically, one classifier is preferred to another if it yields a higher class posterior probability than the other (Web, 2005; Mwitondi, 2003). We have shown that the rate at which the models out-perform each other is data-dependent and so the fitting patterns may provide a good starting point in the search of optimality. If we assume a Bayesian error from a notional population on which the performance of a predictive model is assessed to be  $\Psi_{B,POP} = \Psi_{D,TST}$  (data-dependent error), then the relationship below holds

$$P(\Psi_{D,POP} \geq \Psi_{B,POP}) = 1 \leftrightarrow E[\Psi_{D,POP}] - \Psi_{B,POP} = E[\Delta] \geq 0 \text{ (Equation 8)}$$

We can then measure model reliability by tracking the quantity  $Var[\Delta]$  as an indicator of stability/variability across models. The algorithm, described in the following simple steps, seeks to minimise the risk of over-fitting or under-fitting the data across applications.

Given a set of competing classifiers  $\{C_j\} j = 1, 2, \dots, K$   
 Extract the vectors  $TP = X_i^T$  and  $FP = X_i^F$   
 Set  $D_i = X_i^T - X_i^F$   
 For  $j := 1:K$   
   For  $i := 1:i - 1$   
      $DIFFS = D_{i+1} - D_i$   
      $TP_d = X_{i+1}^T - X_i^T$   
      $FP_d = X_{i+1}^F - X_i^F$   
   End For  
 Store the Differences  $DIFFS$ ,  $TP$  and  $FP$   
 End For  
 Set a long bandwidth vector (typically Gaussian)  $\beta \in (1, 0)$   
 While NOT END of  $\beta$  Do  
   Compute and plot the densities of  $D_i$ ,  $TP_d$  and  $FP_d$   
 End While  
 Examine the resulting plots and choose the one that best separates the groups  
 End.

Graphical illustrations of the algorithm results based for both datasets at different bandwidths are shown in Figure 2. The spiky lines at the foot of each of the four density panels represent a slightly noised univariate vector of TP and FP from left to right. Since the main purpose is to separate each of the two classes we are interested not only in the positioning of the ROC curves as in Figure 1 but also in the sequential differences in the fitted parameters which suggest that DT1 is suitable for the BUPA and DT2 for the PIMA data.



**Figure 2:** Differences between sequential TP and FP values and those of differences between them

The Gaussian kernel was used in approximating the differences in the algorithm. The optimal choice of the bandwidth is estimated as  $b = \left(4\hat{\sigma}^5/3n\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5}$  (Silverman, 1984) where sigma is the standard deviation of the samples  $n$ . Runs based on these optimal bandwidth values yielded similar results to those in Figure 2 at  $b \leq 0.1$  - suggesting DT1

for BUPA and DT2 for PIMA. Since each of the ROC curves in Figure 1 measures the probability that the corresponding model will rank higher a randomly chosen positive instance than it will rank a randomly chosen negative case, the patterns in Figure 2 provide an insight into the level of class separation and can be used to guide model selection.

#### 4 CONCLUDING REMARKS AND POTENTIAL FUTURE DIRECTIONS

Selecting the “best” model from a potentially set of competing models is a conventional challenge in data science and this paper sought to demonstrate an optimisation procedure for making that decision. Guided by the Bayesian rule, ROC curves and the Youden Index we empirically demonstrated the variation of the allocation rule using three decision tree models. For the purpose of addressing specific application requirements - a practical reality in a data sharing environment – we introduced a novel strategy for model selection. Based on graphical visualisation, the strategy seeks to help minimise data over-fitting and performance obscurity while remaining easily understood by non-specialists. The results from this paper serve to highlight the importance of paying attention to the allocation rules in Table 1 and the associated generalising error. The strategy can be adapted to other data-dependent domain-partitioning models such as neural networks and support vector machines. The quantity  $\text{Var}[\Delta]$  can generally be accepted as a measure of performance which, for domain-partitioning purposes, we can align alongside similar measures such as the ROC curves. The proposed strategy can readily be adapted to all applications of a binary nature or those which can be converted into such and our results highlight novel paths towards tackling various real-life challenges in areas such as remote sensing, seismology, oceanography, ionosphere and many others. Since error costing differs across applications, the decisions relating to model selection will typically remain application-specific. However, the proposed strategy provides prospects of interactivity and multi-disciplinary compliancy in a general data sharing framework irrespective of the nature of the applications. We hope that this study will supplement previous studies which have focused on methods for selecting optimal models in our increasingly expanding cross-disciplinary research environment.

