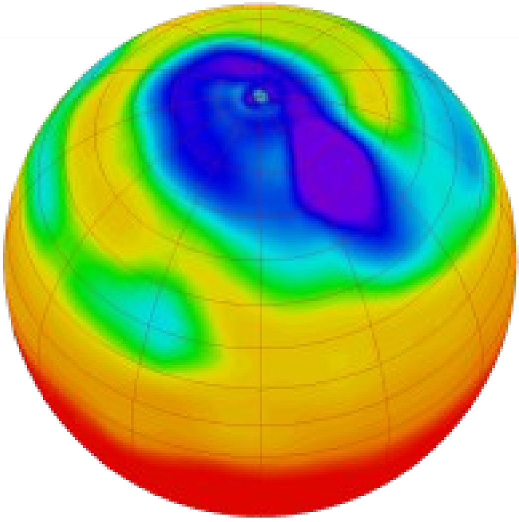


STFC
Centre for Environmental Data Archival (CEDA)
Annual Report 2012
(April 2011-March 2012)

CEDA delivers the

British Atmospheric Data Centre

for the National Centre for Atmospheric Science

and the

NERC Earth Observation Data Centre

for the National Centre for Earth Observation

and the

IPCC Data Distribution Centre

for the IPCC

Introduction from the Director

The mission of the Centre for Environmental Archival (CEDA) is to deliver long term curation of scientifically important environmental data at the same time as facilitating the use of data by the environmental science community. CEDA was formed to host two of the Natural Environment Research Council (NERC) designated data centres: the British Atmospheric Data Centre and the NERC Earth Observation Data Centre, as well as the UK arm of the IPCC Data Distribution Centre. In 2011, the UK Solar System Data Centre joined CEDA. Here we present the fourth annual report, covering joint activities from April 2011 to March 2012 (previously the constituent centres reported independently).

The report itself is in two sections, the first broadly providing a summary of activities and some statistics, and the second a selection of short reports on some specific activities beginning, under way, or completed. This section is intended to provide a taster for the range of activities that CEDA undertakes, rather than a complete report of activities, since CEDA staff are involved in a huge range of scientific and informatics projects, not all of which are appropriate for reporting here. CEDA continues to engage in informatics projects to help improve the provision of: (1) suitable tools to document and manage both high volume and highly heterogeneous data; (2)tooling and services to enable the community to exploit CEDA data holdings, and; (3) fundamental standards. The latter, both to improve the likelihood that others can build standards compliant software we can deploy, and to support interdisciplinary science.

As in the previous year, the 2011/2012 year was dominated by the two major challenges of dealing with CMIP5 (e.g. see page 36) and the establishment of new services under the banner of the International Space Innovation Centre (discussed in the articles on CEMS on pages 27 and 28). However, while those were high profile external activities, issues of scale became dominant internally; the funding report on page 14 summarises some of the issues: of the order of 10^8 files – o(10^8) – using o(petabytes) of disk, on o(300) different computers, split into o(600) datasets on o(100) disk partitions – without a consistent metadata standard or file format across the archive. Despite a decade of effort on metadata systems, and what had been a very efficient computing environment, CEDA was beginning to creak at the seams – with disk failures, insufficient documentation, and complex network issues becoming more and more prevalent. Ongoing growth using the same technical environment would have been a problem.

Fortunately, in late 2011, CEDA received significant capital investment, culminating in the delivery in March 2012 of a new computing system – JASMIN/CEMS – consisting of storage and compute funded both by NERC and UKSA and delivered by CEDA in what was then the e-Science department in STFC (now part of the Scientific Computing Department). JASMIN is discussed on page 26 and CEMS on pages 27 and 28. JASMIN/CEMS are not just about supporting the traditional archival services of CEDA though – they are intended to additionally provide support for high performance analysis of high volume data by the greater NERC scientific community. The physical delivery of these systems is of course just part of the story, in next years annual report we will be discussing the difficulty of migrating data to the new environment, and some of the new services which their advent has engendered.

While we expect the physical system issues to be resolved with the new hardware, issues of documentation still exist – both in terms of the content, and how it is organised. CEDA continues to invest, with both core and project funding, in new metadata developments, aiming to address both issues. Work on data publication and citation is intended both to improve the integrity of the scholarly record, and to provide incentives for the production of good documentation, and work on metadata standards to ensure that we have the information organised fit for automating our environment and scientific use! Many of the one page reports discuss projects in this arena.

I trust that whatever your background, you find something of interest in the material presented here.

Bryan Lawrence, Director

Table of Contents

Summary of 2011/2012

CEDA continues to support the environmental science community in the UK and abroad through the provision of data management and discovery services, and has continued to develop tools and services to aid data preservation, curation, discovery and visualisation.

In this year CEDA delivered in excess of 281 TB of data in over 27 million files from 198 datasets to 3723 distinct users.

CEDA continues to support the fifth Coupled Model Intercomparison Project (CMIP5) through its interactions with the large global collaboration to deliver an "Earth System Grid Federation" under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP). The major international collaboration built around the EU funded Metafor project (which ended this year) has continued to support the community in the form of PIMMS (see page 32) and the CMIP5 questionnaire support system, and work continues on the IS-ENES project to support and develop infrastructure for earth sciences. CEDA staff continue to take leading roles in these and many other initiatives.

Notable Activity

1. The International Space Innovation Centre (ISIC) was launched in May 2011. It drives innovation and enterprise, creating new technologies and developing applications and intellectual property for the benefit of the UK. ISIC is a not-for-profit organisation based in the UK formed between industry, academia and government, with CEDA as a leading academic partner.
2. Esther Conway joined the CEDA science support team and also continues to work on data preservation projects LTDP (Long Term Data Preservation, funded by ESA) and the EU Framework 7 project Scidip-ES (SCIence Data Infrastructure for Preservation – with focus on Earth Science).
3. This year has seen the first assignment of digital object identifiers (DOIs) to datasets held in the CEDA archives, and the development of guidelines and procedures to enable DOIs to be assigned to datasets held in other NERC data centres. Collaborations have continued with such groups as the CODATA-ICSTI Data Citation Task Group, the DataCite Working Group on Criteria for Data centres and SCOR IODE MBLWHOI Library Data Publication Working Group. This year has also seen the launch of a new data journal, Geoscience Data Journal (GDJ), which is a collaboration between Wiley-Blackwell and the Royal Meteorological Society. Sarah Callaghan is an associate editor with special brief for data centre interactions in GDJ.
4. Scientists working at the BADC, NCAS CMS and NCAS Climate have continued to work together to support the web based questionnaire (http://q.cmip5.ceda.ac.uk/) being used to collect information and metadata from the climate modelling groups submitting data for the next Coupled Model Inter-comparison Project Phase 5 (CMIP5). This climate model data will form the basis of the next Intergovernmental Panel on Climate Change Assessment Report (AR5), due in 2013. The questionnaire gathers information about the details of the climate models used, how the simulations were carried out, how the models conformed to the CMIP5 experiment requirements and details of the hardware used to perform the simulations. The CMIP5 model documentation will provide the most comprehensive metadata of any climate model intercomparison project.
5. CEDA continues to run the Discovery Web Service that underpins the NERC Data Catalogue Service. This process involves the continual harvesting of metadata from NERC centres and the ingestion into the underlying catalogue. CEDA provides statistical info to NERC on the number of records present. CEDA has also been involved in discussing a likely successor to the current set-up to provide a more standards compliant system that will be compatible with the requirement for NERC to publish its records to the UK Location Portal as part of their compliance with the EU INSPIRE directive.

6. CEDA continues to harvest metadata and run a Discovery Web Service for the MEDIN (Marine Environment Data Information Network). CEDA has developed and deployed a working prototype updated service and is working in collaboration with GeoData Southampton to ensure this new service can proceed to operational status. CEDA is also in the process of deploying an operational CSW (Catalogue Services for the Web) on behalf of MEDIN to serve as a single point to upload MEDIN records to the UK Location Portal to meet the MEDIN INSPIRE obligations.

Download and Help Desk Statistics

The CEDA Helpdesk (BADC, NEODC and UKSSDC user support) includes responding to user queries and handling of electronic application forms for access to restricted data. 93% of user queries are handled by 2 CEDA Data Scientists while the 7% remaining are covered by other CEDA team members as necessary. An improved data extraction service with supporting documentation for the most popular BADC datasets has lead to a notable reduction in queries received (down by 20.4% on anticipated levels when compared to 2010/11 queries per active user figures.)

CEDA continues to provide prompt and effective support services to the user community at a high priority level, recognised by CEDA users as this selection of unsolicited complements demonstrate:

"Thank you for the very swift solution! I must say that I am delighted with the quality of your support!!" "I just would like to thank you for your great assistance. This really helps my research." "Thanks very much for your prompt and positive reply!"

"I want to give you the positive feedback, since I appreciated very much your quick and rapid action to fix the problems!" "I'd like to thank you for your quick, detailed and very focused reply, things are much clearer now!" "Thousand times thanks for this, you helped me a lot, I am so happy now :) " "Heartily appreciate for your dedicated work and advice." "I don't know the official channels to route this comment to, but if you could pass that I have found the BADC data, website and helpdesk invaluable for my research and have been really impressed with all aspects, especially the customer service I have received when I have emailed for advice."

While 61% of BADC and NEODC users and 20% for UKSSDC are based in the UK (mostly universities), the CEDA data and services have a global user community: for BADC/NEODC 13% of users are from the rest of Europe, 10% from USA and 6% from BRIC countries.

Statistics for period 1st April 2011to 31st March 2012

CEDA Queries closed	3856 (a reduction of 7.4% in absolute figures from 2010/11 volume, or a reduction of 20.4% with respect to the increase in active user numbers) - 92.8% are BADC queries.
Total CEDA registered users (to 03/05/2012)	22204 (16323 in 2010-11) – includes 5881 (3584 in 2010-11) new users in period for BADC and NEODC
NERC funded active users[1] (to 03/05/2012) (An active user has access to one or more restricted datasets)	23% of Total BADC and NEODC registered users (no figures for UKSSDC) have had access to restricted data held by CEDA – compared with 21% of all active users (3723 users).
Identifiable users actively downloading	3723 (3175 in 2010-11)
Total download volume	281.6 Tb (35% increase from 2010/11) in over 27.8 million files (increase of 6.8 M from 2010/11) from 198 datasets

[1] An active user has access to 1+ restricted datasets and/or has registered within 12 months.

Major Collaborations

In 2011/2012, significant national and international collaborations have been continued and/or begun. On the national scale, CEDA itself reflects a collaboration between the earth observation community and the atmospheric sciences community (via NCEO and NCAS). Additionally, CEDA is:

- Working closely with the other NERC centres, under the auspices of the implementation plan for the NERC Science Information Strategy.
- Building the Earth System Grid Federation in partnership with the US Programme for Climate Model Diagnosis and Intercomparison and their US Earth System Grid partners (particularly those at NCAR[2] and GFDL[3]) on software to support the forthcoming fifth Coupled Model Intercomparison Project (CMIP5).
- A leading partner in two major European projects: Metafor (documenting climate codes and their resulting simulations to unprecedented levels of clarity) and IS-ENES (developing an InfraStructure for a European Network for Earth system Simulation).
- Using UK Department of Energy and Climate Change (DECC) funding to lead the delivery of the IPCC data distribution centre (http://www.ipcc-data.org in partnership with the DKRZ[4] hosted World Data Centre for Climate and Center for International Earth Science Information Network (CIESIN) at Columbia University).
- Working with the European Space Agency to extend earth observation metadata standards.
- Delivering a key role in evolution of the Climate Forecast NetCDF metadata conventions via standard name management.
- Providing data discovery services for the Marine Environment Data Information Network (MEDIN).
- Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.
- Working with the European 'Space Weather' community to develop a standards-based system for harmonised access to space weather data in ESPAS (near earth data infrastructure for e-science.
- Developing INSPIRE data specifications with thematic experts from all EU Member States.
- Working with the European Space Agency to review and redefine data archive formats (SAFE, Standard Archive Format for Europe)
- Working with commercial and academic partners in ISIC such as Logica, Astrium, Vega, NCEO – University of Reading and the University of Leicester on CEMS (Climate and Environmental Monitoring from Space).

Publications and Major Conference Presentations

(CEDA authors underlined)

- V Bennett, A Haffegee, B Matthews, S Nagella, A Shaw, J Styles. Building a video wall for earth observation data. Proc. Theory and Practice of Computer Graphics (TP.CG.2011), Warwick, UK, 06-08 Sept 2011
- Sarah Callaghan. Making Data a First Class Scientific Output: data citation and publication by NERC's environmental data centres. 7th International Digital Curation Conference, 5 – 7 December 2011, Bristol, UK

[2] National Center for Atmospheric Research

[3] Geophysical Fluid Dynamics Laboratory

[4] German Climate Computation Centre

- Sarah Callaghan, Roy Lowry, David Walton and members of the NERC SIS data citation and publication project team, Data Citation and Publication by NERC's Environmental Data Centres, submitted to Ariadne magazine, November 2011.
- Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras. GBS 20.7GHz slant path radio propagation measurements, Sparsholt site, British Atmospheric Data Centre, 2003-2005, 1st April 2011, doi:10.5285/E8F43A51-0198-4323-A926-FE69225D57DD
- Sarah Callaghan, Mark Morgan, Eric Guilyardi, Sophie Valcke, Charlotte Pascoe, Bryan Lawrence and the METAFOR Project Team. Supporting the climate community by providing common metadata for climate modelling digital repositories: the METAFOR project. In EGU2011, 3rd – 8th April 2011, Vienna, Austria
- Sarah Callaghan, Nathan Cunningham, Mark Thorley, Roy Lowry, Gwen Moncoiffe and the NERC Data Citation and Publication Project Team. Data citation and publication by the UK's Natural Environment Research Council's data centres. In EGU2011, 3rd – 8th April 2011, Vienna, Austria
- E Conway, D Giaretta, S Lambert, B Matthews. Curating scientific research data for the long term: a preservation analysis method in context. International Journal of Digital Curation 6 (2) (2011)
- E Conway, B Matthews, S Lambert, D Giaretta, M Wilson, N Draper. Managing risks in the preservation of research data with preservation networks. Proc. 7th International Digital Curation Conference (IDCC2011), Bristol, UK, 05-07 Dec 2011, paper in proceedings 7th International Digital Curation Conference – Accepted for publication in the International journal of Digital Curation as full peer reviewed research paper
- E Conway, S Lambert, B Matthews, A Shaon. Managing Preservation Networks: Issues of scale for scientific research assets. Proc. 8th International Conference on Preservation of Digital Objects (iPres2011), Singapore, 01-04 Nov 2011
- E Conway (Chapter 14 Preservation Analysis); Giaretta, D. (2011). Advanced Digital Preservation. (1st Ed). Springer
- Alastair Gemmell, Jon Blower, Phil Kershaw, Stephen Pascoe, Ag Stephens. The MashMyData project – Combining and comparing environmental science data on the web. (Extended Abstract), UK e-Science All Hands Meeting, September 2011
- Alastair Gemmell, Jon Blower, Philip Kershaw, Stephen Pascoe, and Ag Stephens. MashMyData: a gateway for scientific visualization and intercomparison of secure, distributed data. European Geosciences General Assembly, April 2011
- Eric Guilyardi, V. Balaji, Sarah Callaghan, Cecelia DeLuca, Gerry Devine, Sébastien Denvil, Rupert Ford, Charlotte Pascoe, Michael Lautenschlager, Bryan Lawrence, Lois Steenman-Clark, Sophie Valcke. The CMIP5 model and simulation documentation: a new standard for climate modelling metadata. CLIVAR Exchanges Special Issue No. 56, Vol. 16, No.2, May 2011 (http://www.clivar.org/publications/exchanges/Exchanges_56.pdf)
- Juckes, M. N., Spread versus uncertainty: untangling climate reconstructions. Abstract and presentation, Session CL1.15/EG5, European Geosciences General Assembly, April 2011
- Martin Juckes, V. Balaji, B.N. Lawrence, M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser: ExArch: Climate analytics on distributed exascale data archives, invited presentation. SOS 15 Conference, March 2011, Engelberg, Switzerland.

