Centre for Environmental Data Analysis (CEDA)

Annual Report 2016

(April-2015 to March-2016)

Bryan Lawrence, Victoria Bennett, Sarah Callaghan
(Editors)

CONTENTS

INTRODUCTION

The Centre for Environmental Analysis (CEDA) exists 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, and consolidated annual reports have been produced since 2009. This annual report presents key statistics for the year past (2015- 2017) as well as a series of snapshots of activity, expressed as short highlights and short reports. Key data centre metrics are also provided.

This year was characterised by continuing evolution in JASMIN capability (JASMIN is the data intensive supercomputer which provides the infrastructure upon which CEDA and the CEDA services are delivered), following the previous year's major upgrade. Uptake in the community continues to expand and the role of CEDA staff continues to evolve to include support for "big data" tools and algorithms by archive and other users on JASMIN. In addition, as in previous years, CEDA staff are involved in nearly all the major atmospheric science programmes underway in the UK, in many earth observation programmes, and in a wide range of informatics activities. To better reflect the range of services CEDA now provides, we changed our name from Centre for Environmental Data **Archival** to Centre for Environmental Data **Analysis**.

The European Sentinel series of satellites are gathering unprecented volumes of earth observation data. Making these data available for the UK science community has been one of CEDA's main challenges, and achievements this year. Other highlights include the provision of successful training for NERC scientists, improvements to our catalogue and data search systems, expansion of the services offered to JASMIN users, and participation in major international climate data collaborations.

Over the years we have reported our key partnerships, and as before, these revolve around our neighbours on the Harwell site (including the Satellite Applications Catapult, with whom we share delivery of the facility for Climate and Environmental Monitoring from Space, CEMS), the European Network for Earth Simulation (with whom we share the delivery of the European component of the Earth System Grid Federation), and many other project collaborators.

Victoria Bennett, Head of Earth Observation at CEDA

PART 1: HIGHLIGHTS AND MAJOR ACTIVITIES

This first section provides a selection of descriptions of key activities and highlights from the year. It has two chapters: one with short highlights selected to showcase some CEDA activities supported through different funding streams, and a second describing a range of key areas of focus for CEDA staff this year:

1. HIGHLIGHTS

Highlight topics this year focus on Sentinel data, and training for CEDA users.

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA Data/Science Highlight

Sentinel Data at CEDA

Steve Donegan

Provision of a Sentinel mirror archive at CEDA

CEDA hosts a mirror archive of selected Sentinel data products for the UK user community. Sentinel1a (S1A) was launched in late 2014, and since then more than 400Tb of S1A data has been retrieved. Since the launch of S1A, three further Sentinels have been launched and CEDA expects to shortly be retrieving 10Tb a day.

Sentinel 2A (S2A) Multispectral Imager (MSI) L1 products are available from March 2016 onwards. All products from the Sentinel 3A (S3A) SLSTR and OLCI are expected in mid 2016. CEDA expects to archive products from S1B from late 2016 to compliment existing S1A products.

The Sentinel data challenge

ESA makes access to Sentinel data available through a number of dedicated “hubs”. CEDA holds one of the UK ESA member states accounts and thus has access to the Collaborative Data Hub (CDH). During 2015/2016 ESA experienced problems with disseminating the volumes of data through this system and created dedicated mirror hubs to balance the load. Not all products CEDA users required was available on the CDH, thus we were forced to acquire data from the ESA Scientific Data Hub (SDH) alongside other users and this affected data bandwidth.

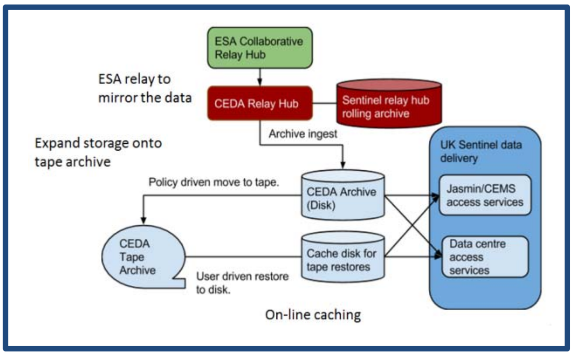
The network of ESA hubs required the development of dedicated download management software that could retrieve data in multiple threads. This was developed to be robust and implement stringent checking and operate on the CEDA ingest parallel cluster as part of the CEDA ingest control system.

The volume of Sentinel data held is now one of the largest CEDA datasets. CEDA is implementing a system where Sentinel data older than 6 months will be moved to the Near Line Archive (NLA) tape facility. Data on the NLA can be reinstated at the users request for a limited period

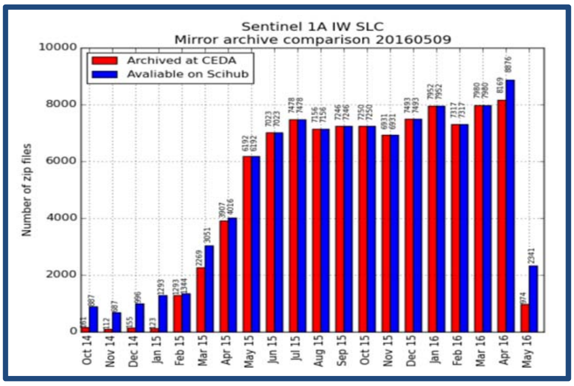
The CEDA Relay Data Hub

In January 2016 CEDA won a contract with ESA to provide an official ESA relay data hub (CDRH). This will be used by ESA to spread the data dissemination loads at the member state level. The CDRH will operate on a dedicated machine outside the STFC firewall.

Steve Donegan
Senior Data Scientist at CEDA, manages Sentinel Data and metadata services/pipelines.


ESA Collaborative Relay Hub
ESA relay to mirror the data
CEDA Relay Hub
Sentinel relay hub rolling archive
Archive ingest
Expand storage onto tape archive
Policy driven move to tape.
CEDA Archive (Disk)
UK Sentinel data delivery
Jasmin/CEMS access services
CEDA Tape Archive
Cache disk for tape restores
Data centre access services
User driven restore to disk.
On-line caching

Above: Schematic of Sentinel data retrieval system and the CEDA Relay Data Hub



Above: State of CEDA mirror archive of Sentinel1a data (May 2016)

Find out more:

- Email steve.donegan@stfc.ac.uk

Return to the CEDA website

Tell us what you think

- How clearly was this article written?
- How interesting or useful was it?
- Do you have any other comments?

Please let us know:
support@ceda.ac.uk

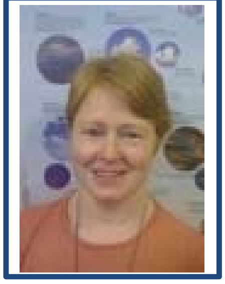
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA Data/Science Highlight

CEDA Training Courses

Alison Pamment[1], Sam Pepler[1], Ag Stephens[1], Steve Donegan[1], James Groves[2], Andy Heaps[3], Alan Iwi[1], Charlotte Pascoe[1], Charles Roberts[3], Dan Walker[2], Louise Whitehouse[2]

([1]CEDA, [2]NCAS Operations Group, [3]NCAS Computational Modelling Service)



Alison Pamment
Environmental Data Scientist specializing in metadata conventions and development of training courses

Training for environmental scientists

CEDA, in collaboration with colleagues from other parts of NCAS, offers training in the skills needed for effective data management and analysis. In April 2015 we ran our one week ‘Introduction to Scientific Computing Course’ at the University of Reading. Twenty-eight PhD students and early career researchers participated in the course which is designed to equip those having little or no previous programming experience with sufficient knowledge to begin automating large or repetitive data processing tasks.

Why is this training important?

