Opportunities for Data Exchange

Ten Tales of Drivers & Barriers in Data Sharing

APA
ALLIANCE FOR
PERMANENT ACCESS

7 SEVENTH FRAMEWORK PROGRAMME

APA
ALLIANCE FOR
PERMANENT ACCESS

UK Office

Dr David Giaretta
Alliance for Permanent Access
2 High Street
Yetminster
Dorset DT9 6LF, UK
+44 1935 872660

Registered office

Alliance for Permanent Access
Prins Willem-Alexanderhof 5,
2595 BE
The Hague
The Netherlands

"The ODE Project is identifying, collating, interpreting and delivering evidence of emerging best practices in sharing, re-using, preserving and citing data, and documenting drivers of change and the barriers impeding progress."

Contents

Introduction

Welcome to this collection of success stories and lessons learned in the area of data sharing, re-use and preservation. These use cases outline the state of play in this dynamic area, and are meant to help stakeholders appreciate the vast potential for innovation as well as barriers to success in the field.

These ten tales, selected by the Opportunities for Data Exchange (ODE) project, are based on personal interviews with leaders in scientific communities, research infrastructures, management and policy initiatives. These unique perspectives look at data sharing from many angles, and provide fresh, first-hand accounts of experience and involvement in the following areas:

- Leading-edge scientific research
- Funding policy
- Coordination of large-scale e-infrastructures
- Researcher access to e-infrastructures
- Extending data infrastructures to meet the needs of the modern classroom.

"Because research in genomics, pharmacology or the fight against cancer increasingly depends on the availability and sophisticated analysis of large data sets. Sharing such data means researchers can collaborate, compare, and creatively explore whole new realms.We cannot afford for access to scientific knowledge to become a luxury, and the results of publicly funded research in particular should be spread as widely as possible."

Neelie Kroes, Vice President
European Commissioner responsible for the Digital Agenda[1]

---

[1] http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/11/596&format=HTML&aged=0&language=EN&guiLanguage=en

Sharing data: the ODE Project

Science is changing. The massive volume and variety of data pouring out of publicly funded science are transforming the face of research. These data belong to everyone. If we manage these precious resources properly, we may tackle the Grand Challenges of our times – even as budgets become more restricted.

It is easy to take for granted that data in the public domain will be protected and remain both available and accessible. Researchers, publishers, policymakers and funders – among many others – have started to appreciate that a robust, sustainably funded infrastructure is absolutely necessary to protect the hard-earned fruits of publicly funded research.

Opportunities for Data Exchange (ODE)[2], a project funded by the European Commission (FP7), is gathering evidence to support and promote data sharing, re-use and preservation. ODE partners are members of the Alliance for Permanent Access (APA) and represent stakeholders with significant influence within their communities. ODE is identifying, collating, interpreting and delivering evidence for emerging best practice in sharing, re-using, safeguarding and citing data. ODE is also documenting drivers of change, and barriers to progress in this important area.

The Alfred Wegener Institute for Polar and Marine Research has produced this booklet on behalf of the ODE Project for dissemination at the 2011 APA Conference, 8-9 November, in London, and beyond.

"It is crucial to support the individual researchers in this community. The community has widely varied experience with research data sharing."
UK Data Archive, Economic and Social Data Service

[2] http://www.ode-project.eu

A User's Guide:
Do's and Don'ts in Data Sharing

"Open data sharing is not always possible for certain datasets. We need to apply specific access controls to enable sharing of confidential or sensitive data."

Libby Bishop and Veerle Van den Eynden (UK Data Archive, Economic and Social Data Service)

The UK Data Archive contains the largest collection of digital social and economic research data in the UK. It acquires, curates, and provides access to datasets and provides the support and technical infrastructure for the community to fulfil the policy requirements set by the funding bodies and research councils. Currently it hosts several thousand datasets. The archive is largely funded by the ESRC, the JISC and the University of Essex. Libby Bishop is Senior Researcher Liaison, and Veerle Van den Eynden is Research Data Management Support Services Manager at the UK Data Archive.

Managing and sharing data
Resources for research and training.

The UK Data Archive deals with research data from academic research, governmental data, and commercial data. We deal directly with the first type of data, produced by individuals and research groups in the domain of the wider Social Sciences and Humanities (SSH).

In social sciences and humanities (SSH), the needs of research data management can be very specialised as data may contain personal information. When it comes to qualitative data for example, some interview data may need scrupulous handling. In this instance, one cannot simply take a dataset and ingest it into a data repository. Further pre-processing is needed to make the research dataset suitable for sharing and for publication, such as anonymizing personal details or ensuring that consent for data sharing or publishing is in place. Data management for this kind of research data requires a lot of engagement with researchers to ensure they pay attention to data preparation, licensing, consent, and access rights during research. We provide this through the Economic and Social Data Service.

What do you do with regard to research data sharing?
In our daily work routine, we have a great deal of hands-on engagement. Researchers who want to share their data in this domain usually need advice from a real person. Many types of research data have special factors that need to be considered before publication (e.g. to preserve anonymity). Much human intervention may be needed, which means automated data processing and ingestion is rather limited. The consultancy work is as diverse as the SSH data; it is important to have specialists in place to deal with it all.

It is crucial to support the individual researchers in this community. The community has widely varied experience with respect to research data sharing; for many researchers it is their first time. They do not know how to share their data. They may know there are vital things to consider before sharing, but they may not know the details, so they need advice. It is also important to note that open data sharing is not always possible for certain datasets. We need to apply specific access controls to enable the sharing of confidential or sensitive data.

With ever more data policies from funding bodies and research councils it is even more important to guide researchers though the do's and don'ts of data sharing, so that they comply with the guidelines and share data in an appropriate manner.

Highlights and challenges
One highlight is the emerging awareness of data sharing throughout the community. Previously, the UK Data Archive organized conference sessions to promote this topic in the community. Now there are more secondary analysis projects, resulting in increased data re-use. This trend comes out of the community, in the sense that people are organizing re-use events independently of the UK Data Archive. The challenge is that the research community is still hesitant when it comes to sharing material. While researchers are busy with research and publishing, sharing research data is often not on their agenda, especially because data preservation and sharing are not considered relevant to career promotion and research assessment.

Currently it seems that it is a case of 'carrots and sticks'. Researchers might preserve and share their data because they are obliged to do so by funding bodies, but they do not really see the benefit yet. This is a long-term development and it is changing, but only slowly. Such change needs more time, more advice and more guidance for researchers.

Any more projects and challenges ahead?
One upcoming project is persistent identification via DOI (digital object identifiers), which will make datasets citable. Now in the discussion phase, it will commence in the near future. A challenge ahead is the financial situation that will impose financial cuts on academia in the UK. This is unfortunate, as data need proper treatment and preparation. Our researchers need the advice provided by the UK Data Archive staff. If one wants to encourage researchers to share their data, one also needs to support this goal with the corresponding infrastructure and services.

"Many subtypes of research data have factors that need to be considered before publication (e.g. to preserve anonymity)."

Libby Bishop
UK Data Archive

Veerle Van den Eynden
UK Data Archive

Interviewer:
Sünje Dallmeier-Tiessen

Making the Best Use of Life Science Data

The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL). EMBL-EBI is located on the Wellcome Trust Genome Campus in Hinxton, near Cambridge in the UK. Associate Director Graham Cameron began working for EMBL in Heidelberg in 1982 and, as second member of staff, played a major role in establishing the institute. He developed and managed the EMBL Data Library, which eventually became the EMBL-EBI and now has more than 500 members of staff. Graham is responsible for several EU projects and oversees EMBL-EBI's vast range of services, particularly the data libraries. He describes himself as a 'data sharer' rather than a classical researcher.