- Martin Juckes, V. Balaji, B.N. Lawrence, M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser: ExArch: Climate analytics on distributed exascale data archives, invited presentation. AGU Fall Meeting, December 2011, San Francisco, USA.
- Philip Kershaw, Rachana Ananthakrishnan, Luca Cinquini, Dennis Heimbigner, and Bryan Lawrence. A Modular Access Control Architecture for the Earth System Grid Federation. In Proceedings of the International Conference on Grid Computing and Applications (GCA11), 2011, pp. 3-9.
- Philip Kershaw, Jon Blower, Alistair Gemmell, Stephen Pascoe, and Ag Stephens. Building a Web Based Environment for the Intercomparison of Distributed Environmental Science Data, Challenges in Access Control and Security. European Geosciences General Assembly, April 2011
- BN Lawrence, CM Jones, BM Matthews, SJ Pepler, SA Callaghan. Citation and peer review of data: moving towards formal data publication. International Journal of Digital Curation 6 (2) (2011)
- Dominic Lowe. Profiling Observations and Measurements for the Weather, Oceanography and Climate communities. EnviroInfo 2011, Ispra, Italy, 5-7 October 2011
- Dominic Lowe. Harmonised access to weather, ocean and climate data using Climate Science Modelling Language and Observations and Measurements. EnviroInfo 2011, Ispra, Italy, 5-7 October 2011
- Brian Matthews, Arif Shaon, Esther Conway. The Preservation of Complex Objects (ebook) How do I know that I have Preserved Software? (Chapter), The Preservation of Complex Objects Symposia Jan 2012
- Alison Pamment, Calum Byrom, Oliver Clements, Steve Donegan, Philip Kershaw, Bryan Lawrence, Roy Lowry. CF Standard Names: Current Status and a New Vocabulary Editor. GO-ESSP 2011, Marriott Hotel, Asheville, N. Carolina, 10-11 May 2011.
- Stephen Pascoe, Martin Juckes, Philip Kershaw, Bryan Lawrence. Ensuring a unified user experience in a federated portal architecture: experiences of the CMIP5 archive. European Geosciences General Assembly, April 2011
- AM Sayer, CA Poulsen, C Arnold, E Campmany, S Dean, GBL Ewen, RG Grainger, B Lawrence, R Siddans, GE Thomas, PD Watts. Global retrieval of ATSR cloud parameters and evaluation (GRAPE): Dataset Assessment. Atmospheric Chemistry and Physics 11 (8) 3913-3936 (2011) [doi:10.5194/acp-11-3913-2011]
- Scheck, L., S. C. Jones and M. N. Juckes, 2011: The resonant Interaction of a tropical cyclone and a tropopause front in a barotropic Model. Part I: Zonally-oriented front. J. Atmos. Sci., 68, 405-419.
- Scheck, L., S. C. Jones and M. N. Juckes, 2011: The resonant Interaction of a tropical cyclone and a tropopause front in a barotropic Model. Part II: Frontal waves. J. Atmos. Sci., 68, 617-637.
- A Shaon, S Callaghan, B Lawrence, B Matthews, A Woolf, T Osborn, C. Harpham. A linked data approach to publishing complex scientific workflows. Proc. 7th IEEE International Conference on e-Science (eScience2011), Stockholm, Sweden, 05-08 Dec 2011
- A Shaon, S Callaghan, B Lawrence, B Matthews, T Osborn, C Harpham. Opening up Climate Research: a linked data approach to publishing data provenance. 7th International Digital Curation Conference (DCC11), Bristol, England, 05-07 Dec 2011

- Ag Stephens, Philip James, David Alderson, Stephen Pascoe, Simon Abele, Alan Iwi & Peter Chiu. The challenges of developing an open source, standards-based technology stack to deliver the latest UK climate projections. International Journal of Digital Earth. Volume 5, Issue 1, 2012. pp 43-62. DOI: 10.1080/17538947.2011.571724.
- Stephens, A., Pascoe, S. & Kershaw, P. (2011). Useful extensions to the OGC Web Processing Service based on a Python client/server implementation. Geophysical Research Abstracts, Vol. 13, EGU2011-7777, EGU General Assembly 2011.

Meeting Attendance

- Graham Parton attended the NCEO early career scientists conference April 2011.
- Matthew Wild attended the National Astronomy Meeting, Manchester in March 2012 and a World Data System meeting in London, also March 2012.
- Members of CEDA attended the annual NCAS Meeting in July 2011 and the NCEO Annual Staff Meeting in September 2011.
- Graham Parton presented the poster “Why archive environmental data?” at the NCEO Annual Staff Meeting in September 2011.
- Dominic Lowe answered questions from UK organisations on INSPIRE data specifications at the UK Location: INSPIRE Data Specification Familiarisation Workshop, DEFRA London. June 2011.
- Dominic Lowe attended the NERC Standards Workshop, Lancaster, October 2011, as CEDA standards representative.
- Dominic Lowe gave a presentation on INSPIRE and SeaDataNet at the SeaDataNet 2 FP7 Kickoff, Athens, October 2011.
- Dominic Lowe gave a presentation on Interoperabilty plans (as Work Package leader) at the ESPAS FP7 Kickoff, Rome, November 2011.
- Dominic Lowe gave a presentation on the INSPIRE consultation process at the SeaDataNet 1st Technical Task Group meeting, Liverpool. November 2011.
- Dominic Lowe participated and part-chaired the INSPIRE Thematic Working Group meeting, Brussels, November 2011 and INSPIRE Thematic Working Group Meeting, Copenhagen, January 2012 meeting as Editor of the Thematic Working Group on Oceanographic Features/Sea Regions.
- Dominic Lowe participated and chaired splinter meetings in this cross-domain meeting as Editor of a Thematic Working Group at the INSPIRE Comment Resolution Workshop, JRC Italy, December 2011.
- Dominic Lowe planned agenda and co-chaired the ESPAS Technical Meeting, Rome, January 2012, and gave a presentation on the Interoperability work package.
- Dominic Lowe participated in the LTDP SAFE review Kickoff meeting, ESRIN (ESA), Frascati, Italy, February 2012:.
- Dominic Lowe participated in the LTDP SAFE Review meeting, ESRIN (ESA), Frascati, Italy, March 2012.
- Dominic Lowe presented an update on INSPIRE at the SeaDataNet 2nd Technical Task Group meeting, Cyprus, March 2012.

- Graham Parton attended the Open Data Master Class workshop on using GIS data to gain experience and share CEDA data availability with fellow delegates from sectors outside usual CEDA user community in June 2011.
- Graham Parton organised BADC trade stand with demonstrations and CEDA delegation at the Royal Meteorological Society 2011 conference. The trade stand highlighted ISIC visualisation service and ability to provide DOIs for archived data
- In October 2011, Graham Parton attended the Royal Meteorological Society special interest group meeting: Transmission, presentation and archiving of meteorological data and presented oral presentation entitled "Environmental Data Archival: Practices and Benefits" explaining data management practices within CEDA from data arrival, through archiving, to end user services

Other Outreach and Knowledge Exchange

- During summer 2011 CEDA took on Ben Sears as a summer placement, under the supervision of Graham Parton, working on ARSF data ingestion and examining the CEDA media library. As part of this latter duty Ben scanned-in various research flight logs that had not previously been arranged in a digital format and placed these into the CEDA Document repository. Ben also prepared a poster entitled "Why archive environmental data?" presented by Graham Parton at the NCEO conference in September 2011.
- CEDA had a stand with handouts and posters at the Royal Meteorological Society conference in Exeter (27-30 June 2011)
- CEDA attended the RSPSOC conference in Bournemouth in September, and shared the NERC ARSF stand, with handouts, posters and leaflets
- Graham Parton gave a talk, titled "Environmental Data Archival: Practices and Benefits", at the Royal Meteorological Society meeting "Transmission, presentation & archiving of meteorological data", held as the British Antarctic Survey, Cambridge, 5 October 2011.
- Charlotte Pascoe and Sarah Callaghan are STEM ambassadors and regularly support science outreach activities in schools and for the general public.
- Sarah Callaghan gave a lunchtime lecture, titled "Improving the foundations of the scientific record: data citation and publication by the NERC data centres", at the British Geological Survey, 25th January 2012
- Sarah Callaghan presented a report on CODATA workshop in August 2011 to SCOR/IODE/MBLWHOI Library Workshop on Data Publication, 4th Session, at the BODC in Liverpool on 3rd and 4th November 2011
- Sarah Callaghan presented "Expanding the scientific record: Data citation and publication by NERC's environmental data centres CODATA symposium" at the DataCite Summer Meeting, 24-25 August 2011,Berkeley, California, USA.
- Sarah Callaghan presented "Data Citation in the Earth and Physical Sciences" at the Developing Data Attribution and Citation Practices and Standards, An International Symposium and Workshop, August 22-23, 2011, Berkeley, California, USA.
- Esther Conway gave a talk about the LTDP survey of standards and initiatives and produced a report for the LTDP Workshop, ESRIN, May 2011.
- Esther Conway hosted and provided training for a group of researchers (Postdoc, Ph.D. and Masters level) from Cranfield University. She is supervising a Master's thesis into cost modelling digital preservation strategies for scientific archives. She presented a student poster at IDCC2011.

- Philip Kershaw hosted the 2nd workshop on Federated identity system for scientific collaborations at RAL 2-3 November 2011. This is one of a series of workshop to gather input from a range of scientific communities on their needs for security and single sign on and write a position paper to present national, pan-national security infrastructure providers to better serve the needs of the scientific community.
- Philip Kershaw contributed to an article for the "International Grid this Week" online journal, "A Single Computing Identity", http://www.isgtw.org/feature/single-computing-identity
- Stephen Pascoe presented at Global Organization for Earth System Science Portals (GO-ESSP) 2011 workshop "Maximising the utility of OPeNDAP datasets through the NetCDF4 API"
- Stephen Pascoe presented at the 13th Workshop on Meteorological Operational systems, ECMWF, "The CEDA web processing service for rapid deployment of earth system data services"
- Stephen Pascoe presented at the Climate Knowledge Discovery Workshop 2011, Deutsches Klimarechenzentrum GmbH, "Climate Metadata Systems in Context"

Science to policy

- Esther Conway has been appointed for the third time as an expert evaluator for the EU Research Executive Agency: call FP7-SPACE-2012-1 proposals
- Esther Conway has been providing management support in the form of scientific reports and presentations to the ESA led Long Term Data Preservation (LTDP) working group
- Dominic Lowe is a member of INSPIRE Thematic Working Group on Observations and Measurements, (European Union), January 2012
- Sam Pepler is a member of the ESA Long-Term Data Preservation Working Group.
- Martin Juckes reviewed 3 chapters of the IPCC 5th Assessment, WG1 First Order Draft.
- Sarah Callaghan and Sam Pepler attended the RCUK Open Data Dialogue – Stakeholder Workshop, 23 Feb 2012

Software Distributions

CEDA has a considerable software infrastructure to support the data centres and projects. While much of the software is customised for internal use, CEDA also releases a considerable amount of software as open source. There are three broad grouping to the software CEDA users and makes public for reuse:

1. Security software which provides implementations of key standards necessary to support federation authentication and authorization (so that CEDA internal systems can be used for federated as well as local applications).
2. Discovery systems software to support the NERC Data Discovery Services (since these are common problems).
3. Data manipulation and visualisation packages (used internally & available for reuse elsewhere).