Whether a scientist is handling “big data” such as that produced by climate models or large numbers of small data files resulting from instrumental observations, appropriate IT skills are essential. For example, a common data management task is to rename all the files in a directory to conform to a particular pattern or convention. One could achieve this manually by renaming the files one at a time but it is far more efficient to automate the task so that many files can be processed in a single command. Students on the Introduction to Scientific Computing course learn to harness the power of Linux commands and Python programming in order to work more efficiently with their own data.

Developing course materials

When drawing together the course materials, our approach was to reuse, wherever possible, training materials that already existed in the public domain. Materials developed by the Software Carpentry organisation (http://software-carpentry.org/) have proven to be especially useful in this context. Additional material was developed by NCAS staff based at CEDA and the universities of Leeds and York.

This work was funded by NCAS.

Main course modules:

- **The Linux shell**
- **Python: programming basics**
- **Python: starting to work with files**
- **Automated logging of instrumental data**

Additional topics:

- **Algorithmic thinking**
- **Software version control**
- **Parallelising your code**
- **Good data management practices**

Above: The course aims to provide all the necessary building blocks for a scientist to start automating data processing and data management tasks. Students require no previous programming experience.

Find out more:

- CEDA training: http://www.ceda.ac.uk/training/
- NCAS training: https://www.ncas.ac.uk/index.php/en/education-and-staff-development
- Email alison.pamment@ncas.ac.uk
- Take a look at the training materials http://www.ceda.ac.uk/ncas-reading-2015

Return to the CEDA website

Tell us what you think
- How clearly was this article written?
- How interesting or useful was it?
- Do you have any other comments?

Please let us know: **support@ceda.ac.uk**

2. SHORT REPORTS

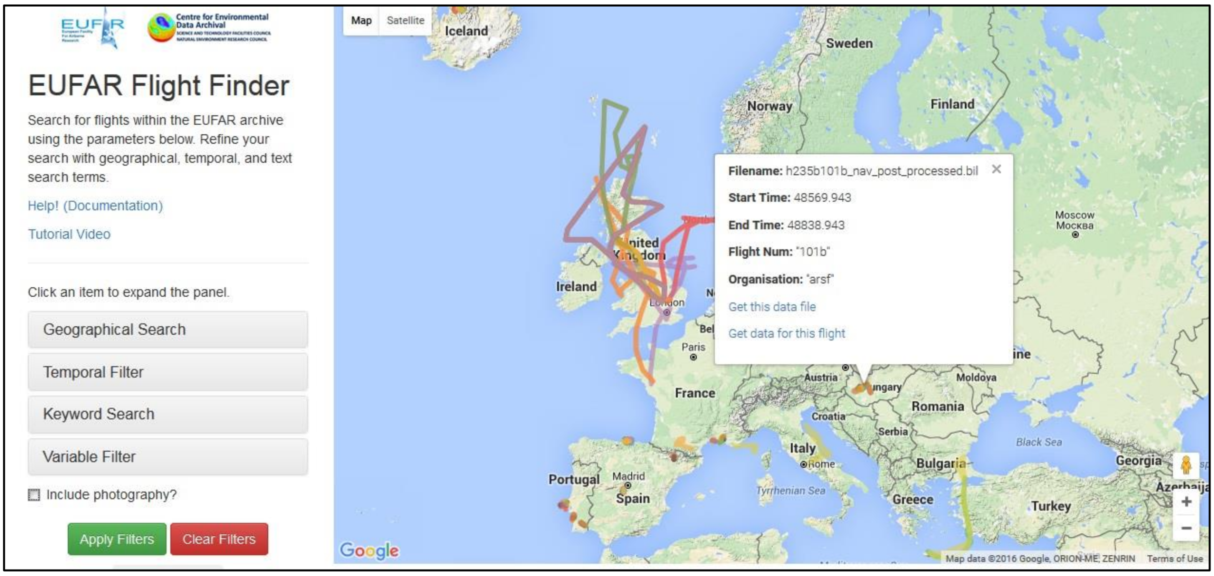
2.1 THE EUFAR FLIGHT FINDER SEARCH TOOL - HTTP://FLIGHT-FINDER.CEDA.AC.UK/

Wendy Garland

The EUFAR Flight Finder (EFF) is a geospatial/temporal/keyword search tool developed by CEDA for the EUropean Facility for Airborne Research for the Environment (EUFAR) project but is also a valuable tool for the FAAM and NERC-ARSF communities. EUFAR, and its successor EUFAR2, both EC FP7 projects, bring together infrastructure operators of both instrumented research aircraft and remote-sensing instruments with the scientific user community. Data collected during the EUFAR(2) Transnational Access (TA) funded flights are accessible through the EUFAR data archive at CEDA - a centralised gateway to the data and metadata collected which links to existing online data archives and provides an archive for otherwise offline data.

The EFF search tool has been developed to provide a search facility based on this archive to maximise discovery and re-use of data. The EFF is comprised of two parts: the first harvests available data and metadata from the data files within the archive including geospatial and temporal parameters as well as auxiliary metadata specific to EUFAR data. Secondly, a web interface uses ElasticSearch functions to query these data and display the results graphically on a map/timeline along with links to the data in the archive.

The first part of developing the search tool concentrated on extracting the relevant information from within the archive structure and data files into a database which could then be queried. Although the data are stored in the community agreed standard formats of NetCDF and ENVI binary, the files to be searched vary in format/content by aircraft (there are different “flavours” from different providers even within these standard data formats). A series of scripts were developed to scan files in the data archive and extract geospatial and temporal coordinates, parameter names and key words (such as project and aircraft names). Associated information, such as flight number, filenames, archive location, data format etc. are extracted from the archive path and system files. Location polygons (hyperspectral) and sub-setted flight tracks (in-situ data) were also calculated. The extracted data are all searchable using ElasticSearch.



EUFAR Flight Finder

Search for flights within the EUFAR archive using the parameters below. Refine your search with geographical, temporal, and text search terms.

Help! (Documentation)

Tutorial Video

Click an item to expand the panel.

Filename: h235b101b_nav_post_processed.bil

Start Time: 48569.943

End Time: 48838.943

Flight Num: "101b"

Organisation: "arsf"

Get this data file

Get data for this flight

Fig 2.1.1 The EUFAR Flight Finder (EFF) allows users to search for EUFAR flight data and link to the archive.

The EFF currently includes all the EUFAR NetCDF format data (FAAM, SAFIRE aircraft). The ENVI binary data, favoured by the hyperspectral community is less rigidly structured and so far files from NERC-ARSF have been included and others will be added later. As CEDA also holds the whole archive for FAAM and ARSF, the EFF has been extended to scan all flights for these aircraft - not just those for the EUFAR project.

This ElasticSearch database is then searchable via a web interface built using HTML, JavaScript and a Google map interface. It consists of a zoomable map area (initially centred on Europe) displaying all flight tracks in that geographic area (see fig 1). In-situ measurement flights appear as lines; hyperspectral flight tracks are much shorter and appear as dots. The search can be further refined using geographical search, temporal filter, keyword search, or variable filter to locate specific variables found from within the data files. Each flight trace is clickable, and doing so gives a pop-up box showing the flight information, date etc., with links to the data in the archive. Data can then be viewed or downloaded.

The EFF was made publicly available in May 2015 (at http://flight-finder.ceda.ac.uk/) and improvements have continued since. It is useful tool for identifying flights of interest and is well received by the FAAM, ARSF and wider airborne research communities.

2.2 CLIMATE INFORMATION PORTAL FOR COPERNICUS

Sarah Callaghan, Martin Juckes

Climate change is real and its initial impacts are being felt around the world. Our society can respond by reducing greenhouse gas emissions (the cause of climate change), but this only reduces the impacts in the longer term. In the short term, Europe has to respond by adapting to the changes.

