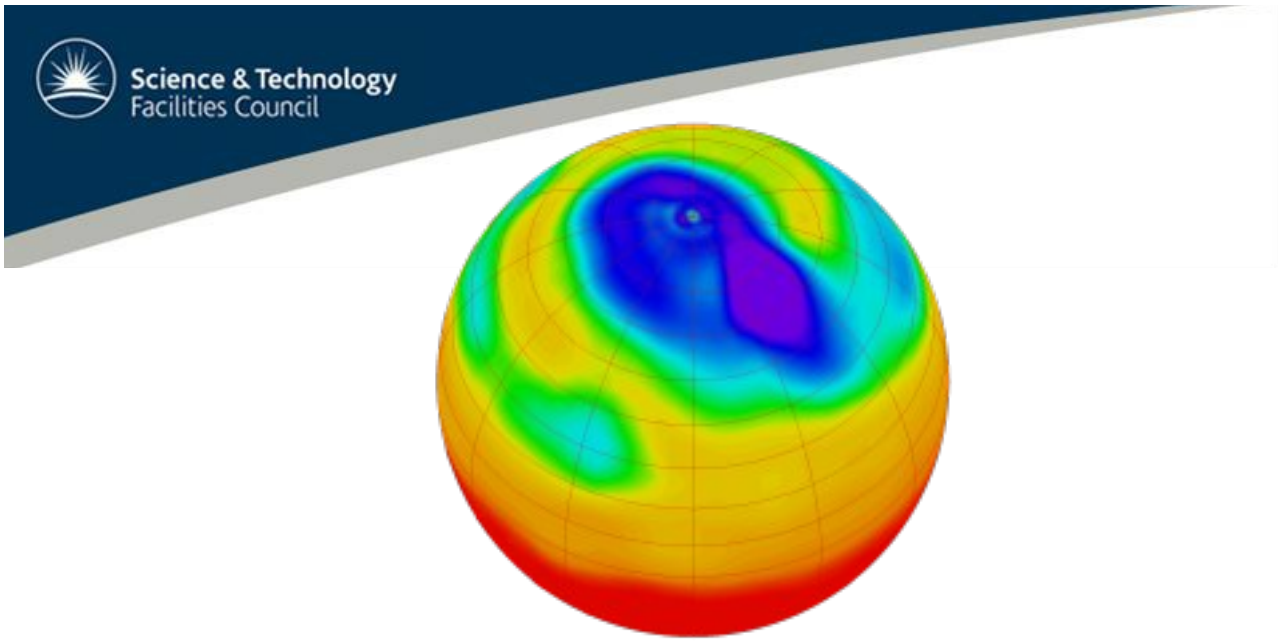


STFC
Centre for Environmental Data Archival (CEDA)
Annual Report
2013
(April 2012-March 2013)

CEDA delivers the

British Atmospheric Data Centre

for the National Centre for Atmospheric Science,

NERC Earth Observation Data Centre

for the National Centre for Earth Observation,

UK Solar System Data Centre

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.

We are pleased to present here our fourth annual report, covering activities for the 2013 year (April 2012 to March 2013). The report consists of two sections and appendices, the first section broadly providing a summary of activities and some statistics with some short descriptions of some significant activities, and a second section introducing some exemplar projects and activities. The report concludes with additional details of activities such as publications, software maintained etc.

This year was dominated by the commissioning and first year of the new JASMIN data intensive computing environment. Nearly all staff were involved in some way or another with the advent of JASMIN since it has transformed both what CEDA could do – both in terms of effective archival and facilitating the exploitation of data. However, it was not trivial to migrate to JASMIN! Some of the issues associated with migrating to JASMIN are discussed in the section on "moving a petabyte" (page 16) alongside a description of JASMIN itself (page 9) and of one of the key services running on JASMIN – the academic component of the facility Climate and Environmental Monitoring from Space (CEMS, page 10). One major early application on JASMIN was the UPSCALE project (page 20) which used JASMIN to provide the analysis environment for what was then the largest climate computing project in Europe (if not the world). JASMIN supports more than just major projects, and the new Group Work Spaces (page 18) provide the facility for new virtual organisations to spring up around shared data resources – and some examples of such activities appear in the description of big data challenges in the earth observation community (page 19).

Traditional data management activities remain important – without data management, data can atrophy and be lost for use, even if the bits and bytes are still preserved. Alongside the operational data centre services (page 7 and 8), CEDA continued to play a leading role in establishing data management support for a wide range of programmes (e.g. those discussed on pages 13 and 31) and in developing and supporting metadata activities such as the Climate Forecast conventions (page 12). CEDA provides key tools for data discovery on behalf of both the NERC community and the wider marine community as described on pages 22 and 28. We continue to believe that formal data publication (page 25) is a crucial part of the armoury of data management, although it is not the only tool in the data management toolbox.

As always, CEDA staff continue to be engaged in a wide range of national and international projects both technical (e.g the ESPAS and CHARMe projects discussed on pages 23 and 27) and scientific (e.g. CMIP5, page 24), and CEDA continues to provide data services under contract to and for government (e.g advice on data handling for agriculture – page 29 – and providing the UK climate projections interface and the IPCC data centre services – pages 21 and 26 respectively). The relationship with the Met Office (page 15) remains important, but it is just one of many important relationships – our peers in NERC and our international partners, particularly those in Europe (e.g. ESA, page 30, are crucial to our ongoing ability to sustain a world class facility.

Bryan Lawrence, Director

Table of Contents

Summary of 2012/2013

The CEDA mission is to support the atmospheric, earth observation and near-Earth environment research communities in the UK and abroad through the provision of data management and discovery services. Over the past year CEDA has further enhanced this role through new tools and services to aid data preservation, curation, discovery and visualisation – adding value for the world-wide user community. In addition to supporting the Natural Environment Research Council's National Centres of Atmospheric Science and Earth Observations (NCAS and NCEO), it also delivers major projects with funding from a range of other bodies, including work for the European Space Agency, JISC, DEFRA and others, as well as participating and coordinating major European projects under the Framework 7 programme.

In the year 2012-2013, CEDA delivered in excess of 410 TB of data in over 4.3 million files from 198 datasets to 3354 distinct users.

CEDA continues to support the fifth Coupled Model Intercomparison Project (CMIP5) through its interactions with the large global collaboration to deliver an "Earth System Grid Federation" under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP). Other major international collaborations include participation in the infrastructure projects (originally IS-ENES and now IS-ENES2) to support the European Network for Earth system Simulation (ENES).

At the time of writing CEDA manages:

- **1748** logical file sets (that is, primary data entities).
- **2.0 PB** of primary data distributed over 3.4 PB of available primary storage.
- **100 distinct disk partitions.**
- **89 million primary files** (in excess of 200 million including secondary and tape copies).

Notable Events

1. The new CEDA website (www.ceda.ac.uk) was launched, providing enhanced visibility of CEDA data centres, projects and services
2. The first year of operations for the JASMIN infrastructure saw CEDA migrating its archive data holdings to new, fast storage within the JASMIN environment. The mammoth task of carefully moving over a Petabyte (PB) of data was completed in less than 6 months, for the holdings of both the British Atmospheric Data Centre (BADC) and the NERC Earth Observation Data Centre (now known as the CEMS Academic Archive). Meanwhile, CEDA has been migrating many of its core systems and outward-facing services to the same infrastructure, and was able to offer new group workspaces and virtual machine hosting to NCAS and NCEO projects. The JASMIN environment currently provides a total of 5 PB fast storage and ~600 compute cores, from which CEDA is able to provide user login access to scientific analysis virtual machines, data transfer nodes and a small processing cluster called LOTUS which has been used to great effect in several projects already: one project calling the facility "game-changing" for whole-mission processing of satellite datasets.
3. The NCAS/Met Office UPSCALE project involved one of the largest (if not largest) ever allocations of compute time for one project. The simulations were run on the German HERMIT supercomputer under the auspices of the European Partnership for Advanced Computing (PRACE), and the data were brought back to JASMIN for analysis. Support for the project included the provision of high performance disk and local computing. Disk usage peaked at 380 TB with an additional 200 TB of backup as over 250 TB was brought across the GEANT European research network from HERMIT to JASMIN – with a further 100 TB migrated onwards

to the Met Office. An UPSCALE scientist said, "We would never have been able to store, nor analyse, that volume of data, without the existence of the [JASMIN] service." With analysis of outputs and the number of collaborators increasing, the UPSCALE data will be relevant for years to come.

4. This year also saw the awarding of £7.4M additional funding for JASMIN phases 2 and 3. Phase 2 (online March 2014) will at least double the JASMIN storage capacity and give a six-fold increase in compute, enabling JASMIN to serve not only an expanded service to NCAS and NCEO but also a wider NERC community via a range of cloud services (e.g providing a cloud platform for bioinformatics, led by the NERC Centre for Ecology and Hydrology, CEH). Additional NERC capital funding was also allocated in both phases for specific components of the JASMIN system to be dedicated to the requirements of some nominated "Big Data projects". Along with the hardware, the two phases of expansion will also include significant investment in software.
5. This year has seen the first publication of a formal data paper for a dataset held in the CEDA archive (Callaghan et al, 2013, DOI: 10.1002/gdj3.2). This work formed part of the NERC Data Citation and Publication project, a cross NERC data centre initiative, aiming to develop a method of citation, peer-review and publication of datasets stored in NERC repositories. (This is of benefit to wide sectors of the NERC community as it provides 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.) Within the project, collaborations have continued with groups with common interests in data citation and publication.)
6. CEDA has deployed a new operational web service to harvest metadata from all the other NERC data centres, both to support the NERC data catalogue service, and the UK Location Portal (thus ensuring the NERC compliance with the EU "INSPIRE directive, http://inspire.ec.europa.eu/). This process involves the continual harvesting of metadata, ingestion into an underlying catalogue, provision of statistics to NERC, and the running of the client service seen by the remote portals.

Major Collaborations

In 2012/2013, 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, the atmospheric sciences community (via NCEO and NCAS) and the space weather community.

Additionally, CEDA is:

- Working closely with the other NERC centres, under the auspices of the implementation plan for the NERC Science Information Strategy.
- Operating and evolving the Earth System Grid Federation in partnership with the US Programme for Climate Model Diagnosis and Intercomparison and a range of global partners in support of the fifth Coupled Model Intercomparison Project (CMIP5).
- A leading partner in many major European projects including IS-ENES and IS-ENES 2 (developing an InfraStructure for a European Network for Earth system Simulation), and OpenAIREplus (delivering an open access infrastructure for scientific publications and data in Europe).
- Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.
- Working with the European Space Agency to extend earth observation metadata standards for climate data products

- Working with the Universtiy of Reading and other partners on the CHARMe project to develop systems for commentary metadata for users of climate data
- Working with commercial and academic partners within ISIC (the International Space Innovation Centre), on the facility for Climate and Environmental Monitoring from Space (CEMS), to support both academic research and opportunities for commercial applications and downstream services from EO and Climate data

Funding 2012/2013

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

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013
NCAS income	970	866	906	883	935
NCEO income	378	389	450	419	445
Other NERC income	788	481	341	527	287
Other income	461	710	1144	1099	1283
Total income	2597	2446	2841	2928	2950

Overall funding for CEDA for financial years 2008- 2009 to 2012-2013

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).

Activities and Highlights from 2012-2013

The following section provides a selection of descriptions of key activities and highlights from the year.

Data services

Sam Pepler

BADC and NEODC provide operational services to acquire, ingest, and catalogue appropriate data from a range of sources. The data are then delivered to users via the web, FTP and other services. Registration, user management and access control services are needed to ensure the right users have access to the right data. This suite of data centre services all need to be maintained and improved.

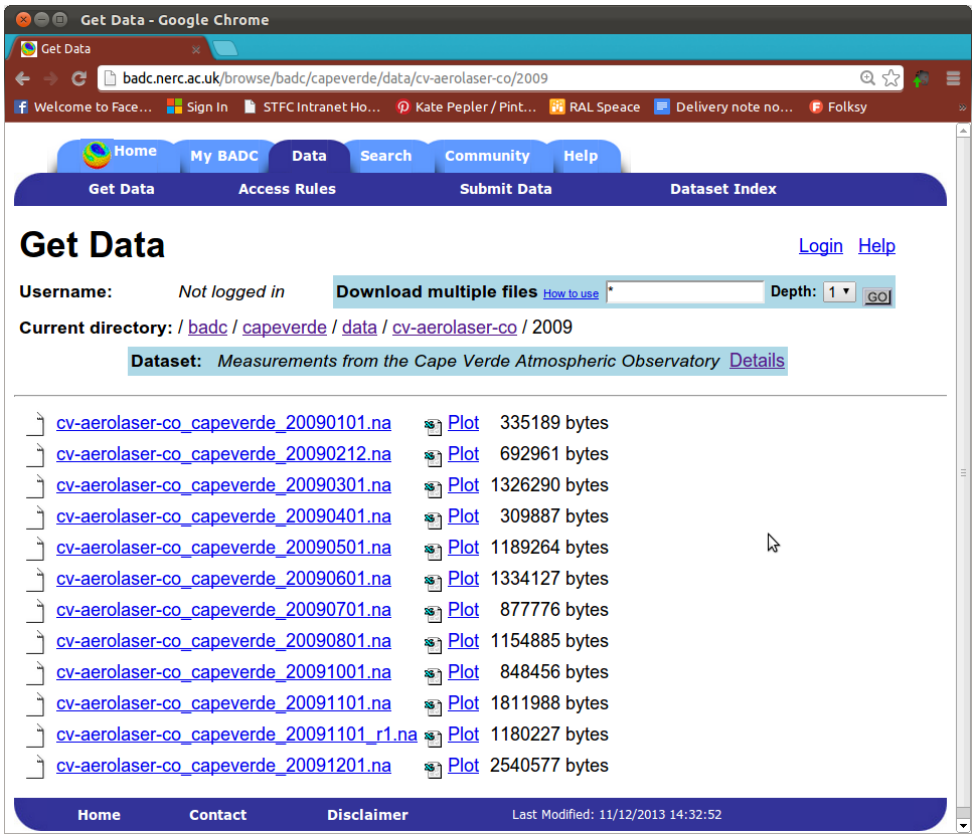


Figure 1: Archive access for the Cape Verde dataset via the web.

This year has been more challenging than most as new problems and glitches were introduced by the introduction of the new JASMIN hardware. The redeployment of services and new networking has required a lot of troubleshooting over the year, but this has resulted in improved services. In the future we are expecting more data, more users and new access methods. This is triggering the initiation of projects to improve the registration service and to deploy a better OpenDAP web service.

The archive is now directly accessible as a file system by users with login access on the JASMIN and CEMS cluster. While this enables users to run programs on large datasets, it was a challenge to put in place suitable access controls that interact with the existing user management and registration system.

Running a data centre requires many different services to interoperate and be robust. While this is difficult we have managed to serve our thousands of users consistently and are making sure we can cope with future use patterns.

Operational User Services

Graham Parton, Anabelle Guillory, Alison Waterfall, Sam Pepler

A key element of CEDA work is to ensure dedicated and timely support to both data providers and archive users. This support is delivered by 1) close liaison with data providers to smooth the process of data preparation and delivery to ensure the long term usability of the archived data, and 2) a comprehensive Helpdesk service which provides an easy access point to data expertise and technical support, by email and phone, for the over 3500 active archive users.