New Software Packages for 2011/2012		
ndg_oauth2_client Version 0.2.0 (Python)	Client-side implementation of OAuth2 security protocol. OAuth enables services and applications access secured resources on behalf of a user.	Due to be added to PyPI repository soon
ndg_oauth2_server Version 0.2.0 (Python)	Server-side implementation of OAuth2 security protocol. OAuth enables services and applications access secured resources on behalf of a user.	Due to be added to PyPI repository soon
Newmoon v1.2 (Java)	A web application generating page to easily submit task to a Fullmoon (which generate a GML XSD schema from a UML model) instance.	Source: http://proj.badc.rl.ac.uk/ndg/browser/mauRepo/newmoon WAR: http://triton.badc.rl.ac.uk:8180/artifactory/webapp/browserepo.html?pathId=libs-releases-local%3Andg%2Fservices%2Fnewmoon%2Fnewmoon-web%2F1.2.0 Running at: http://bond.badc.rl.ac.uk/newmoon
dapbench v0.1	A framework for testing the performance the OPeNDAP data access protocol under serial and parallel workloads. Also verifies security constraints on ESG Federation data access services using the THREDDS and OPeNDAP standards.	http://github.com/cedadev/dapbench [This package replaces the "thredds_security_test" package]
CedaMarkup (Python)	A library implementing the Opensearch specification.	Project at https://github.com/kusamau/cedaMarkup
Software Packages improved in 2011/2012		
ndg_xacml Version 0.5.0 (Python)	Implementation of XACML (eXtensible Access Control Mark-up Language). Enables the expression of access control policies to determine who or what has the rights to access a given dataset or other resource. Also for ESGF and CMIP5. Added functionality for improved CMIP5 support and request from Ocean Observatories Initiative Cyberinfrastructure Project: http://ci.oceanobservatories.org	

ndg_saml Version 0.6.0 (Python)	CEDA implementation of SAML (Security Assertion Mark-up Language) – needed for the Earth System Grid Federation (ESGF) and CMIP5. Added support for SAML 2.0 profile of XACML v2.0. This is in support of work for the CEDA's OGC Web Processing Service.	http://pypi.python.org/pypi/ndg_saml
ndg_httpsclient Version 0.2.1 (Python)	PKI security library for HTTP Python clients. Improves HTTPS support for Python. (Renamed from urllib2pyopenssl)	http://pypi.python.org/pypi/ndg_httpsclient
ndg_security_server Version 2.2.2 (Python)	A complete tool kit to manage access control in a federated infrastructure compliant with the system developed for the ESGF and CMIP5. It includes an implementation of the single sign on technology OpenID and features pluggable components for securing any given Web based application. Enhancements to support new OAuth-based security solution for CEDA OGC Web Processing Service and Web Map Service	http://ndg-security.ceda.ac.uk/
COWS-WPS version 0.2.1	An implementation of the OGC Web Processing service that supports synchronous and asynchronous process execution on grid and cluster resources. (COWS-WPS is the unifying technology behind the UKCP09 User Interface.)	http://cows.badc.rl.ac.uk/cows_wps.html (home page) http://ndg.nerc.ac.uk/dist/ (download)
Maintained Software		
MyProxyClient Version 1.3.0 (Python)	Lightweight python based client to the MyProxy package developed by the US National Center for Supercomputing Applications. It enables users to manage their personal identity tokens using remote token repositories. Continued support of this package for CMIP5.	http://pypi.python.org/pypi/MyProxyClient/
MyProxyWebService Version 0.1.2 (Python)	Enhances the MyProxy service software by adding a HTTP based interface to the server side software enabling any simple Web based client to access it and obtain identity tokens. In use for CMIP5 supporting Met Office users.	http://pypi.python.org/pypi/MyProxyWebService/
CEDA OGC Web Services Framework (COWS) Version 1.6.1 (Python)	A framework for developing OGC Web Services in Python	http://cows.badc.rl.ac.uk/
COWS-server Version 1.6.1 (Python)	An implementation of the OGC Web Map and Web Coverage Services specialising in serving NetCDF-CF data	http://cows.badc.rl.ac.uk/
COWS-client Version 1.7.0 (Python)	A web application for visualisation of OGC Web Map Services.	http://cows.badc.rl.ac.uk/

Funding 2011/2012

CEDA is funded by a wide range of sources, through direct funding via service level agreements and on a project basis.

Financial Summary:

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012
NCAS income	753	970	866	906	883
NCEO income	242	378	389	450	419
Other NERC income	410	788	481	341	527
Other income	160	461	710	1144	1099
Total income	1565	2597	2446	2841	2928
JASMIN and CEMS infrastructure grants					3979

Overall funding for CEDA for financial years 2007- 2008 to 2011-2012

Most of the funding to CEDA comes from a service level agreement (SLA) between the Natural Environment Research Council (NERC) and the Science and Technology Facilities Council (STFC).

Many of the programmes funded by the SLA are multi-year programmes, with funds being allocated in one year, but not spent until some years later. Funds are generally deferred by a combination of three mechanisms: simple accounting carry over from one year to the next, or formal deferment of milestones at either STFC or NERC (in which case the funds remain outside of the CEDA account). Because there are very large fluctuations in income from one year to the next, because large item spends come from accumulating capital, and because staffing is relatively static, there can be considerable carry overs from one year to the next.

As of the time of writing CEDA manages:

- **582** logical file sets (that is, primary data entities that need to be managed independently either for scientific reasons – they are different datasets – or for logistical reasons – they are too big to fit into one disk partition).
- **953 TB** of primary data distributed over 1.3 PB of available primary storage (with 2.2 PB of total storage including space for secondary data copies etc.).
- **93 servers, 30 hypervisors: 265 distinct computers** (including virtual machines).
- **140 distinct disk partitions**
- **89 million**[5] **primary files** (in excess of 200 million including secondary and tape copies).

The 2012/13 budget appears below. Some points to note include:

- BADC:
  - primary data 424TB (partitions allocated for 480 TB, 37% of the total primary storage).
  - 46% of the data centre operation budget comes from NCAS National Capability.
- CMIP5
  - primary data: 288 TB (partitions allocated for 413 TB, 32% of the total primary storage, with at least another PB expected this year).
  - CMIP5 is supported from a wide variety of funding streams, with significant support from NCAS.

[5] This figure from a manual evaluation of script outputs, it's not (yet) automatically generated.

Further points to note:

1. Roughly half the data ingestion is supported from NCEO and NCAS National Capability, and half from other funding.
2. NC is responsible for roughly half of the staff budget – CEDA already does very well for levering funding. NERC does not pay for networks or power, both of which are funded out of NERC overheads by STFC.

(FTE[6] except last row)	NCAS	NCEO	NC	RP	RM	EU	Gov	Other	STFC[7]	Total
Core Data Centre Activity										
Acquisition and ingestion	1.4	0.7	0.1	0.3	1.0	0.2	0.2	0.0	0.1	4.0
Information Management.	1.4	0.4	0.1	0.0	0.0	0.1	0.2	0.0	0.1	2.4
Access and delivery	0.7	0.5	0.2	0.0	0.0	0.5	0.2	0.0	0.1	2.2
Community support	0.7	0.2	0.1	0.0	0.0	1.3	0.9	0.0	0.1	3.4
Standards management	0.7	0.1	0.1	0.0	0.0	1.2	0.1	0.0	0.1	2.1
General Management	1.4	0.3	0.1	0.0	0.0	0.0	0.2	0.0	0.3	2.4
Manage infrastructure	1.5	0.9	0.1	0.0	0.0	0.0	0.7	0.0	0.7	3.8
Non – DC functions	0.0	0.5	0.0	0.0	0.5	1.6	0.0	0.9	0.0	3.6
Total	7.9	3.6	0.8	0.3	1.5	4.9	2.4	1.0	1.5	23.9
Overall Picture (Summary)	NCAS	NCEO	NC	RP	RM	EU	Gov	Other	STFC	Total
General Data centre ops	5.0	1.9	0.5	0.0	0.0	0.9	1.2	0.0	1.3	10.7
Community specific activity[8]	2.8	1.2	0.3	0.3	1.0	2.4	1.2	0.0	0.3	9.5
Non-DC functions	0.0	0.5	0.0	0.0	0.5	1.6	0.0	0.9	0.0	3.6
Total	7.9	3.6	0.8	0.3	1.5	4.9	2.4	1.0	1.5	23.9
Non -Pay (£)	71,002	41,095	9,000	30,000	13,000	9,216	0	0	71,200	234,513

[6] All FTE calculated at £100K regardless of grade.

[7] This includes the backup tape system (£60K+0.6 FTE) and the STFC contribution to the UKSSDC (plus director's salary at 20% FTE). This does not include the “well-founded-lab” support, which includes all electrical costs, and in excess of 10 Gbit/s of bandwidth for CEDA (the cost of which would likely exceed 100K per annum).

[8] 100% of acquisition/ingestion and community support and 50% of access, delivery and standards management.

2012-2013 Detailed Targets

These are the targets that appear in the NCAS and NCEO annual business plans for CEDA activities:

National Capability	2011 Priority Activities
T5.1 Data Ingestion: Acquire, ingest, and catalogue, appropriate data from a range of sources including the Met Office and NERC-funded projects.	1. Continue to maintain existing data streams.
T5.2 Hardware Infrastructure: Maintain and upgrade computing systems and networks to support data holdings and data access.	1: Improve national and international network bandwidth. 2: Operationalise parallel file system for scratch usage. 3: Put hardware on “SLA footing” (clear internal interface with requirements)
T5.3 Software Infrastructure Develop, maintain and upgrade necessary software and information systems to support data curation and data access.	1. Replace perl file browser and security system. 2. Replace website with new Django based system. 3. Integrate web components with common security system. 4. Upgrade ISIC vis system to use security and deploy more widely. 5. Deploy CEDA Web Processing Service delivering data services.
T5.4 Core User Services – Curation: Curate existing information according to best practice principles: create, delete, migrate data and information as necessary.	1. Maintain DOI support for datasets. Continue to support data citation. 2. Deploy MOLES 3 based infrastructure and update information records accordingly. 3. Migration and backup control from single database driven configuration. 4. DMP services as necessary. 5 Support development of UK strategy to support LTDP
T5.5 Core User Services – User support: Provide prompt and effective user support. Provide additional services to users such as a meeting registration service and distributed document management service.	1. User support as necessary. 2. Support met/ocean working group & INSPIRE met features. 3. Support CEDA goals via ESA secondment. 4. Representation on committees and panels, e.g. ARSF SC/NEODAAS SC
T5.6 CMIP5 support: Provide support for CMIP5 by providing a UK data node, a replicated copy of the global core archive, and appropriate interfaces (software, hardware, and networks). Deploy and maintain any additional necessary services (e.g. the CMIP5 questionnaire.)	1. Operation of data nodes. 2. Operation of gateways. 3. Operation of replication. 4. Operation of Metafor gateway and questionnaire. 5. Carry out CMIP5 quality control. 6. Compute service for CMIP5 analysis
T5.7 Activities on behalf of NCEO: Support the UK earth observation community by continuing to provide high speed UK cache archives for ESA, NASA (and other high volume remote data).	1. Continue to maintain existing data streams. 2. Support CF-NetCDF standard names, and on vocabulary management for Earth Observation.
T5.8 Activities on behalf of NCAS: Provide data management services (data management	1. Continue to maintain existing data streams.

National Capability	2011 Priority Activities
plans and formal archives) for NCAS and NCEO themselves (including FAAM, ARSF, UFAM and the NCEO scientific themes).	
T5.9 SIS: Contribute to the implementation of the NERC Science Information Strategy Programme, in particular leading and/or managing the projects dealing with architecture, data centre metrics, data citation and publishing, data value check lists and data policy implementation.	1. Provide architectural advice. 2. Citation activities.
T5.10 ISIC: Work with commercial and academic partners within the NCEO and STFC communities to deliver the ISIC (International space innovation centre) scientific visualisation services. Contribute to the development of a plan for ISIC phase 2**.**	1. Expand the number of datasets visible to the visualisation system (condition datasets appropriately, addressing security policies). 2. Contribute to the further development of the visualisation activities. 3. Continue to investigate options for system sharing. 4. Contribute to the scoping of the CEMS activity.
T5.11 UKSSDC. Continue to integrate the delivery and management of the UK solar system data centre into CEDA, while maintaining services to UKSSDC community.	1. Retire or transfer computing systems into common CEDA computing pool. 2. Migrate information systems onto CEDA Linux VMs
T5.12 NERC Data Catalogue Service Support operation of the system to harvest metadata records from partner data providers. Provide a metadata search facility via a web service: "NERC Data Catalogue Service"	1. Support operation of data catalogue system 2. Provide search facility for DCS
Other	**2011 Activities (detail within project contracts)**
TR1: RP Support: Provide data management support for NERC programmes and research projects consistent with their programme budgets. (Develop data management plans, provide support to scientists to aid delivery of structured data and meta-data consistent with NERC data policy, ingest data into the CEDA archive system.)	•QUEST •CASCADE •RAPID-WATCH •ClearfLO •StormsRiskMitigation •MAMM •SAMMBA
TR2 RM Support: As for TR1 but for grants.	•Amazonica, •ABACUS-IPY, •CASCADE, •MashMyData
TR3 Commercial Contracts: Obtain and deliver research and service projects consistent with developing and/or exploiting CEDA infrastructure, skills and services.	•DECC support for CMIP5 •DECC support for IPCC/DDC •DEFRA support for UKCIP09 •DEFRA support for the Agricultural Greenhouse Gas Inventory Platform •UKMO support for CMIP5.

National Capability	2011 Priority Activities
	•ESA Long Term Data Preservation support. •MEDIN discovery service operation
TR4 European Commission Contracts: As for TR3 but for EC based funding.	•IS-ENES (European infrastructure for earth simulation) •Contrail (Research into cloud computing) •ESPAS (Tools and data services for Near-Earth Space Data Infrastructure for e-Science) •OpenAIREplus (Scientific data citation) •SCIDIP-ES (Developing preservation services for Earth Sciences) •SeaDataNet2 (Aligning SeaDataNet with INSPIRE and ISO standards) •EuroGEOSS (provide access to climate data via GEOSS)
TR5 Academic Contracts: As for TR3, but for RCUK (including NERC) and JISC based funding.	•Valor (Rapid-Watch project assessing the value of the RAPID array observations for prediction) •ISIC Support •ExArch (Exascale data architectures)

Short Reports Describing Activities in 2011/2012

Publishing the Mars Analysis Correction Data Assimilation (MACDA) Dataset

Kevin Marsh

The BADC were contacted by Oxford researchers concerning a new atmospheric dataset, which they were keen to be made available to as many users as possible, and to have properly curated. What made this request different was that the atmosphere in question was not that of Earth, but of Mars. After the initial shock wore off we could see that these data would be of wide interest, and in any case, the title "British Atmospheric Data Centre" doesn't constrain itself with respect to planetary origin of the atmosphere!