The Climate Information Portal for Copernicus (CLIPC[1]) project is developing an integrated platform of climate data services to provide a single point of access for authoritative scientific information on climate change and climate change impacts. This ambitious objective supports the Copernicus Earth Observation Programme for Europe, which will deliver a new generation of environmental data for Europe's citizens, decision-makers in the public and private sector, and academics. CEDA leads this 3-year, 22 partners, 6 million Euro project and is responsible for workpackages on climate data access, project coordination, outreach and communication.

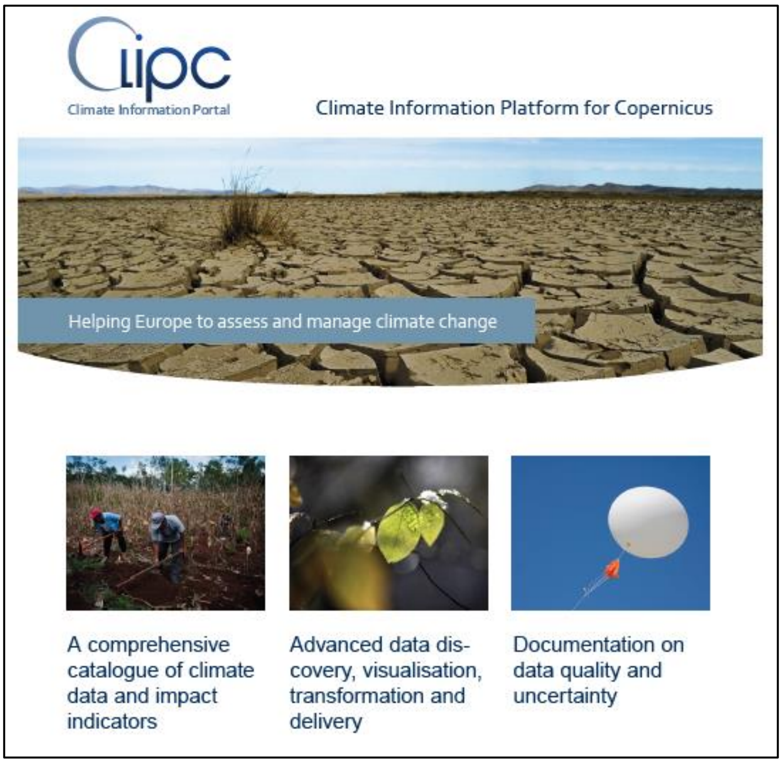

Clipc
Climate Information Portal
Climate Information Platform for Copernicus
Helping Europe to assess and manage climate change
A comprehensive catalogue of climate data and impact indicators
Advanced data discovery, visualisation, transformation and delivery
Documentation on data quality and uncertainty

Figure 2.2.1: CLIPC project overview

Deciding on the most effective and efficient response to climate change

[1] http://www.clipc.eu

requires accurate and relevant information in usable and understandable forms. Information about climate change is currently held in a number of national and international, public and private repositories, and comes from many sources, including satellite measurements, terrestrial observing systems, and computer models. These data are managed by different communities in very different formats software and hardware systems, which can be confusing. Decision makers are not interested in the individual streams but in highly aggregated products which have gone through many layers of analysis to deliver information which is meaningful in the business, social or political context.

The CLIPC project is itself a melting pot, bringing together a diverse range of climate information experts. These are split into two overlapping sectors:

1. Climate scientists and information technology specialists working harmonisation of data and access to climate datasets derived from models, observations and re-analyses (syntheses of all available observations constrained with numerical weather prediction systems).
2. Climate impact researchers developing a climate impact toolkit to evaluate, rank, and aggregate Climate Change Impact Indicators.

CLIPC services will provide:

- A single point of access for the whole range of data on climate and climate impacts allowing quicker, easier and harmonised access in a wide range of data formats;
- Access to climate data from satellite and in-situ observations, re-analyses and climate model projections and simulations on both global and regional scales;
- Integrated access to climate change impact indicators in urban, rural and water thematic areas;
- Integrated support for a wide range of users, from climate scientists to policy and decision makers;
- Supporting information about the data, its limitations and uncertainties, and guidance on how to use it, as part of a knowledge base of authoritative and expert-provided climate information;
- Services to transform and visualize data to suit the needs of different users;
- Climate Change Impact Indicators transparently linked to underlying data;
- Visualisation data transformation tools allowing flexible exploration of the whole range of data products;
- A toolkit to support aggregation of Climate Change Impact Indicators;
- A sandbox allowing users to explore variations in the parameters and data choices involved in the production of Climate Change Impact Indicators

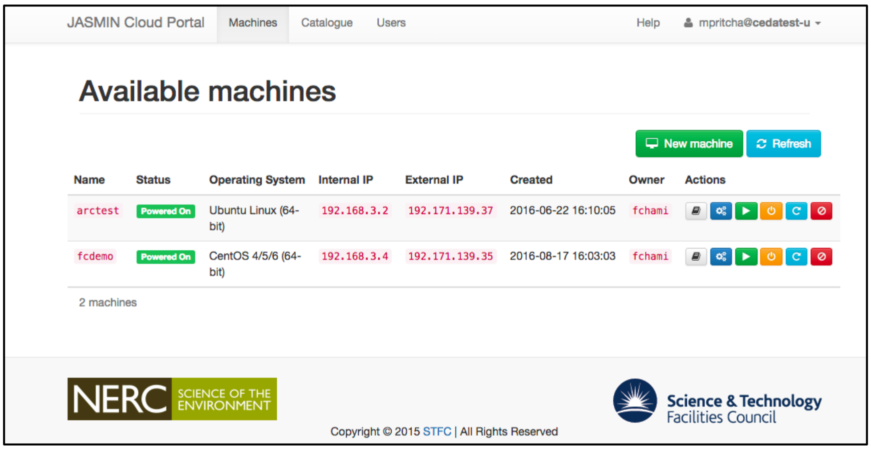
The CLIPC project is developing a portal to provide a single point of access for authoritative scientific information on climate change. This ambitious objective is made possible through the Copernicus Earth Observation Programme for Europe, which will deliver a new generation of environmental measurements of climate quality.

2.3 CONNECTING WITH JASMIN CLOUD TENANTS

Matt Pritchard

CEDA hosted the JASMIN Unmanaged Cloud Tenants' Forum on 3rd December 2015. This event brought together many of the tenants of the newly-launched JASMIN Unmanaged Cloud, and provided an opportunity to share experiences and best practice in working with this innovative new environment.

The JASMIN Unmanaged Cloud provides NERC-related science projects with “Infrastructure as a Service”: a private cloud offering where projects can be given their own “virtual organisation”. Each tenant organisation has its own allocation of compute, storage and network resources and can provision its own hosts for services within that envelope of resources. This is particularly useful for projects which need to host their own outward-facing services or portals to disseminate results or provide a custom interface to scientific work going on inside the main JASMIN infrastructure, but with the autonomy to manage those service themselves.



Available machines

Name	Status	Operating System	Internal IP	External IP	Created	Owner	Actions
arctest	Powered On	Ubuntu Linux (64-bit)	192.168.3.2	192.171.139.37	2016-06-22 16:10:05	fchami	
fcdemo	Powered On	CentOS 4/5/6 (64-bit)	192.168.3.4	192.171.139.35	2016-08-17 16:03:03	fchami	

2 machines

Copyright © 2015 STFC | All Rights Reserved

Figure 2.3.1: The JASMIN Cloud Portal enables projects to have a cloud tenancy in which they can create and manage their own virtual machines via a simple web-based interface.

The day was attended by representatives of several current tenant projects, along with staff from the JASMIN team within CEDA and STFC's Scientific Computing Department. In addition, interested parties from potential tenant projects joined to find out how this type of environment could be of benefit in their own activities.