> "Science is international, and so are databases. It is important to respond to the needs of researchers and build usable interfaces that facilitate re-use."
>
> **Graham Cameron, Associate Director, the EMBL-European Bioinformatics Institute**

Managing research data has always been a challenge, and one that EMBL staff have tackled from its beginnings. In the 1970s, they started to collect data from research projects and in 1981 EMBL established one of the first data libraries in the world for nucleotide sequence data. At first the goal was simply to extract data from journals. But with the acceleration of methods for DNA extraction and the growing efficiency of high-throughput methodologies, the focus shifted to attracting direct data submission by the researchers themselves. Initially, journal editors were rather reluctant to expand their involvement in data extraction and sharing, but over time this has changed.

Similar developments were happening at the same time around the world, notably in the US with GenBank. In 1986, the International Nucleotide Sequence Database Collaboration (INSDC) was signed, kicking off the successful cooperation between the DDBJ in Japan, GenBank in the US and EMBL-EBI's Nucleotide Sequence Database in the UK. These three databases exchange and synchronize their data daily, thus making it easier for researchers to access up-to-date data and information from around the world. Hopefully, the agreement will expand in the next year to include partners in China.

How do you share research data in the domain of molecular biology?

Because research data are published in the public domain, they could potentially be aggregated and sold by commercial users. The decision to place data in the public domain is driven by the demand for easy access and re-use of the information that life science communities need to progress.

Sometimes, data is first submitted and accepted into the database with a delay in the publication date. This is usually driven by the submission and acceptance of a publication in a journal that requires a data accession number at the time of submission. But there are cases when data producers do not want to have their data made available before the publication of their paper.

In the early days, databases only published datasets that were discussed in peer-reviewed publications, in the belief that these data were quality controlled. This has changed because the data are not integral to the classical peer-review process. Data submitted to EMBL-EBI's databases are tested with quality control procedures. This is mainly an automated process but it also requires some "hands-on" curation by human beings, who can contact the data producers directly if questions arise.

What are the challenges associated with data sharing in molecular biology?

Over time, we have come to regard data as an established scientific record. Data access is undoubtedly beneficial to the community. For instance, biomedical data access could accelerate scientific advancements for human wellbeing, while access to molecular forestry data could provide direct benefits to the environment.

In molecular biology, the development of methodologies and data production has accelerated rapidly. For example, the Human Genome Project took ten years to complete; now, that same work could be done in a matter of minutes. This acceleration is happening across the life sciences, and we are now handling a staggering volume and variety of information that requires careful management and integration.

The extension of data storage is a challenge, and there are initiatives working on, for example, data compression. But with the increasing size and complexity of the data being produced, a major bottleneck today is the contextualization and integration of data. A researcher who is interested in a particular topic might want to look beyond one specific analysis to other research that might be related. How can we integrate and display this information?

A new development in molecular biology research generally is the pursuit of projects that concentrate solely on data production – the analysis and interpretation of these data are performed separately. Usually, the data produced in a project like this are submitted to the public database immediately. This facilitates early usage, but it also requires new discoverability tools to make it easier to re-use the massive amount of new material – this is another challenge for bioinformatics.

Commercial data production also poses difficulties. Even though an estimated 15-20% of database users work in commercial enterprises, they hesitate to share their data openly. EMBL-EBI's activities are stimulating data sharing between different commercial sectors. However, issues like patenting are still considered constraints.

Why is the molecular biology community so successful at sharing research data?

This relates to the question of why molecular biology itself is so successful. One answer could be that genes are everywhere. It is obvious to the research communities that public access to the entirety of the scientific record is needed. Everyone needs to share their data; otherwise, what is made available will be of limited value.
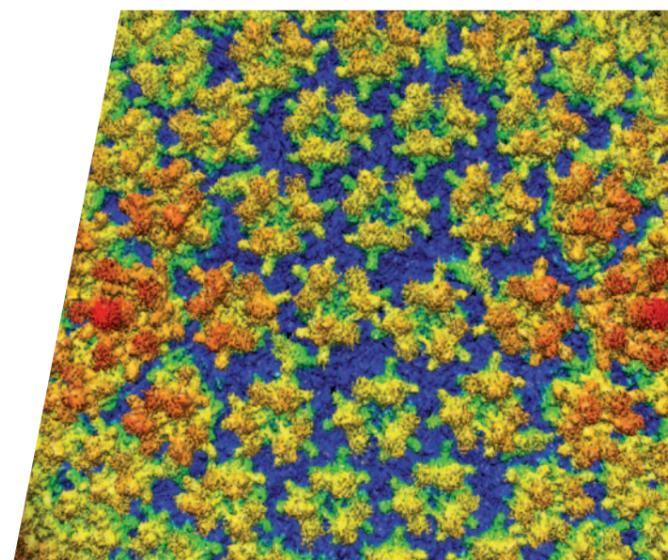
It is relatively easy to work with molecular data. Science is international, and so are the databases. In the past, paper publications were the repositories for scientific results. When journals began to require the data accession numbers for submission, databases became more relevant and the use of the research data increased. This re-use of data is potentially very powerful; just browsing through datasets could lead you to new research areas to explore.

The biggest challenges facing the life science community are access to chemical information and the data deluge. Chemical information is an integral part of bimolecular research and although biological information is shared openly, chemical information is not. Chemical data are often proprietary and access is limited and costly.

As for the data deluge: managing the flood of new data and information is a daunting task, and one that no single organisation – or indeed nation – can manage it alone. Now more than ever, there is a need to integrate diverse life science data from many different databases and make it discoverable. We must respond to the needs of researchers and build usable interfaces that facilitate easy re-use of the material.

EMBL-EBI is coordinating ELIXIR, the purpose of which is to safeguard molecular data by creating a sustainable infrastructure for biological information in Europe. This is a massive undertaking to provide the facilities necessary to support life science research and its translation to medicine, the environment, the bio-industries and society. ELIXIR will effectively help researchers throughout the world to make the best possible use of molecular data, which is the foundation on which our understanding of life is built.

Graham Cameron
EMBL-European
Bioinformatics Institute

Interviewer:
Sünje Dallmeier-Tiessen

Financing the e-Infrastructure to Cope with the Future Flood

Michael Diepenbroek
WDC-MARE

Interviewer:
Angela Schäfer

Interviewer:
Hans Pfeiffenberger

"We need additional financial acknowledgment to develop future integrative data-related e-Infrastructures to cope with the exponentially increasing flood and complexity of data."

Michael Diepenbroek (WDC-MARE)

Michael Diepenbroek is Managing Director of PANGAEA and responsible for the operation of the World Data Center-MARE, based at the Centre for Marine Environmental Sciences (MARUM) at Bremen University and the Alfred Wegener Institute of Polar and Marine Research (AWI) in Germany. Starting in 1992 he worked on the implementation of PANGAEA and was strongly engaged in transforming the World Data Centre system into the new ICSU World Data System, ratified by the International Council for Science in 2008.

What is PANGAEA?

PANGAEA is a data publishing system for Earth & Environmental Science and, as such, partner in numerous European and international projects covering all fields of geo- and biosciences. Its data management services are supplied internationally. Recently PANGAEA has also become engaged in projects supporting spatial data infrastructures, and is a lead partner in the implementation of data portals and infrastructures in several initiatives. PANGAEA has assembled substantial knowledge and practical experience in the implementation of international standards and web technologies.

What drawbacks has PANGAEA encountered in developing scientific data management?

Nowadays the overall aim of PANGAEA is making scientific data available for re-use. In that process we had, and still have, to cope with two separate challenges: technical installation and software management—besides of course running after the data personally, since data storing and sharing is not a standard commitment for all scientists.

