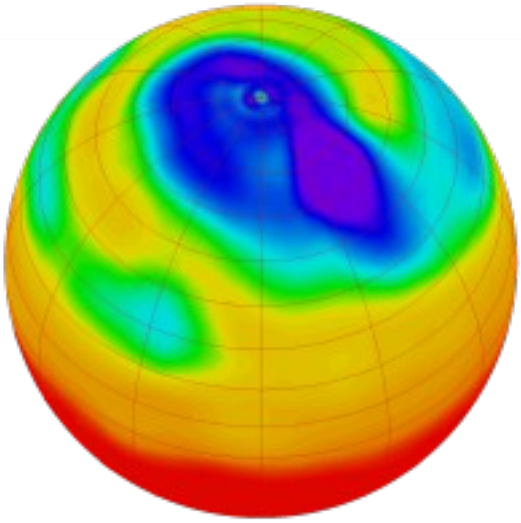
Science & Technology Facilities Council



STFC
Centre for Environmental Data Archival (CEDA)
Annual Report
2011
(April 2010-March 2011)

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

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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 established by the amalgamation of the activities of two of the Natural Environment Research Council (NERC) designated data centres: the British Atmospheric Data Centre, and the NERC Earth Observation Data Centre. The process began with administrative functions (in 2005) and has proceeded steadily since, as new activities have and continue to be accreted into CEDA. Until 2008, the constituent parts of CEDA reported independently to NERC, but in 2009 we produced the first public report for CEDA. We are pleased to present here our third annual report, covering activities for the "2010" year (actually from April 2010 to the end of March 2011). The report itself is in two sections, the first broadly providing a summary of activities and some statistics with some short descriptions of some significant activities, and new this year, a second section introducing some of the staff, and what they do from day-to-day. (Note that although the UK solar system data centre joined CEDA in this year, we have yet to include significant reporting from that activity.)

CEDA staff are involved in nearly all the major atmospheric science programmes under way in the UK, in many earth observation programmes, and in a wide range of informatics activities. The CEDA involvement in informatics is main targeted at achieving three main objectives: (1) Providing suitable tools to document and manage both high volume and highly heterogeneous data both in CEDA and the community; (2) Delivering tooling and services to enable the community to exploit CEDA data holdings, and; (3) Improving the ability of fundamental standards both to improve the likelihood that others can build standards compliant software we can deploy, and to support interdisciplinary science.

While all of these activities are of course aimed squarely at supporting the UK community, of necessity, and like the science programmes in which we work, we could not complete our objectives without both building on and contributing to other activities – both in the UK and abroad. In particular we rely on partnerships we have built with other organisations so that we can leverage the informatics investments elsewhere to deliver solutions to the three objectives above. One of our closest partnerships is with the German Climate Computing Center (DKRZ), but we have strong connections with a range of other institutions, particularly those within the Global Organisation for Earth System Science Portals (GO-ESSP, see http://go-essp.gfdl.noaa.gov). Other important relationships include with the Met Office, our sister data centres in the Natural Environment Research Council community, and the European space data community (in particular the European Space Agency).

In the report that follows it will be seen that many of our activities involve partners from the list above delivering solutions in support of our two main scientific communities: the UK atmospheric and earth observation communities. In 2010 these communities delivered two especially major challenges to CEDA, challenges which are likely to be with us for some years: firstly, how to acquire, document, distribute, and support the massive amounts of data being produced by international model intercomparison projects (and in particular, CMIP5); and secondly, how to develop a strong engagement with the new International Space Innovation Centre sharing our site at Harwell. Clearly ISIC provides us with a vehicle which could greatly strengthen our ability for the data collected by the UK academic community to make greater commercial and scientific impacts. Our highlights section indicates some of the activities we have begun in support of these two challenges, and we might expect to see much more on these in future years.

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

Bryan Lawrence, Director

Table of Contents

Summary of 2011/2012

CEDA continues to support the atmospheric 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 210 TB of data in 21 million files to 3175 distinct users.

Major international collaborations built around two European projects, Metafor and IS-ENES, have been strengthened, even as a significantly larger global collaboration to deliver an "Earth System Grid Federation" to support the upcoming fifth Coupled Model Intercomparison Project (CMIP5) has been further developed under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP). CEDA staff are taking leading roles in all of these initiatives.

In additional to the core remit of serving the Natural Environment Research Council's National Centres of Atmospheric Science and Earth Observations (NCAS and NCEO), CEDA has delivered major projects in support of both Defra and DECC[1] (providing the data systems for the UK Climate Projections 09 and IPCC[2] Data Distribution Centre respectively). A number of other projects with funding from a range of other bodies were also carried out, including work for the European Space Agency, the Joint Information Services Committee and others.

Notable Events

1. The first data from CMIP5 was delivered into the CEDA CMIP5 archive, marking a significant milestone in our support of the international climate community.
2. The NERC SIS Data Citation and Publication project is a cross data centre initiative which aims to develop a method of citation, peer-review and publication of datasets stored in NERC repositories. This will be of benefit to wide sectors of the NERC community as it will provide scientists with academic credit for ensuring that their data is properly documented and archived in a trusted data repository, thereby allowing the data to be more easily discovered and re-used. At this time the focus has been on citation of datasets. The BADC are leading the project and have been working closely with members of the British Library in order to develop the procedures and standards required to assign digital object identifiers (DOIs) to datasets, thereby allowing those datasets to be cited in the same manner as an academic paper. Interactions have also been carried out between project members and 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. Discussions have also been held with Wiley-Blackwell about how to enable the journal Atmospheric Science Letters to publish datasets. Sarah Callaghan (BADC) has been invited to join the editorial board of Atmospheric Science Letters as an associate editor, in order to work closely with Wiley to develop a new format of data paper for that journal.
3. Eduardo D. da Costa joined the CEDA staff as a data scientist specialising in climate model data. His work aims to capitalise on the CMIP5 data by getting it used as widely as possible. This includes the transformation of raw climate model data into end-user friendly formats and producing such processed products which are expected to improve the flow of information and to make a direct impact on the knowledge transfer from research to actual adaptation and/or mitigation policy.
4. Scientists working at the BADC, NCAS CMS and NCAS Climate worked together to produce a web based questionnaire (http://q.cmip5.ceda.ac.uk/) which is currently being used to collect information and metadata from the climate modelling groups who are submitting data for the next Coupled Model Inter-comparison Project Phase 5 (CMIP5). This climate model data will form the

[1] UK Department of Energy and Climate Change

[2] Intergovernmental Panel on Climate Change

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 questionnaire is an ambitious metadata collection tool and will provide the most comprehensive metadata of any climate model intercomparison project.

5. On 14th April 2010 the Eyjafjallajokull volcano in Iceland erupted, spewing clouds of volcanic ash into the atmosphere which was then carried aloft over the UK and Europe. Whilst UK and European airspace was closed for 5 days, and the commercial airline industry grounded, the atmospheric science community was actively making measurements of the ash plume with airborne and ground-based instruments. The BADC team played a central role in this data collection by providing the means to keep track of what was being measured, and collating the data as quickly as possible into a useful dataset. Support was provided in 3 main ways:
   i. Provision of a central location to discover, view and disseminate data and supporting metadata from the science community. This included a secure upload area and the rapid development of and agreement on distribution conditions in order to promote wide usage whilst protecting the rights of data providers.
   ii. Collection and distribution of additional data to support the analysis of the measurements made by the academic community. These data include near-real time meteorological fields and products from Met Office, Met Office operated lidar instruments, near-real time IASI satellite data from EUMETSAT.
   iii. Production of an event log mapping tool, allowing users both to enter details of events and observations related to the eruption, and to view the events in their spatio-temporal context. This tool can be found at: http://cedaapp1.badc.rl.ac.uk
6. As part of the NERC SIB project to revitalise the Discovery Service, CEDA have worked jointly with BGS and BODC to develop and deploy a metadata portal. This web service has been updated to reflect NERC's requirements and requires metadata according to the new NERC standard. It has recently been launched for operational use and is central for searching NERC's data holdings. As part of its efforts to run this system operationally, CEDA have also developed and deployed the Data Providers Web Service (DPWS). The DPWS centralises many previously disparate processes governing the harvesting and ingestion of metadata into the Discovery service and minimises CEDA's management burden in running this service. This enables other data centres and NERC HQ to control their data in the Discovery service and monitor assets without the need for active response from CEDA. Preparation for launch of the data discovery service has included a full review of the CEDA metadata catalogue, which describes all the data held in the CEDA archives and is a key component in building services to permit the wider research community to discover relevant data. The catalogue review ensures that the data are described by spatial and temporal ranges, and that their quality and lineage are described. The DDS brings not only CEDA's 250+ data holdings together with similar catalogues from the other NERC data centres, but also ensures that these entries are all NERC data holdings are described by INSPIRE compliant catalogues.
7. CEDA were also commissioned to develop, deploy and run operationally a Metadata discovery service that underpins the Marine Environment Data Information Network (MEDIN) portal. This portal requires the collection and ingestion of marine metadata according to the MEDIN metadata standard, itself a profile of ISO19115 and UK Gemini and INSPIRE compliant. CEDA have run this service successfully for MEDIN for almost 12 months with few problems and have responded quickly to address and resolve issues for our MEDIN clients.

Major Collaborations

In 2010/2011, 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[3] and GFDL[4]) 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[5] 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.

Publications

Meteorological influences on the design of advanced aircraft approach procedures for reduced environmental impacts"; Liling Ren, Tom G. Reynolds, John-Paul B. Clarke, David A. Hooper, Graham A. Parton and Anthony J. Dore; Meteorological Applications (2010). Published online; DOI: 10.1002/met.206 (Also picked up as a news story in NERC's Planet Earth[6])

Parton, G., Dore, A. and Vaughan, G. (2010), A climatology of mid-tropospheric mesoscale strong wind events as observed by the MST radar, Aberystwyth. Meteorological Applications, 17: 340–354. doi: 10.1002/met.203

Bell, C. J., Gray, L. J. and Kettleborough, J. (2010), Changes in Northern Hemisphere stratospheric variability under increased CO2 concentrations. Quarterly Journal of the Royal Meteorological Society, 136: 1181–1190. doi: 10.1002/qj.633

Thomas, G. E., Poulsen, C. A., Siddans, R., Sayer, A. M., Carboni, E., Marsh, S. H., Dean, S. M., Grainger, R. G., and Lawrence, B. N.: Validation of the GRAPE single view aerosol retrieval for ATSR-2 and insights into the long term global AOD trend over the ocean, Atmos. Chem. Phys., 10, 4849-4866, doi:10.5194/acp-10-4849-2010, 2010.

[3] National Center for Atmospheric Research

[4] Geophysical Fluid Dynamics Laboratory

[5] German Climate Computation Center

[6] http://planetearth.nerc.ac.uk/news/story.aspx?id=820

Major Conference Presentations

Callaghan, SA (2010) Patterns of precipitation: Fine-scale rain dynamics in the South of England. In: EGU2010, 2nd – 7th May 2010, Vienna, Austria.

Callaghan, SA and Hewer, Fiona and Pepler, Sam and Hardaker, Paul and Gadian, Alan (2010) Data

Publication in the Meteorological Sciences: the OJIMS project. In: EGU2010, 2nd – 7th May 2010, Vienna, Austria.

Callaghan, Sarah and Guilyardi, Eric (2010) The METAFOR project: providing community metadata standards for climate models, simulations and CMIP5. In: EGU2010, 2nd – 7th May 2010, Vienna, Austria.

Kershaw, Philip and Ananthakrishnan, Rachana and Cinquini, Luca and Lawrence, Bryan and Pascoe, Stephen and Siebenlist, Frank (2010) A Flexible Component based Access Control Architecture for OPeNDAP Services. In: European Geosciences Union General Assembly 2010, 2-7 May 2010, Vienna.

Kershaw, Philip and Lawrence, Bryan and Lowe, Dominic and Norton, Peter and Pascoe, Stephen (2010) Applying the Earth System Grid Security System in a Heterogeneous Environment of Data Access Services. In: European Geosciences Union General Assembly 2010, 2-7 May 2010, The Austria Centre Vienna.

Lowe, Dominic and Woolf, Andrew (2010) Evolution of Climate Science Modelling Language within international standards frameworks. In: European Geosciences Union General Assembly 2010, Vienna.

Pascoe, Charlotte L and Lawrence, Bryan and Moine, Marie-Pierre and Ford, Rupert and Devine, Gerry M (2010) The CMIP5 Model Documentation Questionnaire: Development of a Metadata Retrieval System for the Metafor Common Information Model. In: European Geosciences Union General Assembly 2010, 3-7 May 2010, Vienna.

Pascoe, Stephen and Stephens, Ag and Lowe, Dominic (2010) Pragmatic service development and customisation with the CEDA OGC Web Services framework. In: European Geosciences Union General Assembly 2010, Vienna.

Other Formal Presentations

Kershaw, Philip (2010) NERC DataGrid Security. In: NERC 2009 Data Management Workshop, 17-18 Feb 2009, Oxford Belfry Hotel.

Pepler, Sam (2010) Why should NERC pay for BADC? In: Research Data Management Forum (RDMF5), Economics of Applying and Sustaining Digital Curation, 27-28 October 2010, Chancellors Hotel and Conference Centre, Manchester.

Alastair Gemmell, Jon Blower, Keith Haines, Hugo Hiden, Philip Kershaw, Bryan Lawrence, Stephen Pascoe, Simon Woodman, The MashMyData project Combining and comparing environmental science data on the web, Extended Abstract, UK e-Science AHM, Sept 2010

Lawrence, Bryan (2010) British experience with building standards based networks for climate and environmental research (Keynote: Information Network Workshop, Canberra, November, 2011)

Lawrence, Bryan (2010) Rethinking metadata to realise the full potential of linked scientific data (Keynote: Metadata Workshop, Gold Coast, November 2011)

Lawrence, Bryan (2010) Provenance, metadata and e-infrastructure to support climate science (Keynote, Australasian e-Research 2010, November 2011)

Meeting Attendance

NCAS Annual Staff Meeting and Atmospheric Science Conference 2010: Most CEDA staff attended the NCAS annual meeting in Manchester in July. Wendy Garland gave a presentation: "New approach for supporting rapid response projects – CEDA's role in the Icelandic volcanic Ash cloud research" and four posters were presented.

6-7 July, ESA LAST Technical workshop (LTDP programme): The Long Term Data Archive Study on New Technologies. Project to evaluate new technologies and best practices for Long Term Archives for the ground segments of existing and future EO missions. Attended by Dominic Lowe from CEDA

2-13 October 2010, NASA JPL/PCMDI CMIP5 meeting: Workshop to discuss EO datasets to be contributed to the CMIP5 archive by NASA. Victoria Bennett attended

21 October 2010, FIRST workshop (LTDP programme):FIRST = "definition of LTDP user Requirements and preservation data set composition", Project to analyse Earth Science user requirements, including accessibility and exploit-ability aspects. Victoria Bennett attended the workshop & presented CEDA expertise "Environmental Science Information Requirements".

14-16 September, first co-location meeting for ESA CCI (Climate Change Initiative): All CCI project teams met to discuss products. Bryan Lawrence was invited speaker on data standards, Bryan and Victoria Bennett attended this meeting.

NCEO Annual Conference 2010: Sam Pepler, Victoria Bennett, Graham Parton attended the NCEO Annual Conference in Leicester and presented 5 posters:

- "EO Data Visualisation at CEDA"; Victoria Bennett et al
- "The Icelandic Volcanic Ash Cloud Research: CEDA's role"; Graham Parton et al
- "Overlay Journals and the path to data publication"; Sarah Callaghan et al
- "The Common Information Model for Climate Modelling Digital Repositories: The Metafor Project"; Sarah Callaghan et al.
- "Standard Names for CMIP5"; Alison Pamment et al.