#### 5 REFERENCES

- Berger, J. O. (1985). *Statistical decision theory and Bayesian Analysis*; Springer-Verlag.
- Breiman, L., Friedman, J. Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*; Chapman and Hall, ISBN-13: 978-0412048418.
- Egan, J. P. (1975). *Signal Detection Theory and Roc Analysis*; Academic Press; ISBN-13 978-0122328503
- Freund, Y. and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting; *Journal of Computer and System Sciences*, Vol. 55, Issue No. 1, pp 119–139.
- Forsyth, R. S. (1990). *PC/BEAGLE User's Guide*; BUPA Medical Research Ltd.
- Mwitondi, K. S. (2003). *Robust Methods in Data Mining*; PhD Thesis; School of Mathematics, University of Leeds; Leeds; University Press.
- Mwitondi, K. S. and Said, R. A. (2011). A step-wise method for labelling continuous data with a focus on striking a balance between predictive accuracy and model reliability; international Conference on the Challenges in Statistics and Operations Research (CSOR); 08th -10th March - 2011, Kuwait City
- NIDDK (1990). *Pima Indians Diabetes Data*; National Institute of Diabetes and Digestive and Kidney Diseases.
- Reilly, P. M. and Patino-Leal, H. (1981). A Bayesian Study of the Error-in-Variables Model; *Technometrics*, Vol. 23, No. 3.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*; Chapman and Hall - Monographs on Statistics & Applied Probability; ISBN-13: 978-0412246203.
- Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005). ROCR: Visualizing Classifier Performance in R; *Bioinformatics* Vol. 21, No. 20, pp. 3940-3941.
- Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32-35.



# DIGITIZATION OF BROMIDE PAPER RECORDS TO EXTRACT ONE-MINUTE GEOMAGNETIC DATA

*N Mashiko<sup>1\*</sup>, T Yamamoto<sup>2</sup>, M Akutagawa<sup>1</sup>, and Y Minamoto<sup>1</sup>*

<sup>1</sup>*Kakioka Magnetic Observatory, Japan Meteorological Agency, 595 Kakioka, Ishioka-shi, Ibaraki 315-0116, Japan*

<sup>\*</sup>*Email: n-mashiko@met.kishou.go.jp*

<sup>2</sup>*Meteorological Research Institute, Japan Meteorological Agency, 1-1 Nagamine, Tsukuba-shi, Ibaraki 305-0052, Japan*

## ABSTRACT

*Many long-term geomagnetic observation results recorded on photographic bromide paper have not yet been fully digitized. To that end, we developed a method to automatically convert photographic records to one-minute digital data. We applied our method to the observation records of Kakioka Magnetic Observatory and confirmed that the resolution of time and amplitude could be greatly improved by numerical conversion compared with conventional data conversion by hand scaling. Our results suggest that highly precise digitization of analog magnetograms is possible.*

**Keywords:** Analog Magnetogram, Digitization, Geomagnetism, Photographic paper, Historical data, Digital data extraction, One-minute data

## 1 INTRODUCTION

In the past, geomagnetic field observations were recorded in analog form on photographic paper (Jankowski & Sucksdorff, 1996) with a silver bromide emulsion, known as bromide paper. Up to now, these analog magnetograms have been read only by hand scaling with low time resolution. Many observatories have analog records covering long observation periods (Iyemori, Nose, McCreddie, Odagi, Takeda & Kamei et al., 2005). At Kakioka Magnetic Observatory (KMO) in Japan, records on photographic paper of geomagnetic observations go back to 1924. However, most of the numerical data available from these magnetograms are hourly values that were digitized by hand scaling. Conversion of these analog records to high-resolution numerical data would make them very useful in investigations of past geomagnetic activity. We therefore developed a method for converting the analog magnetograms into digital data with high time and amplitude resolutions and then examined the conversion accuracy.

