Annual Report 2018
(April 2017 to March 2018)

Victoria Bennett, Sarah Callaghan,
Poppy Townsend (Editors)
CONTENTS

1. Introduction .......................................................................................................................................................3

Part 1: Highlights and Major Activities ..............................................................................................................3

1. Highlights .......................................................................................................................................................3

1.1. Standards for NCAS instrument data .........................................................................................................4

1.2. CMIP6 Data Request ...................................................................................................................................5

1.3. Automated Cataloguing of CMIP5 ................................................................................................................6

1.4. Integrated Data Indexing at Scale ..................................................................................................................7

1.5. Delivering the UK Component of the IPCC Data Distribution Centre (DDC) ......................................8

1.6. Engaging with the Copernicus Climate Change Service (C3S) .................................................................9

1.7. Data management for EUFAR2 ....................................................................................................................10

1.8. Sentinel data retrieval at CEDA ....................................................................................................................11

1.9. CEDA Satellite Data Finder ........................................................................................................................12

1.10. Near-line Archive now operational ...........................................................................................................13

1.11. JASMIN Phase 4 .........................................................................................................................................14

1.12. JASMIN Cloud and Accounts Portal ..........................................................................................................15

1.13. RCUK Cloud Working Group and Workshop ............................................................................................16

1.14. Improving communications with data providers ......................................................................................17

1.15. Data Science Journal ..................................................................................................................................18

1.16. CEDA graduates collaborate with partner organisations ........................................................................19

Part 2: CEDA and JASMIN Summary Information 2017-2018 ...........................................................................20

2. Usage of CEDA data ........................................................................................................................................20

3. JASMIN ..........................................................................................................................................................22

3.1. Storage .......................................................................................................................................................22

3.2. Managed compute Infrastructure ................................................................................................................22

3.3. Network and supporting infrastructure ......................................................................................................23

3.4. JASMIN Cloud ............................................................................................................................................23

3.5. User support and outreach ...........................................................................................................................24

4. Collaborations ................................................................................................................................................25

4.1. Major collaborations ...................................................................................................................................25

5. Funding and governance .................................................................................................................................25

5.1. Annual total funding ...................................................................................................................................25

5.2. Externally funded projects for the year 2017-2018 ....................................................................................26

5.3. Governance ................................................................................................................................................28

Part 3: Metrics and Publications ........................................................................................................................29
1. INTRODUCTION

The Centre for Environmental Analysis (CEDA) is based in the Science and Technology Research Council (STFC)’s RAL Space department. CEDA operates data centres and delivers data infrastructure, primarily for the Natural Environment Research Council (NERC), and undertakes project work for a range of national and international funders. CEDA’s mission is to provide data and information services for environmental science: this includes curation of scientifically important environmental data for the long term, and facilitation of the use of data by the environmental science community.

CEDA was established in 2005, as a merged entity incorporating two NERC designated data centres: the British Atmospheric Data Centre, and the NERC Earth Observation Data Centre. Consolidated annual reports have been produced since 2009. This annual report presents key statistics for the year past (2017-2018) as well as a series short reports showcasing a cross section of our activities. Key metrics are also provided.

JASMIN is the data intensive supercomputer which provides the infrastructure upon which CEDA and the CEDA services are delivered. Usage of JASMIN in the community continues to grow and the role of CEDA staff continues to evolve to include services and support for users of increasingly large and complex datasets. This year saw a considerable investment in JASMIN Phase 4, allowing us to refresh old hardware and significantly increase capacity, in storage and computing, but also in the cloud capabilities offered to users.

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.

Ensuring access for UK researchers to the unprecedented volumes of Earth Observation Data from the European Sentinel satellites continues to be a major focus for CEDA. Similar, if not larger, volumes of climate model data for CMIP6 (Coupled Model Intercomparison Project, Phase 6) will follow soon, and preparations for this are well underway. Other highlights this year include JASMIN’s second user conference, major upgrades to our systems and methods for cataloguing and indexing our data, and a range of external projects which rely on our experience and knowledge in data management, data standards and data services.

As in previous years, our key partnerships include our sister department in STFC, Scientific Computing Department, with whom we deliver the JASMIN infrastructure, and internationally the European Network for Earth Simulation (our collaborators in delivering the European component of the Earth System Grid Federation), and many other project collaborators.