International Conference on Airborne research for the Environment (ICARE) in Toulouse, France & EUFAR annual meeting (October 2010): attended by Wendy Garland

OGC (Open Geospatial Consortium) and INSPIRE: Dominic Lowe and Spiros Ventouras attended a number of OGC and INSPIRE related workshops and meetings: including INSPIRE editor meeting (Ispra, Dec 2010), OGC/GIS workshop (Met Office, Nov 2010), OGC Technical Committee (Meteo-France, Sep 2010)

RSPSOC 2010: NEODC were represented at RSPSoc 2010 in Cork by Graham Parton and Steve Donegan, where they ran the NCEO trade stand.

STFC Environment Futures Workshop: Charlotte Pascoe represented and promoted CEDA (Nov 2010)

STEM Ambassadors: Charlotte Pascoe supervised a work experience student in summer 2010 and following this was trained and approved to become a STEM-ambassador. Sarah Callaghan is also a STEM ambassador.

Other Outreach and Knowledge Exchange

Discussions continue with Infoterra and other parties regarding future collaboration opportunities in the context of ISIC

ITT have contacted us regarding potential work areas of common interest. One particular topic relates to secure data access via OpenDAP and a meeting is being planned.

Discussions are under way with Wiley about publication of a data journal

CEDA are talking to AEA about exploitation of the CMIP5 archive (in the context of MIRP – Models and Impact Relevant Prediction; NERC KE grant)

Help desk and associated services

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. 92% of user queries are handled by 2 CEDA Data Scientists while the 8% remaining are covered by other CEDA team members as necessary.

Statistics for period 1st April 2010 to 31st March 2011

CEDA Queries closed	4164 - 90% are BADC queries
Total CEDA registered users (to 21/05/2010)	16323 – includes 3584 new users in period for BADC and NEODC
NERC funded active users (to 21/05/2010) (An active user has access to one or more restricted datasets)	4% of Total BADC and NEODC registered users (no figures for UKSSDC) - or 22% of all active users
Identifiable users actively downloading	3175
Total download volume	212.5TB (69% increase from 2009/10) - in over 21 million files (increase of 4 M from 2009/10)

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 your prompt response! ... I must admit that I have been thoroughly impressed with all aspects of the service that the BADC provides."

"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 help desk 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."

"I would just like to complement the service that you offer, as it offers a very reliable and easily accessible source for data which I, as a statistician, find a very valuable commodity."

"You are so helpful at BADC!"

"Once again, you are my favourite person :) Thank you so much!"

While 63% of BADC and NEODC users and 24% for UKSSDC are based in the UK (mostly universities), the outreach of CEDA services goes well beyond the UK borders and atmospheric physics as shown below – with increasing interest from BRIC countries.

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 2010/2011		
drslib Version 0.3.0a3 (Python)	Support library for the CMIP5 Data Reference Syntax and data versioning tool.	http://pypi.python.org/pypi/drslib
cfchecker Version 2.0.3 (Python)	Packaging of the online cfchecker tool developed at Reading for easy installation and command-line use.	http://pypi.python.org/pypi/cfchecker
NDG-XACML Version 0.4.0 (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.	http://pypi.python.org/pypi/ndg-xacml/
thredds_security_test Version 0.1.0 (Python)	A framework for verifying access constraints for ESG Federation data access services using the THREDDS and OPeNDAP standards.	http://pypi.python.org/pypi/thredds_security_test
urllib2pyopenssl Version 0.1.0	PKI security library for HTTP Python clients. Improves HTTPS pport for Python.	Available soon at http://pypi.python.org/pypi/urllib2pyopenssl
Software Packages improved in 2010/2011		
1. NDG-SAML Version 0.5.5 (Python)	CEDA implementation of SAML (Security Assertion Mark-up Language) – needed for the Earth System Grid Federation (ESGF) and CMIP5.	http://pypi.python.org/pypi/ndg-saml/
1. 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.	http://pypi.python.org/pypi/MyProxyClient/
1.MyProxyWebService Version 0.1.1 (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.	http://pypi.python.org/pypi/MyProxyWebService/
1. NDG-Security Version 2.2.2 (Python)	A complete toolkit 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.	http://ndg-security.ceda.ac.uk/
3. COWS-Server Version 1.6.1 (Python)	Implementation and deployment package for OGC Web Services built on COWS	http://cows.badc.rl.ac.uk/ (home page) http://ndg.nerc.ac.uk/dist (download)
3. COWS-Client Version 1.7.0 (Python)	Provides a graphical user interface to the COWS server interfaces which can be accessed from a browser. Can also be used to access Web Map Services (WMS) from other data providers.	http://cows.badc.rl.ac.uk/ (home page) http://ndg.nerc.ac.uk/dist (download)
3. COWS-WPS Version	An implementation of the OGC Web Processing service that supports synchronous and asynchronous process execution	http://cows.badc.rl.ac.uk/cows_wps.html (home page)

(Python)	on grid and cluster resources. (COWS-WPS is the unifying technology behind the UKCP09 User Interface.)	http://ndg.nerc.ac.uk/dist/ (download)
3. CDAT-lite Version 6.0rc2 (Python)	CDAT-Lite is a package for manipulating climate science data. It is a subset of the CDAT tools developed at Lawrence Livermore National Laboratory which focusses on data management and analysis distributed in a compact package.	http://pypi.python.org/pypi/cdat-lite
Discovery_Metadata_Ingest Version 4.3.1	Code to insert NERC UK Gemini conformant records into the NERC Discovery Web Service Catalogue. Work is ongoing to ensure that the Discovery_Metadata-Ingest codebase is also applicable to the MEDIN Metadata Ingest. Ongoing development work has seen the addition of validation by Schematron.	http://ndg.nerc.ac.uk/dist/ (download) Homepage at http://proj.badc.rl.ac.uk/badc/wiki/ingest_v4.3.0
2. Discovery WS Version 4 (Java)	Provides a SOAP interface to the Discovery Database based on UK Gemini Metadata ingestion completed by the Discovery Metadata Ingester (v.4.3.1). An additional SOAP service is included in this release: the Data Providers Web Service (DPWS). This provides an interface to provide control over the harvesting, ingestion and reporting of metadata in the Discovery database.	Download from SVN: http://proj.badc.rl.ac.uk/svn/ndg/mauRepo/revitalizationProject/trunk/project DWS Home page at http://proj.badc.rl.ac.uk/ndg/wiki/DiscoveryWS DPWS Home page can be found at http://proj.badc.rl.ac.uk/ndg/wiki/DPWS_API_NOTES
Maintained Software		
2. OAI Info Editor v1 (Python)	Web client that provides a public interface allowing users to initiate an OAI harvest and ingest sequence into a Discovery metadata database. Provides admin role for editing.	http://ndg.nerc.ac.uk/dist/ (download)
2. CF Standard Vocabulary Editor Version 1.0	Web portal and client that allows the editing of the CF Standard Name Table and the update of accepted terms to the NERC Vocabulary Server. The editor provides an interface for the CF community to see and comment on proposed changes and allows accepted updates to be properly versioned with the CF master table held in version control	http://proj.badc.rl.ac.uk/svn/badc/VocabTermEditor/trunk (download)
3. OWSLIB Version 0.4.0 (Python)	OWSLib is a community-based open source project which provides a python API for accessing OGC services and making requests for maps/data/features. CEDA is one of the main contributors to this project.	http://pypi.python.org/pypi/OWSLib
3. NAPPY Version 0.9.9 (Python)	Python input/output package for handling NASA Ames files, including conversion to/from NetCDF.	http://pypi.python.org/pypi/nappy
3. CSML Version 2.7.21 (Python)	The Climate Science Modelling Language (CSML) package provides a set of python modules for reading and writing CSML documents and interfacing CSML with climate data formats such as NetCDF.	http://csml.badc.rl.ac.uk (home page) http://ndg.nerc.ac.uk/dist (download)

Funding 2010/2011

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

In 2010/2011, at CEDA:

- One new NERC grant (MIRP) was begun.
- One new JISC project (ACRID) was begun.
- No new EC projects began, though several proposals were submitted (with results expected in the next year).

Financial Summary:

	NCAS	NCEO	Other NC	Data RP	Data RM	TOTAL SLA	Other
Carry-In	106.2	-125	-36	-180	-49	-283.8	-76
Income	-906.1	-449.8	-11.6	-222.4	-106.7	-1696.6	-1144.2
Spend	1059.6	522.8	19.2	240.2	59.8	1901.6	1195.2

CEDA financial summary for 2010/2011.

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.

The table above tells us for 2010/2011 that:

- There was a significant overspend against direct NCAS funding.
- There was a slight underspend in direct NCEO funding (primarily because NCEO requirements were funded elsewhere).
- Other NERC national capability funding significantly underspent.
- Data management funding associated with NERC research programmes underspent considerably (this is normal, as generally the money arrives before the work, and this indicates that in this year, on average, projects are nearer their starts).
- Data management funding associated with NERC research mode (generally consortium grants) underspent significantly (again, as normal).

2011-2012 Detailed Targets

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

National Capability	2011 Priority Activities
T1 Hardware Infrastructure: Maintain (and develop as necessary) 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. 3: Put hardware on "SLA footing" (clear internal interface with requirements)
T2 Software Infrastructure: Develop, maintain and upgrade 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.
T3 Curation: Curate existing information according to best practice principles (assess, improve, migrate and/or delete as necessary)	1. Add DOI support for some datasets. 2. Deploy MOLES 3 based infrastructure and update information records accordingly. 3. Migration and backup control from single database driven configuration.
T4 Core Data Management: Provide data management services (data management plans, formal archives) for NCAS & NCEO (including FAAM, ARSF, UFAM and the NCEO scientific themes).	1. DMP services as necessary. 2. Support development of UK strategy to sport LTDP
T5 User Support: Provide prompt and effective user support. Promote NERC science by attending conferences and appropriate PR.	1. User support as necessary.
T6 Community Engagement: Contribute to national and international strategy and standards efforts to ensure that outcomes are aligned with NCAS, NCEO and NERC objectives. Support CF NetCDF. Met Office and ESA liaison. Contribute to peer review.	1. Support CF-NetCDF standard names, and on vocabulary management for Earth Observation. 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
T7 Science Research: Maintain validity and relevance of CEDA activities by providing research breaks for staff.	1. Short breaks as requested.
T8 SIS: Contribute to the implementation of phase 2 of the NERC Science Information Strategy Programme.	1. Provide architectural advice. 2. Citation activities.
T9 NCAS External data: Acquire, ingest, and catalogue, appropriate data from a range of sources including the Met Office, NOAA and ECMWF.	1. Continue to maintain existing data streams.
T10 MIP Support: Provide support for CMIP5 and CORDEX by providing ESGF interfaces (software, hardware, and networks) to UK data and global data cache. Deploy and maintain adjunct services (including CMIP5 questionnaire).	1. Production data nodes. 2. Production gateways. 3. Production replication. 4. Production Metafor gateway and questionnaire. 5. Carry out CMIP5 quality control. 6. Compute service for MCIP5 analysis

T11: NCEO External data : Support the UK earth observation community by continuing to provide high speed UK cache archives for ESA, EUMETSAT (and other high volume remote data).	1. Continue to maintain existing data streams.
T12 ISIC: Work with commercial and academic partners to deliver ISIC (International space innovation centre) scientific visualisation services. Contribute to the development and implementation of other ISIC activities.	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.
T13 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
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 & QESDI •APPRAISE •RAPID-WATCH •ClearfLO •StormsRiskMitigation •PAGODA (part of Changing Water Cycle programme)
TR2 RM Support: As for TR1 but for grants.	•Amazonica, RONOCO, Fennec, ABACUS-IPY, COBRA-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 sport for UKCIP09 •Defra support for the Agricultural Greenhouse Gas Inventory Platform •UKMO support for CMIP5. •ESA Long Term Data Preservation support. •MEDIN discovery service operation
TR4 European Commission Contracts: As for TR3 but for EC based funding.	•METAFOR (Documenting climate models) •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) •ACRID (Support CRU data and workflow documentation) •ISIC Support? •ExArch (Exascale data architectures)

CMIP5 support – activity 2010/11

Martin Juckes

The fifth Climate Modelling Inter-comparison Project, (CMIP5) will provide a comprehensive suite of climate simulations and projections which as well as delivering ground breaking science, will provide key inputs to underpin the 5th Assessment Report of the Intergovernmental Panel on Climate Change (AR5).

CEDA has worked with and through many collaborative frameworks (both funded and unfunded) to support CMIP5. Key funders have included the NERC, the European Commission, the Met Office, DECC and Defra. Working with the Global Organisation of Earth System Science Portals (GO-ESSP), CEDA led the specification and implementation of a security infrastructure supporting flexible access (including programmatic access) to the distributed archive by users registered at distributed centres and played a leading role in the METAFOR EU FP7 project which has established new methods for documenting climate models and simulations which has been implemented for the CMIP5 archive. A second major EU FP7 project, "Infrastructure Support for the European Network of Earth System Modelling for Climate" (IS-ENES), is co-ordinating the European component of the global CMIP5 archive; CEDA is leading and co-leading the two relevant work packages. Within the UK, CEDA is using the NERC Knowledge Transfer project "Models and Impact Relevant Prediction" (MIRP) to facilitate access to the CMIP5 archive by users outside the climate modelling community.

Figure 1: The CMIP5 gateway, cmip-gw.badc.rl.ac.uk, provides access to CMIP5 data held at many locations around the world.

Providing support for the CMIP5 archive has involved most, if not all, of the CEDA team to some extent. Some key contributions are listed here, further details are in individual reports: international leadership in negotiating high standards of documentation for the data and flexible access (Bryan Lawrence); working with MOHC to ensure that UK simulations were the first to be published in the CMIP5 archive (Ag Stephens); leading a consortium integrating European data services (Martin Juckes); bringing together a range of complex and evolving software packages to meet aggressive deadlines (Stephen Pascoe); providing a security architecture to enable flexible access to the federated archive (Phil Kershaw); managing a European collaborative project to define and implement meta-data standards (Sarah Callaghan) and negotiating agreements on those standards with the international climate modelling community (Charlotte Pascoe); guiding the debate on standard names for over 800 model variables to a successful and timely conclusion (Alison Pamment); implementing the IPCC approved quality control (Kevin Marsh); providing MOHC CMIP5 data through OGC (Open Geospatial Consortium) interfaces for visualisation at the launch of the Harwell International Space Innovation Centre (Dominic Lowe); installing a 1Petabyte disk archive and associated software (Sam Pepler, Matt Pritchard, and team); working with JANET to isolate network performance issues in order to achieve high data transfer rates to N. America (Dave Terrett).

Active participation in the CMIP5 archive delivery has ensured that UK data is available to the international research community with minimum delay and also that UK researchers have the best possible access to data being produced by other modelling centres.

The CEDA vocabulary editor

Alison Pamment, Steve Donegan, Calum Byrom[#], Oliver Clements[*], Bryan Lawrence, Roy Lowry[*]

([#] Tessella Ltd, [*] British Oceanographic Data Centre, BODC)

A controlled vocabulary provides a standard way of labelling the contents of data files so that they are unambiguously defined and easily retrievable. At CEDA we use a number of controlled vocabularies. The CEDA vocabulary editor is a new tool which streamlines the management of the controlled vocabularies that we published to the scientific community through the NERC vocabulary server at the BODC. The new editor also provides a means of keeping scientists informed of work in progress and an archive of completed work so that they can readily see how a controlled vocabulary is developing over time.