In 2012-13, the Helpdesk service received 1089 user queries concerning data availability (18%), account issues (17%), service issues (10%) and issues related to using the JASMIN/CEMS environments (15%). As well as the queries themselves, the helpdesk team also process applications for restricted resources. In the case of data, such restrictions protect the right of first use for the data providers themselves and, in the case of third party data, any of their license constraints. Services are restricted, either because of limited available resources or due to restrictions on underlying data. In 2012-13 CEDA processed over 2100 resource access applications, with an increasing number related to setting up user access to the new JASMIN/CEMS infrastructure. In particular, the advent of JASMIN meant more interaction with users granting access Group Work Spaces (GWS) and science processing machines.

As well as user support – either direct archive access users or processing environment users – data scientists support data providers helping them to prepare their data for deposit into CEDA archives. Many data providers are NERC grant holders, who are providing data in accordance with the NERC data policy, which requires all of its grant-holders, facilities and research centres, to actively plan the management of the data they create. The policy also requires data created from NERC funding to be openly accessible, so that the data can be used in the broadest of contexts. For activities in the CEDA disciplines, meeting this policy requires extensive work with the CEDA data scientists, who generally work over a period of time with data providers giving advice on archiving formats and metadata standards, ensuring the delivery route is set up and usable and capturing all the supporting material needed to construct data catalogue records about the data to be archived.

Inevitably this work needs planning, from the initial engagement, through to the delivery of data – sometimes over years. CEDA is leading a project in the Science Information Strategy programme to help implement such planning. A system has been developed to partition up all NERC grants to ensure that everyone is allocated to a NERC data centre so planning can be done consistently. The grants process has also been changed to make a simple data management plan mandatory when applying for a grant.

CEDA staff also liaise with third party data providers for access to data deemed to be of significant use to the wider research community, but to which access is often hard or restricted. Such data include Met Office observational data, ECMWF modelling data and satellite data, both as a long term archive and also in campaign support modes where near real time feeds are required. In those cases, CEDA staff themselves become remote data users, and CEDA data providers.

The data ingestion process which forms the “delivery route” forms one of the largest tasks within CEDA, beginning with data management plans, and followed up with software for data management, the conversations and information acquisition for documentation, and the eventual data acquistion. All this work is essential to ensure the utility of the data to end users, whether immediately within a field campaign or several years later.

The direct support to data providers and consumers forms the core CEDA mission: to curate and disseminate environmental data. Often systems work fine and the growing world wide user community are able to access the resources they need for their research. However, when things don't go smoothly, effective user support is crucial, and the CEDA dedication to this part of their job is often recognised by users.

The Joint Analysis System: JASMIN

Matt Pritchard

JASMIN is now the infrastructure on which CEDA provides services to the NCAS, NCEO and wider scientific communities. It provides CEDA with the capability to run not only its pre-existing data centres but also a range of new services including Group Workspaces, general purpose and specialist virtual machines, and the LOTUS processing cluster.

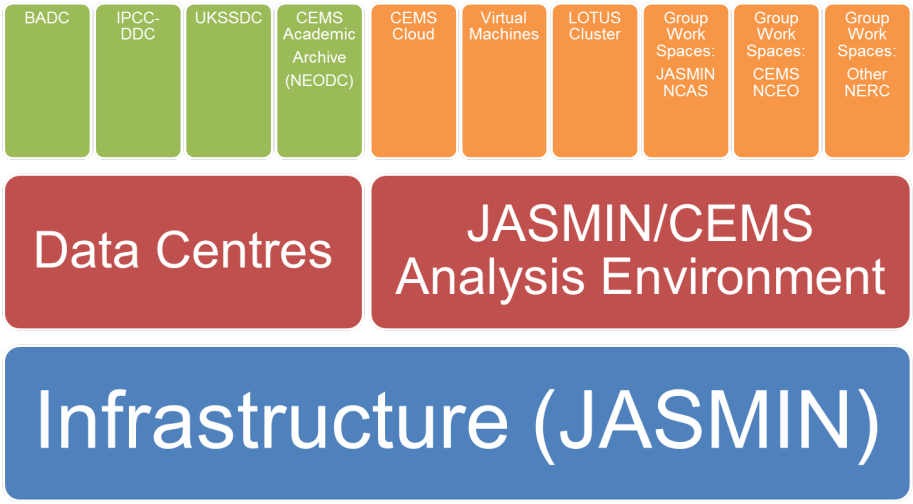


Figure 2: JASMIN infrastructure and CEDA services

Currently general users are provided access to the new facilities via login and science machines – JASMIN for NCAS users, and CEMS for NCEO users. Some communities have their own machines. A steady stream of existing and new users have requested accounts enabled for login access. The helpdesk has helped users through this process, as well as helping to grant access to Group Workspaces and virtual machines. A one-day NCAS research forum on JASMIN was held in September 2012[1] to bring together stakeholders and early adopters to share experience and learn about future plans. Despite relatively little publicity, the JASMIN/CEMS storage system was nearly full by the end of this reporting period and the LOTUS compute facility was often operating at capacity.

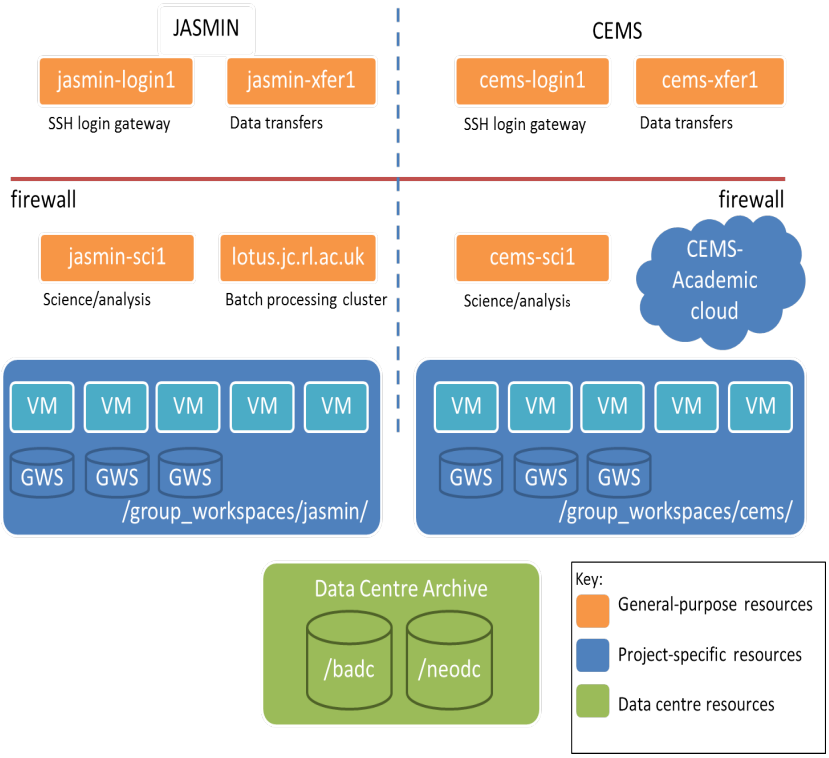


Figure 3: JASMIN and CEMS interfaces, showing general purpose, project-specific and data centre resources.

One of the reasons for relatively little publicity was the initial lack of resources for establishing quality documentation and training materials (an issue being addressed in the JASMIN phase 2 and 3 expansion). Nonetheless, help pages have been published on the CEDA and JASMIN websites, alongside technical advice on how to best utilise the massively parallel analysis environment.

[1] http://www.jasmin.ac.uk/stories/jasmin-forum-september-2012/,

Supporting EO Science with the facility for Climate and Environmental Monitoring from Space (CEMS)

Victoria Bennett

CEMS, the facility for Environmental Monitoring from Space, was established as a joint academic-industrial partnership to facilitate the exploitation of EO data in the commercial sector and to support the Earth Observation (EO) and climate research communities. Through significant government investment in 2011/2012 a joint infrastructure (hardware, software and services) was developed for CEMS on the Harwell site, hosted and operated at interconnected academic and commercial nodes. The academic half of CEMS is fully integrated with the JASMIN hardware infrastructure. CEMS opened for business during summer 2012 and has seen interest and uptake from a wide range of users across sectors, including the NCEO, NERC's National Centre for Earth Observation, science community.

CEMS provides a centralised resource for access to large volume EO and climate datasets, alongside computing infrastructure to process, manipulate and analyse the data. CEMS allows users and organisations to deploy their applications and services alongside the data, but also serves as a hosted processing facility. This facility is primarily being used to support the production of long term climate datasets, or ECVs (Essential Climate Variables), and for validation of EO data.

Feedback from users has shown that benefits to users are the ability to provide a central repository for large volumes of data, connected via a high-bandwidth network to compute nodes, minimising the overheads of data transfer. The multi-node features of CEMS enable rapid, parallel processing of separate observations. The system also supports better dissemination of results, including to potential new users outside academia.

EO science projects underway on CEMS projects include:

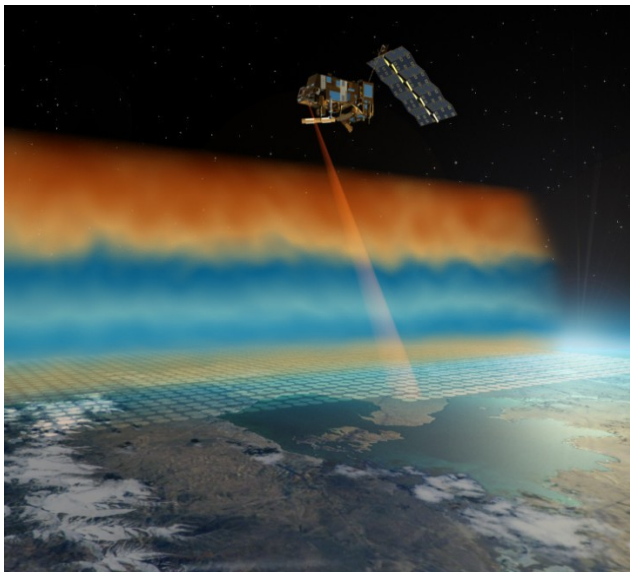


Figure 4: IASI data product profile (C) ESA-AOES Medialab. CEMS is being used to process data from the IASI instrument to produce atmospheric composition maps (Remote Sensing Group, RAL Space)

- Land Surface Temperature processing from AATSR (D.Ghent, Leicester)
- Fire impact processing using MODIS data (J. Gomez-Dans, UCL)
- GlobAlbedo: 100 TB land surface products dissemination and processing (J.P. Muller, MSSL/UCL)
- Sea Surface Temperature processing and data dissemination (C. Merchant, Reading)
- Cloud ECV, 120 TB data storage and access (C.Poulsen, RAL)
- Supporting earth quake strain processing and data storage (T. Wright, Leeds)
- Prototyping a commercially-developed hosted processing environment for diverse applications, including atmospheric composition processing (Magellium, Terradue, STFC)

Improving the ARSF data archiving experience

Wendy Garland

The NERC Airborne Research and Survey Facility (ARSF) aircraft records hyperspectral data, LiDAR measurements of ground height, and aerial photographs of specific survey sites and is an effective means of monitoring terrestrial, freshwater, marine and atmospheric environments. CEDA holds an archive of this data from 1981 to the present. Since 2006 the volume of data stored has risen sharply from <0.5TB to >15TB per year due to the inclusion of new instruments, increase in flight numbers, and the inclusion of raw data as a long-term backup. The previous process for acquiring, rearranging and archiving the ARSF data was not scalable, so it has been redesigned and brought in line with other CEDA aircraft datasets and now runs with minimal intervention. The timely availability of the new JASMIN-CEMS fast disk storage system at CEDA was an essential part of this success with ARSF data being the first to use this new system.

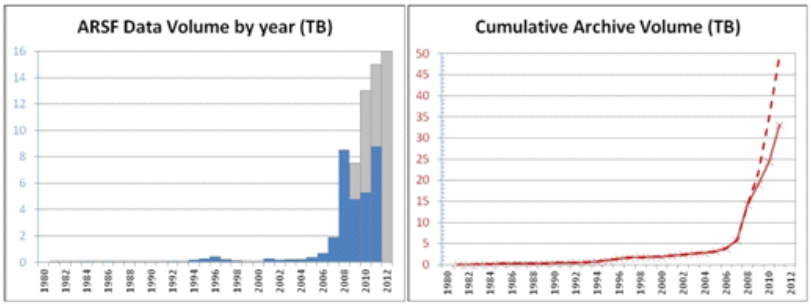


Figure 5: Left -Data archived per year 1981-present. Blue shows data actually in the NEODC archive, grey (2009-2012) outstanding data still expected. Right – cumulative volume in ARSF data archive 1981-present – solid line: actual volumes; dashed line: final expected volume.

The motivation to improve the ARSF ingestion process was the need to make the data available to users as swiftly as possible. The main issue was the size of the files and the time taken to manipulate them compounded by disk-space limitations: originally upload transfers took several days (longer if connections were interrupted), then many hours (sometimes days) to checksum, unpack and rearrange files ready for archive. The disk space available to perform these tasks was simply not large enough to unpack and manipulate the >500GB files; each stage was reliant on human interaction checking available space in order to proceed. A single flight's data often had to be managed piece-meal and recombined in the archive manually. The time taken to transfer, manipulate and archive these large files was becoming longer than the gap between flights- which was unsustainable.

The breakthrough came with the arrival of the new JASMIN storage system which came online in April 2012. This has allowed the whole process to be combined into a single Python script which, when initiated, obtains the flight data folder, identifies the processed data products and packages the remaining raw data into a single file for archival – these are then rearranged into the standard folder template. Useful information (e.g. location of flight and flight log) are extracted and inserted automatically and the correct access control is applied depending on the data type and project funder. Finally, the complete flight folder is deposited into the archive and the appropriate table of the flight data information is updated. This single scripted procedure removes the empty time waiting for a human to check and initiate the next stage, and eliminates human error and issues caused by disks filling up. The new more efficient process still takes several days to complete but email notifications are issued at each step facilitating tracking and freeing up the data scientist to other duties. On top of this, as processing disk space is no longer a bottleneck, several flights can be acquired and archived in parallel.

To keep pace with increasing volumes of data supplied to the ARSF archive, the procedure for acquiring, managing and archiving ARSF data has been updated and improved. Exploiting the extensive processing cache space of the new JASMIN-CEMS system, the new scripted process runs more efficiently with less intervention, thereby ensuring valuable ARSF data are made available to users in a timely manner.

Support for Metadata Standards and Conventions

Alison Pamment, Graham Parton and Stephen Pascoe

CEDA contributes to the development and maintenance of a number of internationally used metadata standards and conventions for the environmental sciences such as CF (Climate and Forecast metadata conventions), OGC (Open Geospatial Consortium) and INSPIRE (Infrastructure for Spatial Information in the European Community). Maintaining and exploiting these standards is crucial to operating an effective archive at scale. CEDA exploitation includes the development of its own software infrastructure – the "Metadata Objects Linking Environmental Science" (MOLES) catalogue.