In the very beginning of our unstructured data management attempts, we concentrated on individual scientific splinter groups and tried to deliver individual solutions. But we could not fulfil both specially defined requirements and generally accepted requirements in one go.

And we could not guarantee sustainability for small groups only since that kind of long-lasting framework was far too large and costly. Yet these small scientific groups demanded data analysis as well as data management, hence scientific interpretation data, analytical result data and derivatives were mixed ineffectively with raw data. Learning from this predicament, we skipped analytical tasks and concentrated purely on the curatorial functions in data management.

What data are worth storing and how can we make data qualitatively fit for storing?

We saw it was inefficient to store uncorrected and unproved raw data therefore we needed to define the principle unit of a data set worth archiving. Very early on it became evident that a data set has to be a publishable and citable entity described by substantial metadata to ensure data-reusability. With our customers (data providers and data users) we assigned a guideline: The original data set that we ingest into the repository should be retrievable as exactly the same fixed and defined unit—open accessibly and fit for re-use! Since data quality has become more of an issue we try to ensure reliability with a defined quality flagging system that depicts outliers, ranges and additional tests of variances. This is all part our plausibility check during data ingest into the information system.

How can we guarantee qualified repository services and true scientific reusability of data?

In the course of storing scientific data from all kinds of multidisciplinary scientific programs and publications PANGAEA became an agent for homogenization of analytical measurements assigned (by the scientific community) to define accepted parameters. These parameter definitions are crucial for data management and data storage. It needs assigned data repositories with trained scientific data curators to assure true scientific parameter homogenization. In terms of data quality, the data submitted originally are not ingested without question, but an assembled data set is sent back and forth between PANGAEA data curators and the author(s) until it is finally quality assured and validated by the responsible author (principle investigator). This is often a time-consuming and tedious task!

Consequently data set editors (scientific data curators) work in-house at PANGAEA—a data publishing system—since the semantic background and expertise has to be assured throughout the whole procedure. To encompass the whole life cycle of data from gathering to storing to reuse, we always operate best internally, within the scientific project itself, first to assure quality and second to assure financing via the same project. In this way, we keep the scientific status quo and we are well embedded in actual science. Normally we participate simultaneously in about 12 major international and national projects, besides the daily contact with our affiliated institutes' scientists or independent requests.

What are the financial aspects of data storing and sharing?

The idea that a data set has to be a publishable and citable entity described by substantial metadata was already appreciated by commercial publishers in 1994, but condemned for not providing a financial profit! Of course, a data archive with such a public assignment to the scientific community cannot work from a pure economic perception. Therefore, we have been cooperating with international publishers over the past 15 years. Our financial pillar is direct participation in scientific projects with the part of funding that recognizes the need of data archiving. But project-based data curation and storage alone does not cover the full cost. We need additional financial acknowledgment to develop future integrative data related e-infrastructures to cope with the exponentially increasing flood and complexity of data. These data are produced by data-intensive sciences that trigger and exploit improved sampling and high resolution sensor technologies. All this happens in international cooperative networks and, of course, everyone wants the data to be integrated, visible, accessible and reusable.

The original data model behind PANGAEA was developed in 1995. In principle it is still the same, but the middleware (the part that breaks down and reassembles the matrices), and the back and front end services had to be created from scratch and adapted continuously. These huge IT-development tasks are not yet fully appreciated by the scientific community or funding machinery.

How do you measure the success of PANGAEA?

PANGAEA is very well known globally in the Earth and Marine Environmental sciences. Our web statistics show tens of thousands of unique users per year, and, on average, nearly 500 datasets are downloaded per day. For the geoscientific and oceanographic community, PANGAEA is unique for its methods developed to handle multifarious interdisciplinary data. Besides data archiving, we deliver synoptic data views of projects for financial and scientific reviewers especially for EU-funded projects.
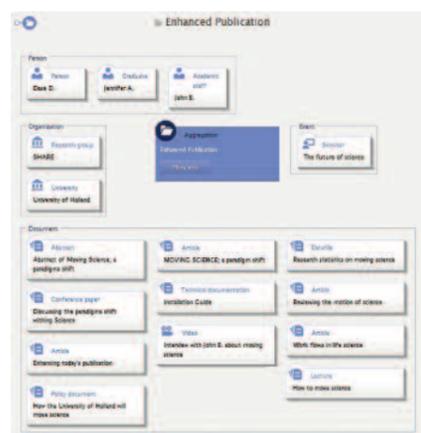
What is the central driver of PANGAEA?

Since our overall aim is focused on the meta-analysis of data (re-use!) we usually participate first hand in projects to cooperate directly with scientists to ensure top scientific quality. We also provide accredited citability and long-term preservation associated with persistent and globally resolved digital object identifiers. As a result we build up reputation and trust – the back bone of good scientific practice.

Exchanging Expertise in Enhanced Publications

The SURFfoundation unites Dutch research universities, universities of applied science, and research institutions. All of these collaborate on innovative projects to improve the quality of higher education and research. SURF acts as a funding body. It established the SURFshare programme, which supports various projects focused on research data. As SURF Project Coordinator responsible for Enhanced Publications, John Doove belongs to the Knowledge Exchange Working Group. Wilma Mossink is SURF Project Manager, Permanent Access to Data, and chairs the Dutch Research Data Forum.

Enhanced Publications (EP) is a core activity in the SURFshare programme. Development began during the DRIVER project, followed by calls for tender in 2008, 2009 and 2011 and now the projects range across disciplines, from the humanities to the hard sciences. The technical infrastructure is similar across the different disciplines, facilitating easy exchange of information across systems. It was very clear, right from the beginning of this model that different disciplines have different habits and needs, for example in archaeology and musicology. To serve these wide-ranging needs, we have customized tools in place, which support the individual workflows.

Currently we are upgrading the repository infrastructure to support the creation, storage, visualization and exchange of Enhanced Publications. We now have a common data model used in the development of customized tools required in the various EP projects (e.g. ESCAPE). Eventually all Enhanced Publications will be aggregated in Narcis, the open access portal for scientific output in the Netherlands.

Another focus of the SURFshare programme is permanent access to research data. SURF started with Enhanced Publications, but we quickly realized that they could not happen without proper data preservation and data access models and we needed to make more of an effort in these domains. That is how Data Preservation and Data Access became individual work packages, following the Treloar[1] silo model (2008) and collaborating closely with Enhanced Publication in SURFshare.

Licensing and related aspects play an important role in data access. We must understand the researchers' habits and needs to launch services that are truly valuable for their workflow. That is why one of reports we commissioned is on what researchers want from research data and it is also why we focus on close cooperation with researchers (e.g. the CARDS project).

> "There is more to share than just the article. Enhanced Publications could be a way to raise awareness of this fact."
>
> **John Doove and Wilma Mossink (SURFfoundation)**

Highlights and challenges in data sharing
One highlight is the Veteran Tapes project on multidisciplinary re-use of digital research files. It produced a quality research corpus of audio clips and transcripts of interviews with research veterans. We integrated the publication of this data in an e-book and the material is actively re-used across disciplines. The Veteran Tapes project was exceptionally successful in making valuable historical documentation and quality interview data available to the public, useable today and re-usable for future generations.

However, the advancement of data sharing remains a big challenge. Researchers seem to be scared of sharing data, they hesitate to publish it. This is a barrier for both national and international initiatives. We need to solve some hard questions: How do you convince researchers to publish their underlying research data? Under what conditions? One proposition could be 'open access where possible, closed when needed'. And what licenses should we have?