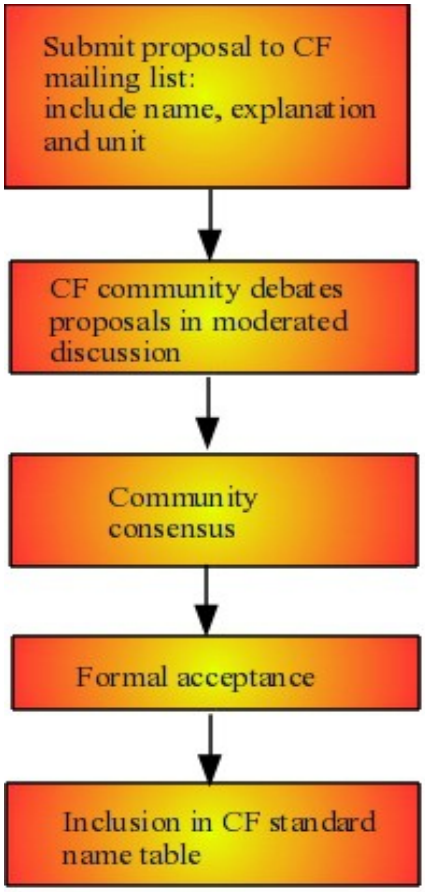


Figure 2: The CF process for creating new standard names.

A controlled vocabulary of particular importance to climate and environmental scientists is that known as the "CF standard names". These form part of the international CF (Climate and Forecast) conventions. The standard names provide precise definitions for geophysical parameters. The CF metadata and the standard names have been required by many large scientific projects, such as CMIP5. A scientist who needs to use CF metadata to describe his or her data can select from a list of over 2000 existing standard names. If, however, an appropriate standard name does not already exist a new name may be proposed to the CF community (this is done by sending an email to a dedicated mailing list). If the proposal is accepted the new name will be added to the published vocabulary. At CEDA we manage the CF standard name process for the international community and we publish the full list of accepted names on both the NERC vocabulary server and the CF website. This requires us to keep track of all planned changes and to publish identical data simultaneously in two different formats and in two locations. The CEDA vocabulary editor was developed as a tool for managing the process of maintaining the published versions of the CF standard names and keeping users informed of upcoming changes. It has been designed for extension to manage any of the hundreds of vocabulary lists held in the NERC vocabulary server.

When a new CF standard name is proposed a secure entry is created for it in the vocabulary editor by means of a web form. Each standard name is entered along with its appropriate units and a paragraph of draft explanatory text. The editor creates a publicly visible web page[7] showing the status of all vocabulary terms that are currently passing through the CF acceptance procedures, and eventually uses a web interface at the BODC to communicate to the vocabulary server when a name is ready for publication. Once a name has been published the information relating to that term is moved from the current work status page to another web page showing completed work which provides a permanent record of changes.

The CEDA vocabulary editor became operational in March 2011 and now that it has been successfully introduced for managing CF standard names we will be extending its use to managing other vocabularies over the coming year.

[7] e.g.: http://ndg.nerc.ac.uk/vocab-term-editor/viewCurrentStatus/http%253A%252F%252Fvocab.ndg.nerc.ac.uk%252Flist%252FP071%252F17

Scientific Visualisation Service for ISIC:

http://isicvis.badc.rl.ac.uk/viewdata/

Victoria Bennett, Dominic Lowe, Richard Wilkinson (Tessella Ltd)

As part of the development of the International Space Innovation Centre at Harwell Oxford, staff at CEDA, the Centre for Environmental Data Archival have developed and deployed a new data visualisation service for Earth Observation data.

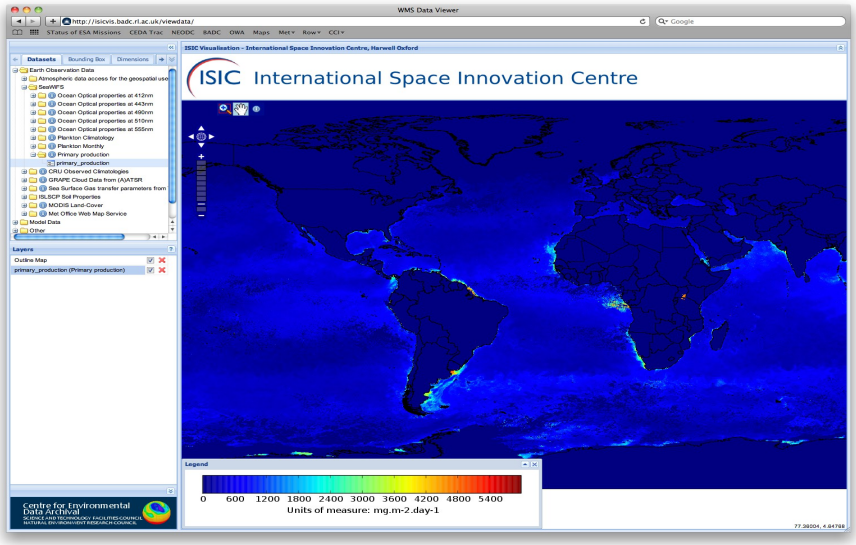


Figure 3: Screen shot of primary productivity data in the SVSeo web interface (data from T. Smyth, PML).

The Science Visualisation Service for Earth Observation (SVSeo) is a web based application to allow users to visualise and make use of Earth Observation data and climate model simulations. Users of the SVSeo can visually explore large and complex environmental datasets from observations and models, view, step through and zoom in to gridded datasets on a map view, export images as figures and create animations. Different views can be easily overlaid, e.g. different parameters in the same data, or different datasets. The SVSeo can also be used at the ISIC facility in conjunction with a large video wall and associated interactive visualisation software developed by partners in STFC e-Science and the University of Reading to create animations on a virtual globe, or multiple, synchronised virtual globes. All images and animations can be exported for viewing and manipulation remotely.

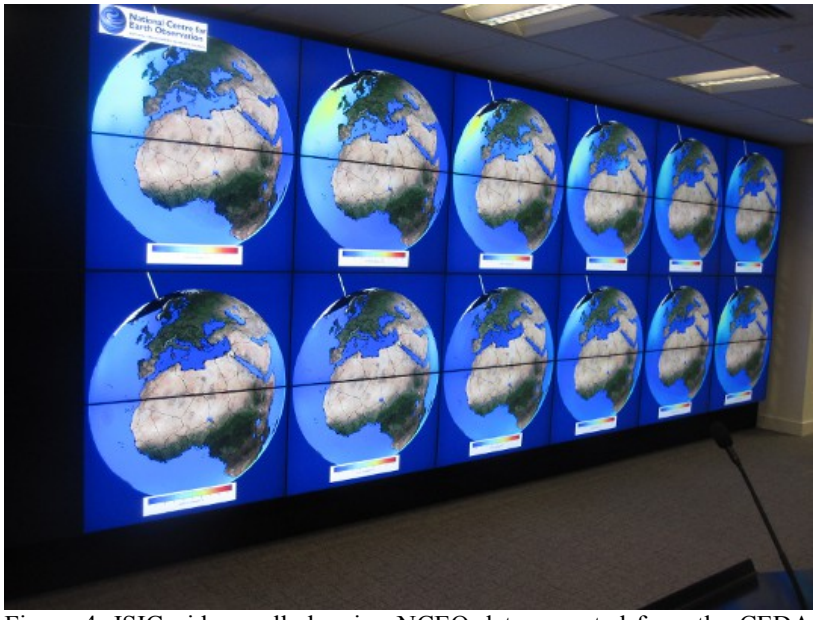


Figure 4: ISIC video wall showing NCEO data exported from the CEDA archives using the SVSeo. The data are significant wave height products from TOPEX, produced by S.Fangohr (NOCS) & D.Woolf (ERI), shown on multiple globes, one for each month.

Many CEDA datasets have been included in the visualisation service, including satellite derived products relating to clouds, plankton, air-sea gas exchange and fire, as well as model output. More datasets will be added as more NCEO datasets are produced and provided to CEDA for long-term archival. The visualisation framework can handle datasets in CF-NetCDF data format, and CEDA staff can provide assistance to data providers who are not familiar with these formats and conventions. It can also handle remote data exposed using an OGC web map service (WMS). SVSeo makes heavy use of a range of CEDA software products, including the COWS framework (CEDA OGC Web Services).

The CMIP5 "Questionnaire" Web Tool

Charlotte Pascoe

The CMIP5 "Questionnaire"[8] web tool captures information about the life-cycle of CMIP5 climate simulations. The information it captures explains what climate simulations were done, and both why and how (see figure 5). The model and simulation documentation captured will broaden access to climate model data, because, for the first time research data users can discover the science encoded in the algorithms of climate models without needing to contact the people who wrote the code. With this new contextual documentation, or metadata, climate scientist are able to analyse more deeply the simulated data produced by different modelling groups.

Before metadata about climate models was collected in a systematic way the only documentation available to everyone was that which made it to the scientific literature. However, the scientific literature focused on the latest and greatest things that models could do, the only standard piece of information published about every model was generally its resolution and this was often expressed in different ways depending on the nature of the coordinate system. The only way to really find out about the science encoded in the model was to contact the scientists who ran the model or better still the people who wrote the code. Without personal contacts, climate models were effectively a black box that people and policy makers were expected to trust.

The Metafor[9] project addressed this issue by creating a new method for describing simulated data based on a new "Common Information Model" or CIM. Figure 5 shows a simplified view of the CIM elements that are populated by the CMIP5 questionnaire: experiments are described as a list of requirements that the simulations must conform to; simulations are made by running models which are themselves made up of software components (which can contain child components). Much of the material is populated using controlled vocabularies which are specific terms, precisely defined that have a common meaning to all climate scientists.

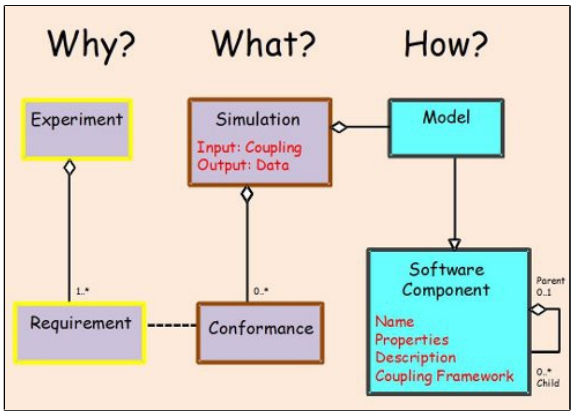


Figure 5: The CMIP5 questionnaire web tool captures metadata about the life-cycle of a climate simulation. Here we see a UML view of the CIM elements that are populated by the CMIP5 questionnaire, they explain why and how the simulated data was created.

These controlled vocabularies were collected by Metafor from interviews with climate modellers aimed at finding out the information that scientists in different climate disciplines needed to be able to compare climate model simulations. The interviews were summarised using mind map diagrams that not only collated controlled vocabularies but also allowed Metafor to build a hierarchy to structure how the information would be collected. The mind maps became the inputs which defined the CMIP5 questionnaire; branches in the mind maps are associated with web forms in the questionnaire and the controlled vocabularies generate drop-down lists. Over 400 specific properties were included in the mind maps and consequentially prompted for in the questionnaire, and the user can add additional properties as required.

The CMIP5 questionnaire is the first attempt to comprehensively describe the science of an earth system model in a manner which can be applied to multiple models. The information which will be collected using it will create an important community resource that helps scientist to do science and builds trust between scientists and policy makers through openness.

[8] CMIP5 questionnaire: http://q.cmip5.ceda.ac.uk

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

Making Data a 1st Class Research Output: Data Citation and Publication

Sarah Callaghan

Scientists are creating larger and more complicated datasets every day. Production of these datasets is a laborious and time-consuming process, which is often undervalued by the wider scientific community. Academic credit is primarily based on the production of peer-reviewed papers published in well respected journals. NERC as a whole wishes to encourage its scientists to submit their datasets to NERC data centres, where the data can be properly archived and curated, discovered and re-used. Data citation and publication, and through them, academic credit, provide an incentive for scientists to submit their data in an appropriate format and with complete and accurate information describing the dataset.

The NERC Data Citation and Publication project brings together all the NERC funded data centres to develop a standardised way of allowing the scientists who archive data in the data centres to cite those data sets, in a way analogous to how scientific journal papers are cited. The project is also collaborating with academic publishers to create a method for peer-reviewing and publishing peer-reviewed data sets. CEDA has taken a leading role in these efforts: Sarah Callaghan (BADC) is the project manager and is also an associate editor of Atmospheric Science Letters with a special brief for data publication. The project has formed partnerships to exploit common activities and achieve wider community buy-in, including with: the SCOR/IODE/MBL WHOI[10] Library Data Publication Working Group, the CODATA-ICSTI[11] Task Group on Data Citation Standards and Practises and the DataCite Working Group on Criteria for Data centres.

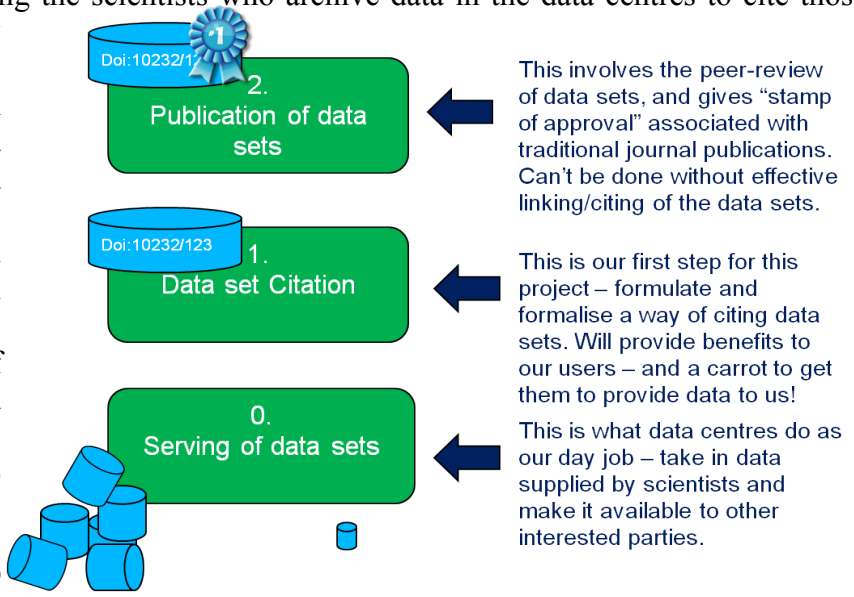


Figure 6: Serving, citing and publishing data.

We use digital object identifiers (DOIs) as a mechanism for citing the datasets held in our repositories because they are actionable, interoperable, persistent links for (digital) objects which scientists are already used to using for citing papers. NERC has been provided with a DOI assignment account through the British Library, who are acting as the UK member institute for DataCite, a member of the International DOI Foundation. This year the first datasets in NERC data centres were issued with DOIs.

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.

[10]Scientific Committee on Oceanic Research (SCOR), International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission (IOC), and Marine Biological Laboratory/Woods Hole Oceanographic Institution Library (MBL WHOI)

[11]Committee on Data for Science and Technology (CODATA) International Council for Scientific and Technical Information (ICSTI)

Security Architecture for the Earth System Grid Federation

Philip Kershaw, Bryan Lawrence, Rachana Ananthakrishnan (Argonne National Laboratory, USA), Luca Cinquini (NASA JPL/ESRL NOAA, USA) & Dennis Heimbigner (UCAR, USA)

The BADC has been closely involved in the development of the Earth System Grid Federation, an international collaboration effort to develop a globally federated infrastructure in support of CMIP5. One key aspect of such an infrastructure is the system for access control needed to restrict access to the large volumes of data generated.

Given the large size of the CMIP5 archive (expected 2-3 PB) coupled with estimates of around ten thousand users, the infrastructure to support faced significant challenges to address in terms of access, distribution and storage. This has necessitated a distributed solution in which the archive is to be globally mirrored: at the BADC and two other main sites (in Germany and the United States). With that distributed model a federated security solution was required. Where traditional models of security express user identity and secured resources within the bounds of a single organisation, a federated system bind organisations together in a 'federation' of trusted parties. By entering into a relationship of trust and sharing a layer of common infrastructure, the user communities they serve are enabled to access data seamlessly across a single Virtual Organisation made up of all the participating institutions.