The main CEDA contribution to the CF metadata conventions is through the maintenance of three controlled vocabularies: the CF standard name table (geophysical parameter names), the "area types" table (earth surface characteristics) and the "standardized region names" (a gazetteer). The standard name table is the largest, which by March 2013 contained almost 2,500 terms together with their definitions and the physical units used to measure each quantity. This year a total of 304 new terms were added and modifications were made to 100 existing terms, e.g. by clarifying the accompanying definition. The newly introduced terms allow labelling of parameters in the following scientific areas: greenhouse gas emissions according to the IPCC (Intergovernmental Panel on Climate Change) 2006 categorization scheme; atmospheric abundances of ozone and land surface characteristics for the ESA CCI (European Space Agency Climate Change Initiative) programme; a number of smaller modelling projects requiring diverse quantities such as cloud cover categorized by cloud type and sea ice classification by age and type. In January 2013 a schedule of monthly updates to the standard name table was adopted, significantly decreasing the time between formal agreement of vocabulary terms and their publication. The move to monthly updates was made possible by the development of a new CEDA vocabulary editor which has streamlined the process of preparing terms for publication.

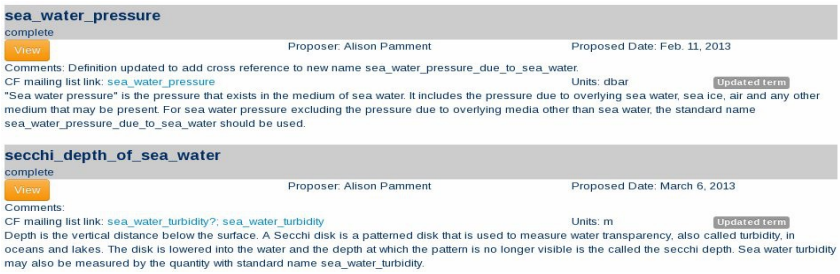


Figure 6: A screen shot taken from the second generation CEDA vocabulary editor showing two proposals that are now published in the CF standard name table.

CEDA staff also lead a Data Standards Working Group (DSWG) for the ESA Climate Change Initiative (CCI) in order to develop agreed standards for data formats and metadata. The CCI aims to produce high quality Essential Climate Variables from Earth observation satellite data across a wide range of scientific areas with final data products as accessible and useful as possible to a wide range of end users. It is currently reaching the end of Phase 1, and preparing for Phase 2, in which long-term global scale data products will be generated based on the development and specification work carried out during the past 3 years. Standards relating to file format and metadata (CF-netCDF), as well as filename and metadata attributes have been formally agreed in the DSWG. These standards have been trialled in Phase 1 and will become requirements in Phase 2 of the CCI programme. The wide variety of data products that will result from the CCI are both a challenge to the DSWG and the reason that its work is so important to maximising the usefulness of the resulting Essential Climate Variables.

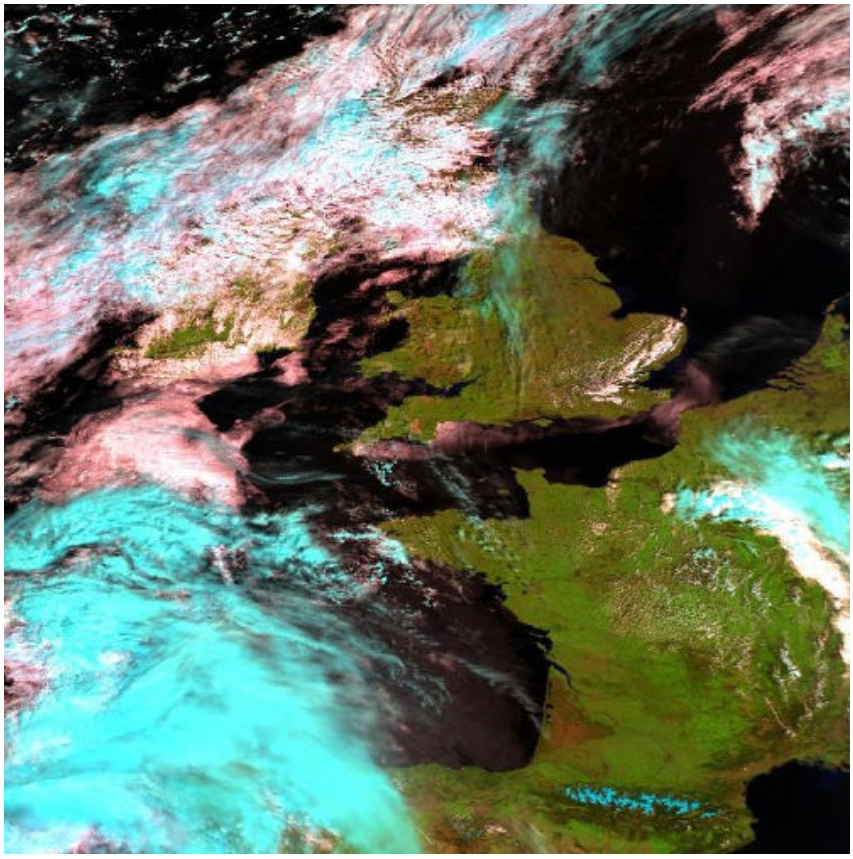
Since September 2012 CEDA has been preparing for the release of the V3.4 MOLES catalogue. This INSPIRE compliant catalogue utilises controlled vocabularies where possible and will permit CEDA to fully curate its datasets by collating dataset parameter details and linking these with controlled lists such as the CF standard names. The bulk of the work has been to migrate and extend existing information from the present (MOLES V2) catalogue, which has been a highly complex and staff intensive task, but which is now nearing completion. The new catalogue should significantly improve user experiences.

NCEO Data Management and Community Support

Alison Waterfall, Victoria Bennett

CEDA scientists are providing data management services and support to the UK National Centre for Earth Observation (NCEO).

The NCEO is a partnership of over 100 scientists from 26 institutions whose mission is to unlock the full potential of Earth observation data. NCEO currently consists of seven scientific themes, and a Data Services theme. CEDA leads the Data Services theme, which provides underpinning support to the other scientific themes. This support consists of both management and archival of the data products created by the research done within NCEO, alongside helping facilitate access to third party datasets, such as those from ESA or EUMETSAT in order to support NCEO research. This also includes providing the academic component of CEMS to enable the research community to access and manipulate high-volume Earth observation data directly, as well as to support the link to the commercial components of CEMS and to its partners, customers and end-users.



The UK and Europe from AVHRR-3 on METOP-A on the 6th of May, 2013. CEDA acquires AVHRR-3 and (other METOP-A instruments) IASI and GOME-2 to facilitate their use by NCEO (and other UK) scientists.

In the year 2012-2013, CEDA (via the NEODC) has continued to acquire data to support NCEO, including acquiring and archiving data from recent flights of the ARSF (Airborne Research and Survey Facility), for which we currently hold ~40Tb of data. New data from the METOP-A satellite (GOME-2, IASI and AVHRR) has continued to be acquired, and is made use of by NCEO scientists, as this provides a simpler access route than going directly via EUMETSAT and can reduce the need for them to store large volumes(currently 85Tb) of this data on their own systems. Similarly CEDA provides access to a selection of data from the ESA ENVISAT satellite (currently ~100Tb), and over the course of this year gaps in our holdings of ESA MIPAS data have been filled to complete the latest reprocessing version (V5) data that we hold. An expert group has been set up to provide advice on future priorities for data acquisition for NCEO and academic CEMS. This identified the provision of MERIS data as an initial priority.

CEDA scientists have attended meetings and the 2012 NCEO annual conference, liaising with scientists and tracking the expected data outputs of the NCEO. CEDA also provides helpdesk support services to the NCEO community; this can cover a range of topics from questions about datasets, advice on data formats, visualisation and use of the data, and also support to academic users accessing CEMS for EO data analysis, manipulation and processing.

UK Met Office Liaisons

Ag Stephens, Bryan Lawrence

The Met Office is a major participant within atmospheric research in the UK. Collaborations and interactions between NERC, NCAS and the Met Office occur on a number of levels, from the high-profile NERC-Met Office Joint Weather and Climate Research Programme (JWCRP) to collaborations between individual scientists. CEDA supports these interactions through the secondment of Ag Stephens (CEDA Head of Partnership) to the Met Office headquarters in Exeter. This enables a great deal of essential technical communication as well as fostering new research opportunities.

A major JWCRP investment has been the joint NERC/Met Office supercomputing facility **MONSooN** which is located at the Met Office in Exeter. As well as aiding in the project oversight via membership of the MONSooN Management and joint Information Infrastructure groups (see Table 1), CEDA provides direct support for MONSooN projects with dedicated JASMIN facilities which include virtual machines and group work spaces.

The UK contribution to the 5th Coupled Model Intercomparison Project (**CMIP5**) has included both data from the Met Office, and infrastructure delivered by CEDA. CEDA runs the UK "Data Node" of the CMIP5 delivery system – known as the Earth System Grid Federation (ESGF) – and both NERC and the Met Office have provided data for that node. The data provision from the Met Office, involved the migration and ingestion of around 50 terabytes (TB) of climate model simulations. To that endC EDA developed transfer, checking and archival tools to manage the workflow.

The **UPSCALE** project, a high-resolution climate modelling collaboration between the Met Office and NCAS Climate, also required high-performance disk, high-speed networks for efficient data transfer, large storage capacity and processing systems. The JASMIN platform was able to deliver these in a timely and flexible manner to enable the project scientists to achieve their objectives.

CEDA has also led a project to run the Met Office Atmospheric Dispersion Model (NAME) on the JASMIN platform, delivering NCAS users with an up-to-date supported version of the model. The use of common platform makes the Met Office support effort more efficient, and allows all parties to collaborate on commons services, such as providing a web-based trajectory service built on top of NAME.

Committee	CEDA Role
Climate Database User Group (Met Office)	Advising on the UK academic requirement for access to long-term meteorological records.
NWP Archive Management Group (Met Office)	Advising on the UK academic requirement for access to an archive of Met Office Weather (NWP) model outputs, including recent forecasts and historical data.
MONSooN Management Group (JWCRP)	Contributing to discussions on the effective management of the joint Met Office - NERC high performance computing facility (MONSooN). Advising on usage and interaction with the JASMIN platform.
Joint Facilities Group (JWCRP)	Representing NERC interests in relation to climate modelling strategy (including the UK Earth System Model), data and computing issues.
Information Infrastructure Group (JWCRP)	Co-chair and secretary of the group; facilitating collaborative working between NERC and the Met Office in relation to data descriptions, transfer, storage, processing platforms and software.

Table 1. Committees attended by CEDA in relation to Met Office interactions

Other activities include a range of other committee engagements (Table 1), and the provision of resources for the regional climate (PRECIS) team to run on JASMIN. CEDA also manage numerous operational data streams that are regularly replicated from the Met Office to the NCAS BADC archive, including the historical UK and Global MIDAS observations and the Global and North Atlantic European numerical weather prediction outputs.

In the current paradigm of expanding computational requirements and tightened budgets it can expected that there will be a continued focus on collaboration. In the next year CEDA expects to improve the integration between MONSooN and JASMIN so that scientists can manage their work across platforms efficiently and quickly.

Moving a Petabyte of precious archive data without anyone noticing

Matt Pritchard

The Challenge

The introduction of high-performance, parallel-access storage presented a unique opportunity to improve the way CEDA archive data can be accessed, but also a huge challenge: that of moving over a petabyte (PB) of valuable science data (some irreplaceable) from legacy storage (on network attached storage, NAS) to the new (Panasas) storage system. The legacy storage held datasets (some fixed in size, some still growing on a daily basis) on an organically-grown plethora of filesystems of between 10 and 50 terabytes (TB), with filesystem boundaries not necessarily aligned with meaningful divisions of data. CEDA had set itself the target of achieving the data migration within a calendar year. In fact, the bulk of it was achieved in half that time, with most CEDA users not even aware that the change had taken place.

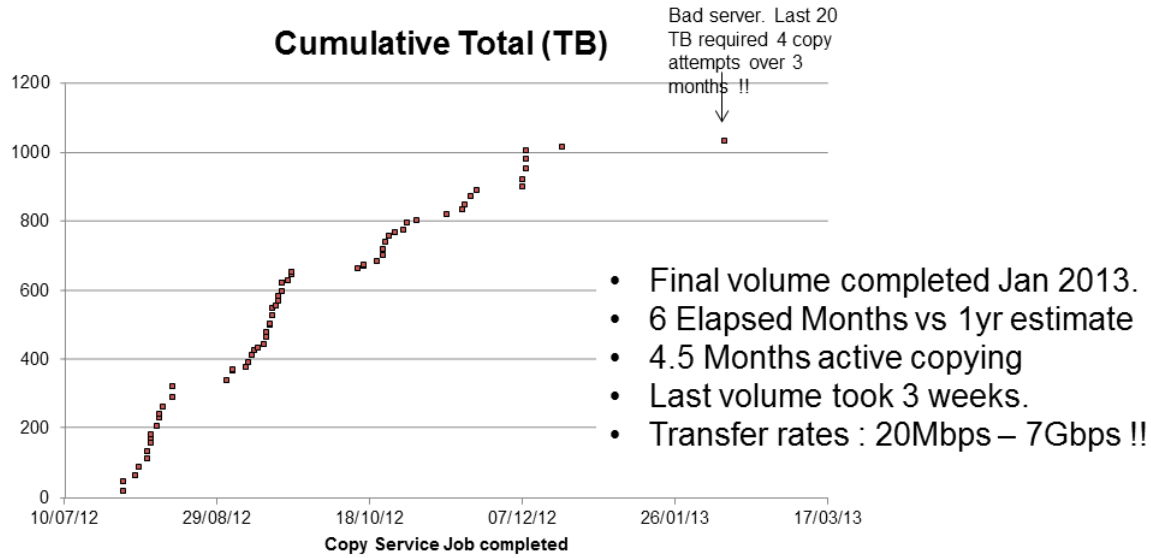


Figure 7: Progress with migration of CEDA archive data from legacy NAS to JASMIN storage.

Parallelisation

Fortunately, the JASMIN infrastructure itself provided tools which would make the job run efficiently. By setting up dedicated virtual machines to act as "worker nodes", several such nodes could be used to make an initial bulk copy of archive data filesystems from legacy storage and to write to the new storage. Parallelising the copy task using the Panasas *pan_pcopy* tool took advantage of the parallel-IO capability of the new storage and achieved overall transfer rates of up to 7 Gbps, unthinkable for previous NAS-to-NAS migrations. However, some legacy hardware proved more difficult – with seemingly unsolvable IO problems holding up transfers (a good reason to remove them from front-line data serving duties).

Several phases of activity were required. The initial copy simply created a new copy of the data on new storage. At this stage, a further "rsync" operation was carried out to "mop up" remaining data (for example, files that had been added to a dataset since the initial copy). At this stage, the archive configuration was modified so that this copy became the primary copy, so that in the case of growing datasets, new data would be written to the new copy rather than the old. One final rsync operation provided additional safety by capturing any missed data. Meanwhile, a near-line tape backup of the old copy was used to perform a bitwise comparison of the dataset at a low level to detect any instances of data corruption. Any issues were highlighted and investigated manually. Only once all these steps had

been completed for each particular dataset, and for all partitions on each NAS server, would a server be disconnected from the network. To date, a total of 30 legacy NAS servers have been disconnected, with final data wiping and power down to follow. Some will be re-tasked as secondary online storage while the majority (particularly those beyond their warranty period) will be decommissioned for disposal.