To solve these problems we need to exchange expertise in research data management on both the national and an international level. That's why the Dutch Research Data Forum was initiated, a national coalition currently consisting of 35 members. SURF is collaborating in many international initiatives, such as Knowledge Exchange, which has a dedicated group for research data. Data publication is on the way, but data are still not considered an independent contribution in scholarly communication. Data still do not count towards promotion or research assessments. The hesitation is apparent across disciplines. There is more to share than just the article. Enhanced publication could be a way to raise awareness of that fact.

It takes continuous development of infrastructures and services and this must always include specifying a discipline's needs because different publication cultures handle material differently. The successful EP model proves the possibility of having one technical data publication backend that can serve a variety of disciplines through specially adapted frontends.

> "It takes continuous development of infrastructures and services and this must always include specifying a discipline's needs because different publication cultures handle material differently."

John Doove
SURFfoundation

Wilma Mossink
SURFfoundation

Treloar, A., & Harboe-Ree, C. (2008). Data management and the curation continuum: how the Monash experience is informing repository relationships. Proceedings of VALA 2008. Retrieved from http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Steering towards Sustainable Data Sharing

Neil Holdsworth has headed the ICES Data Centre since 2007, ensuring that data strategy, policy and implemented business plans reflect the changing needs of the ICES user community. A key partner in the marine network, Holdsworth takes a lead role in setting international data standards. He has wide experience as a data systems analyst, working on making marine data more readily available to scientists and the public, and developing automated online systems to control the quality, validity and format of marine data. In 2008 he was assigned a member of the Marine Observation and Data Expert Group, MODEG advising the European Commission in Brussels.

The International Council for the Exploration of the Sea
Meeting at House of Lords, London, April 1929

> "R&D funding should not only produce immediate short-lived results, but should generate and steer sustainable integrated research efforts. This is still a tremendous task."
>
> **Neil Holdsworth (ICES)**

The International Council for the Exploration of the Sea (ICES) coordinates and promotes marine research on oceanography, the marine environment and ecosystem, and on living marine resources in the North Atlantic. Members of the ICES community include all coastal states bordering the North Atlantic and the Baltic Sea, with affiliate members in the Mediterranean Sea. ICES is a network of more than 1600 scientists from 200 institutes linked by an intergovernmental agreement (the ICES Convention, 1964) to add value to national research efforts and gather information about the marine ecosystem. This information is developed into unbiased, non-political advice. The 20 European and American member countries that fund and support ICES use this advice to help their governments and international regulatory bodies manage the North Atlantic Ocean and adjacent seas.

ICES maintains some of the world's largest databases on marine fisheries, oceanography, and the marine environment, and its Data Centre is part of a global network of distributed data centres. ICES operates an open access data policy adopted by the ICES Council in 2006, which conforms to the IOC Oceanographic Data Exchange Policy. ICES publishes its scientific information and advice in openly accessible reports, publications, its own Journal of Marine Science and on the ICES website.

What was the beginning of ICES – the initial sharing of information and data?

The beginning of ICES goes back to 1902 (Inaugural Meeting in Copenhagen), when a group of dedicated scientists decided to share information and data to know more about fish distribution,
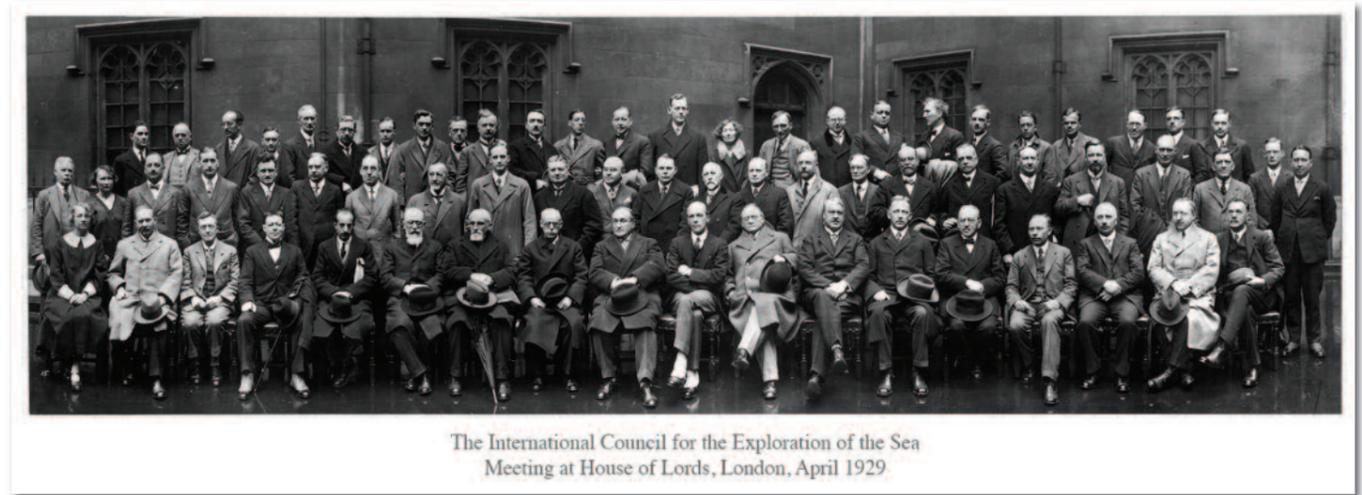
oceanography and the marine ecosystem beyond borders. The founding members were Denmark, Finland, Germany, the Netherlands, Norway, Sweden, Russia and the United Kingdom. The initial exchange of information and data was driven by scientists, not politics! It started with sharing fisheries' logbooks and landings, and with collecting information consistently over a period of time to make more information available, nowadays in digital format. The signing of the 1964 ICES Convention in Copenhagen is an official intergovernmental agreement that finally solidified ICES as an advisory board to add value to national research efforts.

What is the main barrier in sharing data internationally?

International guidelines are too complicated and impractical. People tend to follow traditional rules and standards based on national or federal regulations. These regulations are diverse, hence national conventions can limit the ability for international cooperative data sharing. But we cannot criticize national conventions for not being generally cooperative or homogenized on a European level since the main funding comes from dedicated funding of regional or nationally driven programs.

Why is ICES data sharing today not as good as it should be?

In the period leading up to the 1990's, scientific disciplines such as fisheries and physical oceanography to a degree still worked separately, since traditionally their data had particular uses unique to themselves. These disciplines grew side by side, but separately, in science as well as in

ICES. Biologists in particular are less advanced in wide-scale data sharing. They have a more regional, hence small scale, approach to their research compared to oceanographers or meteorologists. Biologists need to couple their investigations on a higher scale to tackle comprehensive global environmental problems.

Later on, with the new ecosystem approach, a fundamental need for integration and thus data sharing emerged. Different standards, guidelines and distinct traditions still exist today and need to be resolved. In the 1980s scientists and politics still did not meet on a practical level. But since the formation of OSPAR, HELCOM and in the context of the EU, integrated and cross-border environmental data are increasingly needed everywhere. We need more interdisciplinary working and standardization groups and education programmes.

What are the top five strategic barriers in data sharing today?

1. Protection of national interests, resources and political power are causing distinct barriers for international data sharing. National and regional competitiveness still exists. Often national funding interests overrule international integrative approaches and there is still a certain European north-south divide to overcome, not to mention the adaptation of Eastern Europe.

2. Another severe cause restricting open access to data are legal problems on national and international levels such as ownership, copyright and protection of once acquired possession. Slowly we are overcoming obstacles through international interdisciplinary committee work, for instance the Open Access policy adopted by the ICES Council in 2006 conforming to the IOC Oceanographic Data Exchange Policy.