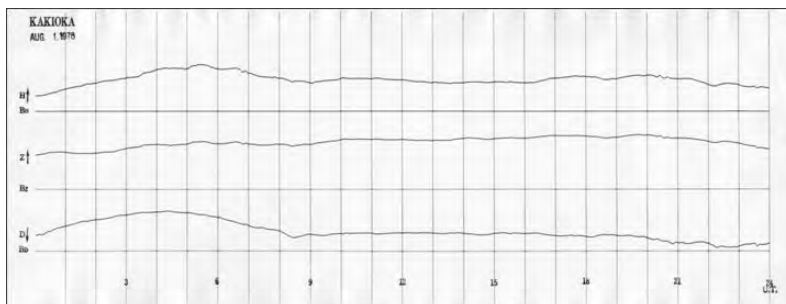
## 2 OUTLINE OF THE CONVERSION METHOD

The first step was to obtain high-resolution scans of the photographic paper records and to store them as graphic files. We then used an image processing program that we developed to distinguish lines and curves in the graphic files. The identified pixels were converted into numerical time and geomagnetic field data. Each step of the conversion is described in detail below.

### 2.1 Photographic paper record and scan specifications

Each record is a sheet of photographic paper about 510 mm long by 195 mm wide (Figure 1). Three components (H, horizontal; Z, vertical; D, declination) of one day's geomagnetic field variations are recorded on one sheet. Time is recorded in the longitudinal direction along three baselines, and 24 transverse lines (time marks) divide the record into 20-mm intervals. Each interval represents one hour, and the time mark indicates the 0th minute of the hour. Each baseline is the zero amplitude line for a curve that records the variation in one geomagnetic field component, and the curve can fluctuate from above to below the baseline with a resolution of about 2.5 nT (H and Z) or 0.29 minutes (D) per 1 mm. These values are average scale values in 1963 and later. Scale values vary depending on the period and detailed values of each year are shown in annual reports of Kakioka Magnetic Observatory (e.g., Kakioka Magnetic Observatory, 1996). In the past, values have been read off the KMO photographic paper records with one-hour and 0.1-mm resolutions by hand scaling, and the read values have been converted into geomagnetic field values by comparing them with baseline, scale, and other correction values. These values, after conversion and correction, have resolutions of 1 hour and 1 nT (H and Z) and 0.1 minutes (D).

In our digitization method, we scanned the photographic records at a relative resolution of 600 pixels per inch, which corresponds to about 4,600 by 12,000 pixels per sheet. This resolution is equivalent to a time resolution of about 7.5 s per pixel, and an amplitude resolution of about 0.1 nT (H and Z) and 0.01 minutes (D) per pixel. The scanned photographic records are stored as 24-bit color bitmap images so that each pixel can be weighted according to its luminosity during image processing.



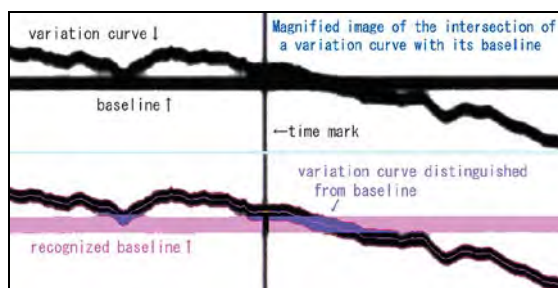
**Figure 1.** An example of one day's analog magnetogram (1 August 1978) recorded on bromide paper

## 2.2 Automatic recognition and numerical conversion of the image

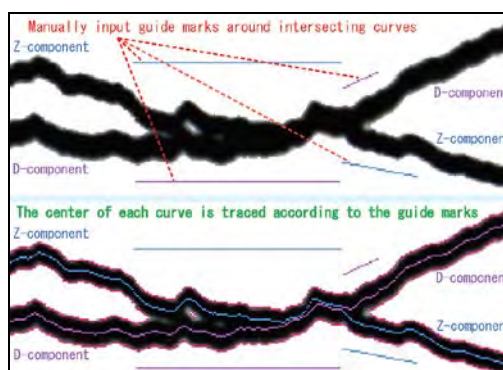
The image processing program automatically converts each scanned high-resolution graphic file to numerical data by successively performing the following steps.