The new storage environment

The Panasas storage hardware consisted initially of 103 shelves each with 11 blade units, with multiple shelves managed together as "blade sets". The Panasas storage management interface allows the administrator to define which shelves belong to each blade set. Typically, one director blade is present in each blade set. Most shelves therefore contain 11 storage blades, but those with a director blade consist of 10 storage blades and one director blade. Each PAS11 blade contains two 3 Tb disks. Initially, most blade sets were configured, after installation and initial testing, as blade sets of approximately 500 Tb for further "soak testing".

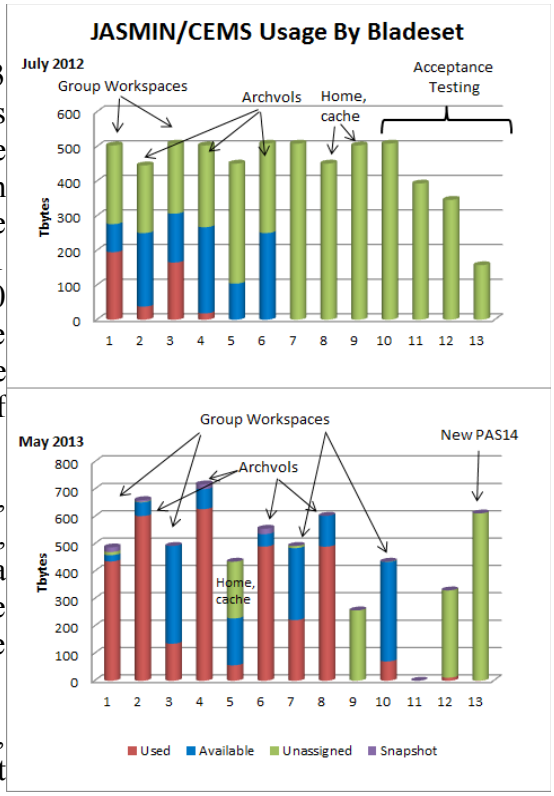


Figure 8: Panasas bladeset configuration in July 2012 and May 2013. Some bladesets (2, 4) have increased in size by "cannibalising" others (9,11).

Although it is technically possible to configure Panasas as one, vast filesystem, this was not desirable in this case. Indeed, from data management and security perspectives, there was a requirement to separate volumes used for long term archive storage from those used for "Group Workspaces" (online workspaces for scientific collaboration).

The task of dividing up the new storage pool (or, conversely, of constructing blade sets) was initially done using a default blade set size of 500 Tb raw size, approximately 400 Tb usable. This represented a compromise between the following factors:

- A blade set should be large enough to deliver the performance benefit of using a large number of blades (since each blade contributes its IO to the effective "data bandwidth" of the system)
- Since volumes cannot span blade sets, a blade set should be large enough to accommodate the largest anticipated volume size.
- Blade sets can be expanded by adding additional shelves, but not decreased in size (without first draining all data from them).
- Volumes (and hence blade sets) should not be so large that, in the event of a blade set failure, reconstruction of blade sets and restoration of data from backup would take prohibitively long. The behaviour and reliability of very large (>800 Tb) Panasas blade sets is not well known.

Clearly, there is an element of guesswork in setting blade set sizes at the outset, however the management functions of Panasas make it relatively easy to adjust these during their lifetime. Figure 5 shows how the usage of CEDA's Panasas bladesets for JASMIN/CEMS storage has changed over the first year of operations. This sort of configuration change would be much harder to achieve with conventional RAID storage.

Providing the NCAS community with a collaborative environment for research.

Matt Pritchard

In its first full year of JASMIN operations, CEDA has brought to the NCAS community a unique scientific analysis environment, working with scientists to provide resources that enable them to manipulate and analyse their own data on systems co-located with the CEDA petascale archive.

Central to the JASMIN analysis environment is the concept of the Group Workspace (GWS), an evolution of the "Project Space" that BADC had offered to its user community for several years. Whereas the project spaces enabled the limited but valuable functionality of being able to upload modest volumes of data via FTP to share with project peers, GWS provide additional capability which many projects are already finding invaluable. Such GWS are created as volumes within the same storage infrastructure as CEDA archives, although archive data is stored on distinct, though identical, hardware.

In particular, GWS provide:

- High volume online workspace
  - High-performance, parallel-IO storage for between a few gigabytes (GB) to tens of terabytes (TB) can be provisioned rapidly in response to user needs (dependent on available disk space)
- Access control
  - A designated member of the project is given the role of GWS manager, with responsibility for approving requests for access to the workspace. Users granted access are joined to a common UNIX group for the GWS, and (under the direction of the GWS manager) encouraged to follow agreed group file and directory permission policies which enable secure data sharing within the group.
- High bandwidth input and output data transfer
  - Shared data transfer servers are available for upload/download of data to/from the GWS. Login is via ssh public/private key-pair, enabling secure transfers using rsync-over-ssh or scp, and parallel data transfer using bbcp.
- Data analysis
  - A shared data analysis environment is available on jasmin-sci1.ceda.ac.uk, with the same access and login policies as the data transfer server. A standard analysis platform is being developed as a Linux distribution, with requests for additional software packages considered.
  - Dedicated virtual machines for data analysis are available on request, subject to resources, and enable projects to construct their own analysis environments, starting from the standard distribution but using specific or license-restricted software.
- "Elastic Tape" near-line storage
  - Currently under development is provision for GWS managers to optimise use of valuable high-performance disk by managing data resources across disk and tape storage. GWS managers will be able to copy or move data from disk to near-line storage for retrieval at a later date. Due to the diverse usage patterns, large volumes of data involved and demands on resources, it is not feasible for CEDA to manage backups of GWS data. Rather, GWS managers will be provided with the tools to enable them to manage their own data .

Helping the NCEO/CEMS community address Big Data challenges.

Matt Pritchard

The JASMIN infrastructure is also used to provide the NCEO/CEMS Academic community with the CEMS-Academic services which have already proved their worth in improving efficiency in some of the large data processing and analysis tasks in the EO domain.

The Impact of CEMS-Academic

Example 1: University of Edinburgh ARC (A)ATSR processing: "CEMS is now central to the way I am now planning to do future projects involving big Earth observation datasets – particularly looking forward to exploiting Sentinel era data for climate science and services. It is an excellent platform for co-operation across the UK which will certainly foster ambitious projects with international impact." Chris Merchant – Science Leader, ESA Climate Change Initiative Sea Surface Temperature

*Example 2: RAL Space Remote Sensing Group Cloud ECV Processing: "*Processing that previously took 2-3 weeks takes 3 days, CEMS + SCARF[2] is a game changer".

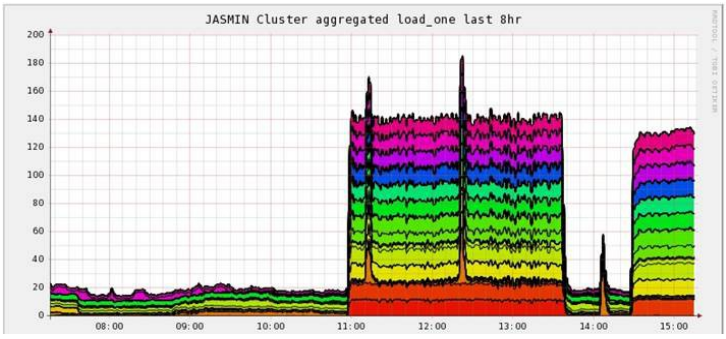


Figure 9: (A)ATSR full mission processing: 140-185 jobs in parallel with no IO issues.

Example 3: RAL Space (A)ATSR Full Mission Reprocessing: The (A)ATSR Flight Operations & Archive Support activity undertaken by RALSpace, is helping to create archive-quality (A)ATSR data products for contribution to a 20-year time series of sea-surface temperature observations from the (A)ATSR instruments. Periodically the algorithms and calibration coefficients are revised and the (A)ATSR science community schedules a complete reprocessing to create new archive data products incorporating these improvements. This group was an early adopter of the infrastructure.

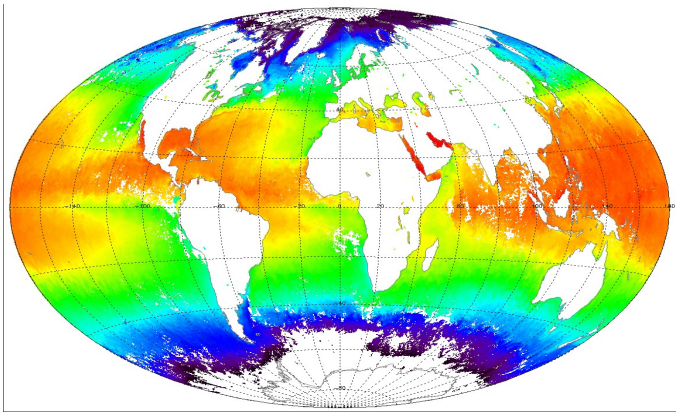


Figure 10: (A)ATSR Full mission processing, typical end result: a monthly averaged Sea Surface Temperature product from ATSR2, August 1996, from the recent (A)ATSR reprocessing on CEMS.

"For the last reprocessing work six years ago it took approximately 3 days for 1 month's worth of data. Now it completes in 12 minutes" Being able to do this on shorter timescales is not only more efficient but makes the whole process of algorithm development and implementation easier: some problems with a revised processor are only apparent when the entire data set is available to inspect after a trial run, and a complete re-run is now not prohibitively expensive. CEMS initially provided a trial virtual machine on which the processor could be tested, before scaling out to use a suite of 12-core Dell R610s each with 48 Gb RAM, enabling over 130 jobs to run simultaneously under the control of the LSF job scheduler. A 10 Tb GWS was provided for output, enabling the generation and storage of up to 22 months of Level 1B and L2 data. With this configuration, a month's L1B processing in 11-12 minutes. The largest run to date has been 17 months, processed to Level 1B in 2hrs 40 mins. UPSCALE: A case-study for supporting the UK atmospheric science community

[2] SCARF is an STFC computing cluster co-located with CEMS, now with read access to CEDA archive data.

UPSCALE: A case-study for supporting the UK atmospheric science community

Ag Stephens

By obtaining over 100 million core-hours on a German Supercomputer, the UPSCALE[3] Project became one of the largest simulation campaigns of recent times[4]. It demonstrated how scientific expertise, coupled with the appropriate technical capability, could achieve modelling, archival and analysis not previously feasible at this scale. The recent JASMIN data and compute platform at CEDA played in a crucial role in enabling the scientists to manage the project effectively.

A collaboration between the UK Met Office Hadley Centre and the NCAS Climate group at Reading developed an ambitious project plan to bring together the high-resolution of operational weather forecasting with the long time-scales of climate models. The capability to do this is only now being realised due to the enormous supercomputing resource required to complete the runs and the transfer, storage and processing capability required for further analysis. The proposal became a reality when 144 million core-hours of supercomputing time were granted by the Partnership for Advanced Computing in Europe (PRACE). This meant porting the code to a German Cray XE6 machine, HERMIT, and generating over 250 Terabytes (TB) of output within a single year.

UPSCALE had a plan for data generation but disk quotas on HERMIT were tight, and time-bound. The project therefore needed a means of efficiently transferring data to the UK, a disk large enough to house the data and computing resource to allow post-processing of the data. At the same time, JASMIN[5], a data and compute platform managed by CEDA, was being commissioned. Funded primarily to support UK atmospheric research collaborations, JASMIN consists of high-speed networking, 4.6 Petabytes of high-performance disk, a range of virtual machines and a processing cluster. UPSCALE became the first major user of the JASMIN hardware, and produced 330TB in 2012 alone. The data were brought back to JASMIN using GridFTP at a rate which varied between 1 and 10 TB per day. The UPSCALE datasets online peaked at around 600TB and is now around 380TB.

Figure 11. An array of PANASAS high-performance disk volumes in the JASMIN computing hall. In total there are 103 shelves, each housing 11 “Storage Blades” .

CEDA set up two virtual machines (each 8 CPUs and 16 GB RAM) to support Met Office and NCAS Climate scientists analysing the outputs. Output generated by the post-processing was written to the allocated UPSCALE “Group workspace” accessible by all collaborators; users of the high-performance disk reported speed-ups from running equivalent analysis on other systems. A wiki was also created to support communication between contributors.

The UPSCALE project showcases JASMIN's potential for providing flexible and scalable solutions to support collaboration within UK atmospheric science. The data are expected to provide a hugely valuable resource for the study of current and future climate complementing previous simulations at coarser resolutions. With increasing emphasis on collaboration between major institutions and NCAS, and the expanding requirements for supercomputing, the role of JASMIN as a hub for data transfer, sharing, analysis and inter-comparison is expected to increase significantly.

[3] UPSCALE: “Unified model on PRACE: weather resolving Simulations of Climate for globAL Environmental risk”.

[4] M.S. Mizielinski, *et al*. 2013. *High resolution climate modelling; the UPSCALE project, a large simulation campaign.* Geosci. Model Dev. Discuss., 7, 563-591, doi:10.5194/gmdd-7-563-2014, 2014.

[5] B. N. Lawrence, *et al*. (2012). *The JASMIN super-data-cluster*. http://arxiv.org/abs/1204.3553

Delivering the latest UK Climate Projections through a web-interface

Ag Stephens

The UK Climate Projections (UKCP09[6]) provide projections for the future of key land and marine quantities over a variety of temporal and spatial scales. The main product is a 25km gridded UK Met Office dataset available for three emissions scenarios for seven 30-year periods up to the 2080s. The probabilistic nature of the data adds a novel and complex component because each dataset is provided as a series of plausible projections, typically expressed as a probability density function. UKCP09 also includes daily and hourly time-series of future weather simulated for a given location and climate scenario, providing potentially billions of unique data products to the user.

CEDA developed and runs a specialised web-based toolkit to enable users to interrogate the data for specific parameters in their area of interest. The user is guided through a set of web pages that provide expert guidance on how to utilise and interpret the climate projections. Custom-built plotting tools allow on-the-fly generation of plots visualising possible future climates (figure 11). Users can also submit large processing jobs capable of producing Gigabytes of output from the Weather Generator.

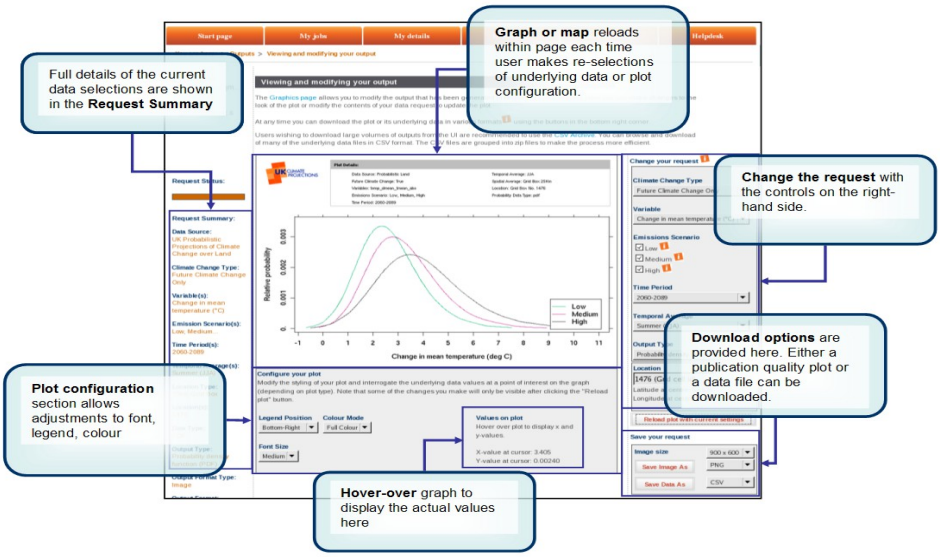


Figure 12. Screen shot of the UKCP09 web-interface developed at CEDA. The plot shows three probability density functions (PDFs) representing 3 possible future climate scenarios.