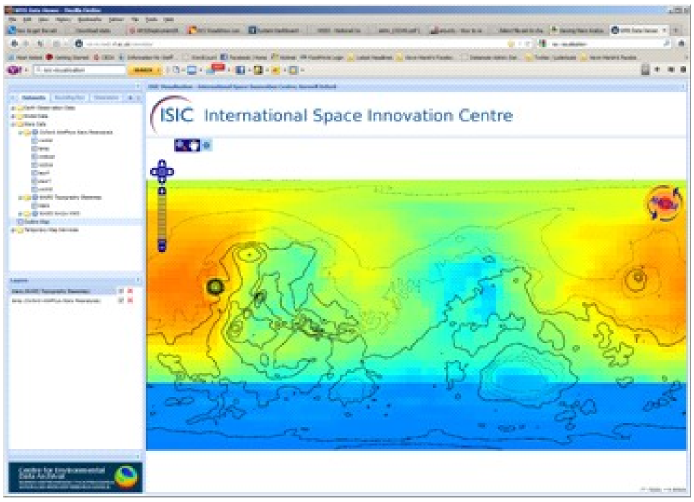


Figure 1: MACDA data visualisation via ISIC

The dataset contains basic gridded atmospheric and surface variables for the planet Mars over three Martian years (1 Martian year is 1.88 terrestrial years), produced by data assimilation using observations made during the science mapping phase of NASA's Mars Global Surveyor (MGS) spacecraft (May 1999 - August 2004). The dataset is produced by the re-analysis of Thermal Emission Spectrometer (TES) retrievals using the Mars Analysis Correction Data Assimilation (MACDA) scheme in a Mars global circulation model (MGCM). The MGCM used is the UK spectral version of the model developed by the Laboratoire de Meteorologie Dynamique in Paris. MACDA is a collaboration between the University of Oxford and the Open University in the UK.

This dataset provided a number of challenges to the BADC. The first was to convert the data files to CF-netCDF format, where the metadata used conformed to the international CF standard. To do this, the expertise of BADC staff who are domain experts in the field were utilised, and detailed discussions took place with the MACDA team to understand the nature of the data we were dealing with. Once the data were ready, they were transferred to the BADC and ingested into the archive. Supporting web pages and documentation were produced in collaboration with the MACDA researchers, and user access rules were determined. A key part of the capability of the support provided was for the data to be able to be viewed on-line by users. The MACDA dataset was selected for inclusion in the new International Space Innovation Centre (ISIC) visualisation service being developed by CEDA. The BADC staff responsible for this service worked closely with the MACDA team to enable the service to be capable of handling the data properly; dealing with the Martian Calendar alone raised a number of issues!

The dataset is now fixed (no further additions to that version are allowed), and a DOI (Digital Object Identifier) has been issued[9] - the DOI provides a persistent, citable means of referring to the dataset, and one which allows the data provider to receive due credit for their work. The MACDA DOI was only the second DOI issued by CEDA, and would not have been possible without the hard work and persistence of all involved!

[9] University of Oxford and The Open University. [Montabone, L.; Lewis, S. R.; Read, P. L.]. Mars Analysis Correction Data Assimilation (MACDA): MGS/TES v1.0, [Internet]. NCAS British Atmospheric Data Centre, 29 November 2011, doi:10.5285/78114093-E2BD-4601-8AE5-3551E62AEF2B

Hosted Processing Facility for Extremely Large Datasets (HPFELD)

Kevin Marsh

The “data deluge” of recent times has meant that data centres are under increasing pressure to hold datasets which are much larger than before. The sheer size of these datasets (often several tens of Terabytes) means that they are also becoming impractical for use. This is especially true for Earth Observation (EO) data from satellites. Users may find that downloading a full copy of the data to their local system is very time consuming, even assuming that they have adequate local storage available. Consequently, the data are effectively unusable by a significant proportion of the user community – a barrier to adding value to a dataset.

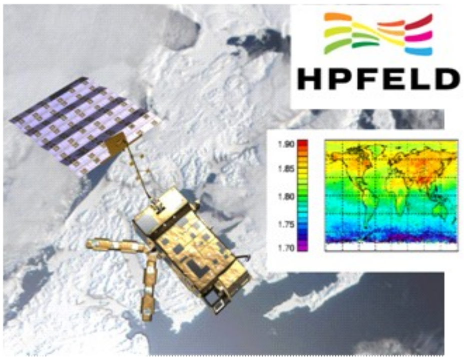


Figure 2: HPFELD prototype will use IASI data.

A more sensible approach would be to allow the data archives to be coupled to processing capability, made available over the network. In this way, remote users could either select a pre-configured processing algorithm to run on the dataset, or upload their own algorithm. The processing would be run on a host system which was ‘close’ to the archive data, and the output data made available to the user to visualise or download to their own system. Other complementary datasets from the data centre could also be incorporated into the processing, such as comparison with model datasets. The host system could use ‘cloud’ technologies, and the HPFELD will provide the environment in which the processing is performed.

HPFELD is an exciting 1-year project to see if existing technologies can be combined and adapted to produce a demonstration system. It is a partnership between STFC (CEDA), and the commercial companies Magellium[10] and Terradue[11]. The aim of HPFELD is to provide an interface where users can discover data and determine what processing should be applied. CEDA are responsible for defining how HPFELD should interact with our archive (e.g. data discovery using OpenSearch, access control, etc.) so that it is easy for users to navigate.

CEDA are also responsible for defining the user requirements for the test system, and have identified a number of Earth Observation climate change user stories from the Remote Sensing Group[12] at RAL. In particular, the reprocessing of Infrared Atmospheric Sounding Interferometer[13] (IASI) data at CEDA from the METOP satellite is one which could significantly benefit from HPFELD. Currently, it can take several days to reprocess data from a single orbit –and this is from a single instrument. With more than 5000 orbits each year, any improvement in performance would be highly beneficial.

Such a hosted facility would be of great benefit to CEDA and it’s multi-Petabyte archive as we look to the future. The Climate and Environment Monitoring from Space[14] (CEMS) facility will be particularly important in providing a suitable host environment, and HPFELD will help CEMS achieve its goals. It may even be possible to offer HPFELD as a subscription based service to commercial EO organisations such as ESA, providing processing capability at a fraction of the cost of supporting a local system. HPFELD represents a significant challenge in terms of the technology required and the relatively short time-scale involved. However, it has the real potential to make data processing capability far more accessible than before and allow the researchers to concentrate on the science behind the data.

[10] http://www.magellium.co.uk/

[11] http://www.terradue.com/

[12] http://rsg.rl.ac.uk/

[13] http://www.esa.int/export/esaME/ iasi.html

[14] http://www.cems-facility.org.uk/

The UK Solar System Data Centre

Sarah James

The UK Solar System Data Centre's holdings include data on the Sun's effect on the Earth and the near-Earth environment, the solar wind and the Interplanetary Magnetic Field (IMF) and ground and space-based measurements of the Sun itself. These data span science areas of interest to the Natural Environment Research Council (NERC) and to the Science and Technology Facilities Council (STFC). Hence the UKSSDC sits partly within CEDA and partly within the Space Environment Group.

Figure 3: Greenwich Photo Heliographic Reports in the UKSSDC archive, some over 100 years old

UKSSDC data holdings include geophysical indices, ionospheric data, solar indices, solar wind and IMF data and solar mission data. The UKSSDC is the primary source for data from the Chilton (RAL) and Stanley (Falkland Islands) ionosondes and for the IF2 and IG monthly ionospheric indices of solar activity. The data are typically time-series data, sometimes over decades. In the case of solar driven processes, decades of data are often needed to understand fundamental processes happening over solar cycles of 22 years. In any case, long data sets have more recently become important in studies of long-term environmental change. On shorter timescales, there is currently much scientific and political interest in Space Weather, that is, the effect of solar events on the Earth and its space environment, particularly on technology - and UKSSDC data is useful there too!

The UKSSDC grew out of the World Data Centre for Solar-Terrestrial Physics founded in 1957. As a result of this long history the UKSSDC has significant holdings of data that have never been digitised and are stored in their original media in our physical archive. For example the UKSSDC holds ionospheric data on prints and film and in reports, from the early days of ionospheric sounding in the 1930s through to the 1990s. The extensive collection of 35mm film is catalogued, but the catalogue itself has also not been digitised and is a card index. The physical archive also holds a collection of solar images taken by the Royal Greenwich Observatory (RGO) and its out stations between 1903 and 1942. The images are a mix of prints, and glass plates. Since acquiring this collection, further solar images have come to light in a warehouse in London. The UKSSDC is investigating taking possession and care of these too. The UKSSDC is preserving the physical archive holdings by assessing and monitoring conditions and by a series of house-keeping tasks to gradually remove potential problems, for example by removing decaying elastic bands and steel paper clips, and by providing suitable storage materials. These actions aim to 'do no harm' and preserve what is there with minimal future deterioration.

In addition the UKSSDC is looking for ways to make these data holdings more accessible to scientists. The first step is to improve the information in our digital catalogues so that users can see in more detail what data are held in hard formats. The second step is to digitise priority data sets. The UKSSDC considers preserving the contents of the reports as digital images a priority and will this year undertake a costing exercise to estimate the resources required to do this.

The UK Solar System Data Centre holds a broad spectrum of data of interest to a range of science communities. The time span of these data sets is of particular importance in answering scientific questions regarding long-term change. The UKSSDC is seeking to make the most of its archive of long-term physical records in response to this need.

CEDA's contribution to INSPIRE[15]

Dominic Lowe, Spiros Ventouras

The INSPIRE (Infrastructure for Spatial Information in the European Community) directive is a legal directive of the European Parliament which requires EU Member States to provide services and data for spatial data relating to the environment in accordance with sets of legally binding implementing rules. INSPIRE came into force in 2007 and is being implemented in various stages, with full implementation required by 2019.

The purpose of INSPIRE is to create an EU wide Spatial Data Infrastructure (SDI), which will enable the dissemination, sharing and access of environmental spatial information between organisations and people in Europe. This will assist in EU-wide and cross-border policy making as well as providing a core infrastructure which can be leveraged for a broad range of applications.

INSPIRE mandates particular types of services that should be provided; discovery, view & download, along with implementing rules for metadata and data content. The data content is described by Thematic Data Specifications defined by Thematic Working Groups (TWGs) which are a mix of domain specialists and technical experts who assess the requirements for a particular environmental theme.

Figure 4: The INSPIRE logo.

CEDA provides technical experts as Editors of three Thematic Data Specifications: Spiros Ventouras is the Editor of “Atmospheric Conditions and Meteorological Geographical Features” while Dominic Lowe is the Editor of “Oceanographic Geographical Features” and “Sea Regions” as well as the Editor of the upcoming “INSPIRE Download Services Technical Guidance”. Furthermore, both Spiros and Dominic are members of the “Cross-thematic Working Group on Observations & Measurements”.

By being actively involved in INSPIRE, CEDA is in an excellent position to help NERC and other stakeholders achieve INSPIRE compliance in the coming years. But what exactly is “INSPIRE compliance”? In actual fact, INSPIRE compliance is not a binary state – there are several levels of INSPIRE compliance. e.g.

- INSPIRE compliant metadata records compliant with Metadata implementing rules.
- INSPIRE compliant metadata services compliant with Network Services implementing rules for Discovery Services.
- INSPIRE compliant view services compliant with Network Services implementing rules for View Services.
- INSPIRE compliant layers in view services compliant with thematic Data Specifications implementing rules.
- INSPIRE compliant download services, compliant with Network Services implementing rules for Download Services.
- INSPIRE compliant datasets in download services, compliant with thematic Data Specifications implementing rules.

Furthermore, the roadmap[16] to compliance is complex, with different themes required to reach different compliance levels at different times over the coming years. So, rest assured, whatever anybody says we are not yet INSPIRE compliant! There is still some way to go. However by engaging with the INSPIRE development process CEDA is in a very strong position for the future.

[15] INSPIRE website: http://inspire.jrc.ec.europa.eu/

[16] INSPIRE roadmap: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/44

Data Standards for ESA's Climate Change Initiative

Sarah James

CEDA staff working in the ESA Harwell office are leading a Data Standards Working Group (DSWG) across the ESA Climate Change Initiative (CCI) in order to develop agreed standards for data formats and metadata. The CCI provides the ESA response to the WMO request for the measurement of Essential Climate Variables (ECVs) as part of the Global Climate Observing System (GCOS). Its purpose is to produce high quality ECVs from Earth observation satellite data across a wide range of scientific areas. The programme aims to make its final data products as accessible and useful as possible to a wide range of end users, in particular to allow easy and intelligent access by the climate community for integration into their models. The CCI is currently in phase 1, which involves scientific user consultation to produce the detailed specifications of the ECV data products and how they are to be produced.

Figure 5: The CCI projects are (from left to right and top to bottom): Fire, Land Cover, Glaciers, Sea Level, Sea Surface Temperature, Ocean Colour, Ozone, Cloud, Green House Gases, Sea Ice, Climate Modelling User Group, Aerosol, Ice Sheets and Soil Moisture.

Each ECV has a different CCI project group, with each group including members from several European institutions and companies. There are 13 groups working to produce the ECVs and a Climate Modelling User Group (CMUG) to provide a climate system perspective to the CCI programme. The ECVs cover a wide and varied range of scientific disciplines from the measurement of atmospheric aerosols, to soil moisture. Each discipline typically has its own practices and standards regarding the production of data and metadata and without intervention there would be a great variation in the nature of the data products produced. The Data Standards Working Group (DSWG) has already come to agreement on the file format and metadata standard for CCI data products, fixing on netCDF 4 classic format and on using the CF (Climate and Forecasting) convention for metadata with CF standard names for the main variables. The DSWG has agreed a minimum set of metadata attributes required to describe a CCI data set, which have been published in draft guidelines for data producers. CEDA staff have been reviewing test data files supplied by the CCI projects to help each project achieve compliance with the guidelines. The guidelines developed in phase 1 of the CCI will become requirements in phase 2 which involves systems development and data product generation.

The DSWG is continues to develop an agreed vocabulary to describe platforms, sensors, algorithms and institutions in the metadata. Work towards a file naming convention across the projects has highlighted huge variations in the timescales of the data products, from epochs of a few years down to data for one satellite orbit, which need to be accommodated by the final filename convention. In addition Sea Surface Temperature have a well defined file naming convention in their field already and the DSWG guidelines aim to permit them to continue to use that whilst meeting the needs of the other CCI projects.