3. The research side and the political advisory side did not develop adequate communication structures, resulting in an imbalance between scientific expertise and political decision-making and lack of cross-border information exchange and data sharing infrastructures. This is being addressed today by international expert groups and interdisciplinary commission work but the outcomes need to be realized more effectively.

4. In the wake of international and national integration programmes the burden of reporting and delivering of data has become huge. There are too many organisations which must be reported to. This seems to be caused by an overall steering problem.

5. National, regional or local funding does not consider international concerns adequately, although it should do so right from the beginning. R&D funding should not only produce immediate short-lived results, it should generate and steer sustainable integrated research efforts. This is still a tremendous task.

How is ICES helping to overcome these barriers?

ICES follows a top down and bottom up approach. On the one hand, we have intergovernmental and political alliances needing special integrated advice. ICES helps to answer their questions. On the other, the scientists in ICES working groups bring up new questions and solutions across disciplines and interact with other groups. In ICES both parties find a meeting and communication platform.
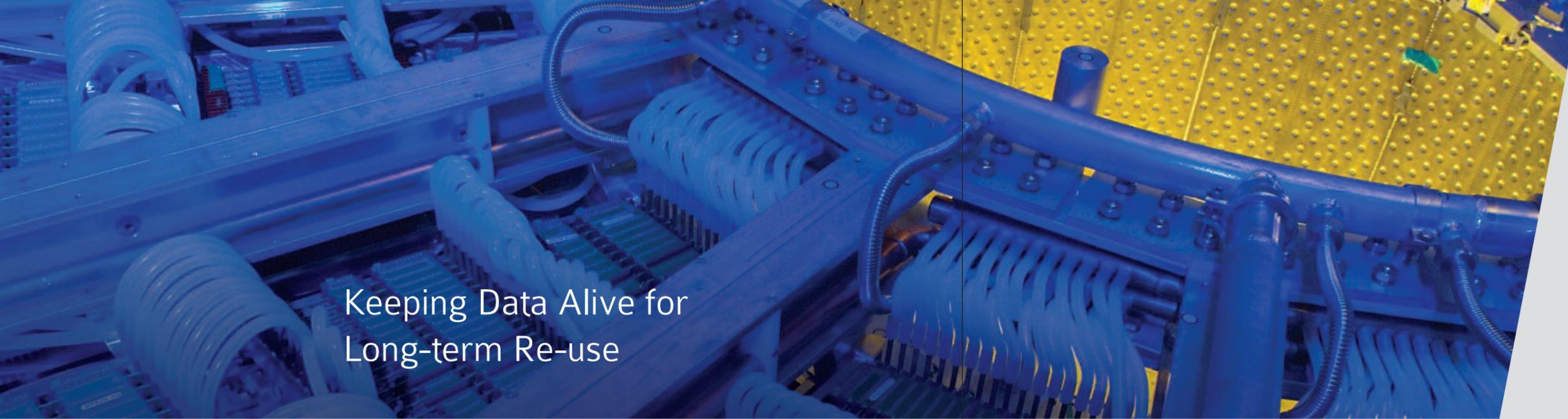
Neil Holdsworth
ICES Data Centre

Interviewer:
Angela Schäfer

Keeping Data Alive for Long-term Re-use

Peter Igo-Kemenes
European Organization
for Nuclear Research
– CERN

Interviewer:
Sünje Dallmeier-Tiessen

> "Keeping data alive is a huge load and it is unlikely that over the long term, experiments alone can provide for this from their research budgets."
>
> **Peter Igo-Kemenes (European Organization for Nuclear Research – CERN)**

Peter Igo-Kemenes, of Hungarian origin, holds a PhD in physics from the University of Leuven (Belgium). After initial positions at Heidelberg University and CERN, he spent two years at Columbia University, then returned to Heidelberg and joined the OPAL experiment on the LEP collider at CERN (pre-cursor to LHC) where he spent the larger part of his scientific career. In the mid-1990s he became the leader of the LEP Higgs Working Group, with the mandate to combine the data of the four big LEP collaborations (ALEPH, DELPHI, L3 and OPAL) in the Higgs boson search. Currently Peter Igo-Kemenes is a professor at Gjøvik University College and advises CERN in questions of open access publishing and long-term data preservation. Recently he participated in two FP7 projects: PARSE.Insight (Permanent Access to the Records of Science in Europe) and SOAP (Study of Open Access Publishing) and helped lay down the foundations of the SCOAP3 project (Sponsoring Consortium for Open Access Publishing in Particle Physics).

The LEP Higgs Working Group was mandated to statistically combine the data from four large-scale experiments with the aim of improving overall sensitivity in the search for the Higgs boson. The 10-year enterprise resulted in several essential publications that mark the end of the LEP era.

Success stories in data exchange

Although LEP ended in 2000, the data have been kept alive, together with the analysis software, and are currently reformatted and stored such that they can be re-used in combination with future search data. The data will be published soon on INSPIRE. Re-analysis is anticipated in the near future, in combination with similar data from the Tevatron accelerator experiments (Fermilab/USA) which will tie up with the subject where LEP left it. Increasing interest in the LEP data can also be anticipated from the LHC experiments which are in their start-up phase.

Another success story is the combined analysis of two datasets produced by two experiments separated by about 20 years. The data were used in a single analysis to determine the energy dependence of a fundamental physical parameter which determines the strength of the 'strong' or nuclear interaction. The results from the JADE experiment at DESY in Hamburg (finished in the early 1980s) were used for the low energy part and the results from the OPAL experiment (LEP, CERN, finished in the year 2000) for the high energy part. During the JADE measurement there was no effort at all to conserve data to make it re-useable for such combined analysis. The success of the combined analysis relied on the dedication of two people from JADE who painstakingly studied old logbooks and computer printouts to revive the JADE data. They eventually became members of the OPAL collaboration for the purpose of the combined analysis. Their archaeological work took several years but the resulting publication became a fundamental document on the subject.

Obstacles to data exchange and preservation

Sociological aspects: the environment of concurrently running similar experiments can be a precarious balance between competition and cooperation. This was indeed the case in the LEP Higgs Working Group consisting of members from the four LEP experiments. Concurrent experiments do not put down all their cards, just the minimum that is necessary to fulfil the common task. Sometimes this is in conflict with the full insight that is needed to produce reliable combined results. Such conflicts will certainly continue when it comes to combining data in the future.

Data preservation: rapidly changing technology is a challenge. For example, the stored LEP data cannot be re-run on currently existing computing platforms without a major revival effort. In general, old hardware and software soon becomes outdated or unreadable. Migration to new platforms and virtualization of the software are some of the efforts that have to be invested in for long-term preservation and re-use.

Conservation of internal knowledge of experimental details: without this it is very hard to analyze old data. Detailed documentation needs to accompany the data. There is a balance to be struck between the levels of detail of the data offered for conservation. On the one hand, a fine granularity of the data requires more detailed knowledge of the exact meaning. On the other hand, a coarser granularity imposes severe limitations on the possibilities of re-use. For HEP experiments, dealing with very complex data, internal knowledge will always be necessary. Although the LEP Higgs data will be open access (with accompanying documentation), one should seek the expert knowledge of former LEP collaboration members, as long as they are available, for successful re-analysis.

Lessons learned

The LEP experiments, which ended in 2000, did not invest in the necessary effort to allow data to be conserved on a large scale for possible re-use. As a result, re-analysis will be limited to very specific domains. Thus far, almost no data preservation took place during the lifetime of experiments which implies a great amount of (sometimes manual) work to revive the data. The JADE/OPAL effort is an illustrative example. To prevent this from happening again, experiments worldwide should try to invest in this effort. Today, an important initiative comes from the Study Group for Data Preservation and Long-Term Analysis in High Energy Physics (DPHEP) gathering some major HEP experiments that have finished data collection (e.g. the Tevatron CDF and D0 experiments, experiments at DESY Hamburg; BaBar at SLAC/US, Belle at KEK/Japan). These and the current LHC experiments might represent the last generation of their kind. Ensuring the possibility of re-using their data at a later stage may therefore become vital.