During the day, presentations covered a number of topics including:

- an overview of JASMIN following Phase 2 and 3 improvements,
- details of the JASMIN network, explaining that the JASMIN unmanaged cloud is situated in a special part of the network with NERC-owned IP space, separate from the rest of the JASMIN infrastructure which sits within the STFC network,
- an introduction to the JASMIN cloud portal, a new tool to enable tenants to provision and manage virtual machines and other resources within their tenancy,
- presentations from tenant projects, showing how they had used the new infrastructure for the benefit of their scientific objectives, and highlighting challenges faced and how these had been overcome,
- a discussion of best practice regarding tenancy management and security

Feedback following the event suggested that attendees found it very useful as a means of finding out first-hand about how to make best use of this new infrastructure, which in turn sparked lively discussion among attendees about future requirements and potential future activities.

The event was streamed live on the internet and an edited video featuring presentations from the day can be viewed here: https://youtu.be/21yYf0avaPE. Further events are planned in the near future, including a wider JASMIN Conference planned for June 2016.

2.4 ENABLING HIGH-THROUGHPUT SCIENCE DATA TRANSFER

Matt Pritchard

JASMIN has made a huge impact on environmental science workflows by providing compute capability at scale, co-located with high-performance access to both the CEDA long-term archive and JASMIN Group Workspace storage. In this environment optimised for data-intensive scientific analysis, the move from the traditional “download and process at home” workflow to the more centralised “bring analysis to the data” workflow might suggest that less data needs to be moved around. Of course, the contrary is true, and CEDA has been at the forefront of efforts to exploit its own and JASMIN's capabilities in helping communities organise their data movement in efficient ways that enable the next generation of science projects to gain access to the data they need. Underpinning these activities are improvements to data transfer architecture, which enable “friction-free” data movement in large volumes over the long distances involved in 21st century international environmental and climate science.

CEDA's involvement with the Earth System Grid Federation (ESGF) predates JASMIN and is the type example of a community coming together to organise its data management, realising that no one centre can hold all the data needed for large-scale climate model intercomparison and analysis. A spin-off of ESGF is the International Climate Network Working Group (ICNWG http://icnwg.llnl.gov), made up of principal modelling and data centres within ESGF, focussed on providing a data transfer infrastructure which can cope with the demands of upcoming challenges presented by CMIP6, which involve efficiently replicating multi-hundred-terabyte datasets between centres distributed around the world.

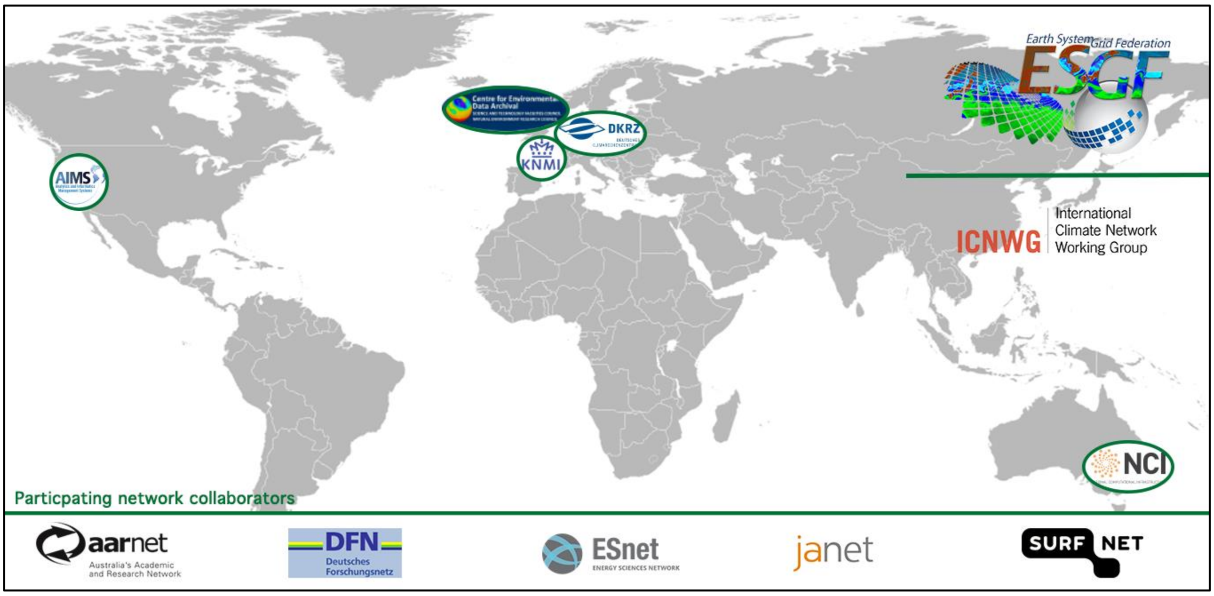

Earth System Grid Federation
ESGF
ICNWG
International Climate Network Working Group
Centre for Environmental Data Archival
DKRZ
KNMI
AIMS
NCI
Particpating network collaborators
aarnet
Australia's Academic and Research Network
DFN
Deutsches Forschungsnetz
ESnet
ENERGY SCIENCES NETWORK
janet
SURF NET

Figure 2.4.1 Partner sites in the International Climate Network Working Group (ICNWG)

Assisted by expert network engineers from ESnet in the US, the group has successfully implemented Esnet's “Science DMZ” concept at the key sites [see http://fasterdata.es.net]. This blueprint sets out a design for a distinct network zone at a site, as close as possible to the site border and hence the high-speed connection to national academic network infrastructure (in JASMIN's case, JANET). It also sets out appropriate security policies within the zone, enabling corporate/business traffic to be separated from science data transfers, so that both may proceed more efficiently: corporate firewalls are not overloaded by scanning petabytes of science data as it passes, and trusted science data flows are not subject to the unnecessary “friction” of passing through the corporate network infrastructure. In addition, the Science DMZ concept proposes monitoring tools necessary for ensuring that packet loss, latency and throughput can be characterised at key places in the network path, enabling initial diagnostics and ongoing assurance of performance for hosting robust services.

Within JASMIN, via the collaboration with ICNWG, CEDA has implemented its “JASMIN Data Transfer Zone” at the very edge of the RAL network and with security policies in place to enable key data transfer services to be located there. High-performance physical servers are used to host these services, ensuring the very best performance possible.

Services implemented so far include:

- CEDA Archive FTP server
    - Enabling download access to the CEDA archive, but from a high-performance server optimally located for best performance
- High-performance SSH-based transfer server
    - Enabling the best performance possible over familiar and convenient interfaces for moving data into and out of JASMIN group workspaces
- High-performance GridFTP transfer server
    - Enabling the sophisticated and reliable managed transfers needed for large-scale institutional data movement, but available for group workspaces users
- perfSONAR monitoring node
    - Continuous monitoring of performance parameters, with a “mesh” published to analyse performance between key sites.

The last 2 of these are prototype components of the infrastructure proposed within ICNWG to support CMIP6 replication. Further services planned for this zone include:

- ESGF data node
    - Dedicated data transfer node for ESGF data download and replication
- Sentinel Data Relay Hub
    - Interface to Sentinel Data holdings shared between key institutions involved in disseminating data from ESA's Sentinel series of satellites.

So far, the JASMIN services have seen dramatic improvements in data transfer capability. GridFTP transfers can now achieve 100-400 Mbytes/second even over international routes, but this is only achievable if the appropriate protocols are used, which are able to fill high-capacity Wide-Area Network network connections efficiently. Work will soon get underway on new documentation for JASMIN users to provide advice in selecting and configuring the most appropriate tools for the wide range of data transfers needed among the CEDA and JASMIN user community. A dedicated workshop is also planned at the upcoming JASMIN user conference in June 2016.