The wide variety of data products that will result from the ESA Climate Change Initiative are both a challenge to the Data Standards Working Group and the reason that its work is so important to maximising the usefulness of the resulting Essential Climate Variables. CEDA staff are playing a key role in the important task of achieving workable data standards across CCI.

Interoperable Infrastructure for Space Weather Data – ESPAS.

Dom Lowe

The ESPAS (Near-Earth Space Data Infrastructure for e-Science) FP7 project aims to provide the e-infrastructure necessary to support the access to observations, modelling and prediction of the Near-Earth Space environment – which includes the plasma and energetic particle environments that surround our planet and the neutral atmosphere above 60 km. These environments are an important target for future research in areas such as space weather and Sun-climate studies.

Space Weather in particular is a topic which has been high on the government agenda recently as a significant space weather event (such as a large coronal mass ejection from the Sun) could cause considerable damage and disruption to our technologically-dependent world; for example it could disrupt power supplies and satellites, which would have a world-wide impact on almost everything including power, communication, transport, financial systems and almost anything else that depends on modern technology!

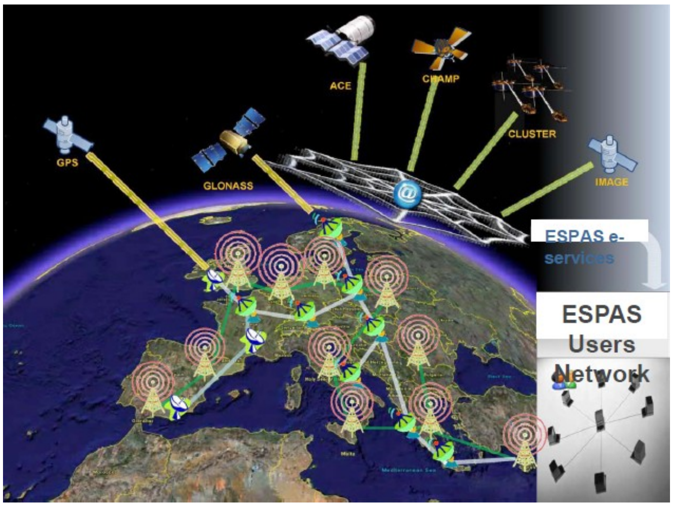


Figure 6: The ESPAS data collection network containing space and ground based instrumentation

In order to anticipate such events with sufficient notice time to put emergency mitigation plans into action (e.g. putting satellites into a safe mode, re-configuring electricity distribution networks etc.) it is necessary that space weather is better understood and greater forecasting capability is developed by the space weather community.

ESPAS aims to support the space weather community by providing greater access to and interoperability between key space weather datasets in Europe. CEDA staff have several roles in this project which reflects the diversity of skills in the organisation:

Dominic Lowe is leading the “Interoperabilty” Work Package, ensuring that the infrastructure keeps up with best practice in geospatial interoperability and standards, so that ESPAS does not become another data silo or a YAP (Yet Another Portal), but instead becomes part of a wider geospatial infrastructure connected to other ecosystems by internationally agreed standards. Meanwhile Sarah James and Matthew Wild from the CEDA UK Solar System Data Centre are using their knowledge of Space Weather data and systems to contribute to both the scientific knowledge in ESPAS and eventually to run and host the core ESPAS systems.

SeaDataNet2: Oceanographic Data and INSPIRE

Dom Lowe

SeaDataNet2 is a European Union Framework 7 (FP7) project which follows on from the successful SeaDataNet project which established a data-sharing infrastructure among dozens of oceanographic centres in Europe.

The intention of SeaDataNet2 is to turn the SeaDataNet infrastructure into an ongoing operational system, aligned with best practice in geospatial standards and policy alignment: culminating in a robust system which provides access to a whole range of ocean and marine environmental datasets.

CEDA's involvement in SeaDataNet comes as a result of our links with the INSPIRE project and also with the BADC sister organisation BODC (British Oceanographic Data Centre). CEDA's role in SeaDataNet2 is as a sub-contractor to BODC, providing advice and guidance on the implementation and use of standards within the SeaDataNet project, particularly standards relating to the development of INSPIRE discovery, view and download services.

CEDA have a staff member (Dominic Lowe) on the SeaDataNet Technical Task Group – a group which addresses all the technical and development issues in the SeaDataNet infrastructure.

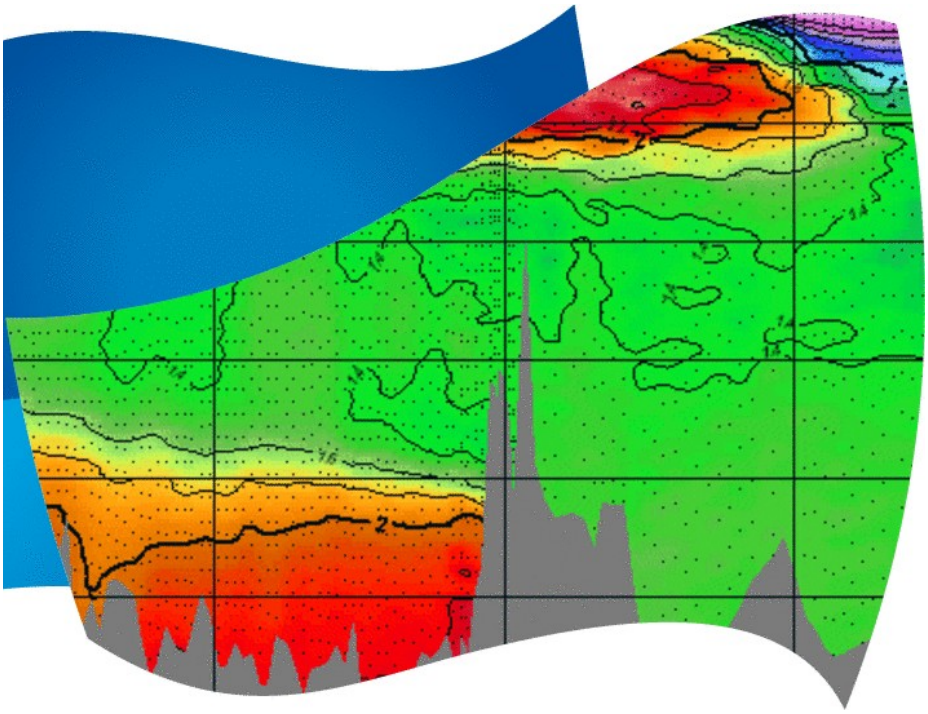


Figure 7: SeaDataNet Visualisation

JASMIN – The Joint Academic System for Shared Data Analysis

Bryan Lawrence, Matt Pritchard, Pete Oliver, Ag Stephens

In early 2011, the CEDA computing system was beginning to creak at the seams, with hundreds of computing systems storing hundreds of millions of files on the order of one hundred storage partitions (primarily using NFS) and connected in a complex local area topology. This reflected a history of growing using small amounts of funding, rather than via planned expansion using capital funding. In late 2011 the situation changed, with significant capital funding becoming available.

Figure 8: A view of some of the JASMIN Panasas storage blades.

In partnership with STFC e-Science, the CEDA team designed, procured, and installed a new computing system in the last quarter. The new infrastructure, consists of two components: JASMIN, a completely NERC funded initiative deployed by the National Centre of Atmospheric Science (NCAS), and CEMS (discussed in following articles), a jointly NERC and UKSA funded initiative deployed by the National Centre of Earth Observation (NCEO). The two components are deployed using the same hardware in a tightly coupled system. The functional components were designed to support and integrate the work-flows for three main goals: the efficient operation of the CEDA curation and facilitation mission; to support data analysis by the UK and European climate and earth system science communities; and to provide flexible access for the climate impacts and earth observation communities to complex data and a set of layered services. As well as generic problems in environmental data analysis, JASMIN is specifically targeted at supporting the joint analysis of weather and climate data by the academic community and Met Office staff.

JASMIN and CEMS together centrally deploy 9.3 PB of storage – 4.6 PB of Panasas fast disk and tapes in the STFC Atlas Tape Store. Over 370 computing cores provide local computation to support both local users and users interacting via web-services. In addition to the central JASMIN resource, remote JASMIN resources were procured at Bristol, Leeds and Reading to provide additional distributed storage and compute configured to support local work-flows as well as work-flows that exploits the central system. JASMIN is intended to become part of a national e-infrastructure to support “big data” problems in environmental simulation, and so fast network links were procured (but not yet in place at the time of writing) to the MONSooN supercomputer at the Met Office, and the HPC environment at the Edinburgh Parallel Computing Service (connecting to HECToR and a new Research Data Facility, also delivered in March 2012). A fast network link to the Royal Dutch Met Institute (KNMI) in the Netherlands is also being deployed, to support the European climate impacts community (with Dutch funding, under the auspices of the European Network for Earth Simulation).

As part of the JASMIN procurement, it was recognised that the providing a new high performance parallel compute environment to the academic community was a necessary condition for advancing the analysis of high resolution data, but it was not sufficient. Some initial work on a software product for data analysis was commissioned from with the JASMIN budget, but further investments in software for parallel data analysis and new tools were, and are needed.

Although the data volumes in earth observation and climate science have become too large for the traditional download and analyse paradigm to suffice in all cases, JASMIN and CEMS together have provisioned CEDA and the academic community to progress well into respectively supporting and exploiting petascale data holdings. However, the challenge for the next few years will be exploiting parallel data analysis.

The Facility for Climate and Environmental Monitoring from Space (CEMS)

Victoria Bennett, Phil Kershaw

CEDA staff are key partners in the recently established facility for Climate and Environmental Monitoring from Space (CEMS). CEMS is a facility within ISIC, the International Space Innovation Centre, which supports both academic and commercial communities by providing access to data, processing and expertise for climate and environmental research, as well as new Earth Observation related technologies and services in the commercial sector.

The partnership to develop CEMS consists of Logica, Astrium-GeoInformation Services, the National Centre for Earth Observation (NCEO) and CEDA. The core partners worked together during 2011 to define early concepts and gather user requirements and were successful in winning £3 million from government investment in e-infrastructure. This funding came to ISIC through the UK Space Agency (UKSA) at the end of 2011 and the project has since then been managed by an Integrated Project Team (IPT) formed of representatives from each partner, which reports to an ISIC CEMS Programme Board. Two CEDA staff members are on the CEMS IPT, Victoria Bennett as Project Scientist and Phil Kershaw as Technical Vision Authority. In addition, much of the technical work to design and deliver the infrastructure, particularly on the academic side, is carried out by CEDA staff.

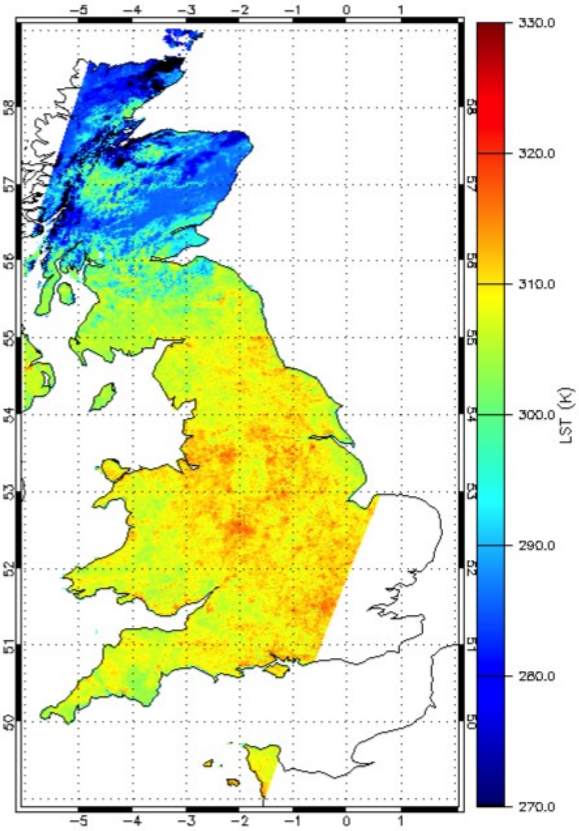


Figure 9: ATSR Land Surface Temperature plot for the UK (John Remedios, NCEO, U.Leicester)

The funding was to be spent within the financial year, i.e. by the end of March 2012, so a very busy period followed with hardware procurements, assessments of user requirements, preparation of demonstration data and applications, software architecture and development and configuration and deployment of kit, mostly all in parallel. The CEMS infrastructure includes significant data storage (1.7 Petabytes) hosted processing capabilities (17 nodes, each with 12 cores), and services for accessing and using EO data in a variety of ways. The system is deployed between two sites, with the academic hardware in the STFC eScience building, and the commercial hardware in the Electron building, both on the Harwell campus and joined by a high speed link. CEMS uses cloud-based technology to enable a variety of working environments next to the data, to support the needs of academic research as well as commercial applications.

The first users in the science community are expected to include research groups processing satellite data for the generation of Essential Climate Variables (ECVs). These users require significant processing capability close to large volumes of source data, in order to reduce the need for local storage and repeated data transfers across networks.

In future, it is envisaged that a range of services for data visualisation, manipulation and analysis will be available in CEMS, as well as tools for scientific data quality and integrity. On the commercial side, CEMS aims to provide an environment to stimulate innovation and growth in downstream applications.