Over 8,500 registered users representing a range of economic sectors have accessed the user interface. The largest representations for a single sector are in flood management and coastal issues (1107), water resources (849), planning (including spatial and sustainable development) (720), biodiversity and nature conservation (640), buildings (533) and energy (507).

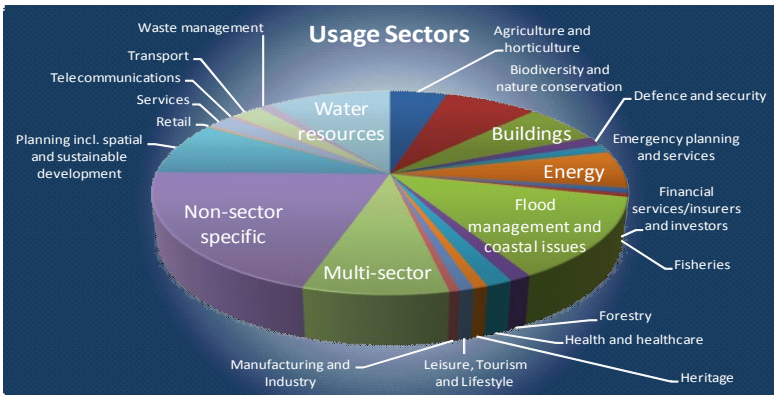


Figure 13. Breakdown of UKCP09 users.

The 2012 Environment Agency review of the existing UKCP09 web delivery systems concluded that the most cost-effective approach was to maintain the existing system - requiring ongoing CEDA management of both the hardware and software, user environments and support. An additional task was the extensive pre-launch checking of a new version of the main projections dataset from the Met Office. The next major task is the migration of the entire system on to CEDA's new JASMIN compute and data platform.

[6] Murphy, J. M., Sexton, D. M. H., Jenkins, G. J., Booth, B. B. B., Brown, C. C., Clark, R. T., Collins, M., Harris, G. R., Kendon, E. J., Betts, R. A., Brown, S. J., Humphrey, K. A., McCarthy, M. P., McDonald, R. E., Stephens, A., Wallace, C., Warren, R., Wilby, R., Wood, R. A. (2009), UK Climate Projections Science Report: Climate change projections. Met Office Hadley Centre, Exeter.

The NERC Data Catalogue Service and the MEDIN Discovery Service

Steve Donegan

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

CEDA supports both these services by running the metadata harvesting, ingestion into the catalogue and the operation of the underlying discovery web service which itself is the interface which portals query for material to provide a response for portal users. Both NERC and MEDIN portals are operated by the BODC.

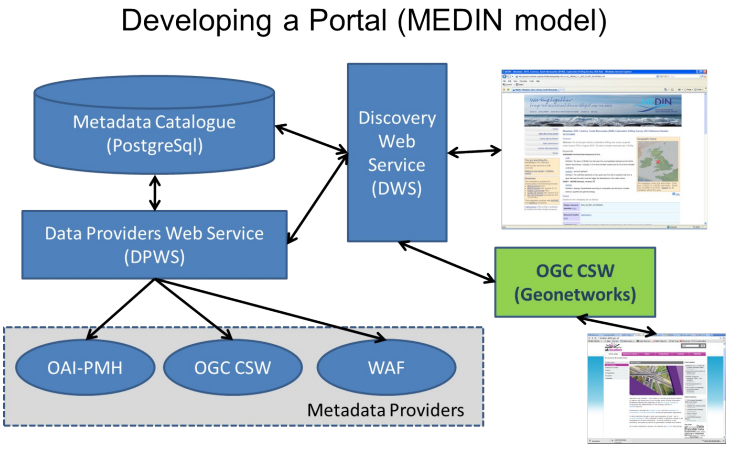


Figure 14: Operation of a bespoke MEDIN discovery web service and catalogue with an additional "OGC" catalogue web service for publishing to data.gov.uk

In addition to the discovery web service and its related backend CEDA provides the data providers web service which allows registered data providers to initiate harvests and ingests into the underlying catalogues for both services.

The data provider web service for MEDIN has been incorporated into a program of upgrades to the MEDIN contract in order to support an improved portal. These upgrades include extra metadata fields harvested from the metadata and made searchable within the catalogue as well as providing a dedicated Open Geospatial Consortium (OGC) catalogue service for the web (CSW). The MEDIN CSW allows records from the catalogue to be published to data.gov.uk so that MEDIN and its providers may meet their legal obligations for the European Commission's INSPIRE directive.

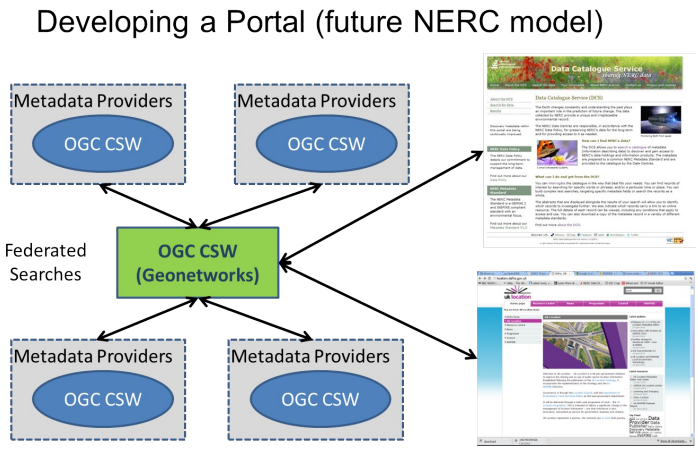


Figure 15: Operation of a single catalogue web service as a "master" node used for publishing to data.gov.uk as well as supporting the NERC catalogue services.

CEDA has also been working with other data centre colleagues within NERC to develop an integrated pan-NERC data catalogue architecture which is compliant with the OGC-CSW tools. Implementing this new system will replace the bespoke systems which have been used up to now.

A new "Geonetworks CSW" will be run by CEDA on behalf of all of NERC, and like it's bespoke predecessors, will harvest metadata from all the NERC data providers in order to provide both an improved portal and a single publishing point for all NERC metadata to got data.gov.uk (and hence meet the NERC INSPIRE obligations).

ESPAS – Near Earth Space Data Infrastructure for e-Science

Sarah James, Spiros Ventouras, Matthew Wild

ESPAS will create a common interface to near Earth space data from ground and space-based instruments, and from distributed data providers. It is an EU Framework 7 project, bringing together a consortium of 22 partners to produce in 30 months an e-infrastructure to allow a more effective and detailed study of the near Earth space environment. RALSpace[7] has the role of Project Coordinator and CEDA is leading the interoperability definition and is involved in the operation and maintenance of the ESPAS system.

The near Earth space environment includes a number of distinct physical domains that surround our planet: the thermosphere, ionosphere, plasmasphere, radiation belts, magnetosphere and near Earth solar wind. Variations in these environments, known as Space Weather, effect many technological systems critical to modern society. Space weather events pose risks to systems including power grids, location and timing systems, satellite communications and civil aviation. The ESPAS platform will serve a range of science and engineering communities that study the near Earth environment: scientists using it as a natural plasma laboratory and seeking to improve our understanding of space weather; and engineering groups seeking to design and test systems working in the near Earth environment.

The ESPAS project started on November 1st 2011, so in April 2012 it was entering its sixth month. During the year since then, the project has produced a "release 1" system capable of searching meta-data and demonstrated this at the first ESPAS High Profile Annual Meeting, held in October in Tromsø, Norway and in the associated CDAW (Coordinated Data Analysis Workshop). ESPAS was subject to its first annual review in December 2012 at the European Commission in Brussels. The first annual review was successful, endorsing the progress to date.

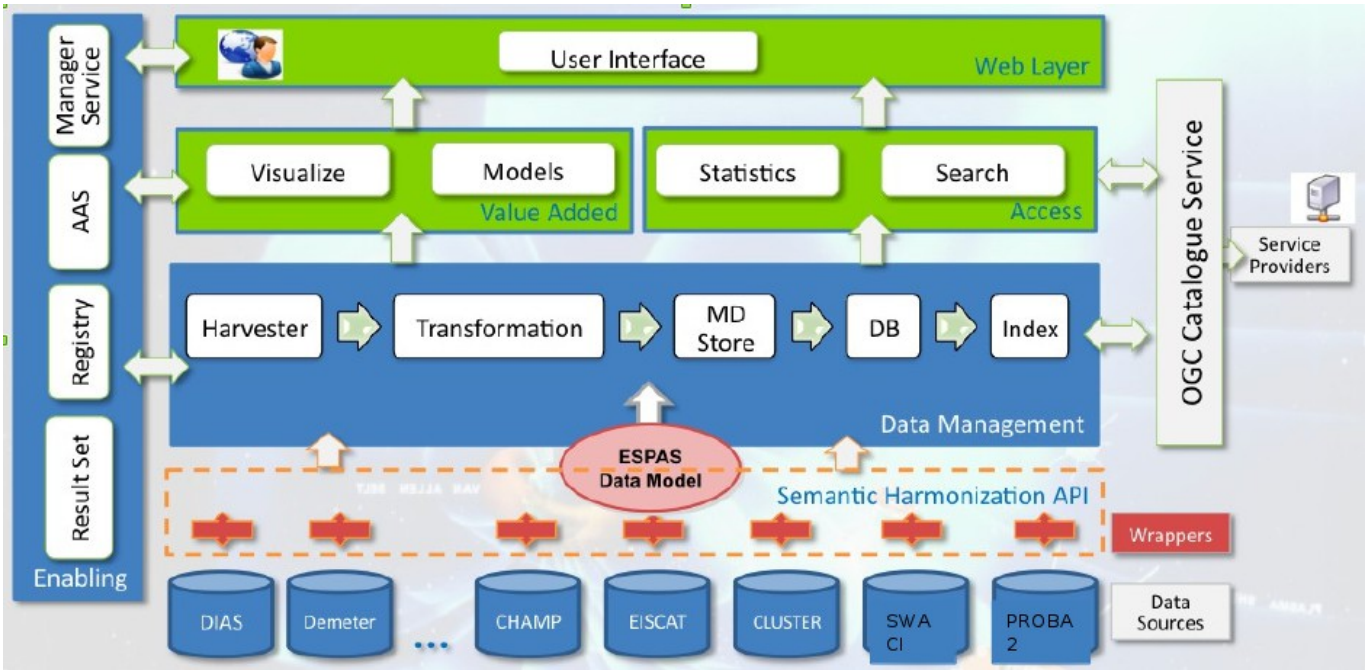


Figure 16: ESPAS Architecture Overview

CEDA has led the work on developing a data model for ESPAS which has been fundamental to the design and implementation of the system. The initial description of the common ESPAS data model was highlighted in the annual review as one of the main achievements of the project in the first year. Planning the move to operations at RAL is now underway. The ESPAS system will be hosted at RAL, alongside the UK Solar System Data Centre (UKSSDC) which already has a great deal of experience in this area.

ESPAS is progressing well towards the creation of a near Earth space data infrastructure for e-science, with CEDA expertise playing a key role in defining the underpinning data model and in providing data centre specialists in near Earth environment data.

[7]RALSpace is the department within STFC that hosts CEDA.

CMIP5: A first petascale dataset for CEDA

Martin Juckes and Stephen Pascoe

CEDA was a key player in the distribution of the data from the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project, Phase 5 (CMIP5), providing one of the data nodes of the Earth System Grid Federation (ESGF). ESGF provides data a distributed archive which is all available through a single search interface, but data transfer rates can be very slow for intercontinental transfers. To provide easier access for UK scientists CEDA has created a petascale copy of the core of the archive. The combination of direct access and a petascale archive of the latest climate projections from all the major climate modelling centres, provides the UK research community with a unique facility.

CEDA has worked closely with the UK Met Office Hadley Centre to ensure that their data was delivered to users efficiently. The UK data was the first data published into the archive in 2010.

The "ESGF" software used to run the distributed archive (also called ESGF) has been evolving as the archive grows. PCMDI have led the development project, but CEDA has supported this process by contributing code, deploying early versions for evaluation and leading discussions on requirements. The ESGF software provides a central search interface for data held at distributed sites. Users can access data anywhere in the federation using a single user identity. Deploying the system turned out to be time consuming, and many centres were not able to run the system cleanly, resulting in a number of anomalies in the published catalogues. Such anomalies cause delays when users try to download the data. The transfer of data to CEDA has taken 2 years, and the issues limiting the rate of transfer have varied over time. The limiting speed of long-distance transfers has been a factor throughout. In mid-2011 there were many problems with access control and stability of data nodes run by the data providers. In late 2011 and early 2012 congestion on CEDA servers became a problem as the previous generation of archive disks filled up. In mid-2012 the transition to the new JASMIN facility at CEDA greatly improved our ability to manage these large data volumes. In late 2012 a completely new version of the archive software resulted in an improvement in transfer speeds from some centres.

All transfers are verified by comparing the checksum of the file which arrives at CEDA with that in the archive catalogue. In some cases the value in the archive catalogue is incorrect, and this can be verified by repeating the file transfer. Most centres will correct inaccurate catalogue entries once notified, but some are evidently lacking resources to deal with such issues. In some cases files have been retracted by a centre and removed from their local disk store without clearing up the catalogue entry. In this case the user simply sees an error message. There are also cases where datasets contain duplicate files or files with overlapping time segments. CEDA is working on a dual approach here, both encouraging those managing components of the distributed archive to clear up their catalogues and working to create a clean copy for use by the UK research community.

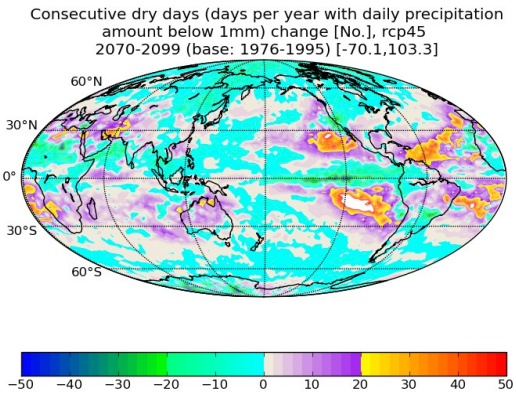


Figure 17. Showing a commonly used climate index, "consecutive dry days", easily calculated from the petabyte archive now held at CEDA.

The clean copy of CMIP5 in the CEDA ardchive not only supports fast UK download of data, but it also makes it possible for users to access data directly using computational resources in CEDA's new JASMIN facility.

The peta-byte CMIP5 archive at CEDA provides flexible access to a vast data resources, with hundreds of data variables available from a huge range of models and experiments. CEDA will use this archive to generate a range of derived products to facilitate access for users who do not want the to deal with the complete dataset, but direct access will also be provided for scientists who wish to explore the archive in full.