An important aspect of data preservation is the fact that within the lifetime of an experiment one never fully exploits the data and only the future can tell us what has been overlooked. New theories for example can generate new interest in old data. The effort within DPHEP is aiming at developing standards, methods and common technologies for data preservation, specifically for HEP. However, DPHEP is interacting with astrophysics where some standards for data exchange are already in place. HEP can learn from astrophysics even though the levels of complexity are not comparable.

The size of the effort of conserving HEP data should not be underestimated and neither should the financial requirements involved. Keeping data alive is a huge load and it is unlikely that, over the long term, experiments alone can provide for this from their research budgets.

Future prospects

It is important to keep in mind that HEP is an exceptional field of science due to the huge size and complexity of the data output. The lessons learned from the past should be taken into account. Data preservation should not be relegated to the very end of the HEP experiments` lifetimes but should be regarded as a parallel effort while the experiments are alive and producing data. The awareness of the problem is already building up within the HEP community but the actions are lagging behind. Good sign: the LHC experiments are currently joining the DPHEP effort.

Establishing a Collaborative Climate for Sharing

Peter Lemke heads the Climate Sciences Division at the Alfred Wegener Institute and is also professor of the Physics of Atmosphere and Ocean at the Institute of Environmental Physics at Bremen University. He has been working on the observation and modelling of climate processes since the mid-1970s, on the interaction between the atmosphere, sea ice and the oceans. He has been on seven polar expeditions, mostly as chief scientist. An active member of the Joint Scientific Committee for the World Climate Research Program (WCRP) 1995–2006, the highest international committee for climate research, Lemke acted as its chair for six years. Now he heads REKLIM, the climate initiative of the Helmholtz Association, in which eight research centres are collaborating on data sharing and model development. Lemke was instrumental in preparing the World Climate Report of the Intergovernmental Panel on Climate Change (IPCC), which was awarded the Nobel Peace Prize in 2007. In June 2010 he was announced as one of the experts for IPCC's Fifth Assessment Report, as Review Editor responsible for the chapter on Earth's cryosphere.

> "The long established data sharing of the national weather services provided the basis for global climate research."
>
> **Peter Lemke (Alfred Wegener Institute for Polar and Marine Research)**

In the meteorological community, data sharing started in 1873 with the beginning of international coordination in weather forecasts by the International Meteorological Organization. Other disciplines in the environmental sciences started data sharing processes with the first Geophysical Year Assembly of 1957/58.

For me personally, data sharing started with my doctoral thesis, for which I had to digitize analogue paper maps (sea ice charts). After completing the task, I submitted the data set to the World Data Centre for Glaciology in Boulder (USA), for use by the wider scientific community. We were taking part in the World Climate Programme implemented by the World Meteorological Organization (WMO) (according to convention by the International Council for Science, ICSU). By 1979 my supervisor was urging me to feed our data into this World Data Centre (WDC), not in the least because of our deep integration in this international programme.

Right from the start, the WCP data sharing endeavour turned out very good at stimulating collaborative science. The WDC glaciology repository digests huge amounts of relevant data globally for research and meteorological services, from ESA and NASA as well. Even NASA is a declared principal data investor in the WDC. Most of the data is open access.

Was data sharing essential in preparing the IPCC report?

The mission of the Intergovernmental Panel on Climate Change (IPCC) is to determine at regular intervals the state of the climate system and its impact on ecosystems and human society and to point out potential political countermeasures. The IPCC was instituted by the WMO and the United Nations Environment Program (UNEP) in 1988 when the possibility of global climate change became evident. The IPCC does not conduct its own research, nor does it provide data. Hence, to prepare the IPCC report, we did not request data directly, and if at all only by means of control or adjustment. Mostly we compiled relevant scientific evidence for comprehensive analysis. The IPCC assessment is mainly based on peer-reviewed, published scientific and technical literature, which is evaluated in a thorough, objective, free and transparent manner.

What kinds of data sharing have you encountered in climate research?

Weather forecasting data have been shared as an imperative necessity for some 150 years: we need to prepare for any weather phenomena in time and of course weather is not constrained by national borders. Very early on, people learnt that it is important to know the weather upwind of London to predict the next day's weather in Hamburg.

Data sharing works basically in these circumstances, because we have had regular fast communication by telegraph (to begin with) ever since the first worldwide operating meteorological service was established. Since meteorological data are naturally distributed worldwide, a centralized weather forecast system was inevitable. The International Meteorological Organization (IMO) lasted from 1873 until it was succeeded by the now well established WMO in 1950. In this field, global data sets are compiled and distributed constantly. So data sharing in meteorology has a long-standing tradition since weather data has been exchanged worldwide taking advantage of the emerging global communication techniques. It works very well compared to other disciplines.

In contrast, experience shows that barrier-free access to hydrological data, for instance, is still causing huge problems. These data are needed to relate collected ground truth data with remote satellite data for evaluation and modelling, especially for disaster risk reduction. Actual hydrological data are subject to state and national administration. If you gain access to these data at all, it is years later, because they are of national strategic importance (resources, agriculture) and are therefore restricted. In this field, international open access data release does not seem to be possible.

International data exchange

However, free access to data from the international World Climate Research Program (WCRP) is the normal case since its establishment in 1980. This very successful program is funded by the WMO, the ICSU and the Intergovernmental Oceanographic Commission of UNESCO. It supports progress in the prediction capabilities of operational centres in extended weather and seasonal forecasts as well as longer-term variability and climate-change projections. Scientists organized in the WCRP provide a major part of the scientific material assessed by the IPCC in its advice to the UN Framework Convention on Climate Change. These activities form the scientific basis for adaptation to climate change and for developing mitigation strategies that are eventually implemented on international and regional levels.

Despite being very well organized internationally, the WCRP does not have its own research money or its own funding. But the program has turned out well as a working platform for meetings and for international data exchange. For example, WOCE (World Ocean Circulation Experiment) was a very successful project, especially in terms of data sharing, as it implemented international databases and created substantial digital world atlases.

But insufficient access to local data for soil moisture, discharge, and suchlike on either the national or the international scale makes essential adjustments to meteorological research models for hydrological data collected on location barely possible. We urgently need to couple global meteorological data with regional hydrological ground truth data to run realistic climate models and predictions, not just for the IPCC report, but also for the bigger picture of worldwide climate research.

Ensuring quality control for data re-usability
In meteorological and climate research, metadata are very important, and generating them always implies great effort. Generally this works out well for the World Data Centre. The German BSH (Federal Maritime and Hydrographic Agency), for example, is also well positioned. At the National Snow and Ice Data Centre (NSIDC) in Boulder, re-usability through appropriate metadata handling works out well.

However, the great effort of handling diverse calibration methods and standard verification procedures hampers data re-usability in meteorology. Even nowadays this is still causing problems for data archiving. It is essential that only suitably expert climate institutions specialize in homogenizing and archiving climate data. For example, the WMO reprocesses and converts historical data to current standards. This organization has the scientific specialists who can interpret these historical data properly and implement international standards.

Quality control, starting right with the individual field measurements, is indispensable for data re-usability. When we were compiling the IPCC report, we noticed data offsets while aligning data from diverse measurement devices. Another

example is overlap. Sensor ranges from more than 20 satellite operators have to be managed and their data needs to be calibrated constantly and with each new satellite sensor. And then the standardization of weather, water and climate data and metadata is essential to ensure orderly and efficient share and use of the information between WMO members, from provider to user. Hence task-expert teams develop and maintain the relevant standards, and develop guidance for their implementation.