CEMS Technical Overview

Phil Kershaw

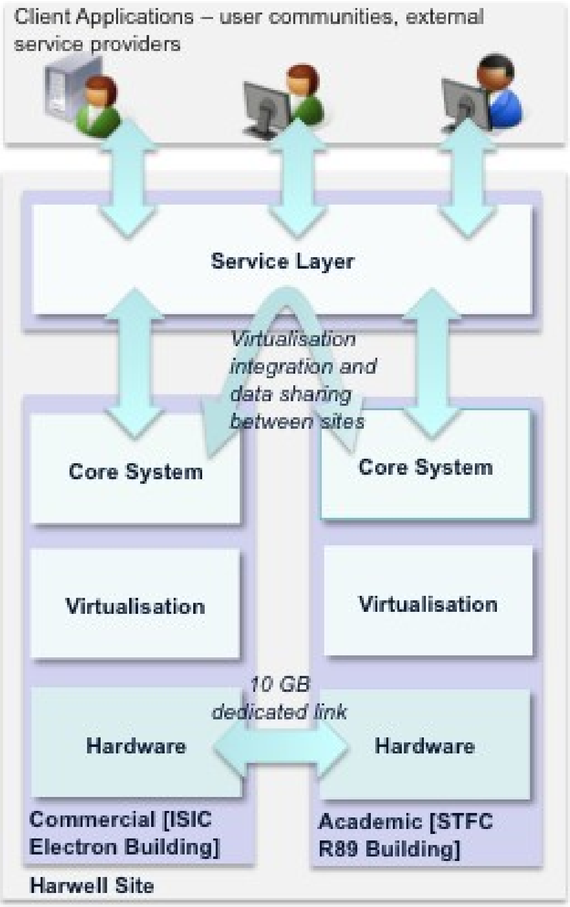

Client Applications – user communities, external service providers
Service Layer
Virtualisation integration and data sharing between sites
Core System
Core System
Virtualisation
Virtualisation
10 GB dedicated link
Hardware
Hardware
Commercial [ISIC Electron Building]
Academic [STFC R89 Building]
Harwell Site

Figure 10: CEMS layered architecture

Three key technical goals can be identified for CEMS: The provision of computing infrastructure for access to large-volume EO and climate model datasets co-located with high performance computing facilities; support for data access and curation, and; delivering expertise and tools for scientific data quality and integrity.

As in many other areas of research, there is a sharp increase expected in data volumes in the EO (Earth Observation) domain over the coming years (e.g. of the order of 20 increase for ESA Sentinels missions). Advances in storage technology have enabled greater capacities but this has not been matched by a corresponding increase in network bandwidth. Consequently, in addition to the provision of services to deliver data to users, CEMS is exploiting cloud-based technology to effectively *bring users to the data*. In order to achieve this, it will provide virtualised environments for users hosted next to the data storage in order to maximise bandwidth between processing software and the storage.

The data access and curation functions of a data centre are also key. Without these, users will not be able to discover and exploit the data effectively for their research goals. CEDA is bringing its extensive experience and expertise to this area and providing access to EO datasets through the new infrastructure. Finally, CEMS will also provide services for scientific data quality and integrity to give users confidence and transparency in its data, services and products. Given the complexity of this subject area, a technical study was undertaken to review requirements and make recommendations for the provision of such services. This has recognised CEDA's expertise in the area of data modelling and also the need for people-based technical consultancy as well as technical infrastructure and software services.

Figure 10 shows the layered architecture for CEMS. CEDA have taken a lead role proposing the this model and in the application of virtualisation technology. The infrastructure and software are deployed across neighbouring academic and commercial sites on the Harwell campus. The hardware is based on the same system as the JASMIN sister project. It uses the Panasas parallel file system to provide resilience, the ability to scale and fast performance eliminating input/output bottlenecks between the processing and storage hardware. The hardware is abstracted through a virtualisation layer based on VMware enabling the consumption of computing resources by third parties via a cloud. The virtualisation layer also hosts the core system which manages the cloud services and the data management and curation functions of a data centre. At the top, a service layer provides the interfaces to external user communities and partner organisations.

Support for the CF Metadata Standard

Alison Pamment

Since 2006 BADC has provided support for the development of the CF (Climate and Forecast) Metadata Conventions[17], primarily by managing the maintenance of the Standard Names[18] controlled vocabulary. This is a list of geophysical parameter names that are used to label observations and numerical model output thus allowing data to be more easily discovered and put to appropriate use. The list of Standard Names grows continuously in response to requests for new parameter names from members of the international climate and atmospheric science communities. During 2011-12 the management of the Standard Name table has benefited from continued improvements to the CEDA vocabulary editor, software that is designed to track all proposed changes to a vocabulary list and allow them to be published easily in the NERC vocabulary server. In 2011 CEDA took formal responsibility for managing two further controlled vocabularies that are embedded within the CF conventions.

Figure 11 shows the range of science domains for which standard names currently exist. 'Ocean biogeochemistry' and 'observing platforms' are both domains that have recently been added in response to CMIP5 modelling activities. An important group of physical oceanography terms have been introduced in the most recent version of the Standard Name table. These relate to TEOS-10, the Thermodynamic Equation of Seawater 2010, which has been adopted by the International Oceanographic Commission (the United Nations body for ocean science) as the recommended method of modelling sea water salinity, temperature and heat content in the next generation of ocean and coupled climate models. In addition to the Standard Names CEDA has taken on the management of the CF 'area types' and 'standardized region name' controlled vocabularies. The former is a list of surface types (e.g. 'sea_ice', 'grassland') whose description is important to scientists modelling the exchange of gases, aerosols and energy between the atmosphere and the earth's surface. The latter is a geographical gazetteer.

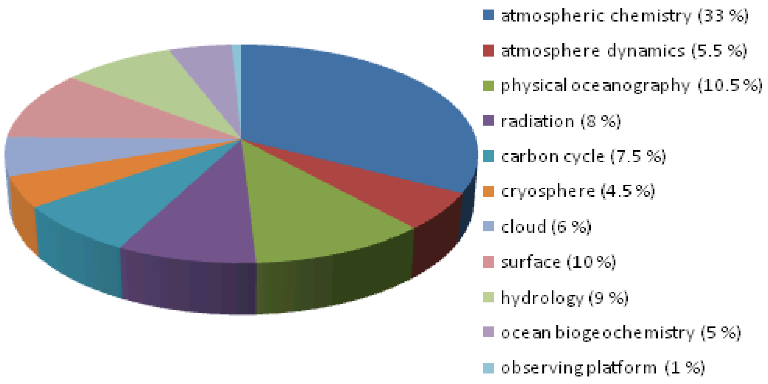


Figure 11: The proportion of vocabulary terms in each science domain for version 19 of the CF Standard Name Table (May 2012). The total number of terms is 2187.

The CF Metadata Conventions continue to be adopted as the basis for metadata standards in many projects requiring international data exchange. The European Space Agency Climate Change Initiative (CCI)[19] is a major project in which ‘Essential Climate Variables’ from a number of ESA satellite missions will be drawn together using consistent metadata. conventions. The recently revised guidelines for data producers stipulate that CCI data products should conform to version 1.6 of the CF conventions and use CF Standard Names for the main data variables. This recommendation has in turn led to further proposals for satellite data related Standard Names. The requirements of the satellite remote sensing community are likely to drive further additions both to the Standard Names and the wider CF conventions during the coming year. The CEDA vocabulary editor provides an efficient pipeline that will allow changes to the Standard Names to keep pace with the requirements of the scientists.

[17] http://www.cfconventions.org/

[18] http://www.cfconventions.org/documents/cf-standard-names

[19] http://www.esa-cci.org/

CEDA Discovery Web Services for NERC and MEDIN

Steve Donegan

The NERC Data Catalogue Service (DCS) and the Marine Environment Data Information Network (MEDIN) Data Discovery Portal (DDP) allow users to search catalogues of metadata harvested from a collection of dedicated providers. In the case of the NERC DCS these are metadata generated by all NERC dedicated data centres describing their available data resources. The MEDIN portal allows the search of metadata from all participating data providers in the MEDIN marine partnership. Search methodologies range from a simple free text search through to complex spatio-temporal and targeted text searching of specific metadata elements.

A key component of these portals operations is the Discovery Web Service (DWS), which provides the interface between the public facing web portal and the metadata catalogue content. Both NERC and MEDIN portals require an individual DWS and associated catalogue and both are run operationally by CEDA. CEDA also runs the Data Providers Web Service (DPWS) – this is a separate web service that controls the harvesting and ingestion of metadata into the catalogue as well as providing statistical functions for reporting purposes. The CEDA DWS infrastructure is comprised of three main components: Metadata Harvest (where metadata is collected from the providers); Metadata Ingest (where harvested metadata is inserted into the DWS catalogues); and a SOAP based interface that marshals requests received from the service portals and formats them as queries to the catalogues returning results according to the appropriate SOAP structure.

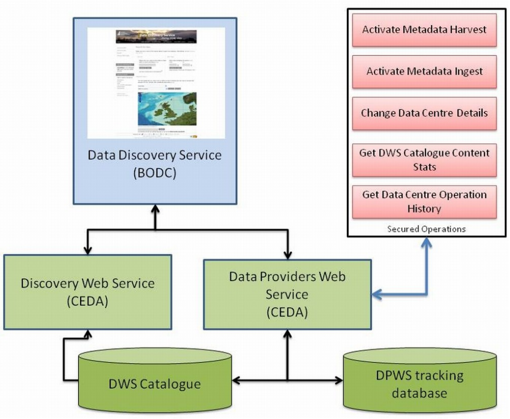

Data Discovery Service (BODC)
Activate Metadata Harvest
Activate Metadata Ingest
Change Data Centre Details
Get DWS Catalogue Content Stats
Get Data Centre Operation History
Secured Operations
Discovery Web Service (CEDA)
Data Providers Web Service (CEDA)
DWS Catalogue
DPWS tracking database

Figure 12: The DWS performs searches on the catalogue in response to requests from the DDS. The DPWS controls the flow of metadata into the DWS catalogue via a secured system from a tab within the DDS itself. Different role types define the operations users of the DPWS can perform.

Both the NERC DCS and the MEDIN DDP use INSPIRE and UK Gemini compliant ISO19115 profile metadata (albeit in differing profiles) to provide information on individual datasets and associated services held by the various data providers.

The DPWS arose from a need to provide a better method for NERC centres to track and control the harvest and ingestion of their metadata into the DCS catalogue. Previously these operations were performed by the use of open-source OAI-PMH software for the harvesting of metadata and a set of separate automatic scripts to place this metadata into the catalogue. Any errors or problems had to be dealt with individually and there was no easy system to provide an ingest or harvest history to the data providers. The DPWS provides an interface to control and track the harvest of metadata, not only by OAI-PMH but also by OGC Catalogue Service for the Web (CSW) as well as the ingest of this metadata into the catalogue. The DPWS additionally allows the querying of the DCS catalogue to provide basic statistics (such as number of records per provider).

CEDA is currently actively involved in upgrading the functionality of the DWS supporting the MEDIN portal (with corresponding extra funding from MEDIN). As part of this work CEDA is also providing additional harvest functionality (Web Accessible Folder) as well as integrating a framework to synchronise the contents of the MEDIN DWS Catalogue with that of a Geonetworks CSW. This allows MEDIN to meet its obligations to INSPIRE and the UK Location Programme.

Release of MOLES version established on ISO standards

Spiros Ventouras, Bryan Lawrence

The Metadata Objects for Linking Environmental Sciences (MOLES) model was originally developed within the Natural Environment Research Council DataGrid (NDG) project to fill a missing part of the ‘metadata spectrum’. It is a framework within which to encode the relationships between the tools used to obtain data, the activities which organised their use, and the datasets produced. With an emphasis on the relationships between entities which describe these things, it has a similar focus to ISO19156 Observations and Measurements (O&M) standard, and work over the last few years has focussed on harmonising MOLES and O&M.

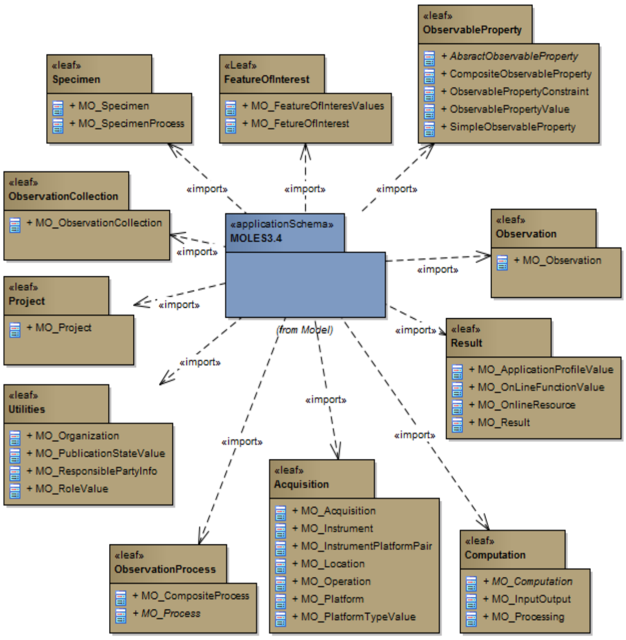


Figure 13:Components of MOLES information model.

MOLES is primarily of use to consumers of data, especially in an interdisciplinary context, to allow them to establish details of provenance, and to compare and contrast such information without recourse to discipline-specific metadata or private communications with the original investigators. MOLES can also be of use to the custodians of data, providing an organising paradigm for the data and metadata (how it is being deployed in CEDA).

A typical sequence of data capturing involves one or more projects/activities under which a number of actions are undertaken, using appropriate tools and methods to produce the datasets. MOLES is not particularly concerned with the details of the datasets structures (i.e. the data specification) but is aimed at capturing the important provenance elements associated with them. Key components of MOLES include:

- Project descriptions;
- the action itself and the processes used to acquire or generate the data;
- the collection of data into groups according to user requirements.

The concepts of MOLES v3.4 are rooted in the ISO 19100 series of standards from the ISO/TC 211 “Harmonised Model”. In particular, the model has been created:

a) following the guidance provided by ISO/TC 211 (i.e. ISO 19101, 19106 and 19109) and has been formalised in the Unified Modelling Language (UML), following the guidance of ISO 19103;
b) by integrating reusable modules of conceptual schemas defined within ISO 19100 series e.g. temporal schema, metadata, etc.

The establishment of MOLES on ISO standards ensures its interoperability and efficiency over a broad area of applications and can be used far beyond NERC boundaries. There is an international interest for MOLES which has lead to international collaborations involving CEDA, e.g. with University of Warsaw.

In 2011/2012 MOLES has also been the base model for national projects such as the CEDA metadata system and the DEFRA funded Agricultural Greenhouse Gas Platform, and the international LTDP (Long Term Data Preservation) project aiming at the development of a preservation strategy targeting the preservation of all European (including Canada) EO space data.