### 2.2.1 Identifying time marks, baselines, and component variation curves

The trace of each baseline and of each time mark is recognized by its luminosity, width, vertical or horizontal trend, and position (spacing) as a linear band, and the pixel of each central point along a band is specified. The traces on the magnetogram show slight distortions due to the daily installation state of the photographic paper, therefore, the bands are not all straight. The variation curve of each component is considered to be a small, elliptical locus that vibrates up and down while progressing in the time direction. The pixels comprising the central points of the ellipse and each end of the locus are specified by scanning pixel by pixel in the time direction. Where a curve crosses its baseline, it is automatically distinguished from the baseline based on its position and width curve before and after the crossing (Figure 2). When two variation curves cross, the presumed widths and central points of both curves are pinpointed automatically based on guide marks manually input into the digital image file that indicate which component trace is above and which is below at the intersection (Figure 3).



**Figure 2.** Distinguishing a variation curve from its baseline



**Figure 3.** Identifying two curves at an intersection

### 2.2.2 Reading time and amplitude

For each central point along the variation curve, the number of pixels in the time direction from the preceding time mark to the point and from the point to the next time mark is counted. Similarly, the number of pixels in the amplitude direction from the baseline to the point is counted. The number of pixels in the time direction is converted into time by interpolating between the time marks, and the number of pixels in the amplitude direction is converted into the actual distance on the original

photographic paper. The sign of the amplitude is positive above and negative below the baseline. These data are then output to a file as a time series of raw amplitude data at time intervals of about 7.5 s (about 8 numerical values per minute), where the amplitude is the distance from the baseline to the central point of the variation curve.

### 2.2.3 Conversion of raw time series amplitude data into geomagnetic field component values

Using the amplitude data sampled about every 7.5 s (8 data points, 4 on each side of the 0th second of each minute), the amplitude at the 0th second of each minute is computed by the least squares method. By applying scale values, ordinate factors, and baseline values calculated from past observations, the amplitudes at the 0th second of each minute are converted into geomagnetic field values. Thereby, one-minute geomagnetic field variation data are output by the program.

## 3 ACCURACY AND PRECISION OF THE CONVERSION

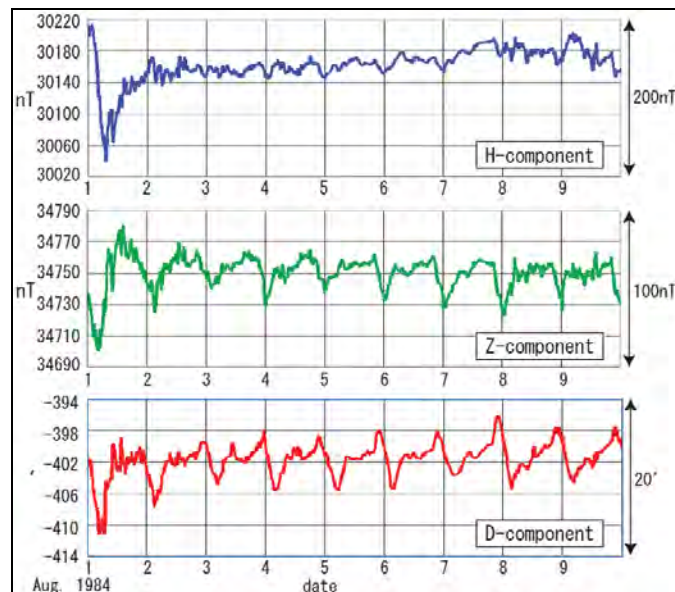
Although the time resolution of these one-minute values is 60 times better than the hourly time resolution obtained by conventional hand scaling, it is necessary to evaluate the accuracy and precision of the digitized values. We describe the evaluation method and the results of the evaluation next.