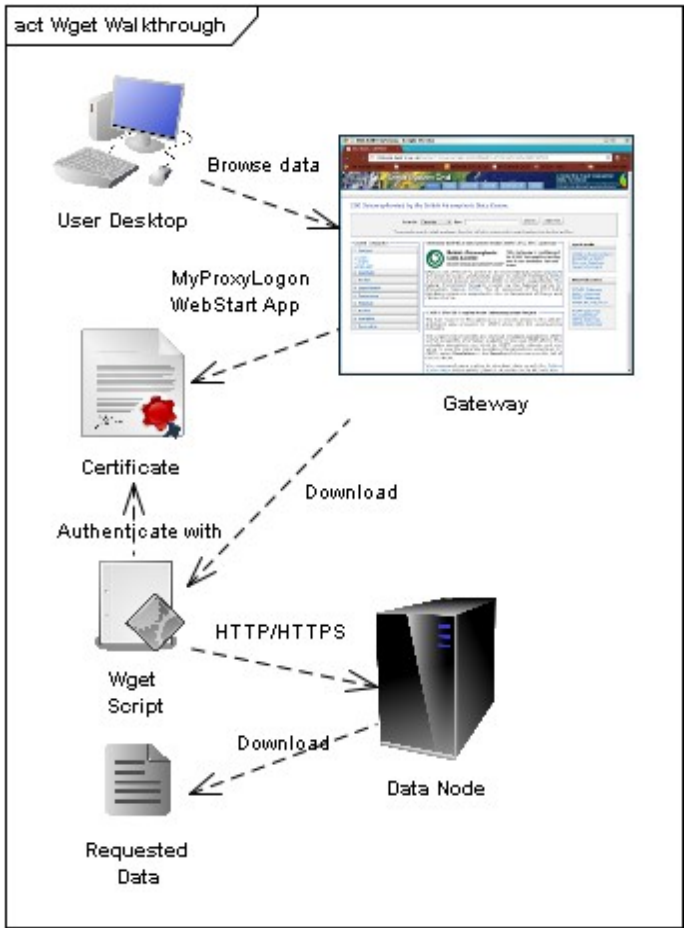


Figure 7: Data services support dual script-based and Browser-based Data Access. In the above, the user finds data using a browser, but then downloads it using a command line script based on the popular Wget program. Federation-wide user identifiers like OpenID and certificates are analogous to a passport, enabling users to access data from anywhere across the federation using a single account from their home organisation.

A number of technologies are available to build such an infrastructure, however any solution must consider interoperability across the most widely used technologies, and within the climate and wider earth system science communities, a wide range of services and software tools are used many of which have no existing means to apply such access control.

To address these diverse requirements modular design principles were adopted applying established standards and protocols to deliver a service oriented architecture. At the level of individual client programs and services, a modular design approach separates access control and the other functionality of application code. By doing so, it has been possible to provide common solutions to secure a wide range of client programs and associated services. This has borne fruit with the development of extensions to the widely used NetCDF[12] software libraries to support the ESGF security protocol. This system has been deployed at a number of sites around the world and a number of projects from related communities have expressed an interest in using and exploiting the security model.

[12] http://www.unidata.ucar.edu/software/netcdf/

Non core data in BADC FAAM archive

Wendy Garland

Data collected on board the Facility for Airborne Atmospheric Measurements (FAAM) BAe-146 research aircraft are archived at BADC. Data for each flight are categorised as core – that produced by a "core" suite of instruments operated by the FAAM team routinely, and non-core – produced by instruments outside of the core suite – such as those operated by Met Office, NERC or external project partners, or new or non-standard products. NERC requires that all FAAM data are archived at BADC to facilitate data access and long-term preservation. Whilst core data is processed and archived promptly and efficiently, some streams of non-core data are currently greatly delayed or not archived at all.

Figure 8 summarises the data processing routes for FAAM data. Core data are processed post-flight and uploaded to BADC quickly, usually within 24 hours. A subset, core-cloud-physics data, are processed separately usually within 7 days. Non-core data are taken away by the instrument operators, processed separately at various locations and uploaded to the archive to be stored alongside the core data for that flight. Non-core data processed by the FAAM team are routinely archived and Met Office instrument operators have developed processing and uploading routines. However, the remaining non-core data from other, mainly NERC, NCAS or FGAM, instruments are often subject to long delays or remain outstanding indefinitely.

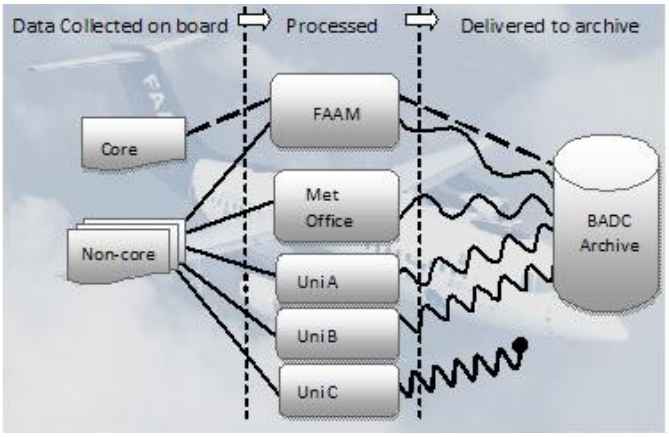


Figure 8: Shows the paths taken for FAAM data to reach the archive. The wigglyness of the path is proportional to the time taken to get to the archive.

Estimating what is missing is difficult: Project funding and instrument configuration vary, but projects funded entirely or partially by NERC/NCAS might reasonably expect to produce some non-core data. It is non-trivial to identify which instruments were operated for each flight, but an estimate can only be made by reviewing the crew lists showing which instrument operators were on board. Even then, not all flights will operate all instruments, several instruments may be combined into one output file, and each operator may oversee more than one instrument. Nonetheless, many projects do not have any non-core data archived several years after flying. There are several potential explanations: it is possible that no useful data were collected, or the processing was never completed. Operators often have "working" versions of the data and then move onto the next campaign or new post and forget to produce "final" data. Some operators may struggle with producing the required formats. In many cases it is simply the institutes' long-time established practices to retain data internally (despite their obligations).

The ramifications of delayed or absent non-core data are that in the short term, data may not be widely available to the project team, and in the longer term, such data are lost to the wider community. Such data could be used for applications outside of the original objective, but if it is not discoverable and accessible then such opportunities are missed. In addition work done to support these projects is wasted.

To improve the situation, BADC will target the instrument operators identified from the crew lists (bottom-up approach, instead of going through the project leaders). BADC will exploit the FAAM team's greater (face-to-face) interaction with project leaders and instrument operators and get more reminders for data timely data submission. Additional liaison with the project leaders at institutional levels will be undertaken to address change in established local storage practices.

Rapid response to the Eyjafjallajokull volcanic ash plume event April 2010

Wendy Garland

The CEDA team played a central role in the coordinated scientific community effort to provide vital measurements of the ash cloud following the eruption of the Eyjafjallajokull volcano in Iceland on 14th April 2010. Clouds of volcanic ash were injected into the atmosphere and carried aloft over the UK and Europe closing the airspace and grounding the commercial airline industry for 5 days. During this time the atmospheric science communities in the UK and Europe were actively making dedicated observations of the ash plume using airborne and ground-based instruments. These measurements were used by both the scientists themselves and government civil contingency planning teams. The CEDA team rapidly developed and deployed a useful mapping tool to keep track of what was being measured, collated the data as quickly as possible into a useful dataset and obtained near-real-time meteorological and satellite products for use by the community.

The CEDA development team quickly, designed and built a geospatial web application using the GeoDjango framework. The resulting application features a data entry tool enabling users to create annotations for events "tagged" with a geographic location and time, and an interactive tool for viewing these annotations. Entries created by a user are stored in a database and exposed as entries in a GeoRSS feed, which is then used by a Google Maps interface coupled with a sliding time-line to provide an interactive view of all the events in the database. It is clear that this tool will have wider application within CEDA activities, providing the capability to add geospatial annotations to data and other resources.

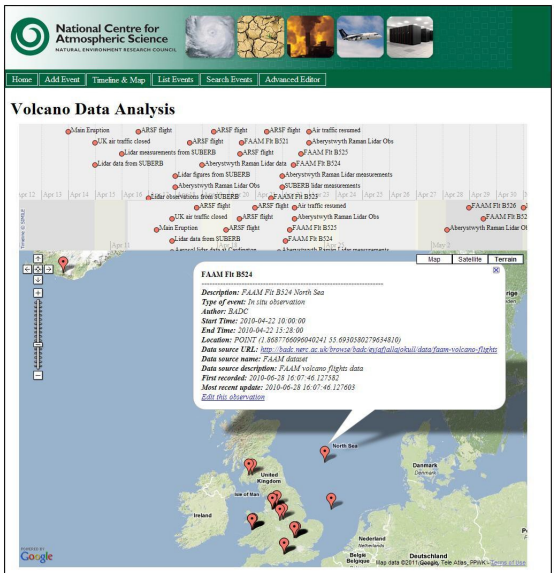


Figure 9: Event log mapping tool showing observations of the ash plume

A dataset was created providing a central location to discover, view and disseminate data and supporting metadata from the science community was provided. A secure upload area was supplied and distribution conditions for each data-stream established in order to promote wide usage whilst protecting the rights of data providers. Initially data access was agreed to be free and open but this was rapidly rescinded when data were misrepresented in the non-scientific media. Initially this dataset was set up for the UK Atmospheric scientists (NCAS) but rapidly extended to include the Earth Observation (NCEO) and the European research aircraft (EUFAR) communities. The analysis of the measurements made by the academic community was supported by the provision of additional data, not generally available to this community in near real time, from the CEDA archive. These data included near real time meteorological products and satellite data, and lidar data from the Met Office and across Europe. Measurements by 7 ground-based lidars at 5 locations nationwide, and airborne in-situ data and flight logs for 17 flights by 4 aircraft including the NERC ARSF and FAAM BAe-146, and 2 other EUFAR aircraft were stored. To encourage rapid uploading and exchange, data was accepted in the operators preferred data format and where necessary converted to standard formats at CEDA. In each case special agreements were made to permit the distribution and usage of data. Close liaison was necessary with the NCAS, NCEO and EUFAR communities and instrument teams throughout the intensive measurement period, with much discussion, electronic, at meetings and conferences.

This activity showed how responsive CEDA can be when scientific events dictate the necessity for flexibility – by acting as a central location and provided a geospatial web application and supporting data-products to facilitate the exchange and analysis of data pertaining to the ash cloud emanating from the Eyjafjallajokull volcano within the academic communities.

Data Scientists and Their Day Jobs

In the remainder of this report we present some one page statements from CEDA staff each of which highlight something about them and their job.

We hope that in doing this, we can both give a snapshot of the variety of activities undertaken by staff and give a flavour of the technical and scientific challenges they face. Each member of staff was asked simply to write a page on some aspect of their job and their involvement with it. We have done a bit of editing for layout and style, and made the titles a bit more declarative, but otherwise this next section is straight from the keyboards of our team.

In 2010/2011 CEDA was broken up into three groups: an operations group, a science support group and a development group – but individuals often have roles in one or more of those groups as you will see from the pages which follow. For that reason we've made no effort to order them in any way, they're basically random! (Nonetheless, some individuals have chosen to expose the multiplicity in their roles, and some of concentrated on one aspect of their roles.)

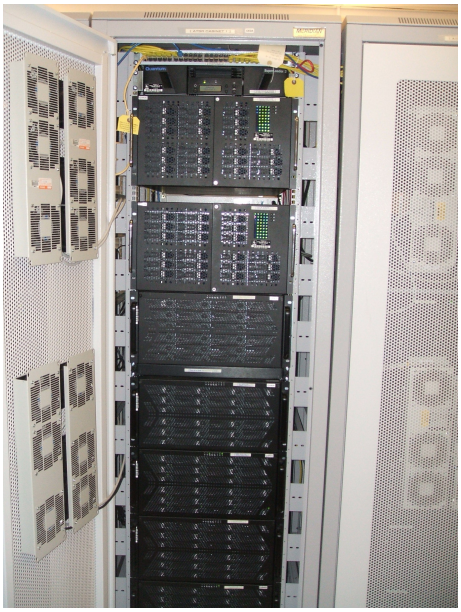


Figure 10: CEDA hardware in operation in the CEDA server rooms.

VALOR – 'VALue Of the RAPID array'

Alan Iwi

The British Atmospheric Data Centre is a partner in the VALOR project, funded by the Natural Environment Research Council and coordinated by the University of Reading, which aims to assess the value of predictions from the RAPID array (described below) for predictions of the Atlantic Meridional Overturning Circulation (MOC).

The MOC is the Atlantic part of a global pattern of ocean circulation known as the Thermohaline Circulation. In the North Atlantic, the northward movement of warm surface water carries up to 1.3 petawatts of heat from the tropics – equivalent to over 3,000 times the total electricity generation of the European Union[13] – and is a key reason why Northwest Europe has a comparatively warm climate for its latitude. This surface water, once cooled, sinks at high latitudes and returns southwards as deep water. Numerical models predict that in the 21st century the strength of the MOC is very likely to decrease[14], mainly because surface water at high latitudes would become less salty and therefore not dense enough to sink readily; paradoxically this could lead to a cooling of European climate. A complete shut-down of the MOC is also possible, although very unlikely to occur during the 21st century.

Since 2004, the strength of the MOC has been monitored by the "RAPID array" of moorings at 26˚N in the Atlantic, from which profiles of temperature and salinity are obtained near the western and eastern boundaries and along the edges of the mid-ocean ridge. These allow calculation of the east-west density gradient across the Atlantic basin and the northward ocean flow. Other RAPID data include transports through the Florida Straits derived from measurements using an electromagnetic cable; combining these with the density-driven flow gives an estimate for the strength of the MOC, after allowing for wind-driven flow near the surface.

A key objective VALOR is to assimilate data from the RAPID array into numerical climate models, i.e. to produce full three-dimensional fields from the climate model that are reasonably consistent both with the supplied profile data and also with the model's internal dynamics. Because of the carefully chosen locations for the RAPID array data, the model state will be one in which the strength of the MOC is tightly constrained by the observations. Work on data assimilation is under way, using both the NEMO ocean model and also the Met Office Hadley Centre's HadCM3 coupled atmosphere-ocean model.

Once data assimilation complete, the resulting fields contained in the numerical model can be used as an initial state from which to run the model without further input data in order to perform both forecasts for future periods, or 'hindcasts' of past periods (for which the actual outcome is already known, as a test of the forecast system). However, even after data assimilation, the initial state for the forecast is at best an approximation to reality. Other initial states which are equally plausible given the coverage and accuracy of the available data can be generated by perturbing these data. Such a slightly perturbed initial states can lead to different forecasts, possibly substantially so, given the chaotic nature of the climate system. By performing an 'ensemble' of forecasts starting from several such initial states, a technique widely used in numerical weather prediction, we can obtain an estimate of the forecast error as the individual forecasts diverge. We can also compare the results with those from a 'no data assimilation' ensemble, which is performed similarly except that it is initialised from a control run of the model, in order to determine the time-scales over which the RAPID data has a statistically significant impact on the forecast of the MOC.

Within BADC, we will support ensemble hindcasts of the MOC for VALOR using the Met Office's HadGEM3 atmosphere-ocean model, which incorporates NEMO as its ocean component. These will use HECToR, the national supercomputing facility based in Edinburgh.

[13]Source: Eurostat (data for 2008).

[14]IPCC Fourth Assessment Report

Supporting scientific research and its community

Anabelle Guillory

CEDA supports the atmospheric and Earth observation science communities in the UK and abroad through the provision of data management and discovery services, and continues to develop tools and services to aid data preservation, curation, discovery and visualisation. My role in this main is delivering high quality support service to the ever growing environment user community, with extra day-to-day activities which include science programme management and associated metadata cataloguing.