I hope this annual report provides an interesting insight into another successful year at CEDA,

Victoria Bennett, Head of CEDA, RAL Space

PART 1: HIGHLIGHTS AND MAJOR ACTIVITIES

In this section we provide a selection of descriptions of key activities from the year. We have included highlights selected to showcase some CEDA activities supported through different funding streams, and a range of key areas of focus for CEDA staff this year.

1. HIGHLIGHTS
1.1. STANDARDS FOR NCAS INSTRUMENT DATA
Ag Stephens, Wendy Garland & Graham Parton

In 2016, the NCAS Atmospheric Measurement Facility (AMF) started a major overhaul of the way that it manages its instrument data. In response, CEDA provided expertise and tooling to support the process to ensure that standards and best practices were followed. Here we describe how this ground-breaking activity will benefit both data providers and end-users for years to come.

Putting Data Management at the forefront of instrument science
By appointing a data facilitator NCAS AMF took an impressive decision to put data management at the core of its science. Whilst scientific projects are often data-centric it is still rare that the full data life-cycle is mapped out from the start. AMF had the foresight to realise that the final data generated by its instruments is the lasting product. The approach taken involved: (1) talking to all instrument/data scientists, (2) discussion with standards experts, (3) development of templates/rules for all data types, and (4) provision of example files.

CEDA was able to provide support in many ways for the emerging “AMF standard”, modelled in a NetCDF file format. We provided detailed input into the data structures, naming conventions and good practice regarding metadata. We were able to advise on the experiences and standards used in other areas and domains (such as FAAM and CMIP5), data interoperability and compliance with the CF Conventions.

One of strongest features of the AMF data project was the use of over 50 structured spreadsheets to define the variables, dimensions and global attributes for each data product. CEDA spotted an opportunity to automate the process of generating a set of controlled vocabularies (CVs) from this information (figure 1). The CVs (figure 2) could then be used to connect to a new compliance-checking framework being prototyped at CEDA.

A set of checks were produced to parse the CVs and to validate data files against the expected contents provided in the initial spreadsheets. The deployment of these checks will ensure that all data archived at CEDA matches the detailed specification of the project.

Although considerable initial input is required from data providers to meet these new specifications, benefits include the ability to develop general tools to aid the process and provision of support at all stages in the data lifecycle. The end result is a set of fully documented, self-contained, format compliant, quality-checked datasets that can be discovered and exploited by the wider community. In turn, the complete and reliable nature of the products will raise the profile of AMF in terms of quality assurance and data reusability.

References
Compliance checking library: https://github.com/cedadev/compliance-check-lib
NCAS AMF Controlled Vocabularies (prototype): https://github.com/agstephens/AMF_CVs

Figure 1: Auto-generation of structured vocabularies

Figure 2: Example JSON vocabulary for “time” variable
1.2. CMIP6 DATA REQUEST

Martin Juckes

The CMIP6 data request provides a consolidated database specifying the data requirements of 22 Model Intercomparison Projects (MIPs), which have been endorsed to participate in the Coupled Model Intercomparison Project 6 (CMIP6). The database gives the technical specifications of around 2000 variables, including detailed metadata specified in the framework of the Climate and Forecast Convention (CF).

Building on our experience of data standards and climate data management, CEDA has developed methods to standardise, and formalise the community’s data requirements. Being involved at this early stage of CMIP6 enables us to prepare for data delivery in advance: data request tools will be reused when data arrives, and we have a reliable insight into the planned data production. This is part of a package of activities through which CEDA contributes to the global effort to provide infrastructure for CMIP: CEDA also benefits from related contributions from peer institutions, coordinated through the WGCM Infrastructure Panel (WIP) and the European Network for Earth System Modeling for Climate (ENES).

CMIP6 follows on from the successful CMIP5 intercomparison project, with the new innovation of incorporating multiple independent science consortia. It has been stipulated that the data requirements from all these independent teams should be consolidated into a single “request” [Figure 3].

The data request provided by CEDA includes a schema specification, defining the structure of the database, an XML document containing the database, a python package providing both a library API and a command-line interface, and a web page allowing the request to be searched and browsed (Juckes et al. 2019, in preparation).