2.5 CEDA'S DATA CATALOGUE – CONTENT IS KEY!

Graham Parton, Kate Winfield, Wendy Garland, Poppy Townsend, Ag Stephens

Between CEDA's vast archives of over 180 million files and an international user community sits the CEDA Data Catalogue – a one-stop-shop helping users browse and discover relevant data for their research. This year a major focus for CEDA was making the catalogue content fit for purpose. Now users can not only find the proverbial dataset needle-in-a-haystack of over 4500 datasets, they can then browse via over 22 thousand connections to discover related data like never before.

Prior to 2015, CEDA's data catalogue content had been hand crafted for around 10 years, though of the 5,500 records it contained - including background records about projects, instruments and models - only the 350 coarser, overarching "dataset collection" records were thought to be fit for purpose. Though somewhat useful to users wishing to discover overall data collections, the underlying datasets with links to the archived data themselves were often of questionable reliability, had missing information or were completely wrong. In fact, in July 2015 only 23% were deemed useful. Thus, CEDA faced an immense task to overhaul the catalogue and more help was needed – which duly arrived in the shape of CEDA's Year in Industry student Kate Winfield - and thus began a year long project to review the entire catalogue and overhaul the dataset catalogue records.

Kate worked, with CEDA team members, though the 350+ data collections and their records in the catalogue. After comparing the dataset records with the parts of the archive they were reflecting, they were reviewed and, where possible, amended. Large numbers of dataset records were scrapped, replaced wholesale or simply updated whilst enhancing connections to re-usable records detailing the "Why" and "How" - such as project and instrument records – which were themselves improved or created as needed.

At the same time as the review work was underway, CEDA also catalogued large parts of the archive that had not been previously represented. This centred mainly around the 950 flights of the Facility for Airborne Atmospheric Measurements research aircraft, each with their own geospatial and temporal, project and instrumentation information that needed to be captured. To do such a task manually would be unfeasible, but thanks to the careful archiving of these data, and crucially, the background information, over the years, this task demonstrated CEDA's new approach to an ever more detailed data catalogue. Well-structured archive content enabled scriptable catalogue record generation – enhancing content reliability, archive coverage and reducing both effort and timescales for their generation. Key to this approach is the increasing ability of CEDA to harvest information directly from the archived data themselves. This uses "Elastic Search" technology, coupled with the power of parallel archive processing within the JASMIN infrastructure is enabling the entire CEDA archive to be scanned and catalogued as never before.

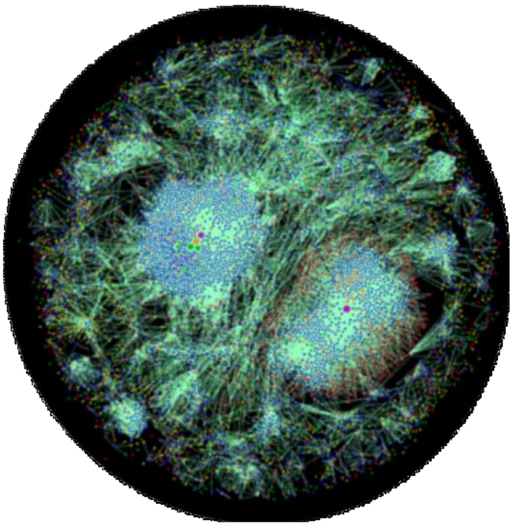


Figure 2.5.1 Visual representation of CEDA catalogue principal records and connections: 7,400 records with 22,000 connections.

So, what is the result of CEDA's catalogue content overhaul? Well, CEDA can now confidently claim that as of July 2016 nearly 99% of the catalogue's dataset records really are fit-for-purpose and with over 65,000 connections within the catalogue's 6500+ records it is one that brings new ways for users to reliably discover data in new ways never previously possible. Add to this the new paradigm of archive cataloguing, drawing on the power of Elastic Search and parallel storage and processing, the future promises an even greater enhanced data catalogue, making sure our global user community has never been so well served!

2.6 CEDA'S DATA CATALOGUE DEVELOPMENTS

Graham Parton, Ag Stephens, William Tucker

2014 saw the roll out of CEDA's new data catalogue - it was new, shiny and built to meet robust ISO-standards, but just as the progress of science goes inextricably onwards, so too do the demands placed on a modern data catalogue. 2015-16 saw some key changes to the CEDA data catalogue to ensure it keeps meeting the needs of a global user community in the big-data world.

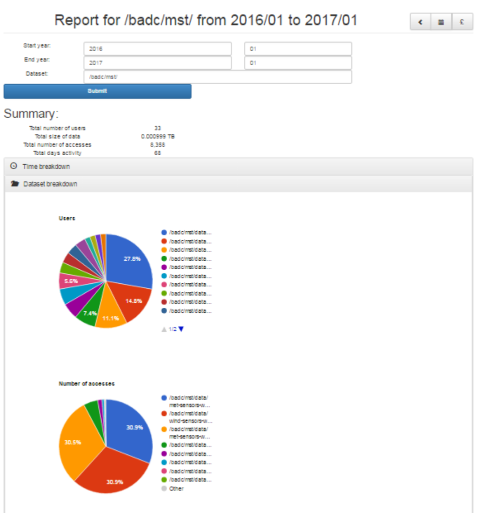
Having an ISO-standards driven catalogue where the content is known to be 99% reliable content has been a great asset for both CEDA and its user community over the past 12 months. More users have been able to discover and access data in CEDA archives, but also CEDA has been able to confidently build on this successful tool to further enhance CEDA's offering to the wider community.

Beyond ongoing adjustments to the catalogue's user experience, significant enhancements have included the integration of the CHARMe[2] annotation tool for dataset records. This brings a powerful tool for the community to annotate the catalogue contents with references to papers, comments on data quality and the like. Additionally, such comments are not confined to the CEDA catalogue, as CHARMe is a wider tool used by other archive and services, allowing greater cross-portal experiences for the user community.

CEDA's catalogue has also been further enhanced to be an underpinning service to external services. Inclusion of SKOS controlled vocabularies has ensured that the CEDA catalogue can now join services such as the data catalogues and faceted search tools within portals such as CLIPC (Climate Information Portal for Copernicus[3]).

Similarly, expanding the catalogue to cover both off-line and external resources has enabled CEDA to support projects such as the ESA CCI (Climate Change Initiative) data portal and enhancing the discoverability of physical archives held by the UK Solar System Data
Centre.

Finally, with a comprehensive coverage of the CEDA archive within the catalogue, CEDA have been able to develop a publicly available view of the archive download statistics on a dataset by dataset basis. This has been an invaluable evidence base supporting various national facilities with their own reports to funders and, when download stats are seen to dramatically improve with a more permissive access policy, has aided the wider adoption of the Open Data paradigm within CEDA's data provider community.


Report for /badc/mst/ from 2016/01 to 2017/01

Start year: 2016 01
End year: 2017 01
Dataset: /badc/mst/

Summary:

Total number of users	33
Total size of data	0.000999 TB
Total number of accesses	8,358
Total days activity	68

Figure 2.6.1 Public download stats pages now available based on CEDA dataset catalogue records

Of course, the development of the catalogue will continue in the next 12 months with plans to connect the catalogue to other CEDA services such as CEDA's Elastic Search indexes and enhanced search interfaces. Such developments will ensure that CEDA continues to deliver enhanced user services and support other data portals in the years to come.

[2] http://charme.org.uk/

[3] http://www.clipc.eu

PART 2: CEDA AND JASMIN SUMMARY INFORMATION 2015-2016

The original 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 access services. As in previous years, CEDA enhanced this role through the development and maintenance of tools and services to aid data preservation, curation, discovery and visualisation — adding value for the world-wide user community.