In this context, confronting the challenge of handling a rapidly increasing range of datasets (covering climate, atmospheric composition, ground, aircraft and satellite based observations, chemical and weather model output) and associated users technical issues (from data file format issues to identifying problems in the Met Office data), means that I have come to develop an all but necessary "Jack of all trades" approach, a key asset for the successful management of the user support service.

"These services have enabled research to be undertaken that would otherwise have been impossible. Please keep up this excellent work."

"Helpful and knowledgeable staff"

"Support have been very helpful with queries."

"Help desk has been extremely helpful in directing me to data I need, and dealing with access issues."

Figure 11: Some results from user survey ran as part of a study by the Research Information Network in January 2010

The CEDA help-desk integrates "responsive" support for atmospheric and earth observation science with "interactive" science support, often funded by specific programmes. One part of this involves controlling user access to data; reviewing applications for data access against the proposed research, user status and the conditions of use imposed by the data suppliers and NERC policies, and liaising with data providers (e.g. the Met Office) on data use policy. Despite the pressure of increasing numbers of registered users (17790 today compared to 286 when I started in 1998), help-desk targets are met: 98% queries still receive an initial response within one working day and where follow-up work is required, the user is kept informed of progress. As a direct result, the help-desk is consistently rated by external users as providing an excellent service. I also take part and at times, initiate and manage projects to improve the CEDA user interface (e.g. restructuring the BADC website in 2000, and more recently, extensive testing and propositions for improving the upcoming Web Processing Service).

In recent years, I have taken responsibility for the management of some NERC science programmes e.g. COBRA (Impact of combined iodine and bromine release on the Arctic atmosphere) and ABACUS (Arctic Biosphere-Atmosphere Coupling at multiple Scales) and assisted with data management tasks for other programmes and projects (e.g. Flood Risk for Extreme Events (FREE) and various Facility for Airborne Atmospheric Measurements (FAAM) Campaigns). The management of science programmes entails not only the provision of a Data Management Plan (DMP) to set up a coherent approach to data issues pertaining to the programme but also the execution of the plan as per the conditions (e.g. metadata standard, file format, file name convention, user access control) and time frame agreed with the science programme manager or principal investigator. Along with the management of science programme, long-term storage of associated metadata is of crucial importance in the age of data archiving. One of my responsibilities is to ensure that CEDA datasets all have appropriate and well documented metadata records in the CEDA metadata catalogue.

In a nutshell, I am a multi-faceted environmental data scientist at the forefront of the interaction between data suppliers and data users and one of the backbones of the day-to-day running of CEDA. I have no doubt that my daily data related activities are contributing, mostly unnoticed, to the advance of cutting-edge research in environmental sciences. Perhaps it is time to spread the word and invite data scientists to come out of the shade.

Managing a controlled vocabulary for environmental data

Alison Pamment

My role at CEDA is to facilitate the development of the Climate and Forecast (CF) controlled vocabulary, of terms for describing environmental data – the CF standard names. A controlled vocabulary provides a standard way of labelling data so that they are unambiguously defined and easily retrievable. CF standard names are an important tool for collaboration because they allow environmental scientists to more easily share and understand one another's data. Finding data is enhanced, because standardised terms can encoded in an 'ontology' which can then be exploited by software tools. For example, a user entering a search term of 'rain' could also be presented with data labelled with related terms such as 'precipitation' or 'drizzle'.

sea_ice_surface_temperature The surface temperature is the (skin) temperature at the interface, not the bulk temperature of the medium above or below. "Sea ice surface temperature" is the temperature that exists at the interface of sea ice and an overlying medium which may be air or snow. In areas of snow covered sea ice, sea_ice_surface_temperature is not the same as the quantity with standard name surface_temperature.	K
sea_ice_temperature	K
sea_ice_thickness	m
sea_ice_transport_across_line Transport across_line means that which crosses a particular line on the Earth's surface; formally this means the integral along the line of the normal component of the transport.	kg s-1
sea_ice_volume	m3
sea_ice_x_displacement	m
sea_ice_x_transport	kg s-1
sea_ice_x_velocity	m s-1
sea_ice_y_displacement	m
sea_ice_y_transport	kg s-1
sea_ice_y_velocity	m s-1
tendency_of_sea_ice_amount_due_to_basal_melting "Amount" means mass per unit area. The specification of a physical process by the phrase due_to_process means that the quantity named is a single term in a sum of terms which together compose the general quantity named by omitting the phrase. "tendency_of_X" means derivative of X with respect to time.	kg m-2 s-1
tendency_of_sea_ice_amount_due_to_congelation_ice_accumulation	kg m-2 s-1

Figure 12: A screen shot of a tiny portion of the current standard name table showing some formal definitions and canonical units.

The list of valid CF standard names is published on the CF website[15] and currently contains around two thousand entries covering a number of science domains including, but not limited to, atmospheric dynamics, atmospheric chemistry, physical oceanography and ocean biogeochemistry. CF standard names form just one component of the metadata needed to fully describe an environmental dataset. (Other metadata include, for example, the name of the model or observing platform that produced the data). CF metadata were originally designed to describe numerical model data such as those produced by climate prediction or weather forecasting models. Increasingly, however, they are being used to describe observational datasets from instruments as diverse as tide gauges and satellite radiometers. The CF metadata conventions are intended primarily for use with the NetCDF file format, although many of the concepts, including standard names, can and are applied to other file formats.

The list of valid CF standard names is not set in stone, as discussed previously (page 16). Guidelines[16] for the construction of standard names are published on the CF website but often new names can be constructed by analogy with existing ones. Scientists make proposals to the CF mailing list for new names, and all members of the CF community can then submit their comments. One of my roles is to act as moderator of these public discussions with the aim of drawing the community towards a consensus decision on the wording of new names, their explanations and units. Once agreement is reached on a new name it can be accepted for inclusion in the published standard name table.

As well as having been adopted as the standard in many large scientific projects, such as CMIP5, CF metadata will also be used to describe satellite data from the next series of NOAA polar orbiting satellites (JPSS) and are even being used to describe data from a model of the Martian climate.

[15]http://www.cfconventions.org/documents/cf-standard-names/standard-name-table/current/cf-standard-name-table.html

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

Infrastructure Management: What does Andrew *really* do?

Andrew Harwood

I have been working as the Infrastructure manager for the Centre for Environmental Data Archival (CEDA) for many years now. So what exactly does an infrastructure manager do? Broadly speaking, this means I look after the various bits of software that keep CEDA operating. Of course, the full story is more complicated.

Figure 13: Andrew after completing the Mont Blanc marathon in 2010. Andrew likes to run marathons in his spare time when he finds CEDA work is not challenging him enough.

Originally the CEDA data centres were separate, but now they have been brought together into one infrastructure, but many services have specific web interfaces. A large proportion of our work consists of making environmental data and metadata available to scientists via the internet. Some of our data has restricted access, so we need to control who can access what data. We also provide help desk support for our users by email and telephone. All of this requires software infrastructure.

The Oxford English Dictionary defines "infrastructure" as "the basic physical and organizational structures and facilities (e.g. Building, roads, power supplies) needed for the operation of a society or enterprise". Let's be clear, I only look after the software infrastructure, I don't take care of the building, sweep the roads or clean the toilets. Our "basic structures" are things such as databases, off-the-shelf software and our own home-grown scripts connecting it all together. Examples of off-the-shelf software includes the Apache web server, Eprints document repository, Mailman mailing list manager and Footprints query handling system. In the beginning most of the scripts that we wrote were perl CGI programs. We now mostly use Python and are moving to using web frameworks such at Pylons and Django.

Before CEDA came in to existence I worked exclusively for the BADC for several years. The data centre was in its infancy then and I helped to develop the databases and software that we use to control access to data and displaying dataset metadata. Back then, our database of choice was Ingres and we wrote our software in perl. A lot of our software consisted of web CGI programs that connected with our database of users or dataset catalogue. Today these systems are still important, although they have been subject to continued improvement. These days we keep our databases in Postgres and write our scripts in Python.

Quite a lot of my day to day time is taken up with fixing things that are not working properly. As our infrastructure has grown, so has the number of things that can go wrong. Often these can be relatively small things, such as a disk not being mounted correctly, but it can sometimes take a while to work out where the problem is. Where possible, I try and reduce the chances of a problem occurring again by making modifications to the system, by adding or improving various regular checks that we make and by improving the documentation in our "Operations Manual". When I'm not fixing things that are broken I also spend my time adding enhancements to our existing system that have been requested by our team members. If I'm lucky, I also get some time to spend on new developments or replacements for existing software components.

Another aspect of my work is installing or updating application software on our systems. The various application software that we use requires regular updating. In addition, members of our team want to try out new software to see if it is suitable for their purposes. At times this can stray a bit into "system management" territory. However, we are fortunate to have a separate team who look after the real hard-core system management of our computers for us. In summary, I spend my time keeping the software that runs CEDA running and helping to develop new and better systems.

Collaborative Projects Manager

Ag Stephens

The role of CEDA Collaborative Projects Manager requires me to coordinate most of the CEDA interactions with UK government, the Met Office and the NERC community. However, the truth is that this role is "part" of my job rather than being a description of what I do.

CEDA expertise lies in the field of data management. Within this remit we specialise in curation of environmental data primarily produced by, and for use by, the research community. We develop and maintain software for data access, manipulation and visualisation. We also make a major contribution to the emerging international standards on data and metadata services led by organisations such as the Open Geospatial Consortium (OGC) and World Meteorological Organisation.

My role spans three main areas: (i) operational data management, (ii) project management and communication and (iii) design and development of software tools. I currently coordinate or contribute to a number of CEDA projects with the major share being taken by the User Interface for the UK Climate Projections (UKCP09, funded by Defra) and its underlying data services, archive and operational systems. Additional activities include the Defra Greenhouse Gas Agricultural Emissions Inventory project and the NERC Model and Impact Relevant climate Prediction projects. Figure 14 shows a matrix of how my day-to-day work maps on to the various projects and funding lines that pay my salary.

Activity → ↓ Project	Project Management and Communication	Data Management	Software Design and Development	System Management and Maintenance	Community and Stakeholder Liaison
UKCP09	■	■	■	■	■
MIRP	■				■
CMIP5 support	■	■	■	■	■
Agricultural GHG Emissions Inventory	■				■
MashMyData			■		
NCEO (MONSooN and HPC activities)	■			■	■
Met Office interactions	■	■	■	■	■

Figure 14: How my day-to-day tasks map to the various funding lines.

Another area in which I and many of my colleagues are currently active is in the development and delivery of the CMIP5 international archive. For this particular project I manage the interactions with the Met Office Hadley Centre in terms of coordinating data transfer and ingestion. I have developed a new ingestion tool with a web front-end providing information about the data transfers and status of the system. This was built for CMIP5 but we have extended its use for other Met Office datasets being transferred to CEDA. I liaise regularly with colleagues at CEDA and the Met Office to ensure that the provider and consumer ends of the system are communicating effectively.

CEDA is active in many areas and collaborates extensively both nationally and internationally. I have a keen interest in making sure that activities with external motivations and outputs should result in some internal gains. My main objective in the first quarter of 2011 is to bring our COWS Web Processing Service (WPS) into operational use within CEDA. This tool was originally built to support the processing requirements of UKCP09 in 2009 and since then its development has been a marginal activity. The WPS will be providing a general capability to run offline jobs and deploy ad-hoc web services quickly and efficiently, delivering both in terms of existing CEDA requirements and in support of CMIP5.

Working within CEDA is both enjoyable and challenging. We are a diverse team of dedicated people working on a range of difficult and novel problems. There is more to it than copy-and-pasting files, with a wide range of skills in data management, computing and communication required!

Software Engineering and Data Modelling at CEDA.

Dominic Lowe

I am a software engineer in CEDA. My role is to make advances in software and data modelling that will be of benefit both to CEDA and the wider Environmental Science community. In particular I am trying to build links between the climate science data community and more general advances in Geographic and Geospatial Information Systems – i.e. recognising that climate science can benefit from integrating with other sources of geographic information and software, such as maps, data services and analysis tools. I am the STFC Technical Representative to the Open Geospatial Consortium – an international community mixing business, academia, government agencies and individuals which seeks to agree on technical standards for the handling, processing and sharing of geographic and spatial information.

One of the main outputs of my work has been the development of software to work with Climate Science Modelling Language (CSML) - a technical framework created at STFC for describing climate science datasets. The software enables the integration of diverse datasets under a single set of tools. So you can, for example, visualise two different datasets in the same application fairly easily (figure 15). This work has fed into a wider community debate about the merits of such a data models, and progress is now being made by several groups towards a community-developed data-model that is heavily influenced by the CSML model.

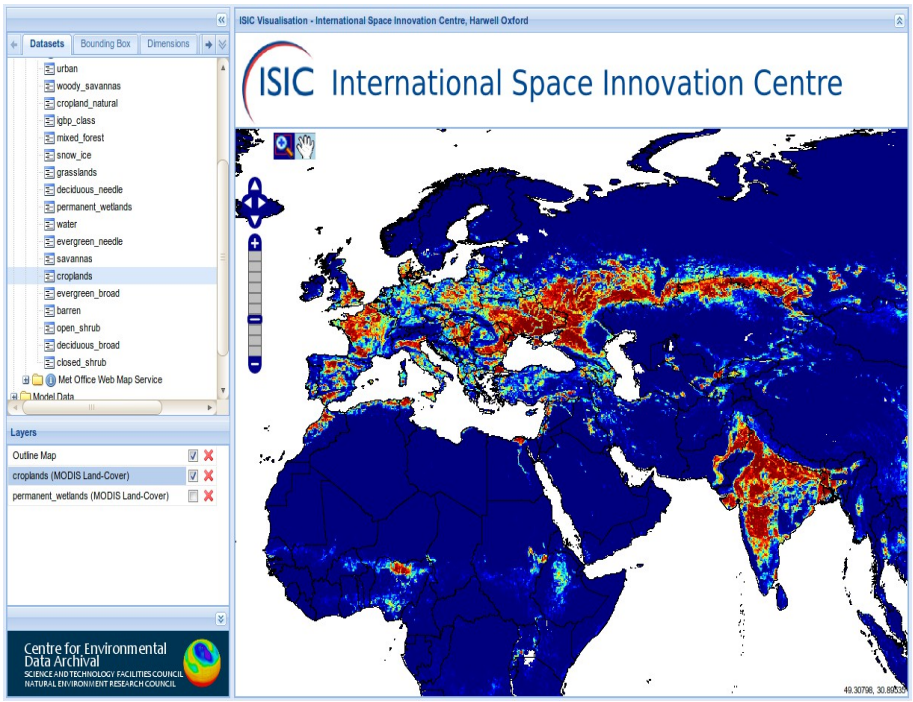


Figure 15: The International Space Innovation Centre (ISIC) Visualisation Software

Through my work with CSML and the OGC, I have become very familiar with standards in geospatial information such as the ISO TC211 international standards., and begun to apply this knowledge to other projects on behalf of CEDA and STFC. For example we have recently developed a data-model for the European Space Agency (The Heterogeneous Missions Accessibility Follow On project) and ire now working on behalf of the JRC (Joint Research Centre) to edit EU data specifications for Oceanographic Features and Sea Regions. These data specifications will form EU implementing rules that data providers (such as CEDA) should follow to be compliant with the recent EU directive on spatial data in Europe (INSPIRE).

I am also managing the technical development of a web-based visualisation for the International Space Innovation Centre (see figures 3 and 15). This is a good example of geospatial standards in action and builds on the work done with CSML and OGC services.

You can see that as a CEDA software engineer, much of my recent work has focussed on the data-modelling aspect of software engineering to build CEDA interoperable services – requiring me to be an active member of the wider community who seeks to develop re-usable solutions to common problems.