A simplified view of the database structure is given in figure 4. The request specifies the variables which should be provided from each experiment to support specified scientific objectives.

The links are made through the Request Item concept, which provides a 3 way link between a MIP, a group of variables and a set of experiments.

There are a number of additional options, such as setting priorities for variables, defining tiers of experiments, and specifying output requirements for limited time slices. These options are designed to ensure that modelling centres can determine the data which they should produce in a consistent way. See http://w3id.org/cmip6dr for more details.

---

1 The CMIP6 Data Request, Geosci. Model Dev, in preparation.
1.3. AUTOMATED CATALOGUING OF CMIP5

Graham Parton, Charlotte Pascoe & Ag Stephens

The 5th Coupled Model Intercomparison Project (CMIP5) involved over 20 modelling centres running more than 50 experiments to generate a multi-petabyte climate data archive. When attempting to catalogue extensive data holdings CEDA were faced with a problem: either create a handful of high-level records (easy, manual, but uninformative) or attempt to generate a detailed set of records that will provide documentation and allow discovery at a level appropriate to scientific usage. One of the oldest questions in Data Management is "what is a data set?". We can perform some analysis based on the directory structure to find this out. For CMIP5, we defined 3 options:
1. ../../cmip5/<product>/<inst>: (~25) per institution
2. ../../cmip5/<product>/<inst>/<model>: (~65) per institution/model
3. ../../cmip5/<product>/<inst>/<model>/<expt>: (~1200) per institution/model/experiment

A Python software library (figure 5) was developed at CEDA to tackle the two sides of the problem. The first dealt with scanning every single NetCDF file in the archive to identify all the institutions, models, experiments, ensemble members and variables. This information was aggregated up to the level agreed for option 3 and written to a JSON (JavaScript Object Notation) cache, to avoid frequent re-scanning of around 2,000,000 files. The second combined a set of known configuration information and vocabularies (common to all of CMIP5), with code to create and update records in the CEDA Catalogue. This part of the library was able to read from the JSON cache and then generate records for the catalogue.

The automated process can be "idempotent", this means that if you run the code again and again you will always get the same outcome. This is a very powerful feature that allows incremental improvements to be made and wholesale changes to be rolled out across multiple catalogue records in a consistent manner. The biggest gain is the scalability of the process. The number of records is irrelevant once the code is set up and configured. Furthermore, CEDA has been able to identify a range of other projects for which this approach is suitable and the library is being adapted and re-used (e.g. for FAAM aircraft records).

Status and the future

The success of this approach is demonstrable on the CEDA Catalogue interface when searching for CMIP5. It has provided CEDA with a mechanism for expanding its metadata infrastructure that can scale with the number and complexity of projects that we work with. The software library is now available as an open source tool, known as MIP-to-MOLES3, which is being adapted to support other large MIPs (Model Intercomparison Projects) including CMIP6 and SPECS. The library includes a common framework that can be configured to meet the needs of new complex projects that CEDA would otherwise struggle to catalogue in detail.

References:
2 CMIP5 via the CEDA Catalogue: http://catalogue.ceda.ac.uk/?q=cmip5
3 The MIP-to-MOLES library: https://github.com/cedadev/mip-to-moles
Vast data archives are a fundamental resource for the research community, but unless users are provided with the right tools to make them usable their true potential can never be fully harnessed. Bridging the gap between what a user wants to find and the sheer scale and pace of growth within the CEDA archive became possible in 2017 thanks to the deployment of ElasticSearch on the JASMIN platform. It was coupled to the CEDA data catalogue and led to dramatic improvements in the search capability.

Decades ago, CEDA staff painstakingly catalogued data held in the CEDA archive by hand - an approach that suffered from incorrect or missing information, becoming quickly out of date and unable to scale with data growth rates which have been increasing year on year. However, more metadata (data about the data) standards have been gradually adopted, both by data providers at the file level and by CEDA through cataloguing practices. Thus, the groundwork has been laid over the years for a new approach, allowing large scale file-to-catalogue information to be harvested, enabling CEDA to deliver detailed cataloguing that is once again fit for purpose.

Underpinning CEDA’s new approach are two key technologies that work in tandem: ElasticSearch and JASMIN.