Portable Infrastructure for the Metafor Metadata System (PIMMS)

Charlotte Pascoe

PIMMS[20] provides institutions with tools to capture information about the workflow of running simulations, from the design of experiments to the implementation of experiments via simulations running models (see figure 14). It is important to know how and why a simulation was performed if you want to make use of another researcher's data. PIMMS uses the Metafor[21] methodology for simulation documentation which consists of a common information model (CIM), a set of controlled vocabularies (CV) and software tools. The initial PIMMS deployment will support climate model documentation, as it is based on the controlled vocabularies collected by the Metafor project in support of CMIP5[22]. PIMMS will extend the CMIP5 controlled vocabularies to cover paleoclimate simulations by partners at the University of Bristol and Limited Area Model simulations by partners at the University of Reading.

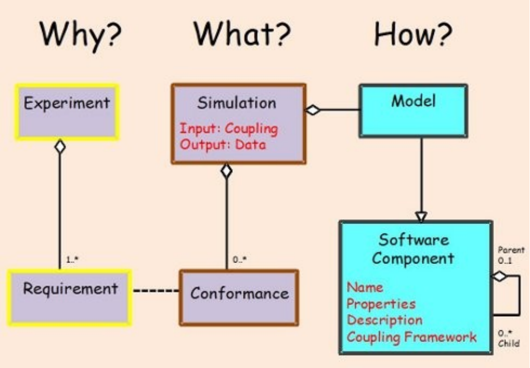


Figure 14: The PIMMS framework captures metadata about the life-cycle of a climate simulation. Here we see a UML view of the CIM elements that are populated by PIMMS, they explain why and how the simulated data was created. It is important to know this if you want to use another researchers data.

PIMMS will:

- Refactor the Metafor questionnaire for CMIP5 for use in university departments
- Create a new tool for describing experiments
- Engage the UK university community with the CIM
- Connect data holdings at the University of Bristol and the University of Reading to Metafor metadata
- Promulgate the Metafor approach to simulation documentation elsewhere in the UK

PIMMS controlled vocabularies are recorded using mind maps. Mind maps not only collate lists of controlled vocabulary but also provide a structure for the way the information is collected in the PIMMS web interface which ensure that a standardised set of metadata is collected. The key to the customisation of PIMMS is the modularity of its tools and the clear separation of structure (CIM) from content (CV). PIMMS will demonstrate this adaptability by creating a new controlled vocabulary to describe Integrated Assessment Models (IAM) in collaboration with the EU ERMITAGE project. The new CV will be used to reconfigure PIMMS to collect metadata about IAMs. PIMMS will further explore how the CV that is used to configure PIMMS may be of further use to our stake holders through the development of the text mining capabilities of the University of Cambridge ChemicalTagger[23] tool.

PIMMS will provide a portal so that research groups can view and search their own content, as well as publish their metadata to institutional, national and international services. PIMMS will also provide data node software so that data documented with PIMMS can be published to the web, both locally and to national and international services.

PIMMS embraces modern methods of communication and dissemination. We regularly blog about progress on the project for dissemination to JISC funders and interested parties who are following progress on the project. PIMMS meetings are held in an online virtual environment provided by Google+, these virtual face to face meetings are great for team bonding and are free.

[20] PIMMS: Portable Infrastructure for the Metafor Metadata System http://proj.badc.rl.ac.uk/pimms/wiki

[21] METAFOR: Common Metadata for Climate Modelling Digital Repositories http://metaforclimate.eu

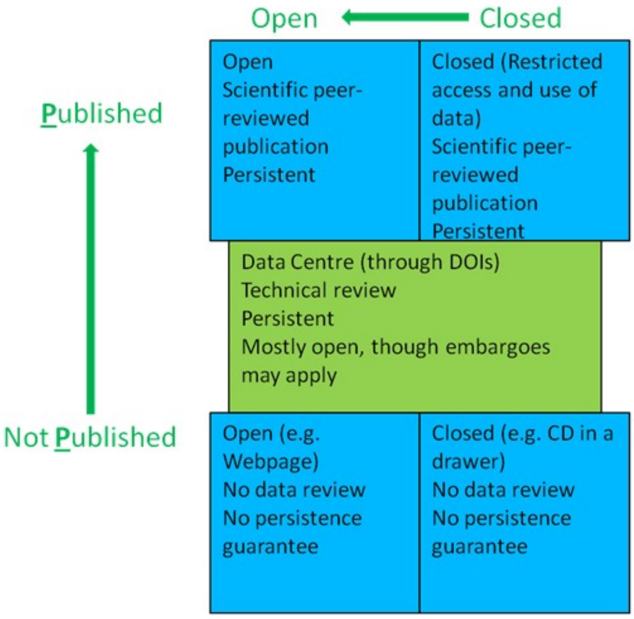
[22] CMIP5: 5th Coupled Model Inter-comparison Project http://cmip-pcmdi.llnl.gov/cmip5/

[23] ChemicalTagger: Natural Language Processing http://chemicaltagger.ch.cam.ac.uk/

Data Citation and Publication in CEDA

Sarah Callaghan

Scientists spend a great deal of time and effort in creating datasets, effort which is, for the most part, unrewarded by the current methods of assigning academic credit. The benefits of sharing data are well known, but without proper incentives, there can be a reluctance on the part of data producers to share their data. The NERC data citation and publication project, managed by Dr Sarah Callaghan of CEDA, aims to provide a mechanism for the citation and publication of data held in the NERC Environmental Data Centres, thereby providing a way for data producers to get the credit they deserve.



Open
Scientific peer-reviewed
publication
Persistent

Closed (Restricted access and use of data)
Scientific peer-reviewed
publication
Persistent

Data Centre (through DOIs)
Technical review
Persistent
Mostly open, though embargoes may apply

Open (e.g. Webpage)
No data review
No persistence guarantee

Closed (e.g. CD in a drawer)
No data review
No persistence guarantee

Figure 15: The tension between open and closed data, and Published and unPublished data.(Note that Published denotes that the data has been through a formal process (e.g. peer-review) which adds value to the data consumer.

The benefits of sharing data include the ability to discover and reuse data which has already been collected, avoiding redundant data collection and saving time and money; and providing opportunities for collaboration. Understandably, research funders are keen to encourage data sharing for these reasons. Unfortunately, with no widely accepted mechanism for data creators to obtain credit for their dataset creation efforts, they often prefer to keep data private until they have extracted all possible publication value.

Data citation and publication are proposed as mechanisms for ensuring that the shared data is of good scientific quality and is suitable for reuse, while ensuring the data creator receives proper attribution and credit. To these ends, the NERC data citation and publication project has developed a mechanism for formally citing datasets through the assignment of Digital Object Identifiers (DOIs) and is collaborating with international organisations such as DataCite, JISC and Wiley-Blackwell to change the scientific culture so that data publication and citation becomes the norm. This year has seen the launch of *Geoscience Data Journal*, a collaboration between the Royal Meteorological Society and Wiley-Blackwell, which came about as a direct result of previous project work[24] done by members of CEDA and funded by NERC and JISC.

Figure 16: Cover image for the Geoscience Data Journal, a collaboration between the Royal Meteorological Society and Wiley-Blackwell.

Data publication (with citation) will ensure that data become first class research outputs: available, peer-reviewed, citable, easily discoverable and reusable. This will facilitate data transparency and scrutiny, enhancing both research efficiency, and the academic status of data producers. Fundamentally, open data is good for science!

[24] Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) http://proj.badc.rl.ac.uk/ojims

OpenAIREplus

Sarah Callaghan

OpenAIREplus (2nd Generation of Open Access Infrastructure for Research in Europe) is a 30 month project, funded by the EC 7th Framework Programme. It will work in tandem with OpenAIRE[25], to support the research work of European scientists by creating and operating a robust, sustainable, and participatory open access infrastructure. This infrastructure will be responsible for the overall management, analysis, manipulation, provision, and cross-linking of a very broad spectrum of scientific publications and a selected subset of related datasets, It will do this through a suite of generic services and technologies, supported by a European-wide network. This large-scale project brings together 41 pan-European partners.

Figure 17: OpenAIREplus logo

OpenAIRE mainly concentrated on peer-reviewed articles as well as other important forms of publications (pre-prints or conference publications). OpenAIREplus will extend this work to other forms of scientific publication, including publication of scientific datasets. This data publication is of direct interest to NERC and STFC, and BADC is already working in this area via the NERC SIS Data Citation project.

The OpenAIREplus project offers us the opportunity to set the standards and develop the tools to be used in data and enhanced journal publications throughout the EU. With 41 partners in the project, there are significant opportunities for networking and collaborating on further research projects. BADC is one of three subject specific scientific partners, the others being EBI-EMBL (health-biology) and DANS (social sciences).

OpenAIREplus have incorporated "separate" communities (distinct use cases, with different metadata research data models) so these communities can help build the required generic information model and tools. The BADC was invited to take part as one of these communities, given our experience with data management and our good relationship with our user community.
More specifically, we are taking part in the following tasks:

- act as “consultants" to the information model (providing user requirements from our specific use cases and domains)
- provide OpenAIREplus with access to our data repository using explicit "connection" mechanisms (e.g., DOI's in publications and data sets)
- help parameterize instances of the OpenAIREplus interlinking tools for our specific disciplines/domains showcasing that the generic infrastructure will be able to work with specific cases.
- help in designing the end user tools for enhanced publications visualization and interaction (usability assessment included)
- participate in a study for researchers or data infrastructures publication- data linking practices

BADC have been given 9 months of effort in this project (worth 119 kEuro), spread out over the project’s 30 months run. Our involvement with OpenAIREplus will allow us to influence the future direction of enhanced publications throughout the EU.

[25] http://www.openaire.eu/index.php

Interactions with the European Space Agency

Esther Conway

CEDA has enjoyed another productive year forging strong collaborative links with the European Space Agency and UK Space Agency. The Long Term Data Preservation (LTDP) working group was established in response to the urgent need for a coordinated and coherent approach for the long term preservation of the existing European Earth Observation space data. The group was formed in January 2008 to establish a European common Long Term Data Preservation policy and a cooperative scheme for the implementation of a European LTDP System, with the aim of guaranteeing the complete European EO space data set preservation. Sam Pepler represents the UK Space Agency on the LTDP working group ensuring the UK space agency has adequate input to and awareness of the group's activities. Esther Conway also presented the LTDP surveys of standards procedures and Initiatives at the May 2012 LTDP workshop in ESRIN.

Phil Kershaw represented the UK Space Agency as a cloud computing expert for the Ground Segment Coordination Body (GSCB) meeting held at CNES in Toulouse, February 2012. This meeting enabled ESA to present their cloud computing strategy and to gather input from the member states on their plans in this area. Victoria Bennett has also continued to engage with GSCB presenting ATSR data management strategies at relevant workshops.

In addition to contributing to ESA's strategic aims in a voluntary advisory capacity, CEDA also supplies direct support to ESRIN through a range of consultancy, research and development activities, funded by a LTDP management support contract. Victoria Bennett also continues her successful secondment to the ESA Climate Office working on Data Standards for the Climate Change Initiative Programme and links between ESA, NCEO, ISIC and CEMS; with support from Sarah James in CEDA.

Figure 18: Envisat

October 2011 also saw the start of the ESA-lead SCIence Data Infrastructure for Preservation with focus on Earth Science (Scidip-ES) FP7 project, of which CEDA/STFC is a major contributing partner. The project aims to deliver generic infrastructure services for science data preservation that address the issues of persistent storage, access and management. The work supports setting up a European Framework for the long term preservation of Earth Science data through the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of the generic infrastructure services in the Earth Science domain.

CEDA also has an important role in the distribution of ESA data to scientists as it is able to provide efficient data access and acquisition for the UK research community through the NEODC data management infrastructure and network capacities. During 2011/12 CEDA has demonstrated that it is a centre of expertise in the management of environmental data, with skills and knowledge in demand by both the UKSA and ESA. It has enhanced its proven track record in collaborating effectively with large national and international organisations; leveraging its significant experience for the benefit of both the UK and wider European research communities.

CEDA and the Earth System Grid Federation infrastructure for CMIP5

Stephen Pascoe.

What does it take to provide policy makers with realistic predictions of the Earth's future climate? It requires the coordination of climate modelling groups situated around the world to generate, exchange and archive petabytes of simulation data and then distribute it to the scientists who will analyse the data and publish their results. CEDA has taken a central role in this endeavour for the 5th Climate Model Inter-comparison Project (CMIP5) thus ensuring NCAS researchers have state of the art access to the data that allows them to contribute to the IPCC 5th Assessment Report (AR5).

Many data intensive areas of science are moving into the Petabyte scale but CMIP5 is novel in its need to transport peta-scale data from globally distributed producers to equally globally distributed consumers. Simply transporting this data across continents has required CEDA to invest in dedicated light-paths across the UK and to continental Europe and optimise existing network links, both internally and externally, to enable terabytes of CMIP5 data to be transferred in and out of our servers each day.

The CMIP5 archive is built upon a software infrastructure developed by the Earth System Grid Federation (ESGF), a collaboration of data centres tasked with developing and maintaining a global data infrastructure for Earth System Science. As a founding member of ESGF, CEDA has developed several of the critical software components used across the globe to archive CMIP5 data.

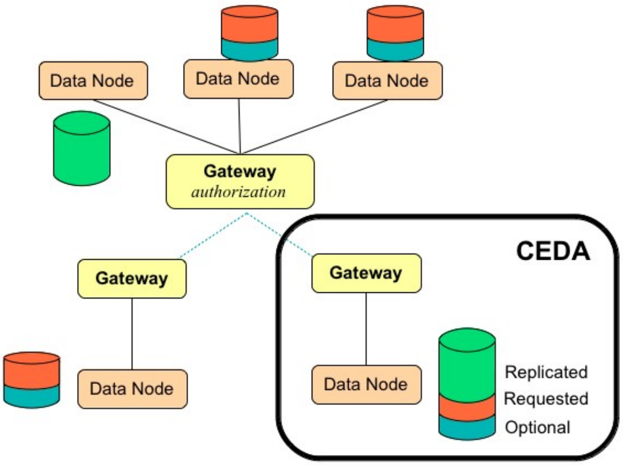


Figure 19: Topology of the first generation ESGF infrastructure with distribution of different classifications of CMIP5 data

CEDA's role in the ESGF infrastructure is multi-faceted. We run one of the ESGF hub Gateways where users can search for and download CMIP5 data from wherever it is stored globally. We store CMIP5 model outputs from the UK Met Office and also publish data for our European partners such as Institut Pierre Simon Laplace and the EC-EARTH group. As one of the hub Gateways we replicate data from other continents, thus providing European scientists with faster access to global data.