Peer REview for Publication & Accreditation of Research Data in the Earth sciences (PREPARDE)

Sarah Callaghan

The Peer REview for Publication & Accreditation of Research data in the Earth sciences (PREPARDE) project aimed to investigate and provide guidance on the policies and procedures required for the formal publication of research data, ranging from ingestion into a data repository, through to formal publication in a data journal. It also addressed key issues arising in the data publication paradigm, such as how does one peer-review a dataset, what criteria are needed for a repository to be considered objectively trustworthy, and how can datasets and journal publications be effectively cross-linked for the benefit of the wider research community.

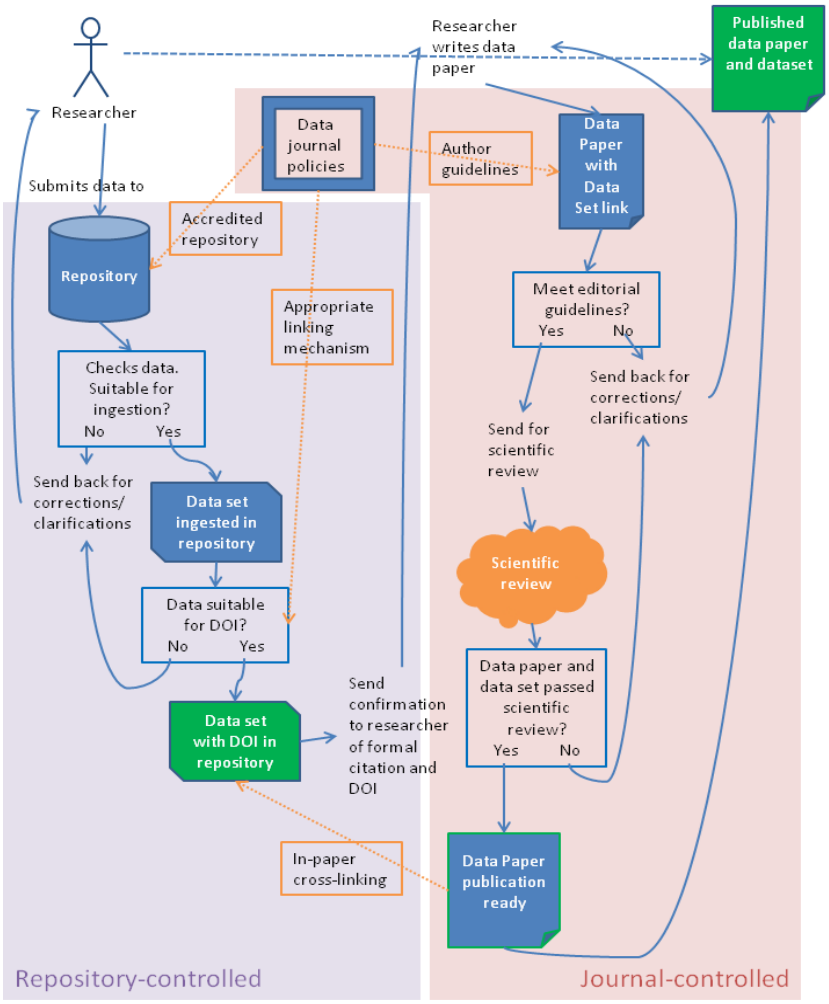


Figure 18: Simplified data publication workflow showing the areas of

PREPARDE brought together a wide range of experts in the research, academic publishing and data management fields both within the Earth Sciences and in the broader life sciences. It produced general guidelines applicable to a wide range of scientific disciplines and data publication types for the accreditation of data repositories for formal data publication and for the peer-review of data. It also captured and analysed the workflows required to publish data, both from a repository and journal standpoint.

Stakeholder interaction was a major part of the project. In total, the project organised three workshops on the three main project themes: workflows and cross-linking, repository accreditation and peer-review of data. It also created a data-publication jiscmail mailing list where people interested in data publication can discuss issues and share information (which has over 300 subscribers at the time of writing).

The project's main conclusions revolved around the standardisation of metadata schemes and the role of a centralised, third party registry and metadata broker in data publication to simplify the process of passing information and metadata between data repositories and journals. Instead of the journal having to set up a line of communication with every data repository hosting data mentioned in the journal article (and vice versa), the journal could set up one way of communicating with the central registry. The repositories could then do the same, reducing the effort involved in maintaining multiple links to many different sources.

The project's guidelines for peer-review of data and the accreditation of repositories for data publication will be published in the next year.

Re-styling the IPCC Data Distribution Centre

Martin Juckes and Charlotte Pascoe

The Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre[8] (DDC) provides data, documentation and guidance for a broad user community working on many aspects of climate change. The DDC is run jointly by 3 data centres: the British Atmospheric Data Centre, the German Climate Research Centre (DKRZ) and the Center for International Earth Science Information Network (CIESIN). The DDC attracts over 50 thousand visitors each year, mainly from the United States, UK and Germany, but increasingly also from India and China.

The DDC acts as a reference archive, providing authoritative information for users from the scientific community as well as for decision makers and their advisors. In the past year several changes have been made to the site to prepare for new users, in particular to prepare for the IPCC 5th Assessment.

The structure of the site has for many years been designed around the domains of the 3 managing data centres (BADC, DKRZ and CIESIN) and around the mandates of the 3 IPCC working groups. As the user community has broadened to include many who have no direct experience of the IPCC working groups or the individual data centres it became clear that old structure was confusing for many. A new configuration was designed, starting from a simple division into 3 generic categories:

- Information about the site;
- Guidance on the use of climate data;
- Access to data;

It is likely that a 4th category will be added to provide a direct link to discussions of the story-lines which underpin much of the IPCC Assessment Process and provide a means of linking between a broad range of socio-economic issues and the complexities of the climate projections.

Figure 19. Word-cloud of phrases used within the IPCC Data Distribution Centre which will be covered by the new glossary.

The data access section is further sub-divided into observational, simulated (including projections) and synthesised (i.e. combining observations and simulations) data collections.

The re-organisation required the adaptation of hundreds of pages, but now provides a cleaner view for users.

The site contains a confusing array of specialist terms from different areas of the climate and socio-economic sciences. CEDA staff have been working with TGICA and representatives of the IPCC working groups to put together a comprehensive glossary which will use definitions prepared for the IPCC 5th Assessment Report where these are available and add definitions for terms which are not covered in that report.

The scope of the site has been broadened as the IPCC DDC is run on limited resources equivalent to approximately one full time staff member at each contributing centre. In order to cover the rapidly expanding scope of climate sciences a streamlined process for assessing data sets has been developed. In general datasets will not be hosted by the DDC, but the DDC will provide links to datasets which pass the assessment.

CEDA has supported work on additional changes implemented by our DDC partners: CIESIN has added new pages to describe the revised scenario process being adopted for the IPCC 5th Assessment cycle, replacing the old "SRES Scenarios"[9] and DKRZ have prepared new pages to provide access to quality controlled elements to the new CMIP5 archive.

[8]IPCC Data Distribution Centre: www.ipcc-data.org

[9]IPCC Special Report on Emission Scenarios, Nebojsa te al., 2000 [http://www.ipcc.ch/ipccreports/sres/emission/].

The CHARMe Project: Commentry Metadata for EO Datasets

Philip Kershaw, Victoria Bennett, Maurizio Nagni and Kevin Marsh

A major impediment to enabling the wider use of EO and climate data is how users can judge if these data are 'fit for purpose', particularly as these data are now being used for increasingly diverse applications. Different users (and uses) require different kinds of supporting metadata.

Lawrence et al[10] provide a taxonomy of different kinds of metadata employed in data infrastructures, and 5 types were defined: Archive (A), Browse (B), Commentary (C), Discovery (D) and Extra (E) metadata. The boundaries between these types are not well defined, but in general A, B, D and E metadata are intrinsic to the dataset and hence known to the data provider. 'C-metadata' are normally produced after the dataset has been published and reflect real use in the community, and is therefore extrinsic to the dataset itself and includes both quantitative and non-quantitative metadata, such as:

1. Post-fact annotations, e.g. citations, ad-hoc comments and notes;
2. Results of assessments, e.g. validation campaigns, intercomparisons with models or other observations, reanalysis, quantitative error assessments;
3. Provenance, e.g. dependencies on other datasets, processing algorithms and chain, data source;
4. Properties of data distribution, e.g. data policy and licensing, timeliness, reliability;
5. External events that may affect the data, e.g. volcanic eruptions, El-Nino index, satellite or instrument failure.

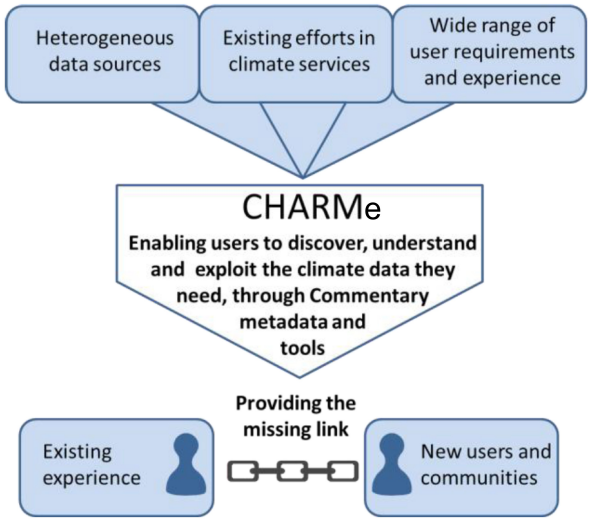


Figure 20. Aims of the CHARMe project (J. Blower)

As yet, there is as yet no robust and consistent mechanism to link 'C-Metadata' to the datasets themselves. The goal of the 2-year EU FP7 CHARMe project ("Characterization of metadata to allow high-quality climate applications and services") is to provide these essential links, by creating a repository of 'C-metadata' plus a set of interfaces through which users can interrogate the information over the Internet. This is intended to provide robust and reusable frameworks for linking datasets to 'C-metadata', reusable software tools that allow climate scientists and users to exploit this information in their own applications, and improved search, intercomparison and time-series analysis tools. A linked data approach is being used along with the Open Annotation standard. There will be a collection of CHARMe nodes to store 'C-metadata', which will have many different clients. It will draw on the experience CEDA gained from the development of the ESGF METAFOR / ES-DOC system for climate model, which uses a lightweight plugin to existing sites. The CHARMe system will also include a Significant Events viewer, a Faceted Search tool and an Intercomparison tool.

The project consortium encompasses data providers, scientists, and developers of future climate services, who participate in major European investments such as GMES, ERA-Clim, ESA's Climate Change Initiative, the Climate Satellite Applications Facility and EURO4M. This will ensure that the CHARMe system is suited to the needs of diverse EO user groups, and facilitate closer links between the project partners. CEDA staff are involved in almost all of the CHARMe work packages, and are leading the development of the data model which will underpin the CHARMe system, the faceted search system and the deployment of a CHARMe enabled ESGF node to demonstrate the value of capturing 'C-metadata'.

Acknowledgements: Jon Blower (University of Reading). This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 312641. The authors are very grateful to all members of the CHARMe Consortium (http://www.charme.org.uk), and the project's advisory board.

[10] Lawrence, B., Lowry, R., Miller, P., Snaith, H., Woolf, A.: Information in environmental data grids. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367 (March 2009) 1003{1014

Portable Infrastructure for the Metafor Metadata System

Charlotte Pascoe

PIMMS[11] captures metadata about models and simulations and the reasoning behind them. PIMMS brings together the technologies used by Metafor[12] to create the CMIP5[13] metadata questionnaire into a single interface that allows users to generate their own metadata questionnaires. PIMMS does this by providing users with tools to describe experiments which record the reasoning behind simulations, to upload lists of questions and controlled answers about model software (controlled vocabularies) and to select experiments and controlled vocabularies to configure bespoke metadata questionnaires. PIMMS tools produce metadata in the Metafor Common Information Model (CIM[14]) format.

PIMMS supports:

- Model Intercomparison Projects (MIPs) where a standard set of questions is asked of all the models which all perform the same set of experiments
- Disciplinary level metadata collection where a standard set of questions is asked of all models but the experiments they perform are specified by the users
- Bespoke metadata creation where the users define questions about both models and experiments
- Model development by providing a metadata framework for model configurations that fail to run

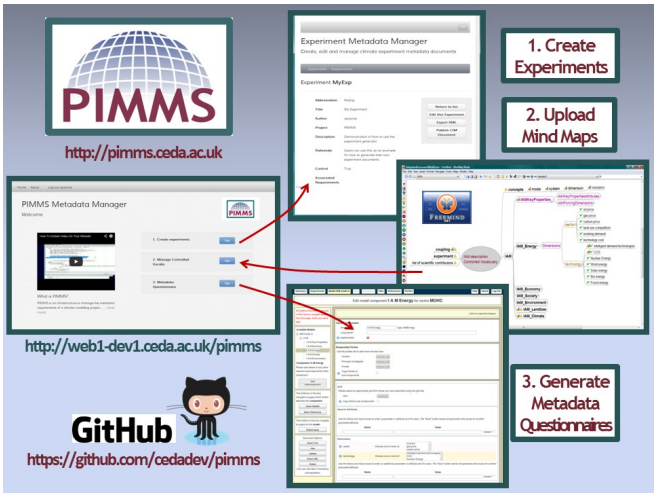


Figure 21 PIMMS tools capture metadata about models and simulations and the reasoning behind them. PIMMS tools produce metadata in the Metafor Common Information Model (CIM) format.

PIMMS controlled vocabularies are recorded using mind maps[15], which not only collate controlled vocabularies but also provide a structure for the way the information is collected. The PIMMS web tools provides mind map checking and mind map to xml translation services. PIMMS also showed that controlled vocabularies can be used in combination with the text mining capabilities of the University of Cambridge ChemicalTagger[16] tool to create CIM documentation from journal article texts.

The PIMMS methodology assumes an initial effort to document standard model configurations. Once these descriptions have been created users need only describe the specific way in which their own model configuration is different from the standard. Thus the documentation burden on the user is specific to the experiment they are performing and fits easily into the workflow of doing their science.

PIMMS metadata is independent of the data produced by simulations and as such is ideally suited to documenting model development. PIMMS provides a framework for sharing information about failed model configurations for which no data are kept, i.e. the negative results that do not appear in scientific literature. PIMMS can be downloaded from the PIMMS GitHub[17] repository and installed on your local computing system. We are also alpha testing a centralised PIMMS web service[18].

PIMMS began life as a project funded by the JISC MRD (Managing Research Data) programme with partners at the University of Reading, The University of Bristol and STFC. You can keep up with PIMMS news on twitter #pimmsMRD and via the PIMMS web site (http://pimms.ceda.ac.uk).

[11]PIMMS: Portable Infrastructure for the Metafor Metadata System http://pimms.ceda.ac.uk
[12]METAFOR: Common Metadata for Climate Modelling Digital Repositories http://metaforclimate.eu
[13]CMIP5: 5th Coupled Model Inter-comparison Project http://cmip-pcmdi.llnl.gov/cmip5/
[14]CIM: Common Information Model http://www.geosci-model-dev.net/5/1493/2012/gmd-5-1493-2012.html
[15]Mind Maps: PIMMS mind maps are generated with freemind http://sourceforge.net/projects/freemind/files/freemind-deb/0.8.1/
[16]ChemicalTagger: Natural Language Processing http://chemicaltagger.ch.cam.ac.uk/
[17]PIMMS GitHub: https://github.com/cedadev/pimms
[18]PIMMS Web Service: http://web1-dev1.ceda.ac.uk/pimms

Advising on Preparing Data Files for Archival in the Agricultural Greenhouse Gases Platform

Ag Stephens, Spiros Ventouras & Charlotte Pascoe.