ElasticSearch is a “no-SQL” indexing and search technology widely used by industry (e.g. VW, Ebay and Rightmove) with components designed to work with geo-spatial information, making it a perfect tool for CEDA. To populate a file-level index we have built a suite of tools able to analyse files in the CEDA archive, collating basic file information (size, location etc) as well as harvesting details such as parameter information from within the files themselves.

However, populating the index was another matter; with over 180 million files needing to be scanned and more arriving by the hour! To tackle this, CEDA leveraged the JASMIN infrastructure’s parallel access and processing power, within which the CEDA archive resides. Thus delivering on a previously unthinkable task of completing and maintaining a whole-archive scan within a matter of days. This now provides over 5,500 CEDA catalogue datasets with accurate, timely and detailed information such as parameters, total volumes and file format information, with a proven path to enable harvesting richer details still: accurate geo-temporal information.

To complete the overhaul of the CEDA catalogue, the search service was replaced with one also utilising ElasticSearch. This brings a fast and accurate search tool to users with capacity for future development too, building naturally on ElasticSearch’s main purpose: powerful, faceted geo-temporal searching.

With great advances already achieved in 2017 and additional search facets to come in 2018 the CEDA archive index and catalogue is now delivering services fit for users in the 21st Century.
1.5, DELIVERING THE UK COMPONENT OF THE IPCC DATA DISTRIBUTION CENTRE (DDC)

Charlotte Pascoe, Martin Juckes & Ag Stephens

CEDA helps to support and deliver the UK component of Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre (DDC). We provide data curation services and expertise for the IPCC through data access and support services via the DDC web site, supporting the development of related policy documents. CEDA has carried out this activity since 2007; some statistics and highlights from this year are shown below:

Web site usage has remained steady in 2017 with 175,088 page views, a marginal 4% decrease compared to 2016 usage [1,2,3]. Figure 7 shows the geographical distribution of sessions in 2017 and reveals substantial interest from Asia (more details in the latest DDC usage report [3]). Progress has been made on the rationalisation of the web site, moving information into the CEDA catalogue pages for robust long term curation and migrating static web pages into a cloud provider (github) to ease collaborative maintenance. Dynamic pages (e.g. the visualisation at www.ipcc-data.org/maps/) will be moved to a separate server maintained at CEDA within the JASMIN facility for ease of maintenance and efficient access to data files held in the CEDA archive.

Stakeholder engagement developed through many telephone conferences and 3 meetings during the reporting period. This included the IPCC AR6 Data Workshop, Hamburg, September, 2017, where coordination between DDC partners and the technical Support Units of the IPCC Working Groups was discussed. It was concluded that two DDC representatives, one each from CEDA and the German partner DKRZ, should attend the “Intergovernmental Panel on Climate Change (IPCC) Expert Meeting on Assessing Climate Information for Regions”.

A virtual server is being provided on the JASMIN cloud for dedicated use of IPCC authors [4]. The server will give users direct access to large climate datasets (including projections, re-analysis and Earth observation) and provide a collaborative workspace where they can share their results. This service is being developed in collaboration with German DDC partners who will provide a similar service at the ICSU World Data Center Climate (WDCC) from CMIP and many other archives. Demand for this service is particularly strong from the many authors who work in countries lacking facilities to deal with large data volumes.

References:
1. The usage report for 2016: https://zenodo.org/record/998064
2. Supplement: https://zenodo.org/record/998060
1.6. ENGAGING WITH THE COPERNICUS CLIMATE CHANGE SERVICE (C3S)

Ruth Petrie, Ag Stephens, Phil Kershaw, Victoria Bennett

The Copernicus Climate Change Service (C3S) is operated by the European Centre for Medium Range Weather Forecasting (ECMWF) on behalf of the European Union. C3S will deliver a Climate Data Store (CDS) containing geophysical data within a homogenous interface and will provide users with easy access to distributed data and analysis services. CEDA is engaging with the CDS via a number of high-profile activities funded via C3S. These include the delivery of quality-controlled global and regional climate projections, the provision of geographically load-balanced Earth System Grid Federation (ESGF) services, systems to deliver in-situ observations, support for processing Essential Climate Variables, and consultancy on the design and interfaces of the CDS itself.