Developing the data sharing ethos
The early installation and improvement of WMO's Global Telecommunication System (GTS) has enabled worldwide usage of all weather service data. It plays a vital role in facilitating the flow of data and processed products to meet requirements in a timely, reliable and cost-effective way, ensuring access to all meteorological and related data, forecasts and alerts. This secured communication network enables real-time exchange of information, which is critical for forecasting and warnings of hydro-meteorological hazards.

Since meteorologists are strongly involved in global joint projects, they are a priori interested in sharing knowledge and data. Otherwise weather and climate research would hardly be possible. Hence data sharing became an implicit commitment, even without personal control. A data sharing ethos developed very early due to the instantaneous need for action preceding natural weather hazards. And of course prediction of any weather condition implies global information and data exchange. In summary, the long established data sharing of national weather services provided the basis for global climate research.

"Actual hydrological data are subject to state and national administration. If you gain access to these data at all, it is years later, because they are of national strategic importance (resources, agriculture) and are therefore restricted. In this field, international open access data release does not seem to be possible."

The Astronomical Importance of Discoverability

Carolin Liefke (1981) has been fascinated by the night sky ever since she was 13 years old. After studying physics at Hamburg University, specializing in astronomy, she worked on stellar activity and X-ray astronomy at the Hamburger Sternwarte for her PhD. An enthusiastic amateur astronomer, in March 2010 Carolin turned her passion into a profession and joined the Haus der Astronomie, Heidelberg's centre for astronomy education and outreach. She maintains the German version of Galaxy Zoo and other citizen projects in the Zooniverse, where large amounts of scientific data are handed over to laymen for special analysis, tasks that require a human brain to solve, such as classifying galaxies, searching for exoplanet transits, or finding unknown asteroids.

What is your personal experience of research data?
While studying physics, I coded tools for data re-use. I'm a real research data re-user and I've searched for and integrated lots of existing research into my projects. Although data sharing is well advanced, I have encountered problems with discovering data. In some cases I only found out later (after a project had ended) about other datasets that could have contributed to my findings. Some of my research could have been improved or accelerated by better data discoverability. I've heard similar stories from friends and colleagues and so I'm glad that the challenge of discoverability is now being worked on by the Virtual Observatories (VO) initiative.

It's also important to address the definition of data sharing. At Galaxy Zoo, data sharing is limited in the sense that participants do not play an active role in the sharing process. They are presented with pre-processed data and a very special task.
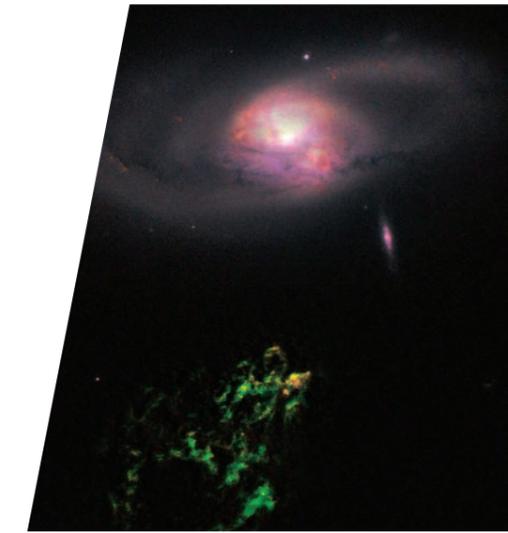
However the raw data the project is based on are shared among the scientific community.

What are your views on data sharing in astronomy in general
There is lots of data sharing in the dynamic field of astronomy. Research information is handled very openly. Data management is usually run by the institutions. In the first year after its production, access is limited to the researchers who proposed and participated in the particular project, but after that, the data becomes open access. The challenge lies not so much in data preservation, but rather in discoverability. The ongoing VO initiative will facilitate easier data discoverability, more sophisticated data mining, and more complex automated analysis.

What are the major challenges in your opinion?
One major challenge is lost data, or data that appears to be lost, and that is being tackled by the VO project. VO is also taking care of old datasets

from projects which have finished, preserving and making them available via their interfaces. The major challenge for the coming years is data management, presenting huge projects, and along with that managing the data deluge. The latter usually requires advanced automated processing and selection for the data archive.

"The ongoing VO initiative will facilitate easier data discoverability, more sophisticated data mining, and more complex automated analysis."
Carolin Liefke (Galaxy Zoo, Heidelberg)

Peter Lemke
Alfred Wegener
Institute for Polar
and Marine Research

Interviewer:
Angela Schäfer

Carolin Liefke
Galaxy Zoo, Heidelberg

Interviewer:
Sünje Dallmeier-Tiessen

Image top: Courtesy of NASA and ESA

Setting Course for a Data Sharing Culture

"Without the infrastructure that helps scientists manage their data in a convenient and efficient way, no culture of data sharing will evolve."

Stefan Winkler-Nees
(Deutsche Forschungsgemeinschaft)

Stefan Winkler-Nees is Programme Officer at the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The DFG is the central, self-governing research funding organization in Germany. In the Scientific Library Services and Information Systems unit, Stefan Winkler-Nees is responsible for strategic activities in the context of permanent access to research data. His background is in marine geosciences and climate research. After completing a number of postdoc projects in Europe and overseas, he worked for a software company. Six years later he returned to science with a position at the DFG.

The DFG Committee for Scientific Library Services and Information Systems released a seminal position paper in 2006 on the future of digital information. Since then, the DFG has been developing strategy for the means and measures to improve the data management in the future. A major topic in information management is 'developing infrastructures for primary research data provision'.

The committee has been examining the challenges and opportunities of this topic in a series of discipline-specific round-table discussions. Stefan Winkler-Nees reports, "The feedback from representatives of the different disciplines varies to a large degree. I find it interesting to see that not just the STM disciplines see the relevance of data sharing." A main finding from the interdisciplinary consultations is that funding for substantial infrastructure has to grow with the implementation of policy activities. Winkler-Nees says, "Without the infrastructure to assist scientists to manage their data conveniently and efficiently, no culture of data sharing will evolve."

Common strategy
Winkler-Nees finds consultation with scientists and infrastructure representatives of prime importance. "As a research funding organization, we have to promote dialogue between the actors, especially in this regard." The DFG has begun improving the dialogue with other funders, cooperating with partners in the European Knowledge Exchange Initiative on joint strategies in the field of research data management. Winkler-Nees explains, "In disciplines working internationally, we must develop common strategies on data sharing." He points out that permanent access to scientific data is a challenge across all disciplines without exception.

In 2009 the DFG sub-committee on information management published Recommendations for the secure storage and availability of digital primary research data. One recommendation states, "Every scientist shall make his primary research data freely available beyond his institution whenever possible." This paradigm, with respect to disciplinary particularities, is guiding the DFG's activities.

Since 2008, the DFG has also been part of a leading digital information initiative by the Alliance of German Science Organizations that has agreed to coordinate activities to ensure the long-term availability and integration of digital information into virtual research environments. The partner organizations agreed to align their funding programmes in the area of research data and, when necessary and appropriate, to merge or harmonize them.

Winkler-Nees says, "Promoting data sharing is such an important task that cooperation and communication between all stakeholders plays a major role." In 2010, the Alliance released Principles for Handling Research Data. In this document, the partner organizations declare their support for "open access to data from publicly funded research", adds Winkler-Nees. "However, this kind of policy paper needs to be followed up by the development of an appropriate infrastructure that will support the implementation of the data sharing culture. And to achieve this, working cooperation between scientists, librarians, IT and information management specialists is essential. In principle all research data management services must be adapted to scientific requirements, but they also need to be operated with the necessary information management expertise."