When writing scientific data there are infinite ways to structure and format the output files. However, use of common formats, conventions and standard vocabularies enables data to be discovered, used and analysed efficiently by a large community of scientists. CEDA provides guidance to the data providers contributing to the Agricultural Greenhouse Gas Research Platform[19]. Along with structured metadata, the data files must be submitted to the archive system (managed by the Freshwater Biological Association) in an agreed data format.

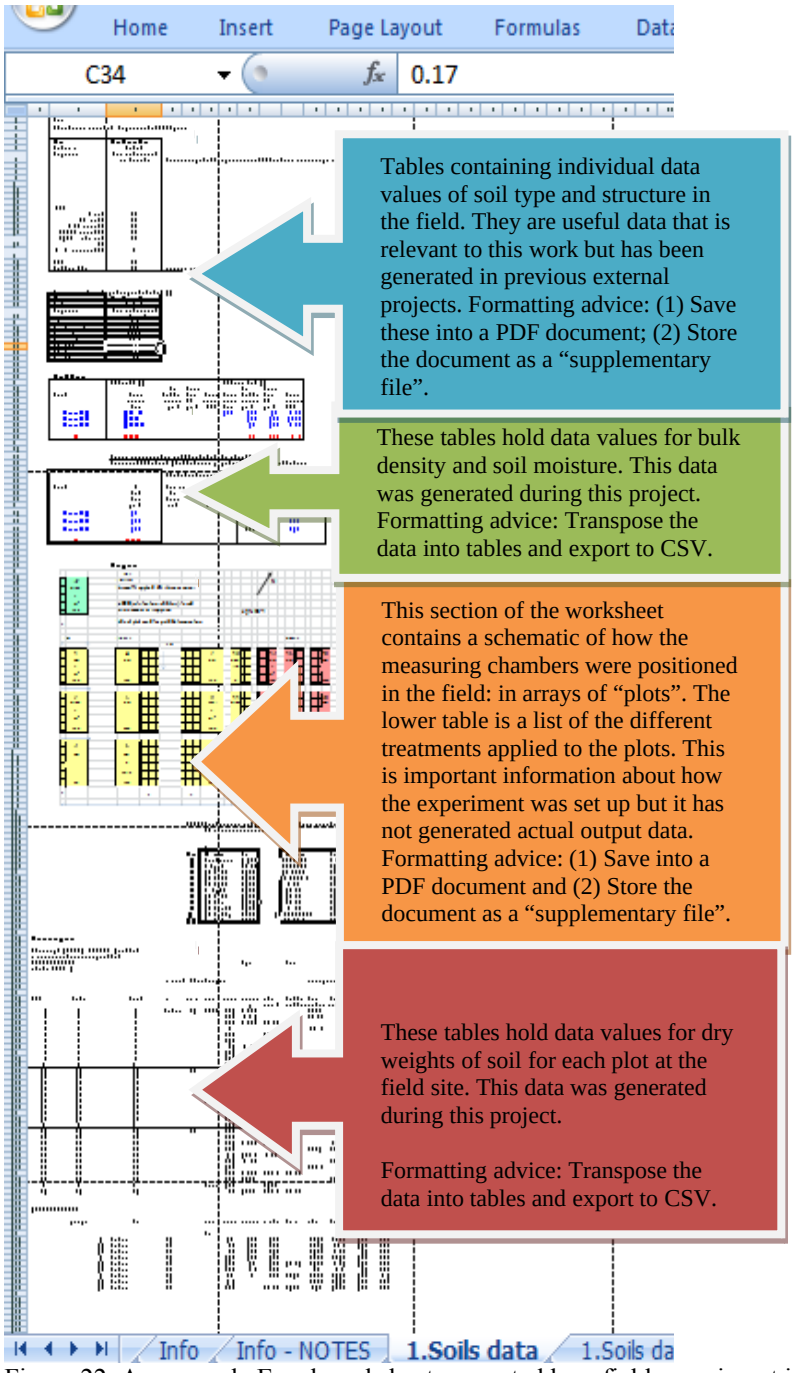


Figure 22. An example Excel worksheet generated by a field experiment in the Greenhouse Gases Platform. The coloured boxes indicate how the various sections of worksheet will be re-formatted into content suitable for the final archive.

In order to make output files that can be understood by both people and computers we promote the use of a very simple file format. The basic constraints are below:

1.Only include one table per CSV file. This means avoiding nested tables and multiple tables next to, or underneath, each other.

2.Remove all annotations. CSV format will not keep any software-specific annotations such as Excel comments, diagrams, colours, special formatting or cell formulas.

3.Remove any non-ASCII characters. Characters such as mathematical symbols and Greek characters (such as "µ") may not be exported correctly into CSV files.

4.Use the top 3 rows to define what is in each column. Row 1 should contain the scientists own label to describe the phenomenon in that column. Row 2 should contain the "term" defined in the Controlled Vocabulary and Row 3 should contain the units for that phenomenon.

5.Separate out "compound phenomena" into separate columns. For example, the phenomenon "Applied Nitrogen to Wheat in England 2010" can be separated into four separate columns: (1) Applied Nitrogen Rate; (2) Crop Type; (3) Country and (4) Year.

Figure 23 shows a worked example in which an existing Excel worksheet is broken down into various components. Each component will either be transposed into a table and exported to CSV, or saved as a PDF document and provided as a "supplementary file" to the archive.

[19] The Agricultural Greenhouse Gas Research Platform aims to improve the accuracy and resolution of the reporting system by providing new experimental evidence on the factors affecting emissions and statistics relevant to changing farming practices in the UK. See: http://www.ghgplatform.org.uk/

CEDA Collaborations with International Space Agencies

Esther Conway, Victoria Bennett

CEDA has enjoyed another year creating strong synergies with the European Space and UK Space Agencies which have yielded positive results. Sam Pepler has continued his work with the LTDP (Long Term Data Preservation) working group representing the UK Space Agency. This group provides long term archival expertise and consists of representatives from Europe and Canada. This group saw the launch of the GEO LTDP guidelines at the GEO symposium in June at the World Meteorological Organisation Headquarters in Geneva which was attended by Esther Conway. These guidelines incorporate original research developed within CEDA on preservation analysis workflows and information models for describing representation information networks.

Figure 23. ATSR instrument collecting data on board the ERS 1 satellite

The LTDP management support project led by Esther Conway has been extended into 2013 allowing us to undertake exciting new collaborative work with ESA. The emphasis of the project has now shifted from the analysis of preservation technologies to the archival process, information views and decision making which support sustainable long term Earth Observation archives. This work will produce a formalisation and extension to core preservation models in order to facilitate systems and process design in the next phase of the LTDP programme along with risk analysis and cross archive reporting to support key archival decision making.

Victoria Bennett continued to support the ESA Climate Office at Harwell and the Climate Change Initiative (CCI) Programme for 1 day/week with advice and coordination on data standards for CCI data producers. Victoria leads regular telecoms and email discussions and works with CCI project teams to ensure data are produced in a consistent manner, for ease of use by climate and other data users. This year she finalised the CCI Phase 1 guidelines for Data Producers, and issued the CCI Phase 2 Requirements for Data Producers.

Steve Donegan successfully managed the task of reprocessing the ATSR1 and 2 v2.1 data from the original UBT data in the CEDA archive. This data will sit alongside the AATSR v2.1 data from ESA. The JASMIN/CEMS infrastructure was used to process the two missions in a significantly quicker time than the last processing, with the resulting data placed in bulk in the archive. These products have been made available to the validation team within a dedicated JASMIN/CEMS workspace environment so the data can be assessed before general release.

The FP7 SCIDIP project has also made good progress during its 2nd year. This is an ESA led collaborative project with Sprios Ventouras from CEDA playing a pivotal role in the development of harmonised metadata for the broader European Earth Science community including our NERC sister archive at the British Geological Survey.

During 2012/13 CEDA has demonstrated that it is a not only centre of expertise in the management of environmental data from space but can take leading role transferring skills and knowledge to the broader Earth Science community

Data Management for NERC Research Projects (RP) and Research Mode (RM) Grants

Wendy Garland

BADC provides data management support to a range of NERC programmes and research projects including: QUEST, CASCADE, ABACUS, RAPID-WATCH, Amazonica, ClearfLO, Storm Risk Mitigation, MAMM, the CMIP5 workshop, SAMBBA, and the JISC Impacts project. For each project/programme, the CEDA science support team work closely with the project participants to draft and agree a data management plan. We then provide the data archive framework and infrastructure; document and distribute data (observation, synthesis and model data); and provide support and advice to data providers and users. In addition to our primary role of data archiving and curation, we also attend science/project meetings (to interact with data providers), coordinate workshops, acquire third party data and provide secure online collaborative workspaces for use by the project teams.

Figure 24. BT tower, Central London, location of a bank of instruments measuring the air quality in the capitol for ClearFLo

For example, this year support has been given to ClearFLo, (Clean Air for London) a NERC collaborative project involving the Universities of Reading, Manchester, KCL, Birmingham, York Leeds Hertfordshire, Leicester, East Anglia, Bristol and Edinburgh. Data has been collected from a large range of instruments situated in several locations in and around London including sites in North Kensington and on the upper levels of the BT tower. Some instruments have been deployed for long-term monitoring and additional instruments brought in for short intensive operational periods. Measurements have also been made by instruments on board the FAAM BAe-146 aircraft and the NERC Dornier aircraft.

Data from these instruments has been processed by the project teams and archived in the ClearFLo archive at BADC (see http://badc.nerc.ac.uk/data/clearflo/). Along with archiving the final data, the CEDA Science Support team has provided filenameing and data formatting support to help the instrument operators to achieve a consistent, well formatted and well-documented re-usable archive.

BADC has also supported the DIAMET (Diabatic influences on mesoscale structures in extra-tropical storms) project – part of the Storms Risk Mitigation Research Programme. Data have been received and archived from ground-based radar, the FAAM BAe-146 aircraft and radiosonde instruments. In addition, substantial volumes of Met Office data (Unified Model, radiosonde, and radar) data has been specifically located and extracted and made available to the project team by BADC to support the project team.

Previously data from NERC RP and RM grants have each been considered and managed on a project by project basis. This year, in an initiative joint with all NERC Data Centres, a more systematic approach has begun to assess the data management needs of all NERC Grants starting with those beginning in the financial year 2010-11. All grants starting during this period have been examined and allocated to a data centre depending on subject area. All those projects allocated to CEDA have been reviewed and the PIs contacted with a view to producing a customised DMP for each project and for data management to follow. In this way, the data needs of all projects are addressed in a more structured manner and efficiencies of effort will be made. This procedure will then be extended to include subsequent years' grants to facilitate the complete archival of appropriate NERC data.

Appendix 1: Additional details of 2012/13 activities

Specific Collaborations and Partnerships

- Collaboration with the University of Oxford on development of the JASMIN community inter-comparison suite. Stephen Pascoe is providing technical leadership and management of the contracted development team.
- Earth System Grid Federation: Stephen Pascoe is contributing to planning new components in the ESGF system which underpins the CMIP5 archive and will be the basis of data archival for the CORDEX and CCI projects.
- Victoria Bennett is working with the International Space Innovation Centre, Logica and Astrium GEOInformation Services to develop and operate jointly the CEMS facility
- Victoria Bennett is collaborating with the University of Reading and other project partners on FP7 project CHARMe
- Victoria Bennett is working with ESA HArwell Climate Office on data standards for Climate Change Initiative Programme
- Charlotte Pascoe is PIMMS project manager for a collaboration with the University of Reading department of Meteorology and the University of Bristol school of Geographical sciences.
- Charlotte Pascoe was invited to become an official stake holder for the ERMITAGE project as part of the PIMMS collaboration with the ERMITAGE.
- Martin Juckes led the FP7 GMES Climate Information Portal (GMESCLIP) Proposal.
- Graham Parton and Sarah Callaghan are collaborating with Elsevier to promote data publication.
- Graham Parton and Sarah Callaghan collaborated with Victoria University and Charles Beagrie Ltd on a project to determine the impact of the BADC as a scientific data service.
- Esther Conway is part of the SCIDIP project and is part of the LTDP WG, LTDP SRR projects.
- Sarah Callaghan collaborated with Wiley to launch the Geoscience Data Journal and is an associate editor.

Publications

- **Bennett, V.L.,** Busswell, G., Clifford, D., Curtis, M., Hilton, R., **Kershaw, P.,** Pechorro, E., Raper, I., Remedios, J.J. and Timms, G, "The Facility for Climate and Environmental Monitoring from Space (CEMS)" PROCEEDINGS OF THE REMOTE SENSING AND PHOTOGRAMMETRY SOCIETY CONFERENCE 2012 "Changing how we view the world". University of Greenwich, London, 12-14 September 2012
- **Callaghan, S.,** Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S., Whyte, A., "Processes and Procedures for Data Publication: A Case Study in the Geosciences" 2013, International Journal of Digital Curation, Vol. 8, No. 1, pp. 193-203 doi:10.2218/ijdc.v8i1.253
- **Callaghan, S. A.,** Waight, J., Agnew, J. L., Walden, C. J., Wrench, C. L. and **Ventouras, S**. (2013), The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK. Geoscience Data Journal. doi: 10.1002/gdj3.2
- Cinquini, L.; Crichton, D.; Mattmann, C.; Bell, G.M.; Drach, B.; Williams, D.; Harney, J.; Shipman, G.; Feiyi Wang; **Kershaw, P**.; **Pascoe, S.**; Ananthakrishnan, R.; Miller, N.; Gonzalez, E.; Denvil, S.; Morgan, M.; Fiore, S.; Pobre, Z.; Schweitzer, R. "The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data", IEEE 8th International Conference on e-Science, 2012.
- **Conway, E,Ventouras, S,** Albani, M, Leone, R, Interoperability and standardization aspects in the data preservation domain, in IEEE Geoscience and Remote Sensing Symposium 2012 (IGARSS 2012), Munich, Germnay, 22-27 Jul 2012, (2012).http://epubs.stfc.ac.uk/work-details?w=63740