**CP4CDS / CORDEX4CDS**

CEDA leads the Climate Projections for the Copernicus Data Store (CP4CDS) project and works in close partnership with French and German colleagues on a similar regional-scale project (CORDEX4CDS). A subset of the existing 5th Coupled Model Intercomparison Project (CMIP5) will be fully quality-controlled and provided to the CDS through dedicated ESGF hosts. The homogenised set of climate projections will be integrated with the CDS and its associated toolbox, thereby reducing the overhead for using the data in climate impact studies (figure 8).

Collaboratively, CEDA are working with the “Birdhouse” web processing service (WPS) to provide a robust “compute node” that can operate on climate projections and observational data.

For both projects CEDA is responsible for the system integration and operations. A key requirement for both projects is to deliver a resilient service with a 99% uptime. A solution has been devised in which services and data are mirrored across the three participating sites using a load balanced configuration. For individual site deployments the use of container and container orchestration technology is being employed.

**Figure 8: Schematic of CP4CDS data services linked to Copernicus Climate Change Service Climate Data Store (C3S CDS)**

**Services to deliver a Global Land and Marine Observation Database (GLAMOD)**

CEDA is also developing a platform to host and deliver a harmonised database of land and marine observations across the globe. The GLAMOD project, led by the National University of Ireland Maynooth, will ingest data from a range of sources into a common data model. Services to query and extract the data are being developed using geospatial technologies such as GeoServer.

**Supporting ECV Production**

The C3S is incorporating a wide range of Essential Climate Variables (ECVs) derived from satellite observations into the Climate Data Store. Several UK teams producing these ECVs are using the JASMIN infrastructure, where CEDA provides up to date satellite data and auxiliary data (e.g. meteorological observations) as inputs to their processing, and JASMIN itself offers compute and storage capability.

**Providing expert advice on data infrastructure and software**

CEDA has provided consultancy support to ECMWF and the CDS development team in the development of the architecture for CDS and the specification of the interfaces between CDS and the dependent projects such CP4CDS which are providing data through CDS and CDS Toolbox.
1.7. DATA MANAGEMENT FOR EUFAR2

Wendy Garland

The European Facility for Airborne Research in Environmental and Geo-sciences (EUFAR2) project - an EC FP7 project - came to an end in January 2018. EUFAR2 was the latest project of a consortium of 24 European institutions operating 18 instrumented research aircraft and a further 3 remote-sensing instruments. Since 2008, CEDA has provided a single gateway to the EUFAR funded flights data by linking existing archives (such as FAAM and NERC-ARF) with an accessible online archive for otherwise offline data. Advice and support on data matters has been given and the EUFAR Flight Finder (EFF) developed – a geo-spatial temporal search tool to facilitate discovery and re-use of flight data.

Archive
The archive of data collected during EUFAR Transnational Access (TA) flights - now exceeds >20TB for 169 flights made by 13 aircraft during 63 campaigns. As each aircraft team, instrument operator and project group operate independently, heterogeneity in flight naming, contents, and data arrangements and formatting (even within the community agreed standard data formats) was an ongoing issue. Dedicated effort at CEDA channelled these data into a (more) homogeneous state for user navigation, and to enable tools to scan the archive. Flight-level catalogue records were produced linking the unique data path to the relevant aircraft, instrument, project, people, geographic-temporal information etc. using python code developed for the task. Digital Object Identifiers (DOIs) have been published for 113 flights so far.

EUFAR Flight Finder tool (EFF)
Searching for data collected on moving platforms, such as research aircraft, is difficult as it is often not obvious where or when a flight took place, for what purpose or with which instruments. The EFF extracts flight paths, temporal information, keywords from the aircraft data files into an Elasticsearch repository which users can query through a graphical web interface and link directly to the data (See Figure 9).

Several challenges were encountered due to non-compliant or heterogeneous file formats, erroneous and incomplete data and other inconsistencies. The EFF was launched in May 2015 with data from 3 aircraft operators. Flights have continued to be added since, and a final push during 2017-2018 to include more format variations and non-compliant data has achieved 95% inclusion of EUFAR flights.