### 3.1 Evaluation method

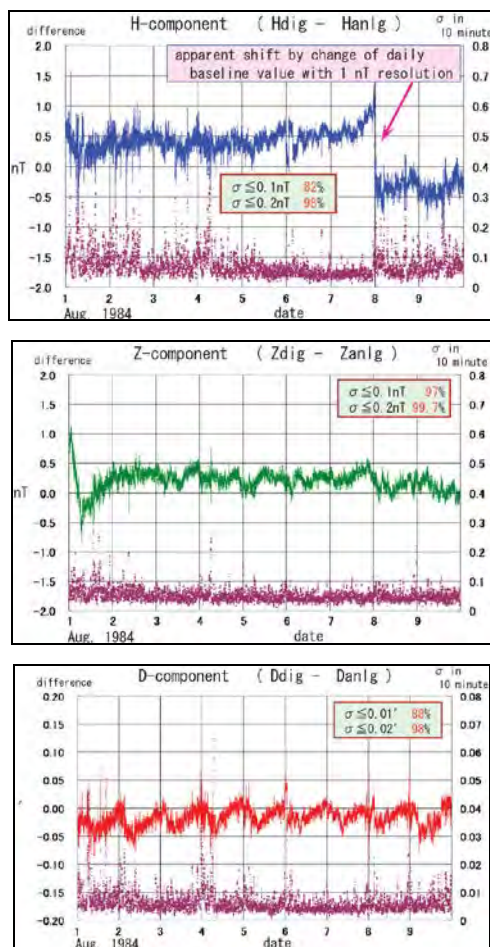
Beginning in 1976, when digital recording magnetometers were installed at KMO (Kakioka Magnetic Observatory, 1978), geomagnetic field measurements were recorded digitally in parallel with the analog recording. Therefore, simultaneous records of observations are available for the same phenomena. We compared the one-minute values of the digitized analog records with the data obtained simultaneously by the digital magnetometers. Here, we show the results for each of the three geomagnetic field components between 1 and 9 August 1984, a period during which the accuracy of the digital observation results was considered adequate.

### 3.2 Accuracy and precision of the conversion results

The one-minute values of the three geomagnetic field components for the period 1–9 August 1984 converted from the photographic paper records are shown in Figure 4. The range of the H-component is about 180 nT. Figure 5 shows the differences between the digitally recorded one-minute values and the converted analog one-minute values. The differences indicate a daily variation of about 0.5 nT (H- and Z-components) or 0.05' (D-component). We attribute these diurnal variations to a difference in the temperature coefficient between the two magnetometers. The shift of about 1 nT in the H-component at around 00 on 8 August (Figure 5, top panel) is attributed to a 1-nT change in the baseline value between 7 and 8 August. The daily baseline values were determined by hand scaling during the conventional calculation of hourly values. We infer that the short-term variations in the differences include reading errors and inherent differences between the two magnetometers. Because at least 98% of the standard deviations of each component were less than 0.2 nT or 0.02', we consider these reading errors to be small.



**Figure 4.** One-minute values of the three geomagnetic field components during 1–9 August 1984 converted from analog records on photographic paper



**Figure 5.** Differences in each component between the digital one-minute values and converted analog one-minute values (upper trace in each panel) and standard deviations computed using one-minute difference values determined over each 10-minute interval (lower trace in each panel)

## 4 CONCLUSIONS

We introduced a method for converting past magnetograms into digital data with a one-minute time resolution and evaluated the accuracy of the converted data. We confirmed that we could convert most analog records obtained during a relatively calm period in terms of geomagnetic activity automatically into one-minute values with an accuracy of better than 1 nT, which is the conventional minimum unit. To enable automatic conversion of all past analog magnetograms, we plan to improve our image processing algorithm so that it can handle two or more traces intersecting intricately and blurry images.

## 5 REFERENCES

- Iyemori, T., Nose, M., McCreadie, H., Odagi, Y., Takeda, M., Kamei, T. & Yagi, M. (2005) Digitization of old analogue geomagnetic data. Retrieved December 1, 2011 from the World Wide Web: <http://www.ukoln.ac.uk/events/pv-2005/posters/iyemori.pdf>
- Jankowski, J. & Sucksdorff, C. (1996) *Guide for magnetic measurements and observatory practice*, Warsaw: International Association of Geomagnetism and Aeronomy
- Kakioka Magnetic Observatory (1978) *Report of the Kakioka Magnetic Observatory geomagnetism Kakioka 1976*, Tokyo: Japan Meteorological Agency
- Kakioka Magnetic Observatory (1996) *Report of the Kakioka Magnetic Observatory geomagnetism Kakioka Memambetsu Kanoya Chichijima 1995*, Yasato-machi, JP: Kakioka Magnetic Observatory

## Conference Organizers

### Conveners of the Conference

MINSTER, Jean-Bernard (Scripps Institution of Oceanography, UCSD)  
WATANABE, Takashi (Solar-Terrestrial Environment Laboratory, Nagoya University)  
IYEMORI, Toshihiko (WDC for Geomagnetism, Kyoto, Kyoto University)