Partnering with the Deutcher Klimarechnung Zentrum Hamburg, CEDA will maintain the official AR5 version of the CMIP5 archive so that future research can utilise the same data that was used for the AR5 literature. This stable archive will be citable via Digital Object Identifiers (DOIs).

CEDA's leadership in data management for CMIP5 increases NERC's preparedness to meet future Big-Data challenges. We are currently rolling out ESGF's next generation “P2P” architecture with a completely re-engineered user interface and more powerful search capabilities. ESGF P2P allows us to deliver multiple projects through the same infrastructure such as the CORDEX regional modelling project and the Obs4MIPS satellite observation project. ESGF P2P provides programmable interfaces to search and download that enable applications to access the data. Within the IS-ENES EU FP7 project we are working with the Dutch meteorological institute KNMI to create a visualisation and analysis portal specifically targeted at the Climate Impacts community using data exposed through the ESGF infrastructure.

Defining a Data Model for the Agricultural Greenhouse Gases Platform

Ag Stephens, Charlotte Pascoe & Spiros Ventouras.

Current estimates suggest that agriculture contributes about 8% of the UK Greenhouse Gas emissions, primarily from methane (livestock and their manures) and nitrous oxide (soil and livestock excreta). Emissions are generated using empirical measurements and modelling, and reported as part of the annual UK Emissions Inventory. The Agricultural Greenhouse Gas (AGHG) Research Platform[26] aims to improve the accuracy and resolution of the reporting system by providing new experimental evidence on the factors affecting emissions and statistics relevant to changing farming practices in the UK.

CEDA is leading a data modelling activity to support the archive and metadata system for the AGHG Platform. The formal data model will describe the storage and documentation of the experimental, survey and model- based data collated by this scientific community. There are 5 key stages to the development of this Data Model:

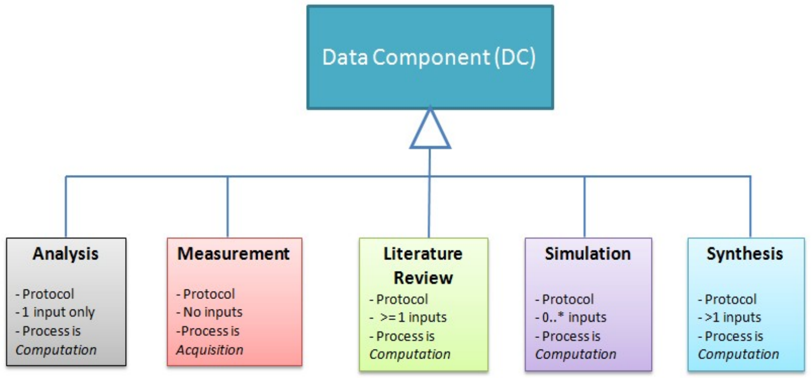


Figure 20: The main Data Component classes of the AGHG Platform Data Model. The specialised classes are distinguishable by the number of inputs and the type of process used to generate the results.

1. Engaging domain experts: Managing workshops and communication with the data providers and potential users, and scoping stakeholder requirements such as the high-level emissions reporting.
2. Profiling of the main data types: Identifying the common types of data and information that are important to this community are essential in terms of producing an appropriate model.
3. Building on existing standards and models: By building on top of existing work, especially international standards, the final data model should be interoperable with other catalogues and services. It should also be compatible with legislative requirements such as INSPIRE[27].
4. Communicating the model to domain experts: The language used to communicate the complex concepts of a standards-based data model must be carefully managed. Data modelling jargon obstructs an effective dialogue between the model developers and the scientists they support.
5. Validating the Data Model with real datasets: The final test is to put real-world data and metadata into the classes and properties defined within the Data Model. At this stage the data providers should begin to appreciate the value of the process and any deficiencies and ambiguities in the model can be highlighted and addressed.

The primary classes of the Data Model are the **Activity**, **Dataset** and **Data Component**. The concept of the Data Component was introduced because a Dataset is typically composed of a number of sub-components. Figure 20 shows how the concept of the Data Component is specialised into 5 classes that reflect the range of data-production methods that take place within the platform. The final Data Model will be published later this year after being validated against numerous example datasets.

[26] Agricultural Greenhouse Gas Research Platform: http://www.ghgplatform.org.uk/

[27] Infrastructure for Spatial Information in the European Community: http://inspire.jrc.ec.europa.eu/

Evaluating the *Iris* Software Package – tools for data manipulation

Ag Stephens, Alan Iwi & Stephen Pascoe
with Guy Griffiths, David Hassell & Andy Heaps, and Neil Massey.

Those working in the earth system sciences will be familiar with at least one data manipulation package. Some packages are well-written and well supported but come with a price tag – this leads to problems when sharing code with collaborators who have paid for a rival tool kit. This is very important to the research community; as collaboration requires sharing of data and analysis code. Many researchers will only work with code that has unrestricted access, allowing joint-development regardless of location.

So much data analysis is taking place that it is easy to imagine the thousands of hours being wasted in duplicated code using different tool kits. This is why many of us continue to search for a package that combines functionality, performance and is licensed in a way that promotes easy collaboration. However, any new package must also compete with the *status quo* in which (1) most scientists have a package of choice that they are comfortable with and have written a set of scripts/libraries for; (2) many organisations have financially bought into one or two packages that they "support" internally.

The **Iris data manipulation software** currently being developed by the UK Met Office aims to provide a range of functionality that is required by both scientists and data managers. The architecture, consisting of a Python package built upon NetCDF4-python, Numpy and MatPlotLib, is a good fit for the data analysis stacks currently in use and envisioned by NERC. With an Alpha release available, and the Met Office showing an interest providing a licence for research use, we developed an evaluation process to inform NERC and NCAS about the suitability of Iris for scientific usage and tool-building. A group of NERC software developers were brought together for a 2-day evaluation workshop at Reading University in March 2012. A set of tests and reviews were carried out to investigate the installation, source code, data model, test suite, documentation, functionality and potential evolution of Iris. Iris was compared to existing tools in terms of the amount of code, clarity of code and time taken to undertake a given task. The following table provides a summary of the results[28].

Criterion	Comments
Usability	Strongly object-oriented API may present a barrier for some scientific users but may become a strength once familiar. Documentation is clean, well-formatted, thorough and usefully tied in to the code-level documentation strings. Relatively straightforward to install.
Fitness for purpose	Most basic functionality expected for this type of tool appears to be in place. In some tests Iris was slower than other packages but performed reasonably. Some tests could not be run using Iris but we note it is still in Alpha release. The "cube" model within Iris maps broadly onto the CF data model and NetCDF outputs from Iris generally scored well in terms of CF-compliance.
Sustainability and maintainability	The Iris team develops the code base on top of sound software engineering principles and clear requirements. The source code is well structured and it is relatively easy to analyse and test, showing that it can be understood, changed and extended by those outside the Iris Team. Iris is built upon technologies that can interoperate with a range of tools and operate within many environments. The future uptake of Iris may hinge on whether it can be licensed in a manner that allows collaborators to share code.

The evaluation team believe that Iris has significant potential for use in both data management and scientific analysis. The code, interface and development process are well thought-out and implemented. However, at Alpha version, the software is not yet complete enough for operational usage. From the point-of-view of the NERC community, the potential usage and uptake of Iris are likely to depend on the licensing and governance processes – both of which are under active discussion at the time of writing.

[28] *Thanks to all contributors to the evaluation process. The final report is available on request from* ag.stephens@stfc.ac.uk

Software development in CEDA

Maurizio Nagni

CEDA take an active interest in maintaining existing software as well developing new software following the evolving requirements of CEDA and its external partners. CEDA primarily uses Java and Python and both “in-house” and external commercial and Open Source tools. All this requires an efficient software life-cycle as projects become more and more complex, and so CEDA takes advantage of best practices in software development (those well established in the commercial sphere as well as in academia) such as agile software development and continuous integration.

The agile paradigm breaks tasks into small increments with minimal planning that does not include long-term planning. Each iteration involves a team working through a full software development cycle, including planning, requirements analysis, design, coding, unit testing, and acceptance testing when a working product is demonstrated to stakeholders. This minimizes overall risk and allows the project to adapt to changes quickly. The agile methodology and continuous integration help to improve both the communication between developers and users and the quality of the final result.

A continuous integration environment was made available to the internal software development group in December 2011 and is now instrumental in many projects such as the Discovery Web Service, Data Providers Web Service, CedaEditor, HPFos, Newmoon. Its importance is expected to grow in the future, as confirmed by the improved continuous integration system inside the JASMIN-CEMS facility. All projects should now complete a periodic review of activities in order to verify how suitable the proposed solution is to the project aim.

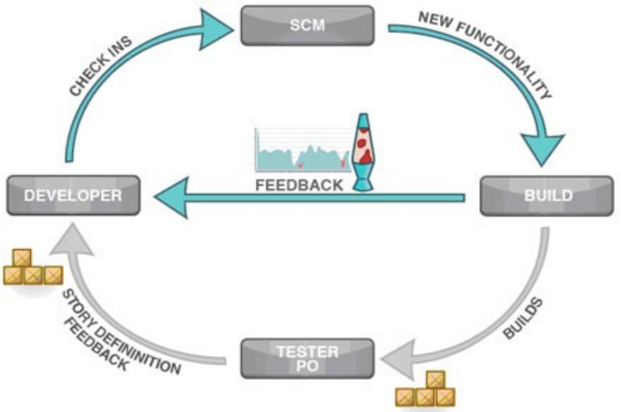


Figure 21: Agile and continuous integration methods connect in a circular   incremental way all the actors which are involved in software development

The CEDA environment is naturally a dynamic one, having to follow both advances in technology and to synchronize its service with other external entities in order to maximize the final user experience. The importance of the “Software as a Service” (SaaS) concept is acknowledged in order to directly offer to our users through an Internet connection, raw data, processed data products and to enforce the use of standards.

Enterprise-class architecture such as Java Enterprise Beans and application servers like the JBoss reliability and robustness allow NERC Data Catalogue Service (DCS) and the Marine Environment Data Information Network (MEDIN) Data Discovery Portal (DDP) to use the Discovery Web Service (DWS) and Data Providers Web Service (DPWS) **web services** as data search engines. This step perfectly fits with the SaaS paradigm offering rapid scalability as well proven reliability.

Positive feedback was received for the new CedaEditor (which manage the new CEDA MOLES3 metadata catalogue) during the European Geosciences Union General Assembly (EGU 2012, Vienna). Similarly the HPFos (an HPFELD sub-project implementing a python library to exploit the Opensearch specifications) show the quality and the potential of the CEDA investment in software development. This and other feedback demonstrate the agile and continuous integration approaches have already shown good results and have great potential to further improve CEDA capabilities.

OWSLib[29]: Open Source development encourages innovation.

Dominic Lowe

OGC web services underlie a number of the CEDA user facing services; for example, providing the tools which support map based visualisation. When CEDA began exploiting OGC web services (mid 2000's), there were no software implementations which supported large model and earth observation datasets. Nonetheless, there were good reasons to work with OGC web services: a lot of design thinking had gone into what to do and how, and the technology provided a route to supporting interoperability between the datasets typical at BADC, and a user community exploiting geographical information systems for environmental consultancy etc. Hence, CEDA invested in building a stack of OGC compliant software: the CEDA OGC Web Services (COWS), which could manipulate and expose CEDA data.

While building this software, we spent some time extending an Open Source python library for accessing Open Geospatial Consortium services, such as Web Map Service, Web Coverage Service, Web Feature Service (services used in INSPIRE and other projects). At the time OWSLib was a very small, somewhat dormant, Open Source coding project with a couple of other developers. Much of the work done within the scope of the CEDA work was to make the library easier to use and exploit to take the hard work out of building tools to access the OGC services. Additional work was done on extending it to support Web Coverage Services and Web Feature Service 2.0 (both necessary for dealing with our large datasets). This reinvigorated the OWSLib project and it began to gain traction within the python geospatial community.

Over time, a few people and organisations began to integrate the OWSLib code in several widely used geospatial products, such as QGIS[30] (an Open Source geographical information system) and GeoNode[31] (a platform for the management and publication of geospatial data). About a year ago, the OWSLib team began to mirror the OWSlib source code on the 'GitHub' platform. GitHub is a social network for developers; programmers can take copies of each others' code, make their own changes, and can request that their changes are considered for integration back into the original codebase. This allows software to be developed by many people at the same time, without requiring initial permission to try new things, which creates an innovative environment for development. This model has been so popular that we have now moved the code entirely to GitHub (so it is no longer just a mirror of code stored elsewhere).

Since being more visible on GitHub an increasing number of developers have extended OWSLib to provide python-based access to other types of OGC services, such as Sensor Observation Services, Catalogue (CSW) Services and Web Processing Services (WPS), making OWSLib into a very well rounded tool kit. Also within GitHub, it can be seen that developers are extending OWSlib for experimental reasons. These extensions can be brought back into the core codebase if they prove useful. The value of the software is currently estimated to be around £100,000, for an investment which was probably a tenth of this. In June 2012, OWSlib was included in the official openSUSE Linux distribution as part of the geospatial applications repository. Build packages for Scientific Linux, RedHat, CentOS are underway, so soon this code will be available to nearly all scientific computer users on the planet.

None of this would have happened without a need that was met by exploiting external open source software, the skill set to exploit and develop that code “in-house”, and a corporate willingness to contribute back to the project. The OWSLib source code is available on GitHub at https://github.com/geopython/OWSLib with an easy installation python 'egg' at: http://pypi.python.org/pypi/OWSLib/.

[29] OWSLib = An Open Source python library for accessing Open Geospatial Consortium (OGC) web services.

[30] http://www.qgis.org

[31] http://wwww.geonode.org