The coverage of the EFF has been widened to include the full catalogue of flights from the FAAM Bae-146 aircraft and the NERC-Airborne Research Facility (formerly ARSF) stored in the CEDA archive, and now covers 120TB of data and nearly 2000 flights. The concept was also adapted and reconfigured to facilitate the discovery of satellite data in the CEDA Satellite Finder. As described in section 1.9, this enables geographic and temporal search capability for 2 million Sentinel 1, 2, 3 and Landsat 5, 7, 8 scenes.
CEDA retrieves satellite data, from the Copernicus Sentinels, to create a mirror archive for users containing products from Sentinel 1 (A & B), Sentinel 2 (A & B) and Sentinel 3 (A). This covers a variety of multispectral and radar data as part of the Copernicus program and in the last year the total volume has increased to over 3PB. The data can be accessed directly using JASMIN or by download using FTP or the Satellite Data Finder. This mirror archive is a great asset for UK users as it locates a significant volume of Sentinel data next to the fast processing resource provided by JASMIN.

In order to create a true mirror archive, CEDA uses the ESA Open Access hub (OAH) as a reference and aims to retrieve all products of interest listed on this hub. The OAH is accessed by a large number of users globally with each user granted a maximum of two download sessions at a time. For these reasons it is impractical to create a mirror of the OAH using data from the OAH.

Fortunately, ESA have made available a set of Collaborative data hubs with access limited to ESA member states to support efficient data distribution. The UK has two accounts for fast access to Sentinel data, one of which holds by CEDA, for the purposes of providing data to the UK academic community. ESA has also created a network of relay data hubs to support the first tier of main Collaborative nodes. All relay hubs and collaborative nodes use the same Data Hub Service (DHuS) software and this allows data to be synchronised between nodes. Currently there are nine data hub relays, one of which is run by CEDA under contract to ESA. The CEDA Sentinel Relay Hub (SRH) retrieves a subset of Sentinel data as directed by ESA and relays this to other hubs within the network. Figure 10 shows the relationship between the two UK relay hubs and the primary tier of Collaborative hubs and the preceding upstream data services supplying raw data from the Sentinels themselves.

CEDA queries the data availability on the Collaborative hubs as well as the SRH and compares to the reference OAH. Where possible data is retrieved preferentially from the SRH to minimise download duplication and to ease network loading for ESA. Once Sentinel data is retrieved to CEDA, ancillary metadata and previews are extracted and all products are deposited on the archive. Content of the OAH is consistently monitored to ensure any reprocessed products are retrieved and gaps arising from download issues are closed. Completeness diagrams are generated for users to assess the current status of the mirror archive. This is useful to show data in the cache (awaiting archive space allocation), missing data and data moved to the Near Line Archive (NLA).

The Sentinel data archive at CEDA is widely used, primarily by the UK Earth Observation academic community processing and analysing the data on the JASMIN e-infrastructure, addressing global climate and environment research questions.

---

2 The other UK account is held by Airbus and is used for the UK Satellite Applications Catapult SEDAS portal.
CEDA SATELLITE DATA FINDER

Richard Smith

The CEDA Satellite Data Finder was developed as a tool to enable users to search data from some of the satellite missions, in the CEDA archive, using a map interface. At present, the Data Finder searches the Landsat (5,7,8) and Sentinel (1,2,3A) missions and provides several filters to help users refine their search (figure 11). The results are plotted on the map preferentially, newest scenes on top, and clicking on any of the drawn scenes reveals more information and details about how to access it. The Satellite Data Finder and Flight Finder are both map based tools built on top of our Elasticsearch infrastructure and were the first CEDA applications to trial the Elasticsearch backend.

The Satellite Data Finder started out as a direct clone of the EUFAR Flight Finder but it soon became clear that this tool would benefit from different filters. The variable and keyword filter were removed and two panels added to enable filtering by satellite mission and a tool to draw a rectangular region of interest.

One challenge with the Satellite Data finder is how to display nearing 5,000,000 satellite scenes, a number which is increasing daily. This is an area which still requires attention but in the first implementation we limit the number of satellite scenes rendered on the map. This is currently 1000 when searching the full map and 100 when using the rectangular region of interest to search. Users are encouraged to refine their search parameters, using the filters, to get the best results. This is not a perfect solution and it is still possible to get a very busy screen, especially when querying Sentinel-3A products as the images cover a greater geographic area. Using filters to constrain the search net, it is possible to find satellite scenes in your region of interest at a specific time and download, link to the archive directory or export the current selection as file paths/URLs for use in scripts.