In recent years this role has further expanded to include support for the JASMIN data analysis environment, so this section of the annual report has been reorganised to present not only summaries of CEDA usage, but also of JASMIN.

3. USAGE OF CEDA DATA

CEDA delivers the British Atmospheric Data Centre (BADC) for the National Centre for Atmospheric Science (NCAS), the NERC Earth Observation Data Centre (NEODC) for the National Centre for Earth Observation (NCEO), the UK Solar System Data Centre (UKSSDC) and the IPCC Data Distribution Centre for the Intergovernmental Panel on Climate Change (IPCC).

(Note that additional metrics also appear in the data centre metrics section, chapter 7).

Annual CEDA Usage: April 2015 to March 2016	
Total number of users	11,231
Total data downloaded	759 TiB
Total number of accesses	13,405,658
Total days activity	63,044

Table 3.1: Summary figures for usage by CEDA consumers during the reporting year

These figures can be broken down by month showing that while usage in terms of number of users and accesses has remained static, there is a general trend upwards in the amount of data downloaded (table 3.2 and figure 3.1)

Month	Users	Methods	Datasets	Number of accesses	Size (GiB)	Activity days
2015 04	1168	8	795	1,152,934	31,023.137	3,678
2015 05	1267	7	1,480	666,381	34,477.167	5,134
2015 06	1291	8	850	887,870	34,031.600	4,200
2015 07	1291	8	673	1,236,670	60,871.947	4,108
2015 08	1118	9	744	2,039,387	56,686.726	4,689
2015 09	1182	8	791	1,486,022	39,621.881	5,394
2015 10	1398	9	962	1,205,939	79,100.321	7,381
2015 11	1308	9	1,571	1,144,812	111,372.573	6,376
2015 12	1133	10	1,476	618,801	93,034.070	4,621
2016 01	1538	10	1,126	1,099,511	83,278.529	5,220
2016 02	2047	11	929	999,559	90,682.352	7,002
2016 03	1597	10	1,056	867,772	63,698.921	5,241

Table 3.2: Monthly summary figures for usage by CEDA consumers during the reporting year

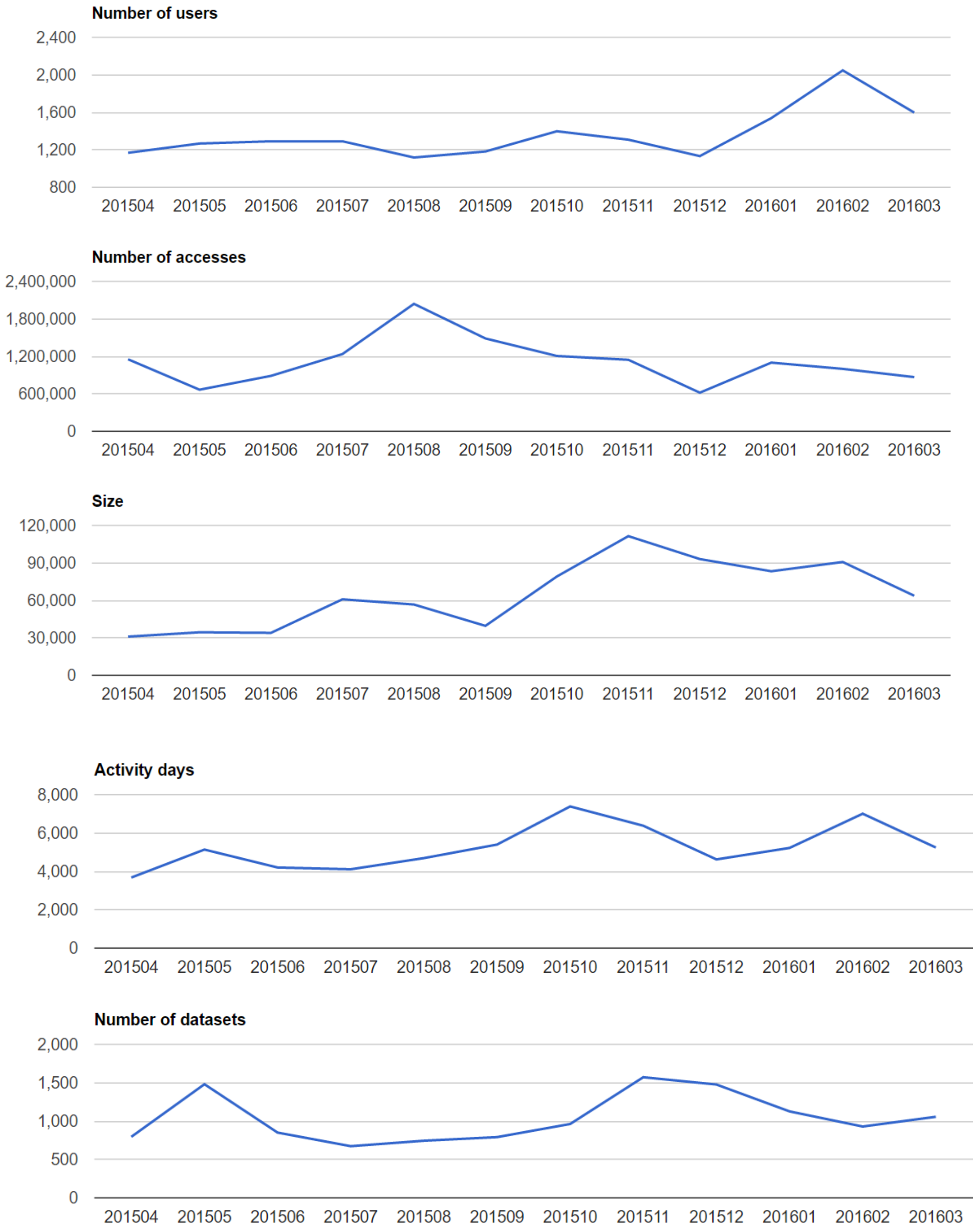


Figure 3.1: Breakdown of CEDA usage by month

4. JASMIN

JASMIN is a custom integrated data intensive super-computer providing four main functions: storage, batch computing, managed infrastructure compute and a private cloud to provide Infrastructure-as-a-Service (the JASMIN “un-managed cloud”).

4.1 STORAGE

During this year, Panasas storage from JASMIN procurement Phase 3 was brought into production increasing usable capacity of “fast disk" by 4.4 PB, to a total of 15.1 PB. In addition, 0.9 PB bulk storage (NetApp) was made available for use, primarily for the JASMIN cloud infrastructure. 50TB Dell hybrid iSCSI storage was also brought online to provide better performance for database applications.

At the start of this reporting year there were 69 Group Workspaces for JASMIN and 32 for CEMS with allocations totalling 3.4 PB and 2.2 PB respectively, serving a total community of 719 users. By the end of the period this had increased to 79 JASMIN and 34 CEMS group workspaces in use with allocations of 4.6 PB and 2.4 PB respectively, serving a total community of 914 users.

The Elastic Tape service, a near-line tape storage facility available to JASMIN and CEMS Group Workspace managers, continued to provide a flexible and scalable complement to “fast disk” storage, enabling GWS managers to optimise their use of online disk resource.

4.2 BATCH COMPUTING

Nearly 400 compute cores with associated 9TB RAM were brought into operation in April 2015. In addition, two super-high-memory (2TB) nodes were deployed as LOTUS nodes for specialist demanding applications. An ingest mini-cluster was made available for some of CEDA’s own data ingestion processes, although this has yet to be brought in to service.

In an effort to improve “fair-share” scheduling on LOTUS, a one-hour default run time was introduced in August 2015. The maximum run time is now 7 days but users need to specify if the anticipated job run time will exceed 1 hr.