Models and Impact Relevant Prediction (MIRP)

Eduardo D. da Costa

The MIRP project is about exploiting the BADC position as one of three global repositories for model inter-comparison data to facilitate information and data flow between the earth system modelling community and the impacts and adaptation communities, including both those who "research" impacts and adaptation, and those at the front end of planning in government and industry. The main aim is to have in place the ability for that information flow to be appropriate and available during and soon after the next IPCC assessment phase. I have been recently recruited for 3 years to work on this project.

My first task is to work to insure we capitalise on the data by getting it used as widely as possible. That not only means that the most recent data should be made available faster and more easily accessible, but also includes the transformation of raw climate model data into end-user friendly formats. My role is to produce such processed products which are expected to improve the flow of information and to make a direct impact on the knowledge transfer from research to actual adaptation and/or mitigation policy.

Figure 16: Usable honeycomb grid connection

Given the diversity of data (sea surface temperature, sea level, atmospheric constituents, clouds, land variables and so on) and user communities, as well as specific needs of the users involved, reaching the agreement about which processed products we should provide is likely to be quite challenging. My role is to engage with the user communities showing them which products we can actually offer. Recently I have computed Weather-regimes and some Atmospheric and Oceanic Indices as the North Atlantic Oscillation Index (NAO) from the CMIP5 data. This is a high profile activity: the analysis of this data will be a key component of the 5th IPCC assessment on Climate Change.

My work is in many ways analogous to scientific research on atmospheric data. However here the focus on the user is a distinctive characteristic which is mainly absent in more traditional scientific research.

Supporting Climate Science

Charlotte Pascoe

My job is about releasing the potential of scientific data. I work on new ways of describing data so that it is easier to find and easier to share. If we describe data well it can be used by other scientists doing the same sort of things. However, if we also describe why and how the data was created it becomes much more accessible and different types of scientists doing different sorts of things can make use of it. With appropriate contextual information or metadata scientific data can be used and understood by anyone who cares to look for it.

The focus of my current job is about working out standard mechanisms for describing climate models and the simulations they produce. Much of that work is currently bound up in the CMIP5 questionnaire described earlier (page 18). Part of my job is to build working relationships with other projects and institutions who want to make that happen. This year I will be working with the University of Bristol to extend the CMIP5 controlled vocabulary to support paeleoclimate and with the University of Reading to embed the CMIP5 questionnaire infrastructure at their Meteorology Department. I am also working with the Ermitage project at the Open University to build a new controlled vocabulary to to capture information about Integrated Assessment Models. The documentation of Integrated Assessment Models will bring transparency to the predictions of the socio-economic impacts of climate change.

Help, Ingest, Review, Deliver

Graham Parton

At the heart of the work carried at CEDA is the mission to assist the wider science community to archive, discover and use environmental science data. In achieving this mission CEDA assists scientists in producing results faster, more efficiently and with greater impact than they would do otherwise. Key to to this success is CEDA community engagement.

Underpinning all of science is the basic tenant that all results should be reproducible. While in the past published papers would contain all the underlying data to enable the work to be transparent, the volume of results generated these days makes this untenable. Thus, the curation of data in well managed archives becomes essential to the scientific process. Sadly, unless early and ongoing engagement with project participants takes place with those responsible for the data archives, the final archiving of the resulting data can all too easily be forgotten or simply left too late for the required information to be captured.

One of my primary roles is to engage with projects early in their life cycle to raise issues on data formats, metadata conventions, data delivery schedules and conditions of use to ensure timely and dedicated agreements. This brings a host of benefits both to the project, though guidance and assistance with data preparation, and the wider community through capturing of essential supplementary information, such as flight details, model run documentation and, crucially, preparation of metadata. Correct metadata records allow archived data to be discovered, published and, therefore, citable by the science community: all essential steps to ensure that scientific outcomes can be scrutinised and are repeatable, but with the added benefit data re-usability beyond the work of the original project.

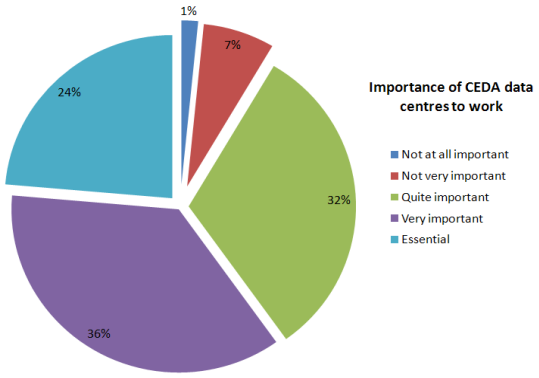


Figure 17: Importance of CEDA data centres to users' work - from 2010 User Survey. Results show the high impact that CEDA data centres have. With nearly a quarter of users saying that CEDA data centres are essential to their work.

Beyond interaction with data providers I take responsibility for controlling ingest processes for the incoming data (mainly for observational data from the Met Office and NERC facilities and operational NWP data) and periodic reviewing of existing datasets. This crucial work brings the latest data to the community swiftly and efficiently, while discovery records and documentation remains concurrent and usable through the review process I've instigated. Handling data ingests and knowing project requirements can bring additional benefits: I can set up dedicated near-real time feeds of meteorological data for campaign planning purposes, while my work with our operations and development teams brings more robust and reliable data delivery and ingest of data.

While much work takes place to support the preparation of data for archiving with CEDA, issues may still arise for the end user. This can range from simple user account issues, questions about specific parts of the archive or assistance with the finding useful data. I ensure such issues receive dedicated support through the CEDA help-desk; where a dedicated team carry out timely triage and onward issue handling to quickly resolve issues and thus enable users to continue with their science as soon as possible – a service much praised in all our user surveys.

Essentially, my role within the wider CEDA activities can be summed up by the title - “Help” both data providers and users; secure and reliable “Ingest” of valuable environmental data; ensure that archive content, practices and services remain fit-for-purpose through regular “Review”; and all this to support the CEDA's raison d’être: to “Deliver” quality environmental data products in usable formats and in a timely manner to the wider science community, facilitating and enhancing their scientific investigations.

What a Data Scientist Does

Kevin Marsh

Experience over many years has shown that data cannot manage themselves. Proper data curation demands dedicated data scientists. The staggering growth in the volume of data produced every year means that this role is becoming more important than ever.

Currently, I am primarily responsible for managing data support for the NERC RAPID-WATCH project, which is looking at the likely effects on European climate of a change in the North Atlantic ocean circulation. Another significant part of my work involves recovering climate and weather model data from The UK Met Office, and ensuring these data are made available to researchers as efficiently as possible. I am also in charge of ensuring that all of the climate model data held by CEDA for CMIP5 has been quality controlled. The volume of data involved (more than 1 PB) is larger than the rest of the current CEDA archive, and so presents a considerable challenge. I also look after a number of smaller datasets (including CRU TS3, which is one of our most widely used datasets), deal with user queries, provide advice to users and bodies such as NERC on good data management practice, and help to resolve various CEDA operations issues which arise. Often, these tasks can be highly time consuming

The terms “data management” and “data curation” sound deceptively simple and straightforward. In order to perform proper data curation, the truth is considerably more involved. My role as a data scientist at CEDA means that I am involved at every stage of the curation process and have to utilise a variety of skills and have a wide breadth of knowledge.

Throughout the duration of the project, I interact with the data providers to build good working relationship, and promote good data management practices to the wider research community. Experience has shown that those data providers who engage with the data scientists throughout the duration of their projects are the ones who:

a) submit the data products originally intended in the recommended format (and on time!),

and

b) submit the most useful associated documentation (metadata).

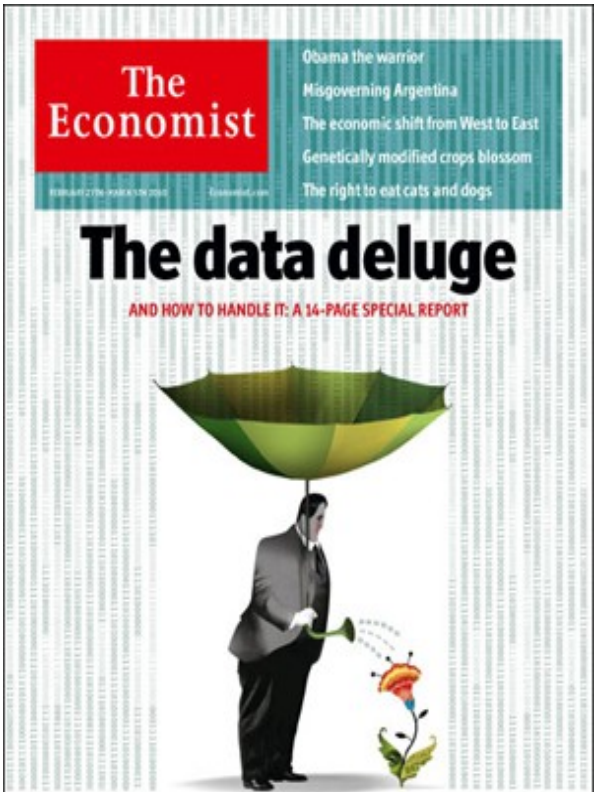


Figure 18: Data Scientists don't just hold umbrellas ...

The data from these projects is often the easiest for other users to utilise and understand, so this is time very well spent.

I am involved in many aspects of CEDA, from helping to shape NERC data policy, providing data management advice and identifying new opportunities, to resolving issues with the CEDA archive and services, right through to helping a desperate user get that vital piece of data for their work.

Our team of data scientists are involved in all aspects of CEDA, from the large (such as shaping NERC data policy, providing data management advice and identifying new opportunities), through to the smaller scales such as helping a desperate user get that vital piece of data for their work. And often that is what makes the job worthwhile.

Infrastructure in support of Earth System Modelling

Martin Juckes

My activities for the past year have centred on leading the data archives component of an EU Framework Program 7 project which is developing infrastructure to support Earth system modelling activities in Europe. The central and immediate challenge is to deliver the European component of the archive for CMIP5: this archive will underpin much of the science that goes into the next assessment report of the Intergovernmental Panel on Climate Change. In addition, I have continued to manage the data management for the NERC QUEST (Quantifying and Understanding the Earth System) programme, and run the associated QUEST Earth System Data Initiative, managed the IPCC (Intergovernmental Panel on Climate Change) Data Distribution Centre and the MIRP (Models and Impact Relevant Prediction) knowledge transfer project. Further, I have engaged with the academic community through supervision of a successful Ph.D. student (doctorate awarded Feb. 2011), work on the statistics of climate reconstructions and multi-model means. Finally, I led a successful bid for funding from the G8 Research Councils joint funding initiative.

The "Infrastructure Support for the European Network of Earth System Modelling" (IS-ENES), led by the Institut Pierre-Simon Laplace (IPSL) provides an umbrella for the work of the European Network of Earth System Modelling. The IS-ENES project also provides a forum for discussion of the strategic priorities for Earth system data services in Europe. Active participation in these discussions helps to keep the UK at the front of such developments, and had a strong influence on the successful bid to the G8 joint funding initiative.

CEDA efforts to manage data for the QUEST program continue to be limited by the relatively small amounts of data provided by the projects. Through engagement with the scientists we have sought to ensure that all available data is appropriately archived and documented.

Exarch

- Strategy and engagement
- Workshops
- Linking with NASA, ESA and GCOS
- Governance
- Accessibility
- Software
- Robust meta-data
- Query management
- Near archive processing
- Quality control
- Scientific Diagnostics

Figure 19: The G8 exascale project "Exarch": activities (See http://proj.badc.rl.ac.uk/exarch.)

The support gained from the G8 Research Initiative for the "ExArch: Climate analytics on distributed exascale data archives" project (see box) will strengthen CEDA's collaborations both within Europe and with North America. The ExArch provides a framework (incorporating a strategy, prototype infrastructure and demonstration usage examples) for the scientific interpretation of multi-model ensembles at the peta- and exa-scale. ExArch focusses on the challenges of the CMIP5 archive. In addition, ExArch will exploit the data resources generated by the CORDEX (Coordinated Regional Downscaling Experiment) project, which will push even beyond CMIP5 in resolution, albeit on the regional scale.

Research results have been presented at the general assembly of the European Geophysical Union, the MUCM (Managing Uncertainty in Complex Models) conference on uncertainty and the "Climate 1K" workshop on reconstructions of the climate of the last thousand years.

My work this year has spanned a wide range of projects, but the central theme has been support for international model inter-comparison projects (CMIP5 and CORDEX). These projects combine huge complexity (both managerial and technical) with immense socio-economic significance. It is very satisfying to be part of a team at CEDA which is providing leadership in this area of data management.

Development Manager

Matt Pritchard

My role exists to coordinate development activities undertaken within the Centre for Environmental Data Archival (CEDA) in support of its mission : the long-term curation of scientifically-important environmental data, and facilitating use of the data by the environmental science community. CEDA's development activities cover infrastructure maintenance and improvement in addition to research and implementation of informatics solutions for secure, federated access to geospatial data resources.

Central to my role is support and communication within the CEDA development team. Regular development team meetings are held, giving an opportunity for developers to present and discuss their work in the CEDA context, sharing solutions to technical problems and stimulating discussion of best practice within the group. In addition, I am the line manager of several members of the team and hold regular feedback meetings to foster good communication and support them in their work.

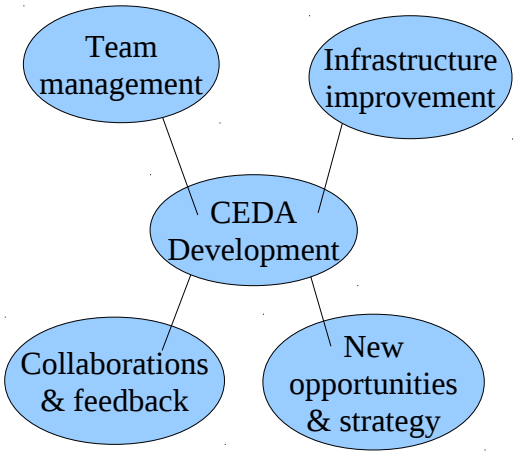


Figure 20: The role of the CEDA Development Manager.

"Behind the scenes" of the many front-line CEDA services lies a considerable infrastructure of storage hardware, systems for data ingestion and user support, and tools for data scientists to do their job of curating the data resources themselves. All these require constant attention and regular maintenance to ensure that CEDA can deliver on its commitments to provide reliable service to the science community. As new technologies feed through from collaborative projects, tools and services evolve to "fold in" these new developments.

CEDA is not just a data store: it is actively engaged with the UK, European and international science communities. In many of these collaborations, development work helps push the boundaries of geospatial informatics to meet the needs of 21st-century environmental science. Here, it is important to coordinate the implementation and deployment of resulting technology in a way that maintains a stable operating environment within CEDA, while enabling "bleeding edge" technology to be tested and used.

New opportunities emerge from all angles and with a wide range of partners, customers and stakeholders. It is important to maintain an information exchange between those bringing in new work and those responsible for delivering development work, to ensure that proposed activities fit with CEDA's strategic goals and are feasible on the required time scales and with available resources. Ideally, development should build on or improve existing components, so that improvements can be fed back into CEDA's operational infrastructure for wider benefit. Where possible, work should aim to adopt, implement or even contribute the development of, industry standards to ensure interoperability.

My role as CEDA development manager sits at the heart of CEDA, coordinating a team of talented developers involved in bleeding-edge geospatial informatics, while overseeing the maintenance and improvement of operational systems required to provide robust and reliable services to the environmental science community.

Software Development in CEDA

Maurizio Nagni

As a Software Developer my duty is to maintain the existing software as well develop new software following the changing requirements of CEDA department and its external partners, using Java and Python as primary languages. Moreover, as a commercially experienced developer, sometime, I try to introduce commercially widely used and well recognized Open Source tools instead of some more "in-house" solutions.