Clicking on one of the polygons drawn on the map (figure 12), will open a window with more detailed information about the selected scene and links to download, preview or browse the data. Due to the large data volume associated with satellite imagery, older datasets in the Sentinel missions are archived to tape. Where the data for the selected scene is stored on tape, this info-window provides a link to information about how to gain access.

At CEDA we are always looking for ways to make accessing and searching our data holdings quicker, easier and more effective. The Satellite Data Finder, coupled with Elasticsearch, provides another way to explore and interact with CEDA hosted datasets.

1 http://geo-search.ceda.ac.uk
2 http://flight-finder.ceda.ac.uk
1.10. Near-Line Archive Now Operational

Neil Massey, Ed Williamson, Sam Pepler and Steve Donegan

The size of the data archive maintained by CEDA has grown year on year, with over 8PB now stored and made available to the public via the CEDA data catalogue browser, FTP and directly via JASMIN. Storing all of this data on disk is very expensive, in terms of hardware, electricity and maintenance costs. Therefore, we have been moving data from both a copy on tape and disk, to having a single copy on tape, plus another backup copy stored off site.

In order for the users of JASMIN to still be able to access this tape-only data, the Near-line Archive (NLA) system has been designed and implemented at CEDA. NLA is, in essence, a user-driven hierarchical filesystem. Via a command line tool, users request files to be restored from the tape-only data. These files are restored to a cache area of disk, and a symbolic link is created to the path where the original data resided. The user can then access the data in the same way they did before it was moved to tape only. The data will remain in the cache for 20 days before it is deleted, although users can extend this length of time.

NLA is an operational service with 41 active users. Although this seems a small number of users, many of these are group workspace managers and retrieve data for their entire group.

The main driver for implementing NLA is the multi-PB volumes (almost 10TB per day) of Earth Observation data from the Copernicus Sentinel program. CEDA aims to keep a rolling archive of the latest data available instantly on fast disks while older data is moved to NLA for users to restore as and when required. CEDA have already moved roughly 2PB of older Sentinel data to NLA, all of which is still accessible to users via the command line tool.

Users have already restored over 50TB of Sentinel data at one time and as the volume of data on tape goes up this is likely to increase in the future. Figure 13 shows the number of files that users have pulled back from the NLA for one month of Sentinel 3A SLSTR data. The data stored on the NLA is important for users to access allowing them to process a full time series to monitor environmental change.

Figure 13: Chart shows the number of files that users have pulled back from the NLA for one month of Sentinel 3A SLSTR data
JASMIN PHASE 4

Matt Pritchard

JASMIN is a globally-unique data analysis for environmental science. As well as providing the infrastructure from which CEDA operates the CEDA Archive, it provides storage and compute facilities to over 170 scientific projects across research topics ranging from earthquake prediction and oceanography to air quality and climate science.

This year saw an £8M capital investment by The Natural Environment Research Council (NERC) to provide JASMIN’s Phase 4 upgrade, which will double the available storage to over 44 Petabytes by the end of 2018 and begin the transition to new types of storage more suited to future needs. In addition, around 40% will be added to the processing capability, with 11,500 cores on nearly 600 nodes. The internal network will be refreshed to ensure maximal I/O performance and JASMIN’s private cloud capabilities will be improved, in particular “cluster-as-a-service”, as well as the data access services.

All this means that the 1,700 registered users of JASMIN will be able to process more data, more efficiently and collaborate more effectively to carry out cutting-edge environmental science.

JASMIN has now been in operation over 6 years and has revolutionised data access for the environmental science community in the UK, largely by providing co-located storage and compute to enable scientists to work on very large datasets in-place rather than having to download them to work on elsewhere. Storage from the first phase of JASMIN now needs to be retired and removed, so work has been underway to migrate from old hardware to newer systems. A total of 5.5 Petabytes of data has now been migrated, with further consolidation planned once new storage from Phase 4 is fully operational.