### Scientific Organizing Committee (SOC)

MINSTER, Jean-Bernard, *Chair* (Scripps Institution of Oceanography, UCSD)  
CHEN, Robert (Centre for International Earth Science Information Network, CIESIN)  
CLARK, David (National Geophysical Data Centre, NOAA)  
DIEPENBROEK, Michael (World Data Centre for Marine Environmental Sciences)  
GENOVA, Françoise (Strasbourg Astronomical Data Centre, CDS)  
HARRIS, Ray (ICSU ad hoc Strategic Coordinating Committee for Information and Data)  
HORTA, Luiz (Large Scale Biosphere-Atmosphere Experiment in Amazonia, INPE)  
MOKRANE, Mustapha (International Council of Science, ICSU)  
NEILAN, Ruth (Central Bureau of the International GNSS Service, JPL)  
RICKARDS, Lesley (Permanent Service for Mean Sea Level, BODC)  
WATANABE, Takashi (Solar-Terrestrial Environment Laboratory, Nagoya University)  
YAN, Baoping (Computer Network Information Centre, CAS)  
ZGUROVSKY, Michael (National Technical University, Kiev Polytechnic Institute, Ukraine)

### Local Organizing Committee (LOC)

IYEMORI, Toshihiko, *Chair* (WDC for Geomagnetism, Kyoto, Kyoto University)  
WATANABE, Takashi, *Vice Chair* (Solar Terrestrial Environment Laboratory, Nagoya University)  
ASHINO, Toshihiro (Toyo University)  
ISHII, Mamoru (National Institute for Information and Communications Technology)  
KADOKURA, Akira (National Institute of Polar Research)  
KANAOKI, Masaki (National Institute of Polar Research)  
KITAMOTO, Asanobu (National Institute of Informatics)  
KUNISAWA, Takashi (Tokyo University of Science)  
MURATA, Takeshi (National Institute for Information and Communications Technology)  
MURAYAMA, Yasuhiro (National Institute for Information and Communications Technology)  
NOSE, Masahito (WDC for Geomagnetism, Kyoto, Kyoto University)  
OGINO, Tatsuki (Solar Terrestrial Environment Laboratory, Nagoya University)  
OHISHI, Masatoshi (National Astronomical Observatory)  
SHINOHARA, Iku (Japan Aerospace Exploration Agency)  
TOH, Hiroaki (WDC for Geomagnetism, Kyoto, Kyoto University)  
TSUBOI, Seiji (Japan Agency for Marine-Earth Science and Technology)  
TSUDA, Toshitaka (Research Institute for Sustainable Humanosphere, Kyoto University)  
YAMAGIWA, Juichi (Graduate School of Science, Kyoto University)

### Regional Organizing Committee (ROC) at Kyoto University

IYEMORI, Toshihiko; HAYASHI, Hiroo; KOYAMA, Nobuyuki; MATSUMURA, Mitsuru; NOSE, Masahito;  
ODAGI, Yoko; OSHIMAN, Naoto; SAITO, Akinori; SHIBATA, Kazunari; SUGIYAMA, Junji; SUZUKI, Shingo;  
TAKEDA, Masahiko; TAKEUCHI, Noriko; TOH, Hiroaki; TSUDA, Toshitaka; and UENO, Satoru

### Sponsor, Co-sponsors

#### *Sponsored by:*

ICSU WDS International Programme Office  
ICSU WDS Scientific Committee  
Science Council of Japan  
Graduate School of Science, Kyoto University

#### *Co-sponsored by:*

Disaster Prevention Research Institute, Kyoto University  
National Institute of Information and Communications Technology  
National Institute of Polar Research  
Research Institute for Sustainable Humanosphere, Kyoto University  
Solar-Terrestrial Environment Laboratory, Nagoya University

#### *Supported by:*

Commemorative organization for the Japan World Exposition `70



**ICSU**  
International Council for Science



ICSU-WDS International Programme Office  
c/o NICT, Koganei, Tokyo 184-8795, Japan  
Tel. +81-42-327-6395  
Fax. +81-42-327-6490  
ipo@icsu-wds.org  
www.icsu-wds.org