CEDA is not a traditional "Software House" nonetheless the software development has a great part in it. An incremental/flexible development process allows in most of the cases, the extension of the existing software towards new needs; a constant brainstorming among the group developers (say the Python/Java permanent discussion) enhance the possibility to find a solution for a specific problem. Moreover this dynamic expedites the introduction of state-of-the-art software which exploits both new development concepts and established market solutions.

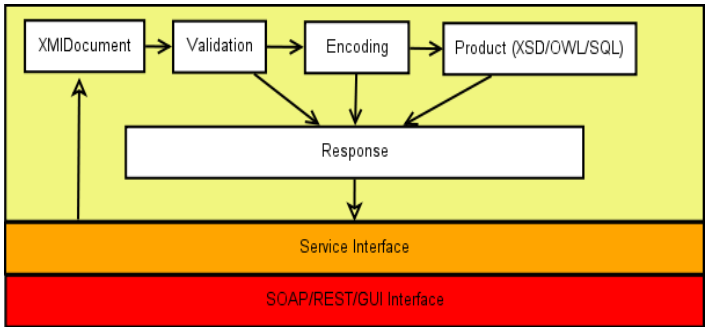


Figure 21: NewMoon is a CEDA project aimed to support the Model-driven architecture software design strategy

The use of the "Software As A Service" (SaaS) concept, introduces several advantages, which are also recognized by the Joint Information Systems Committee (JISC). Apart from the internal management, the CEDA external interface makes our image, and I am not alone to consider the SaaS as crucial concept in CEDA activity. It's a huge step over the simpler "common/shared software library" concept. The SaaS describes which features we can directly offer to our users, through synchronous or asynchronous services, accessible from anywhere with an Internet connection; such services may expose raw data, processed data products or help the user to "transform a document".

One example of the latter is the NewMoon project. NewMoon provides a web interface that allows a remote user to upload a UML document designed using the CSIRO "HollowWorld"[17], and get downloaded an XML schema document compliant with the "Geographic Markup Language". This is one part of the tool-chain necessary for standards compliant metadata to support data interoperability between institutions and disciplines.

Specific tasks, as the Discovery Service for the Marine Environmental Data & Information Network, the NERC revitalization project or the NewMoon project, allowed me to introduce commercially recognized environment like the JBoss. As J2EE certified Application Server, JBoss allows to go over the limitations of the previously used pair Tomcat/Axis2 both reducing the number of components required to deploy a service and strongly improve the environment's features (SOAP, Object-to-DB mapping, JMS, Timer, etc.) available to the specific deployed service. This step perfectly fits with the SaaS paradigm offering rapid scalability as well proven reliability.

The challenges, the innovating projects as well satisfactory results are a constants driving force. Using the SaaS as "guide line", the users requests and/or critics as constructive product's feedback, the attention to the new technologies as opportunities for further improvement, and the team coordination as productivity tool, all these things, surely give strength to the services that CEDA offers to the environmental science community.

[17]NewMoon is available at http://bond.badc.rl.ac.uk/newmoon/faces/pages/encodeXMI.xhtml; HollowWorld is documented at https://www.seegrid.csiro.au/wiki/AppSchemas/HollowWorld

Restricting access to data to make it more widely available

Philip Kershaw

My role is as a technical specialist in the area of access control and security developing software and designing systems to perform this function. In much the same way as a Gatekeeper might stand guard and allow only trusted people through, access control software forms a protective layer applied at key entrance points – the software services which serve data from within an organisation to the outside world over the Internet. It is one of the first points of contact for users. Whilst it should prohibit those who are not authorised, this layer shouldn't at the same time make it difficult for those who have legitimate access. To a large extent, access control should do its job but never be seen. If it is seen, there is invariably something wrong.

Think of the word security and it is likely to conjure images of padlocks, safes, fences to name but a few examples – all means of preventing access to something or at least making it more difficult. In the context of CEDA, security, or more precisely access control, provides a means to restrict access to the data we host through Web based services. This could at first sound odd as a core function of CEDA is to make data available to the user communities it serves. In many cases however, licenses or terms of use for datasets can be a barrier to access: without ready means to enforce such agreements, data providers cannot or will not open access to their data. There is a paradox then and this explains the apparently contradictory title above: provide a means to secure access, and such data can be made available to the benefit of a much wider community of users. This is one illustration but there are other drivers for providing access control: the need to protect finite computing resources from that unintended request to download half of the archive, or the ability to keep a record of user access requests to feedback to data providers and project sponsors.

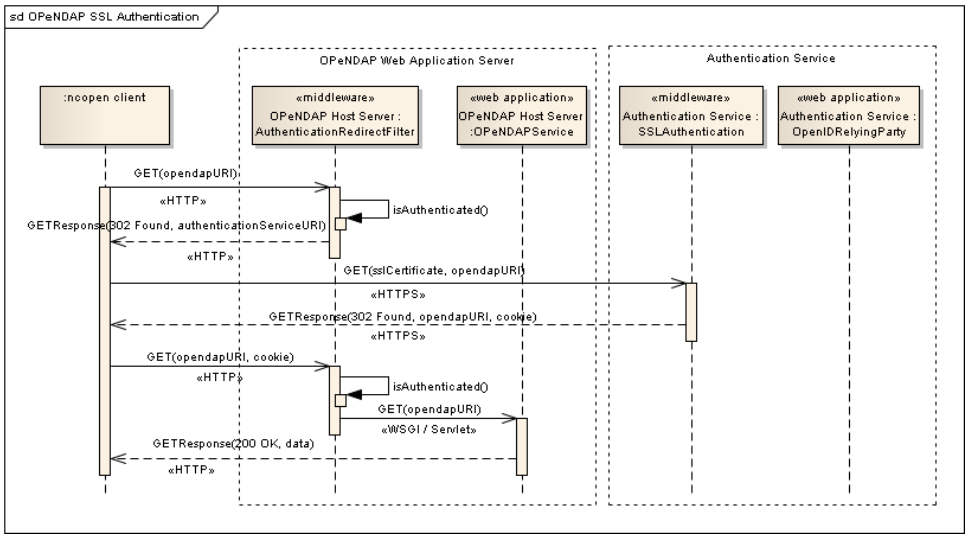


Figure 22: Example sequence diagram showing how an OpenDAP data access request can be secured for authentication. (See http://centaur.reading.ac.uk/25087/ for details.)

Access control software manages access across a given boundary and traditionally this has meant the confines of a given organisation. However, with the advent of Grid and Cloud Computing paradigms these boundaries have been blurred – a brief introduction of these issues in the context of CMIP5 appears on page 20.

Collaborating with colleagues in my field and sharing research and best practice at conferences and other meetings are important aspects of the work I do. Another output is contribution to the Open Source software development community. The software developed for these projects is free and open for anyone to use and modify. At CEDA we make use of such software and I have developed a number of security related software packages which are available as Open Source.

Identity, security and trust are challenging concepts to apply in a virtual environment. For any given problem there is almost certainly no one best solution and this makes this field all the more interesting.

Project Manager

Sarah Callaghan

The work of a project manager is as varied as the projects that they manage. CEDA, as well as being a world-class data management facility, also collaborates internally between groups and with external organisations on projects as diverse as developing information models and producing web-based data portals. As a project manager, I am responsible for ensuring that projects run according to time, produce outputs of an acceptable quality and keep costs within budget. I am also responsible for communicating with various funders via regular progress reports and by responding to specific requests for information.

At this time, I am project managing three significant and distinct projects, all on operating on different time-scales, and financed by different funding bodies (with different reporting requirements and processes): Metafor, the NERC citation and publication project, and ACRID.

Metafor aims to document climate models and the software that produce them in a standard way. The project team comprises of over 40 members of staff from 12 institutions spread out across Europe and the United States. The total budget for this project is 2.85 million Euro. Metafor is now approaching its final six months of operation. As well as project management, I lead the work package on disseminating the project results, and so am responsible for coordinating the project workshops, meeting and conference attendance.

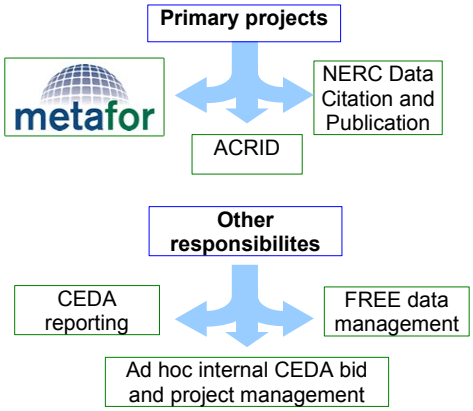


Figure 23: Sarah Callaghan's projects and responsibilities.

The NERC data citation and publication project brings together all the NERC funded data centres to develop a standardised way of allowing the scientists who archive data in the data centres to cite those data sets, in a way analogous to how scientific journal papers are cited. This project is beginning its second year of a three year lifespan, and has 18 project members. As part of this work, I am a member of the CODATA-ICSTI Task Group on Data Citations and the DataCite Working Group on Criteria for Data centres.

The Advanced Climate Research Infrastructure for Data (ACRID) project aims to implement a linked-data approach for sharing some example climate datasets. It is a collaboration between the University of East Anglia, CEDA and the Met Office. The budget is just under £188,000 and the project will be completed in July 2011.

I also act as a project manager for the ad hoc internal CEDA projects, such as the recent CEDA time audit, and coordinate CEDA responses to invitations to tender and calls for project proposals. I still do a little data management: liaising with the principal investigators of the NERC Flood Risk and Extreme Events (FREE) programme to ensure that the data they produce is properly documented and archived.

On a daily basis, I keep track of all the work being done in my projects, who's doing it, and how much time and money they're spending on it. I ensure that the CEDA annual and quarterly reports and requests for information get dealt with, and that all the project reports and other deliverables (e.g. software) are finished and sent to the customers on time. This involves reading and writing a lot of email, making phone calls, attending meetings and teleconferences, and keeping track of schedules, documents and people. Unsurprisingly, this job requires significant organisational, communication and juggling skills, but it's always varied, and rarely boring!

Curation and Science Delivery

Sam Pepler

My job is to keep the data centres operated by CEDA running effectively. These data centres are the main business of CEDA and provide a foundation for many of the other projects undertaken. The data centres source, ingest, store, document and provide access to data to enable environmental scientists to more effectively reuse and share information. The three NERC data centres run by CEDA are the British Atmospheric Data Centre (BADC), the NERC Earth Observation Data Centre (NEODC) and the UK Solar System Data Centre (UKSSDC). As these names suggest these data centres look after data from different environmental science disciplines.

The fundamental reason for spending so much effort on data management is to enable reuse of the data to make science more effective and more impactful. This is not just about some abstract future where data may or may not be useful in some hitherto unthought-of research, but reuse by people doing research in the here and now. Much of the activity of data centres is around sourcing and ingestion of data. We support scientists as they transform data to allow reuse, document the data, and formalise any usage restrictions that may apply to the data. I manage the data scientists responsible for dealing with this incoming data to ensure that we have a consistent approach that will lead to the best outcome for science.

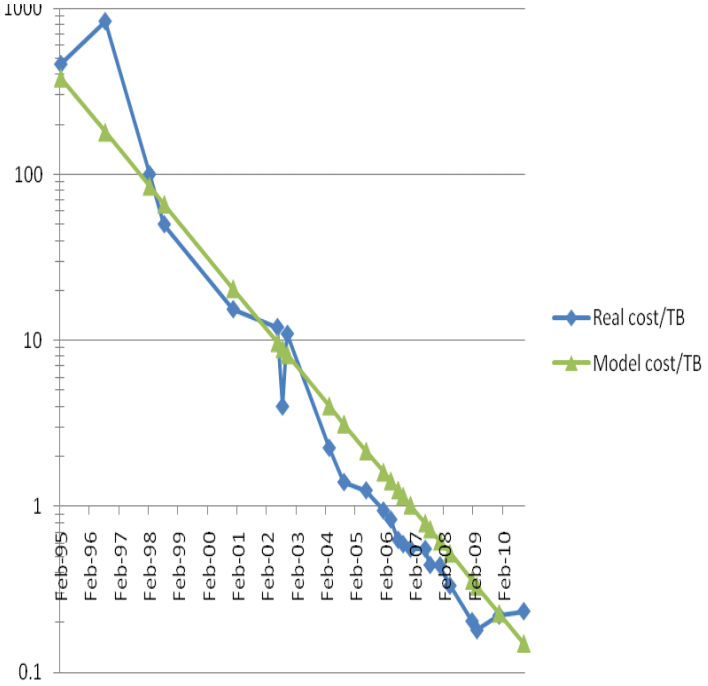


Figure 24: The erratic drop in storage cost. The blue curve shows the actual prices we have paid for storage systems in pounds per terabyte (£/TB) and the green line is a simple log model.

Figure 1. The erratic drop in storage cost over the life time of the BADC. The blue curve shows the actual prices paid for storage systems in pounds per terabyte (£/TB) and the green line is a simple log model.

As the data is ingested it is deposited into a storage infrastructure. I responsible for the integrity of the data in this storage. At the moment we have at least two copies of most data: one on disk, using Network Attached Storage, and one on tape (using shared infrastructure with other sciences[18]). I have to coordinate changes and expansion of this storage, and ensure the data repository has adequate backup, performs integrity checks etc. It's not all about file management; there are also metadata and data access services, and documentation to coordinate.

Ingesting the data and supporting the infrastructure to store and access the data are the backbone services on top of which many other project within CEDA build. All these projects need financial and human resource. I maintain an overview of all the CEDA projects in order to spot and forecast problems. Adapting to the cost implication of the ever changing world of IT is a perennial problem (see figure 24).

Running data centres to store research data is a relatively new and evolving area. Other data centres and data service providers are forming a new digital curation community. CEDA is at the forefront of this community and we are now key members. I'm on the review board of the International Journal of Digital Curation and I have also sat on a British Library's committee developing strategy for the digital future.

Data management is now an integral part of the research process. There are more and more projects and services dealing with data. CEDA is going to be busy for some time to come.

[18]The Atlas Petabyte Tape Store: http://www.stfc.ac.uk/e-Science/services/atlas-petabyte-storage/22459.aspx

Information Management

Spiros Ventouras

My role within CEDA is twofold: interaction with data providers and users for archiving and maintaining scientifically important data, and; development of information models to support data curation and data access.

In the first instance, I maintain regular contact with the Centre's customers by liaising closely with the Principal Investigators, and with members of the scientific teams of each programme. I provide support for issues such as data collection, data processing, file naming, formatting and submission of their data to the BADC. I also manage dataset access according to agreed guidelines for each programme, such as those set out in the relevant Data Management Plan. From a technical perspective, I also set up and manage the archiving procedure. Finally, I prepare reports on data management issues which I present to the steering and management committee of each programme with which I deal.

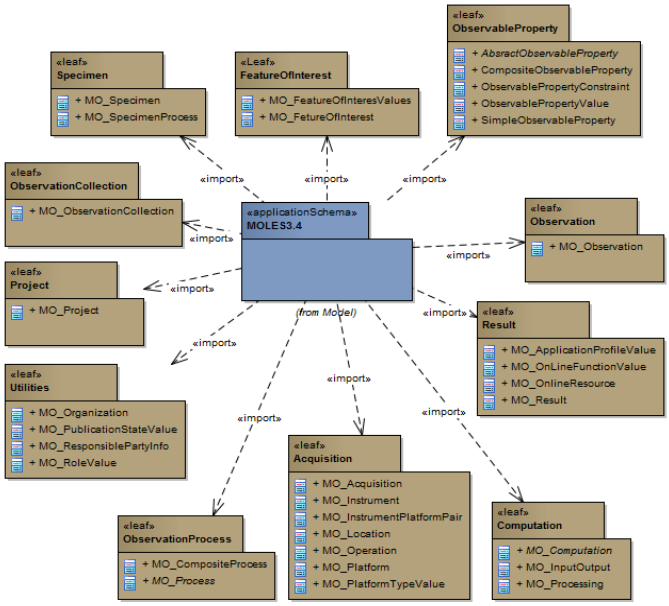


Figure 25: The MOLES information model

In my role as information modeller, I have been involved in the wider spatial information infrastructure activities of CEDA at both national (NERC) and international level. These activities are based on the integration of the concepts of geographic information with those of information technology. Their purpose is to increase the understanding and usage of geographic information, by the development of appropriate interoperable conventions and standards. The aim is to increase the availability, access, integration and sharing (interoperability) of geographic information.