- Guilyardi, Eric, and Coauthors, 2013: Documenting Climate Models and Their Simulations. Bull. Amer. Meteor. Soc., 94, 623–627. doi: http://dx.doi.org/10.1175/BAMS-D-11-00035.1
- Jones, B., Broeder, D., Kelsey, D., **Kershaw, P.**, Lueders, S., Lyall, A., Wartel, R., Weyer, H.J, "Federated Identity Management for Scientific Collaborations", Extended Abstract (accepted), TERENA Networking Conference, May 2012
- **Kershaw, P.**, Curtis, M., Pechorro, E., "CEMS: Building a Cloud-Based Infrastructure to Support Climate and Environmental Data Services", ESSI2.10, European Geosciences General Assembly, April 2012
- **Lawrence, B.N.**, Balaji, V., Bentley, P., **Callaghan, S.**, DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R.W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M.-P., Murphy, S., **Pascoe, C.**, Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A., and Valcke, S.: "Describing Earth system simulations with the Metafor CIM", Geosci. Model Dev., 5, 1493-1500, doi:10.5194/gmd-5-1493-2012, 2012.
- Long Term Preservation Of Earth Observation Space Data: European LTDP Common Guidelines http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue2.0.pdf
- **Nagni, M., Ventouras, S.**, "Implementation of UML Schema in Relational Databases: A case of Geographic Information" (to be published in International Journal of Distributed System and Technologies)
- **Pascoe, S.**, Wilkinson, R., **Kershaw, P**. "Benchmarking OPeNDAP services for modern ESM data workloads", ESSI2.7, European Geosciences General Assembly, April 2012
- **Pascoe, S**. "The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data". Submitted to IEEE Fifth Generation Computer Systems special issue
- **Pascoe, C.**, "The PIMMS project and Natural Language Processing for Climate Science" Proceedings of Open Repositories 2013
- Shaon,A., **S Callaghan**, **BN Lawrence**, B Matthews, T Osborn, C Harpham, A Woolf. Opening Up Climate Research: A Linked Data Approach to Publishing Data Provenance. International Journal of Digital Curation 7 (1), 163-173, 2012, doi:10.2218/ijdc.v7i1.223.
- Shaon, A., **Conway, E.**, Giaretta, D., Matthews B., Crompton, S., Marelli, F., Yu, J., Giammatteo, U.D., Marketakis, Y., Tzitzikas, Y., Guarino, R., Brocks, H., Engel, F., Towards a Long-term Preservation Infrastructure for Earth Science Data. In International Preservation Conference 2012 (iPres2012), Toronto, Canada, 1-5 Oct 2012, (2012).http://epubs.stfc.ac.uk/work-details?w=64387
- Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [**S.Ventouras, S.A.Callaghan**, C.L.Wrench] (2012): ITALSAT radio propagation measurements at 50GHz in the United Kingdom. NERC British Atmospheric Data Centre. http://dx.doi.org/10.5285/10.5285/597C906A-B09E-4822-8B60-3B53EA8FC57F
- Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [**S.Ventouras, S.A.Callaghan,** C.L.Wrench] (2012): ITALSAT radio propagation measurements at 40GHz in the United Kingdom. NERC British Atmospheric Data Centre. http://dx.doi.org/10.5285/10.5285/ 4a60ee2f-0fd1-4141-9244-7bebf240bb49
- Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [**S.Ventouras, S.A.Callaghan,** C.L.Wrench] (2012): ITALSAT radio propagation measurements at 20GHz in the United Kingdom. NERC British Atmospheric Data Centre. http://dx.doi.org/10.5285/10.5285/ 3158D138-D592-4045-ADE4-B76CF9F42129
- Wilson, MD, **Conway, E,** Shaon, A, Bunakov, V, Planning Digital Preservation for e-Health using Preservation Network Models In eChallenges e-2012 Conference (e-2012), Lisbon, Portugal, 17-19 Oct 2012, (2012). http://epubs.stfc.ac.uk/work-details?w=63733

Further Funding

- Stephen Pascoe submitted the "GMES Climate Impacts Platform" bid for EU FP7 call SPA.2013.1.1-04 as a work-package co-leader.

- Ag Stephens was awarded the project: Delivering the UK Component of the IPCC Data Distribution Centre (DDC) and climate impacts data services Funder: DECC Period: August 2012 – March 2015 Value: £699,795 (of which £224,000 will be sub-contracted to Met Office Hadley Centre)
- Ag Stephens was awarded the project: NetCDF Abstraction Tool for CEH Funder: CEH Period: January 2013 – March 2013 Value: £25,000
- Ag Stephens was awarded the project: Support for the UKCP09 User Interface Funder: EA Period: April 2012 – Sept 2013 Value: £107,000
- Victoria Bennett was awarded the project: CEMS: helpdesk setup , value of 16k
- Victoria Bennett was awarded the project: HPFELD, value of 50k
- Victoria Bennett was awarded the project: CHARME FP7
- Victoria Bennett was awarded the project: ESA LTDP
- Victoria Bennett was awarded the project: STFC (support to ESA): value of approx 50k
- Victoria Bennett was awarded the project: GMES PURE FP7: £150k (shared with RSG)
- Martin Juckes secured EU FP7 funding for IS-ENES2: Infrastructure for the European Network for Earth System Modelling (participating as WP leader)
- Esther Conway secured a 60 K Euro extension to the LTDP managements support contract/
- Sarah Callaghan secured funding for the JISC and NERC funded PREPARDE project
- Sarah Callaghan secured funding for the JISC funded Impacts of a science data service project

Dissemination/Communication

- Stephen Pascoe organised the NERC JASMIN forum to bring together stakeholders in the JASMIN infrastructure and discuss development of the JASMIN service and applications.
- Victoria Bennett presented to the EO data workshop in ESRIN: https://wiki.services.eoportal.org/tiki-index.php?page=Models+for+scientific+exploitation+of+EO+Data
- Victoria Bennett presentedan introduction to EO and data to DEFRA staff.
- Victoria Bennett wrote an article in NCEO magazine "Blue Marble" about CEMS
- Phil Kershaw gave a presentation to NCEO Conference, Sept 2012: CEMS: The Facility for Climate and Environmental Monitoring from Space
- Charlotte Pascoe wrote 16 Blog posts for the PIMMS project http://proj.badc.rl.ac.uk/pimms/blog
- Charlotte Pascoe presented at the JISC MRD workshops April, October, December 2012
- Charlotte Pascoe presented at Open Repositories - May 2012
- Charlotte Pascoe presented at the NCAS Staff meeting - June 2012
- Charlotte Pascoe presented at the ERMITAGE workshop - September 2012
- Charlotte Pascoe presented at the PIMMS dissemination workshops (2 in Bristol and 1 in Reading) - January and February 2013
- Martin Juckes co-ordinated a workshop for the international ExArch project, with partners from USA, Canada, France and Germany and guests from the Netherlands.
- Martin Juckes attended a meeting of the IPCC Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA).

- Martin Juckes attended and presented at 2nd EU-Australia Workshop hosted by EU Commission, Brussels, 26/27 June 2012
- Wendy Garland gave a talk to school children about weather
- Maurizio Nagni presented a poster at EGU 2012 about the implementation of UML schema
- Graham Parton developed and formalised the use of CEDA social media
- Sarah James presented posters advertising physical archive of solar images at 2 Royal Astronomical Society community meetings, and one of our glass plate solar images appeared on Sky at Night.
- Esther Conway presented 3 posters on SCIDIP at the following: NCAS staff meeting, NCEO staff meeting, EVO International Conference, London 16 May 2012 http://www.evo-uk.org/evo-cloud-services-portals/get-involved/evo-international-conference-london-16-may-2012/
- Esther Conway co-authored presentation on LTDP Data Lifecycle and Preservation process Pin - Pérennisation des Informations Numériques with Danielle Boucon of CNES; given by Danielle 4th January 2013 Paris
- Esther Conway attended the GEO Air Quality Metadata Workshop Dublin 2012 http://wiki.esipfed.org/index.php/Air_Quality_Metadata_Workshop_Dublin_2012
- Sarah Callaghan blogs at http://citingbytes.blogspot.co.uk and the PREPARDE project blog at http://proj.badc.rl.ac.uk/preparde/blog
- Sarah Callaghan attended and presented at the OpenAIRE Interoperability Workshop held at the University of Minho, Portugal, on the 7/8 February 2013. She blogged about the workshop at and her slides are available at the same place. A video of her presentahttp://proj.badc.rl.ac.uk/preparde/blog/OpenAIRE_Interoperabilitytion can be seen at http://vimeo.com/channels/openaireworkshop2/59741033
- Sarah Callaghan was part of the panel at the British Library DataCite workshop "Making citation work: practical issues for institutions" to be held on the 8th March.
- Sarah Callaghan attended JISC-British Library DataCite Workshop #3: Managing and citing sensitive data on the 29th October. She also attended the DataCite client meeting on the 19th November
- Sarah Callaghan participated on the panel in the Data Reuse session at the SpotOn London conference http://www.nature.com/spoton/event/spoton-london-2012-data-reuse/
- Sarah Callaghan presented at AGU in the following sessions: "Data Stewardship, Citation With Confidence, and Preparing Next Generation of Data Managers", and "Publishing Research Data: Peer Review, Data Center Accreditation, and Linking".
- Sarah Callaghan attended and presented at the European Association of Science Editors conference, and the DataCite Summer meeting, both in June 2012. She also attended the NordBIB conference, OpenAIREplus workshop and the CODATA working group on data citation meeting during the same week.

Software

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 2012/13		
esgf-pyclient	A Python library which enables access to Earth System Grid Federation services, including search, login and download. esgf-pyclient underpins several operational systems maintaining the ESGF infrastructure and is used internationally by end users and institutions to access ESGF data	http://esgf-pyclient.readthedocs.org/
Jasmin-CIS	The JASMIN Community Inter-comparison Suite (CIS) is an analysis tool specialising in inter-comparison of model, observation and satellite data. CIS enables co-location, visualisation and extraction of specific datasets within these science domains and is optimised for use on the JASMIN infrastructure.	http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/CommunityIntercomparisonSuite
Jasmin-JAP	The JASMIN Application Platform (JAP) is an extension of the RedHat family of Linux distributions specifically designed to provide the consistent set of tools desired by the Atmospheric and Earth Observation science community.	http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/AnalysisPlatform
PIMMS	PIMMS brings together the technologies used by the Metafor project to create the CMIP5 metadata questionnaire into a single interface that allows users to generate their own bespoke metadata questionnaires.()	http://www.ceda.ac.uk/projects/pimms/
CEDA Markup	CEDA Markup is a library for exposing search services through the OpenSearch API.	https://github.com/cedadev/cedaMarkup
FATCAT	FatCat is a search system which enables discovery of individual files based on file-level metadata and geospatial queries. FatCat has been used to supply the search backend in the HPFeld project and is being further developed to enable file-level search of the CEDA archive.	
Improved Software		
CEDA-MOLES	CEDA Moles is an implementation of the MOLES3 information model which applies ISO-compliant modelling practices, such as OGC Observations and Measures, to the description of atmospheric science archive catalogues. CEDA-MOLES provides a web interface for cataloguing scientific datasets and will underpin the next generation CEDA catalogue.	

DRSlib	DRSlib enables the maintenance of data archives compliant with the Data Reference Syntax metadata specification adopted by CMIP5 and subsequent projects using the ESGF infrastructure. In particular drslib maintains a version-controlled directory structure which avoids data duplication and enables multiple versions of datasets to be available simultaneously.	http://esgf-drslib.readthedocs.org
Maintained Software		
CF-checker	The CF-conventions checker is developed at Reading University and CEDA maintains a web service which enables it use online. As part of our maintenance of the web service we also collaborate with Reading University in maintaining the software and adapting it for use on the JASMIN infrastructure.	
cdat_lite -	Cdat-lite is a Python package for managing and analysing climate science data. It is a subset of the Climate Data Analysis Tools (CDAT) developed by PCMDI at Lawrence Livermore National Laboratory. Cdat-lite aims to compliment CDAT by focussing on it's core data management and analysis components and by offering a radically different installation system to CDAT.	http://proj.badc.rl.ac.uk/cedaservices/wiki/CdatLite

Appendix 2: 2013-2014 Detailed Targets

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

Science Programme Support
NCAS JASMIN Research Collaboration: Develop and deliver services, carry out appropriate research, to assist in delivering the science projects with JASMIN group workspace and compute access.
NCEO CEMS Research Collaboration: Develop and deliver services, carry out appropriate research, to assist in delivering the science projects with CEMS group workspace and compute access.
Data Programme Support
NCEO data management: Curate data archive (NEODC) and provide data management services including data management plans and data acquisition for NCEO facilities, staff and collaborators (including ARSF).
NCEO Community Support: Support the UK earth observation community by continuing to prioritise the provision of high speed UK cache archives for ESA, Eumetsat (and other prioritised high volume remote data).
NCAS data management: Curate data archive (BADC) and provide data management services including data management plans and data acquisition for NCAS facilities, staff and collaborators (including joint facilities such as FAAM and UKSSDC).
NCAS Support: Support the UK atmospheric science community by continuing to prioritise the provision of high speed UK cache archives for Met Office, CMIP5, NASA (and other remote data services)
RP and RM data management: Provide data management support for NERC programmes and research projects and grants (data management plans, data acquisition etc).
General Activities
CEDA Management: Staff project and programme management.
Data Ingestion: Acquire, ingest, and catalogue, appropriate data from a range of sources as identified in data management plans associated with the funded CEDA projects.
Hardware Infrastructure: Maintain and upgrade computing systems and networks to support data holdings and data access. This includes the JASMIN/CEMS infrastructure.
Software Infrastructure: Develop, maintain and upgrade necessary software and information systems to support data curation and data access.
Operational User Services - Curation: Curate existing information according to best practice principles: create, delete, migrate data and information as necessary.
Operational User Services – User support: Provide prompt and effective user support. Provide additional services to users such as a distributed document management service and administration for Group Work Spaces (GWS) and science processing machines.
Operational User Services including Data distribution: Provide effective information systems to support data and metadata access services including visualisations.
Data Management Planning: Evaluate projects, develop data management plans, provide advice to the community**.**
Specific high profile activities
CMIP5 support: Provide support for CMIP5 by providing a UK data node, a replicated copy of the global core archive, and appropriate interfaces (software, hardware, and networks). Deploy and maintain any additional necessary services (e.g. the CMIP5 questionnaire.)
SIS: Implementation of the NERC Science Information Strategy, in particular leading and/or managing projects on architecture, data centre metrics, data citation and publishing, data value check lists and data policy implementation.
CEMS: Work with commercial and academic partners within the Catapult, NCEO and STFC communities to deliver the CEMS Facility (Climate and Environmental Monitoring from Space).
UKSSDC. Continue to integrate the delivery and management of the UK solar system data centre into CEDA, while maintaining services to UKSSDC community.

JASMIN: Work with the e-Science and e-Infrastructure communities to provide a state-of-the-art high performance data analysis facility, and support the projects using it.
NERC Data Catalogue Service: Support operation of the system to harvest metadata records from partner data providers. Provide a metadata search facility via a web service: "NERC Data Catalogue Service"
ESA support: Work with ESA climate office at Harwell to develop data standards for the climate change initiative.
CF and other metadata standards and conventions; development and support: Maintain CF name table. MOLES development. Contribute to OGC, INSPIRE and other geospatial data activities**.**
Met Office Liaison: Support NERC- Met Office interaction and the development of joint projects.
ENES Support: Carry out networking and joint research and deliver data services in partnership with the ENES community.
UKCP09 User Interface: Deliver an operational UKCIP09 User Interface for the Environment Agency.
IPCC Data Centre: Deliver the core node for the IPCC data distribution centre, take part in IPCC activities and provide leadership as appropriate to the IPCC data centre community.
Project specific
Commercial Contracts: Obtain and deliver research and service projects consistent with developing and/or exploiting CEDA infrastructure, skills and services.
European Commission Contracts: Obtain and deliver research and service projects consistent with developing and/or exploiting CEDA infrastructure, skills and services.
Academic Contracts: Obtain and deliver research and service projects consistent with developing and/or exploiting CEDA infrastructure, skills and services.