Close cooperation
Linking scientists with infrastructure professionals is part of the DFG's strategy for research data management. Winkler-Nees says, "Close cooperation between our LIS unit and all the different disciplinary units helps us to promote the dialogue on the opportunities and challenges of data sharing." This communication has also had a positive influence on awareness of this topic. Internal discussions have increased attention for the significance of data sharing.

In April 2010 the DFG published revised Guidelines for Proposals. Now applicants must indicate what measures they plan to both secure collected data and facilitate its re-use. This requirement is intended to encourage applicants to share their data and to raise a general awareness to this issue. "Due to the diversity of disciplines, we have decided to take small but effective steps. Some disciplines, such as the geosciences, are already demanding further steps, such as mandatory data management plans. But we must take the needs of all disciplines into account." Winkler-Nees emphasises the need to avoid data bureaucracy. "We mustn't forget those disciplines where data sharing is impossible or difficult due to legal aspects."

To encourage the development of infrastructure, in 2010 the DFG released a Call for Proposals on information infrastructures for research data, which generated enormous interest in a wide variety of disciplines. Winkler-Nees notes, "The huge number of applications shows the importance of the topic. All the proposals were reviewed by infrastructure experts and by scientists, to ensure their relevance to the related discipline. Now, with funding of 9.9 million Euros for 28 infrastructure projects, we hope to facilitate research data sharing and set a foundation stone for the future infrastructure of scientific information."

International framework
In future, the DFG will promote data sharing in the international framework. A joint statement by a group of major international funders of public health research serves as an example. True to their motto 'Sharing research data to improve public health', in January 2011 17 signatories, including major public funding agencies, charitable foundations and international organizations committed to cooperate increasing the availability of data emerging from funded research.

"From the perspective of the scientists working internationally and particularly at European level, we have to set the course for a culture of data sharing. We have to adjust our activities with other significant stakeholders." says Winkler-Nees. "We have to develop the financial and legal frameworks for national activities by working together with all the relevant partners and organizations. And while we are funding data sharing infrastructures, such as repositories, we have to develop sustainable funding solutions. That is the challenge."

"True to their motto 'Sharing research data to improve public health', in January 2011 17 signatories, including major public funding agencies, charitable foundations and international organizations committed to cooperate increasing the availability of data emerging from funded research."

Stefan Winkler-Nees
Deutsche Forschungsgemeinschaft

Interviewer:
Heinz Pampel

Convincing Incentives for Sharing Data

Heather Piwowar is a researcher associated with the DataONE and Dryad projects at the National Evolutionary Synthesis Centre (NESCent). Her PhD focused on biomedical data sharing and now she is very much interested in patterns of data re-use.

"If researchers could see that they are cited and attributed for their data publication and that their sharing is considered in their promotion committees, this would be an important incentive."

Heather Piwowar (NESCent)

My interest began when I tried to re-use some data for a project and failed to find what I was looking for. So then I started to study data sharing and so far I've mainly focused on gene expression microarray data. I chose it because the data sharing standards and infrastructure in this discipline were already well established, but the archiving practices were not yet universal. It is an interesting datatype, perhaps more typical of investigator-driven research than the data stored and handled, for instance, via GenBank. Gene expression microarray data are collected under a range of experimental conditions, on a variety of incompatible platforms, and undergo variable processing steps. I have been studying data sharing over time. The proportion of gene expression microarray datasets that have been deposited into public archives increased between 2000 and 2009, but the rates seem to be plateauing at about 45%. My analysis suggests that the NIH policy requiring a data management plan for large grants is not associated with an increase in public data archiving.

One project I am working with is Dryad, a data repository that accepts data of all formats associated with published research in biology. To serve more communities and to facilitate easy re-use of the materials, it will 'handshake' with the main existing databases in biology such as GenBank. One of the important issues is sustainability. The project is funded under an NSF grant; we need to be financially sustainable beyond the end date. Quite likely Dryad will begin charging journals for data submission.

Citation benefit associated with data sharing
My co-authors and I investigated journal data policies and practices in the environmental sciences. We looked at 500 articles across six journals and saw that data availability policies are rarely articulated or standardized. This research suggests one barrier – the lack of data policies on data availability – as journal policy is strongly correlated with data sharing behaviour. But I think the key barrier is the researcher's hesitation to share their material. Now there is lots of guesswork, few real numbers are available, and to convince researchers to begin sharing I think it is important to show them some compelling numbers. For instance, our study of 85 papers showed that publishing openly available research data was associated with a citation benefit of 70%.

Interestingly, researchers who have already shared their data once are more likely to share their data again. Researchers who publish in open access journals are also more likely to share their data. If all researchers could see that they are cited and attributed for their data publication and that their sharing is considered in their promotion committees this would make an important incentive. Thus one of the core activities in coming years should be the provision of evidence of research data re-use.

Heather Piwowar
NESCent

Interviewer:
Sünje Dallmeier-Tiessen

Image left: Courtesy of Steve Jurvetson
Published under CC-BY

Glossary

APA Alliance for Permanent Access

AWI Alfred Wegener Institute for Polar and Marine Research

BSH Federal Maritime and Hydrographic Agency

CARDDS project: http://www.surffoundation.nl/en/projecten/Pages/CARDS.aspx

CERN European Organization for Nuclear Research

DDBJ DNA Data Bank of Japan

DESY Deutsches Elektronen-Synchrotron

DFG German Research Foundation

DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

Dryad www.datadryad.org

EBI European Bioinformatics Institute

ELIXIR European life science infrastructure for biological information

EMBL European Molecular Biology Laboratory

EP Enhanced Publications

ESA European Space Agency

ESCAPE project: http://www.surffoundation.nl/en/projecten/Pages/ESCAPE-Enhanced-Scientific-Communication-by-Aggregated-Publications-Environments.aspx

ESRC Economic and Social Research Council

Genbank: http://www.ncbi.nlm.nih.gov/genbank/

GTS Global Telecommunication System

HELCOM Helsinki Commission

HEP High-Energy Physics

ICES International Council for the Exploration of the Sea

ICSU International Council for Science

IMO International Meteorological Organization

INSDC International Nucleotide Sequence Database Collaboration

IOC Intergovernmental Oceanographic Commission

IPCC Intergovernmental Panel on Climate Change

JISC Joint Information Systems Committee

LEP Large Electron Positron Collider (at CERN)

LHC Large Hadron Collider (at CERN)

LIS Library and Information Science

MARUM Centre for Marine Environmental Sciences

MODEG Marine Observation and Data Expert Group

Narcis: www.narcis.nl

NASA National Aeronautics and Space Administration

NESCent National Evolutionary Synthesis Centre

NSF: National Science Foundation (US)

NSIDC National Snow and Ice Data Centre

ODE Opportunities for Data Exchange Project

OSPAR Convention for the Protection of the Marine Environment of the North-East Atlantic

PANGAEA Publishing Network for Geoscientific and Environmental Data

REKLIM Helmholtz Climate Initiative Regional Climate Change

SCOAP3 project: Sponsoring Consortium for Open Access Publishing in Particle Physics

SOAP: Study of Open Access Publishing

SSH Social Sciences and Humanities

STM Science, Technology, Medicine

SURF: www.surf.nl

UNESCO United Nations Educational, Scientific and Cultural Organization

VO Virtual Observatories

WCRP World Climate Research Program

WDC World Data Centre

WDC-MARE World Data Centre for Marine Environmental Sciences

WOCE World Ocean Circulation Experiment

WMO World Meteorological Organization

Acknowledgements