At the moment, I am involved with two important information modelling activities: MOLES and INSPIRE.

I have made significant contributions to the conceptual design of The Metadata Objects for Linking Environmental Sciences (MOLES). This project has been initiated within NERC to fill a missing part of the 'metadata spectrum'- data about data. Through my involvement with MOLES I became familiar with the concepts and principals of conceptual modelling, particularly through interactions international colleagues in the context of the ISO19156 Observations and Measurements standard.

The INSPIRE (Infrastructure for Spatial Information in the European Community) framework aims to enable the sharing of environmental spatial information among public sector organizations and better facilitate public access to spatial information across Europe. I am the editor of the INSPIRE Data Specification of the combined Thematic Working Group (TWG) on Atmospheric Conditions and Meteorological Geographical Features, two of 34 spatial data themes. My role involves the development of the conceptual models and the documentation of the data specifications for these two themes. I also advise the TWG on data modelling issues, ensure consistency with other themes and ensure that ISO TC 211 standards are applied.

My natural preference for analytical thinking (vital to information modelling), means that I feel more suited to the work I do as an information modeller than a data support scientist.

Scalable Software Specialist

Stephen Pascoe

There is a huge difference between an exciting technology that could revolutionise the way scientists use data and reliable IT infrastructure that realises that technology's potential. The former involves innovation, abstraction and the breaking down of assumptions whereas the latter requires pragmatism, a grasp of resilience technologies and a focus to drive a prototype towards a usable system. Balancing these two aspects is essential to improving CEDA's ability to provide cutting-edge informatics support to scientists. This is my speciality within CEDA's software development group.

CEDA participates in a broad spectrum of scientific informatics projects from improving internal systems to international collaborations on informatics research and infrastructure. I work within many of these projects to ensure that the software they develop is suitable for inclusion in CEDA's operational services. Operational services are required to meet high standards of availability, security and resilience, and achieving these standards involves more than simply deploying the service on appropriate hardware. Such services must be designed from the outset to run on multiple servers simultaneously, whilst appearing as a single application to the user, thus maintaining availability and responsiveness during periods of high demand. At CEDA we achieve this by building our software on solid software engineering principles such as Service Orientated Architecture (SOA). Using SOA we can compose our software components in the areas of scientific metadata, web security, geospatial visualisation and on-demand data processing into scalable web applications that meet the specific requirements of the project. Components can be replicated on multiple servers to increase service resilience.

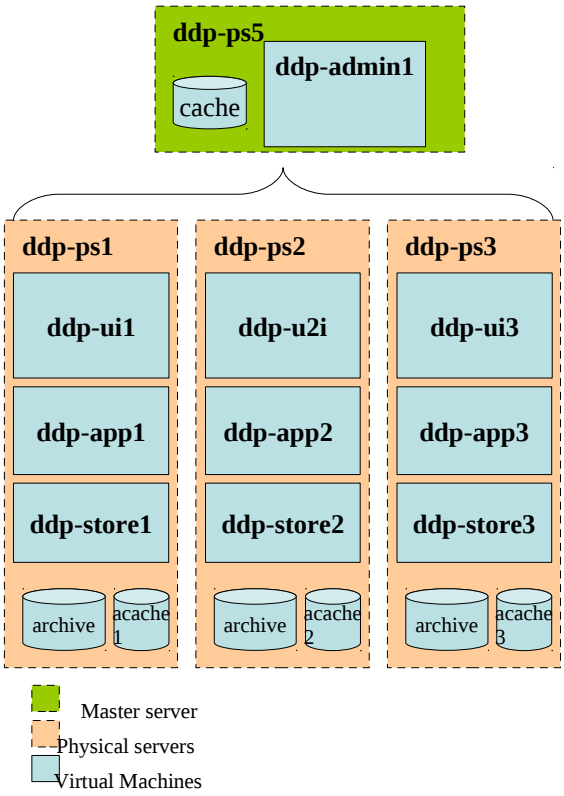


Figure 26: The UK Climate Projections User Interface serves over 3,000 registered users. During launch the server cluster was scaled-up to meet user demand from the 7 virtual servers shown here to 22 servers.

It was this careful adherence to scalable software engineering principles that enabled the UK Climate Projections user interface, an interactive data-driven web application developed at CEDA, to be scaled-up to over twice the number of servers in response to the high demand during it's launch period, therefore achieving a successful launch of this high-profile service (see Figure 26).

The experience I have gained in deploying scalable web applications at CEDA is now being applied to several international collaborations with the aim of building a global infrastructure for climate and satellite data. The European Union IS-ENES project, Earth System Grid Federation and G8 ExArch project are all ongoing efforts to build a reliable IT infrastructure capable of serving petabytes of data to a global community of scientists. These highly ambitious projects are ground breaking for climate science informatics and their success will depend in large part on a balance between innovation and pragmatism we have adopted within the CEDA deployment group.

The Data Providers Web Service (DPWS) API

Steve Donegan

The NERC Data Discovery Service (DDS) allows users to search a catalogue of the datasets held at all the NERC data centres. This service is a vital component in the NERC web presence as it allows users to search disparate NERC data resources for specific matches to their particular requirements. Search methodologies ranging from a simple free text search through to complex spatio-temporal and targeted text searching are all supported. For the DDS to operate successfully all NERC data centres must produce and publish timely metadata. All NERC data centres have variously developed different methodologies to collect and store metadata about their own data holdings and also to publish this information. As part of the a project to revitalise the DDS, CEDA provides the operational service to collect such metadata from all NERC centres and insert them into the DDS catalogue in addition to running the web service that underpins the DDS itself. In order to help manage this process, CEDA has developed the Data Providers Web Service (DPWS).

The DDS consists of 3 main components: ingest, responsible for collection and insertion of metadata into the discovery database catalogue; search service, the discovery web service (DWS) that searches the catalogue; and the web portal itself that users use to conduct searches (a client of the DWS). One of the remits of a recent project to revitalise the DDS was to spread development effort and the hosting of these services across the NERC centres. The new DDS portal was developed[19] and hosted by BODC. The DWS and metadata catalogue was to be a further development of the CEDA service provided for the Marine and Environment Data Information Network (MEDIN).

The original DDS had a number of problems, it was difficult for NERC centres to track and control the inclusion of their metadata, and 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. It used a customised interface that didn't support standards compliance.

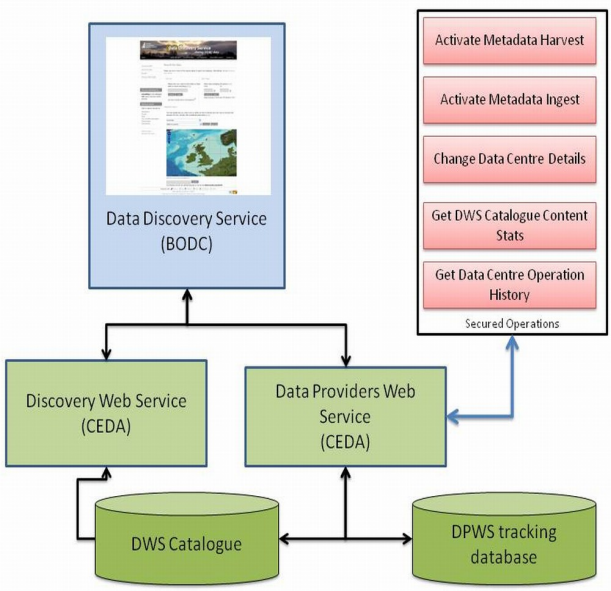


Figure 27: The DWS performs searches on the catalogue in response to requests from the DDS. The DPWS controls the flow of metadata into the DWS catalogue via a secured system from a tab within the DDS itself. Different role types define the operations users of the DPWS can perform.

CEDA proposed and implemented a system that would address these issues: the DPWS, to operate alongside the DWS at CEDA and a new portal was to be developed by BGS[20] to provide a visual interface for data providers to use. This new portal would simply take the form of an extra "tab" within the new DDS portal run at BODC inheriting the DDS style and visual cues.

In the provision of the DPWS CEDA has shown that it can develop and integrate new services in coordination with other NERC developers. The development of the DPWS has not only had the bonus of easing the management oversight required for the various processes needed to support the DDS and DWS, but has also simplified the systems required at CEDA to perform these tasks.

[19]Based on the MEDIN Data Discovery Service which is Open Source.

[20]As of March 2011 this has not been incorporated into the main updated Discovery portal, however the DPWS back end itself is fully functional and capable of supporting the operational harvest and ingest of metadata into the DWS.

Earth Observation Data: NCEO, ISIC and ESA

Victoria Bennett

CEDA is part funded by NCEO, the UK National Centre for Earth Observation. As such, CEDA is tasked with providing data management support for NCEO participants, and for the wider community of earth observation data users. One of my roles in CEDA is to make sure we carry out our NCEO work programme, and report on progress to the NCEO Directorate. The NCEO work programme encompasses a wide range of data centre activities carried out by a number of people, but specific tasks on which I have worked this year have been 1) archival of NCEO products, 2) the ISIC (International Space Innovation Centre) scientific visualisation service, and 3) supporting the The European Space Agency (ESA)'s Climate Change Initiative (CCI).

NCEO Products: The NCEO comprises 7 science themes, with around 120 scientists in over 30 institutions working on diverse projects, but in many cases the aim is to generate novel scientific data products from satellite data. These derived products, often global long-term datasets, need to be archived and made available to other users. I work with the data providers to ensure the data are in a suitable format, with metadata as agreed in the NCEO Data Management Plans, and any additional information needed to document the products in the CEDA catalogue.

Provider ID	Created	Type	Title	Subtype
neodc.nerc.ac.uk	2010-02-16	Activity	SeaWiFS primary production	Deployment
badc.nerc.ac.uk	2010-02-16	Activity	SeaWiFS ocean properties	Deployment
badc.nerc.ac.uk	2009-10-16	Data Entity	Ocean properties from SeaWiFS/SeaStar (QUEST/CASIX, 2009)	Measurement
neodc.nerc.ac.uk	2010-01-20	Data Entity	CASIX - SeaWiFS primary production	Measurement
neodc.nerc.ac.uk	2007-09-12	Activity	ARSF - 03/18 project	Activity Data Campaign
badc.nerc.ac.uk	2008-10-06	Data Entity	QUEST - Marine Biogeochemistry and Initiative in QUEST (MarQUEST)	Measurement
neodc.nerc.ac.uk	2007-09-08	Activity	ARSF - 04/13 project	Activity Data Campaign
neodc.nerc.ac.uk	2010-02-16	Activity	Centre for observation of Air-Sea Interactions and fluXes (CASIX)	Activity Data Collection

Figure 28: Entries in the CEDA catalogue that are returned in response to a search for SeaWIFS.

ISIC Scientific Visualisation Service: During the last year, CEDA was a partner in the creation of the ISIC Visualisation Centre, and developed the “EO Science Visualisation Service” (SVSeo), as a component part. SVSeo is discussed on page 17. I was responsible for the CEDA contribution to this work, setting up the subcontracts, liaising between partners etc. Figure 28 shows catalog entries which include the data shown in Figure 3 - making the link between the fundamental work acquiring and documenting the data, with the visualisation services deployed by CEDA – both internally and in partnerships like ISIC.

ESA Climate Change Initiative: The CCI is a large (75 million Euro) international programme, funding European consortia to produce long-term high-quality datasets of Essential Climate Variables (ECVs) from satellite observations. The aim is to generate climate datasets with high visibility and usability across a range of user communities, targeting mainly, but not uniquely, the climate modelling community. To achieve this, it's important to not only put effort into high quality, well calibrated measurements and state of the art processing algorithms, but also into the definition of "rules" for end products to ensure maximum uptake of the data. From January 2011 I have been seconded into the ESA Climate Office at Harwell to advise and coordinate the ECV projects on these "data standards". My first task is setting up a Data Standards working group with one member from each ECV project. The aim of this group is to agree on guidelines for the entire programme's data products, taking into account data formats, metadata within the files, metadata describing the datasets, and how this will all be managed. Given the diversity of data products (land cover, glaciers and fire disturbance, sea surface temperature, sea level, ocean colour, atmospheric constituents, clouds and aerosol) and user communities, reaching agreement is quite challenging. My role is to make sure we do as good a job as possible at harmonising the data, and adhering to existing relevant data standards.

Science support for aircraft data and NERC measurement-based research projects

Wendy Garland

The CEDA science support team role is to ensure that observations and measurements collected during NERC and other projects are properly curated and available to the widest audience to maximise usage and potential. Real observations and measurements are essential for understanding the complex atmospheric processes and are vital for designing, validating and verifying new instruments and the many complex models used for analysing, forecasting and future prediction. Atmospheric measurements can be made either *in situ* – by aircraft, airborne instruments or at ground level; or remotely –from the ground, satellite or airborne platforms. My role focuses on managing data collected by projects using research aircraft often including supporting measurements by ground-based instruments.

CEDA manages and archives the data from the NCAS-Met Office FAAM BAe-146, the NERC ARSF Dornier, and data collected under the EUFAR project from the 20+ European research aircraft. Operating aircraft platforms is expensive and often cannot be repeated and so it is important that the data is stored and documented to extract as much science out of each flight as possible.

Managing aircraft data is a complex issue. The data are often voluminous and files per flight are numerous. Many data types are collected for each flight which require specialist post-flight processing by instrument operators and therefore can be delivered to the archive by many routes, on different time scales. In addition, data is collected by these aircraft facilities on a project-by-project basis, each being very protective of its data in the first instance, so the data access conditions vary per data-type and per project. A secure access system needs to be implemented for each project (sometimes more than one) with specific access conditions, time scales and requirements. The complexity is illustrated thus: for the year 2010 the FAAM aircraft made 66 flights for 15 projects, each with 4+ data types:- raw data, processed data (core, cloud-physics, dropsondes, video) and non-core data – raw and processed (for both NERC and Met Office instruments), NERC-ARSF had 23 flights for 10 projects and EUFAR had 18 projects using 12 aircraft. In order to keep track of these many and complex data and to manage the archive, tools have been developed to ingest files and to display the archive contents at a glance.

Regular communication with all aircraft operators, data providers and project teams at all levels (PIs to instrument operators) is essential throughout the projects. This includes producing a data-management plan (DMP) at the beginning of the project which defines what data is expected to be produced and how it should be formatted and archived. Useful meteorological forecast products can be provided for use by the project teams during intensive campaigns and helpful email lists and online work-spaces to facilitate planning and the exchange of preliminary findings supplied. Once measurements have been made, support is given to ensure it is correctly formatted (into community- agreed standard formats – namely NetCDF and HDF) and documented, and to assist in the upload to CEDA where it is then logged, checked and archived.

Competent data management is essential so that valuable data is not lost or rendered inaccessible, however it is often a low priority to busy scientists eager to press on with interesting research. To ensure it is given appropriate consideration CEDA is represented on project management/advisory teams, the Steering/operating committees of the NERC-ARSF and FAAM aircraft and is a EUFAR partner.

The CEDA science support team are actively involved in throughout measurement projects and support the project teams in the planning of observations and the formatting and archiving of their data. Data from these campaigns, especially those involving aircraft, are varied and complex and require careful handling in order to ensure they are archived in a manageable way. By doing this, these data are discoverable, preservable and widely accessible and their potential usage and value is maximised.