4.3 MANAGED COMPUTE INFRASTRUCTURE

Two further 2TB nodes were deployed as (physical) scientific analysis servers for general use: jasmin-sci3 and cems-sci2, helping to ease the increasing workload on existing servers but also providing an environment for test runs of high-memory code.

Additional services were added to JASMIN’s “Science DMZ” or “Data Transfer Zone”: this is an area of network situated as close as possible to the RAL site border, with high-performance hardware and appropriate security policies to support high-throughput data transfer services. This already included a high-performance version of the standard transfer server, but work during this period has added a second instance tuned for high-latency (long path) transfers for intercontinental transfers, and a gridftp server to support secure transfers including those managed via Globus Online.

Other improvements included upgrading the main uplink from JASMIN route to RAL site from 2x10Gbit/s to 2x40Gbit/s (Sept 2015), and bringing the Icinga2 monitoring & alerting system in to production, providing a framework more able to cope with the scale of JASMIN and its component services.

4.4 JASMIN CLOUD

A new JASMIN "dev/ops" software developer, Matt Pryor, joined CEDA in July 2015 to complete work on the JASMIN could portal and support cloud tenants. The cloud portal was brought into operation over in summer/autumn 2015 and provides tenancy managers with a simplified interface with which to manage virtual machines in their tenancy, acting on top of the more complex VMWare vCloud director.

A JASMIN "Unmanaged Cloud Tenants' forum" was hosted at RAL in December 2015, attended by current & potential cloud tenants. This event provided an opportunity for the JASMIN team to demonstrate the infrastructure & management interfaces, and advise on best practice for system admin in the cloud environment, while enabling current tenants to share their experiences of using the JASMIN cloud within their projects. Further similar events are likely to be planned as a means of fostering closer interaction with and between cloud tenants.

4.5 USER SUPPORT AND OUTREACH

A new JASMIN user support scientist, Fatima Chami, joined CEDA in December 2015. Her work includes support for users of batch and interactive compute, and initial tasks included a review of JASMIN public documentation to coincide with the move to the "HelpScout" helpdesk management system. In addition, her work will focus on monitoring and improving LOTUS queues, providing advice about parallel compute, and helping with logistics and the preparation of training materials for the upcoming JASMIN user conference in June 2016.

The Harwell Open Days in July 2015 saw over 15,000 members of the public visit RAL on the public day (Saturday), with numerous school and stakeholder visits in the preceding week. CEDA staffed a stand with information about its activities including its use of JASMIN, and STFC's Scientific Computing Department gave a machine room tour (including the JASMIN infrastructure) to over 700 people, which included the opportunity to handle sample components in a "computer petting zoo".

5. COLLABORATIONS

CEDA continues to support international climate modelling community 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).

5.1 MAJOR COLLABORATIONS

In 2015/2016, 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:

1. Working closely with the other NERC Environmental Data Centres, as part of the NERC Data Operations Group.

2. 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 sixth Coupled Model Intercomparison Project (CMIP6).

3. A leading partner in many major European projects including IS-ENES 2 (developing an InfraStructure for a European Network for Earth system Simulation and CLIPC (Climate Information Portal for Copernicus).

4. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.

5. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) portal.

6. Working with commercial and academic partners and the Satellite Applications Catapult, 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.

7. CEDA is part of the UK Collaborative Ground Segment for Sentinel data (with UKSA, Airbus, Satellite Applications Catapult) with the role to provide Sentinel data mirror archives and data processing capability for the UK academic community.

6. FUNDING

In addition to supporting the National Centres of Atmospheric Science and Earth Observations (NCAS and NCEO, research centres of the Natural Environment Research Council, NERC), CEDA also delivers major projects with funding from a range of other bodies, including work for the European Space Agency (ESA), JISC, DEFRA and others, as well as participating and coordinating major European projects.

6.1 ANNUAL TOTAL FUNDING

	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16
NCAS income	970	866	906	883	935	829	829	808
NCEO income	378	389	450	419	445	392	390	393
Other NERC	788	481	341	527	287	272	600	621
Other income	461	710	1144	1099	1283	1486	1394	1505
Total income	2597	2446	2841	2928	2950	2979	3213	3327

Table 6.1: Overall funding for CEDA for financial years 2008 — 2009 to 2015 — 2016 (in £k)

Most of this funding comes to CEDA via a service level agreement (SLA) between the Natural Environment Research Council (NERC) and the Science and Technology Facilities Council (STFC). This SLA now covers both CEDA and JASMIN support explicitly.

6.2 EXTERNALLY FUNDED PROJECTS FOR THE YEAR 2015-2016

CEDA participates in projects funded by ESA, EC, Met Office, DECC and others. Funding sources and figures for this year are shown below.

Name	Description	Funder	Start date	End date	Value
CEMS support to ESA CCI SST		ESA	01/05/2014	31/03/2017	£72.9k
CLIPC	Climate Information Portal for Copernicus	FP7	12/12/2013	31/12/2016	£676k
ESA CCI Data Portal		ESA	01/05/2015	31/03/2018	£195.3k
ESA Optirad	OPTIRAD (OPTImisation	ESA	01/04/2014	31/01/2018	£22k

	environment for joint retrieval of multi-sensor RADiances) aims to advance the state of the art in EO data assimilation in land surface processes.				
EUFAR-2	The European Facility for Airborne Research (EU-FAR) aims at coordinating the operations of the European fleet of instrumented aircraft in the field of environmental research in the atmospheric, marine, terrestrial and Earth sciences.	FP7	01/02/2014	31/01/2018	£174k
EUSTACE	EU Surface Temperature for All Corners of Earth	H2020	01/01/2015	30/06/2018	£142.9k
FIDUCEO	Fidelity and uncertainty in climate data records from Earth Observations http://www.fiduceo.eu/	H2020	01/02/2015	31/03/2015	£102.5k
BACI	Towards A Biosphere Atmosphere Change Index http://baci-h2020.eu/index.php/	H2020	01/01/2015	31/10/2018	£38.5k
IS-ENES2	InfraStructure for the European Network for Earth System Modelling	FP7	01/04/2013	28/02/2017	£551k
JASMIN Support		Met Office	01/06/2015	31/03/2016	£40k
SeaDataNet2	CEDA is contributing expertise with INSPIRE and ISO/OGC standards to enable the SeaDataNet infras- tructure to become compliant and interoperable with wider initiatives.	EC FP7	01/04/2012	30/09/2015	£41k
SPECS Data Archive	Data archive for the SPECS FW7 seasonal prediction	Institut Catala De Ciencies Del Clima	01/08/2013	31/10/2016	£121k
UKCP09 User Interface 1516			01-Apr-15	31/03/2016	£61.1k
Forestry TEP		ESA	01/01/15	31/07/17	£125.2k
Monsoon Overflow 2	Providing support for MONSooN users on JASMIN	Met Office	01/04/14	31/03/17	£199.6k
Optirad-2 SY-4Sci Synergy		ESA	01/04/15	31/12/16	£10.9k
EO4CDS Trial Plan Support		UKSA	01/12/2015	31/03/2016	£7.2k

PRIMAVERA		H2020	1/1/2015	31/10/2019	£141k
ACCLIMATISE SHELL	Climate Consultancy to generate a high-quality set of climate statistics from the CMIP5 and CORDEX	commercial	31/3/2015	31/3/2016	£59.1k
DECC Climate Archive Support Extension		DECC	1/4/2015	31/3/2016	£260.5
EO4CDS Phase 2		UKSA	21/10/2015	31/3/2016	£45k
ESA Sentinel Data Hub Relay		ESA	1/5/2015	30/10/2017	£311.2k
MEDIN Metadata services		MEDIN	1/4/2015	31/3/2016	£16.3k

Table 6.2: Externally funded projects for 2014-2015 (non-core NERC)

PART 3: METRICS AND PUBLICATIONS

7. ADDITIONAL DATA CENTRE METRICS

CEDA is required to provide metrics quarterly in a number of categories. Some additional metrics to those provided in Chapter 2 are provided here.

Note that a considerable amount of use of CEDA is by users on JASMIN, who would not be measured in most of these statistics because the data is directly available on the file system.

7.1 ACCESS RELATED METRICS

We can break down the users accessing registered datasets by geographical origin and institute type.

Area	Q1		Q2		Q3		Q4	
UK	2775	65%	2869	64%	2931	63%	2903	62%
Europe	462	11%	488	11%	516	11%	528	11%
Rest of the world	983	23%	1053	24%	1137	24%	1162	25%
Unknown	71	2%	74	2%	78	2%	83	2%

Table 7.1.1: Users by area

Institute Type	Q1		Q2		Q3		Q4	
University	3059	71%	3176	71%	3281	70%	3271	70%
Government	694	16%	722	16%	760	16%	767	16%
NERC	201	5%	230	5%	242	5%	253	5%
Other	248	6%	263	6%	278	6%	283	6%
Commercial	46	1%	48	1%	51	1%	55	1%
School	39	1%	41	1%	46	1%	43	1%

Table 7.1.2: Users by Institute type

7.2 DATA HOLDINGS

Data Centre	Q1	Q2	Q3	Q4
NEODC	778	796	811	1330
BADC	2381	2385	2379	
UKSSDC	30	30	30	

Table 7.2.1: Number of “dataset” discovery records held in the NERC data catalogue service. (Note that in Q4 CEDA stopped differentiating between data centres for records pushed to the data catalogue service.)

	Q1	Q2	Q3	Q4
datasets	2942	2963	2714	2202
collections	309	315	315	338

Table 7.2.2: Number of “dataset collections” and “datasets” identified by CEDA and displayed via CEDA catalogue.

7.3 HELP DESK RESPONSIVENESS

	Q1	Q2	Q3	Q4
Received	415	386	393	362
% Closed within 3-days	71	62	63	59
Closed	421	396	407	374

Table 7.3:1 Help desk queries received and closed by quarter, including the three-day closure rates. These queries cover all aspects of data support except dataset access issues.

	Q1	Q2	Q3	Q4
Received	932	872	1111	1096
% Closed within 3-days	84	82	93	92
Closed	838	868	1108	1101

Table 7.3.2: Help desk queries specifically about access authorisation for restricted CEDA datasets and services received and closed by quarter, including the three-day closure rates.

8. PUBLICATIONS AND PRESENTATIONS

Bennett, V. L., Conway, E. A., Waterfall, A. M., Pepler, S.; 'Between Oais and Agile a Dynamic Data Management Approach' American Geophysical Union, Fall Meeting 2015, abstract #IN22A-07

Bennett. V.L. , M. Juckes, P. Kershaw, B. Lawrence, A. Stephens, M. Pritchard, S. Pepler, 'Providing an Analysis Environment with Access to High-Volume Simulation and Observational Data for Climate Science Bennett', American Geophysical Union, Fall Meeting 2015, abstract #IN31A-1752

Bennett, V.L. , S. Donegan, J. Walker, A. Waterfall 'Mirroring Sentinel Data for UK Researchers'; Presentation at "Earth Observation in the Sentinel ERA" RSPSOC, NCEO and CEOI-ST Joint Annual Conference, 8-11 September 2015, Southampton UK

Clarke, H., E. Pechorro, V. Bennett, C. Farquhar, J. Blower, 'CCI Open Data Portal', Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.374

Clifford, D., R. Alegre, V. Bennett, J. Blower, C. DeLuca, P. Kershaw, C. Lynnes, C. Mattmann, R. Phipps, I. Rozum; Capturing and sharing our collective expertise on climate data: the CHARMe project; Bulletin of the American Meteorological Society 2016 97 (4), doi: http://dx.doi.org/10.1175/BAMS-D-14-00189.1

Donegan, S.; Bennett, V.; Waterfall, A.; Kershaw, P.; Williamson, E., 'Sentinel, Climate & EO datasets and the JASMIN "super-data" cluster at CEDA' ; ESA Living Planet Symposium, Prague, 9-13 May 2016, Paper 1713

Juckes,M., Rob Swart, Lars Bärring, Annemarie Groot, Peter Thysse, Wim Som de Cerff, Luis Costa, Johannes Lückenkötter, Sarah Callaghan, Victoria Bennett, A Climate Information Platform for Copernicus (CLIPC): managing the data flood, EGU General Assembly Conference Abstracts Vol.18, 2016/4, p15396, http://adsabs.harvard.edu/abs/2016EGUGA..1815396J

Juckes, M. , Rob Swart, Lars Bärring, Annemarie Groot, Peter Thysse, Wim Som de Cerff, Luis Costa, Johannes Lückenkötter, Victoria Bennett, Sarah Callaghan, Communicating across the disciplines to support climate

services: the CLIPC portal, EGU General Assembly Conference Abstracts Vol.18, 2016/4, p16255, http://adsabs.harvard.edu/abs/2016EGUGA..1816255J

Kershaw, P., Jonathan Churchill, Stephen Pascoe, Matt Pritchard, Bryan Lawrence, JASMIN Cloud, ESGF Face to Face Meeting, Livermore CA, Dec 2014 Kershaw, Philip, Bryan Lawrence, Jose GomezDans, and John Holt, Cloud hosting of the IPython Notebook to Provide Collaborative Research Environments for Big Data Analysis, EGU2015, Vienna, Apr 2015

Lawrence, B., Victoria Bennett, Jonathan Churchill, Martin Juckes, Philip Kershaw, Sam Pepler, Matt Pritchard, and Ag Stephens, Beating the tyranny of scale with a private cloud configured for Big Data, EGU2015, Vienna, Apr 2015

Parton, G., Making FAAM Flights Discoverable. Presentation 28/04/2016. NCAS Data Meeting on archiving and visualising observations. University of Reading.

Parton, G., CEDA Developments. Presentation 07/11/2016. NCAS Data Meeting on archiving and visualising observations. University of Leeds

Parton, G.A., Steven Donegan, Stephen Pascoe, Ag Stephens, Spiros Ventouras, Bryan N Lawrence. MOLES3: Implementing an ISO standards driven data catalogue. International Journal of Digital Curation, 2015, Vol. 10, No. 1, pp. 249-259. doi:10.2218/ijdc.v10i1.365

Pritchard, M. and Jonathan Churchill, JASMIN : Petascale storage and terabit networking for environmental science, Presentation at JANET Networkshop 43, University of Exeter 31/3/2015 https://networkshop.ja.net/events/networkshop43/programme/7689/1415

Pritchard, M., and J Churchill, JASMIN: petascale storage and terabit networking for environmental science. Oral presentation at JISC Networkshop43, 31 March 2015 2 April 2016. https://www.jisc.ac.uk/events/networkshop43-31-mar-2015/programme

Pritchard, M. , J Churchill. Enabling high-performance access to big data from space, Oral presentation at ESA 2016 conference on Big Data from Space 15-17 March 2016. https://ec.europa.eu/jrc/en/publication/proceedings-2016-conference-big-data-space-bids16

Schutgens, N., Philip Stier, Philip Kershaw, and Stephen Pascoe, Comparing apples and oranges: the Community Intercomparison Suite, EGU2015, Vienna, Apr 2015

Waterfall, A., V. Bennett, S. Donegan, M. Juckes, P. Kershaw, R. Petrie, A. Stephens, and A. Wilson; ’Big Data Challenges indexing large-volume, heterogeneous EO datasets for effective data discovery’; Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.29