However, it’s not simply “out with the old and in with the new”. Phase 4 introduces new types of storage which, as well as being more cost effective, open up new possibilities for different usage and workflows in the future. By the end of 2018, JASMIN will have:

- 8.2 PB of parallel file system (of the type which has dominated JASMIN storage until now)
- 30 PB of “Scale out Filesystem” software-defined storage which is cost effective to increase in scale
- 5 PB of high-performance Object Store, for use with cloud services
- 0.5 PB of high-performance solid-state disk, already in operation for user home directories and a high-performance scratch for use with the LOTUS processing cluster
- Access to the UK JANET network via an improved RAL site network link (increasing from 40 to 100 Gbit/sec)

Object Storage is considerably cheaper than POSIX (i.e. file system) storage, so will feature prominently in JASMIN’s storage mix in the future. For now, most analysis workflows rely on file systems, so a transition to Object Storage at volume is not yet feasible. However, a significant deployment of Object Storage now enables us to start work towards that goal, initially by providing an S3 (Amazon Simple Storage Service) interface to object storage from tenancies within the JASMIN unmanaged cloud, which also helps meet a specific need for improved data access in that area.

As of the end of 2017/18, all procurements were completed successfully and the hard work of integrating the new components into operational use had just started, to roll out the benefits of the upgrade to end users... while planning ahead for further upgrades to follow in Phase 5!
The cloud and accounts portal are two applications which have a key role in making resources available to users on JASMIN. A core design principle for JASMIN is to make computing resources available alongside data, and a Cloud computing paradigm fits closely with this model. JASMIN runs its own Community Cloud to enable users to effectively borrow chunks of the overall pool of computing resources offered by JASMIN and use them for their own purposes. In practice, this means the ability to create their own virtual computers (Virtual Machines - VMs) and install their own software as part of a tenancy - their share of the overall cloud’s resources. Cloud users have used this capability on JASMIN in all kinds of novel ways, from running software to run virtual training courses, to operating websites, portals and providing virtual computer desktops.

Cloud computing is better known though for the commercial or public providers of these services such as Amazon Web Services, Microsoft Azure and Google Cloud Platform. However, even a cursory look at the web interfaces reveals the range and complexity of functionality offered by these services. This is also true for the underlying cloud technology used for JASMIN. From early in the development of JASMIN’s cloud, a custom web portal was developed; with the goal of simplifying access, making it friendlier to use and more intuitive for users who may not be familiar with cloud computing. From the portal, users can easily select the kind of virtual machine they would like and provision it themselves.

Over the last months, a major upgrade has been carried out to JASMIN’s cloud; replacing the existing software used to manage the cloud with a new system. We are excited to bring a solution based on OpenStack, an open source cloud implementation which has gained widespread adoption in the research community during the past years. CEDA works closely together with the Scientific Computing Department in STFC to develop and operate the cloud service. At this stage, one of the challenges has been to manage the upgrade whilst minimising disruption to users with the existing service. A new version of the Cloud portal has been developed to work with OpenStack but also, much work has been carried out behind the scenes to integrate the system with the rest of JASMIN’s infrastructure. Over the past months, users’ existing cloud tenancies have been migrated from the old to the new system and from now on all new tenancies will be deployed on the new platform. OpenStack will also allow us to scale-up the JASMIN cloud, it will introduce deployment flexibility thanks to the new API and at the same time it will allow novel applications which were not possible with the previous implementation.
In January this year the Research Councils UK (now UK Research and Innovation) Cloud Working Group\(^3\) held its third annual workshop for the UK research community. Over 160 delegates attended at the Francis Crick Institute in London. This meeting enables researchers to meet, network and learn from colleagues about how they can make use of cloud computing technology to help their research. There was representation from large scientific centres like CERN, a panel session including representatives from the large public cloud providers (Amazon, Google and Microsoft) and participation from many other companies and other groups involved in facilitating the use of cloud for research.

First developed by Amazon over ten years ago, cloud computing provides an alternative to owning and setting up your own computer hardware by instead renting from a third party cloud provider. This model has been driving a revolution in the IT industry with companies like Netflix and Slack using cloud to grow and build their businesses. Cloud has important features that the research community can take advantage of, particularly in addressing the challenge of Big Data: rather than users downloading large amounts of data to their local computer they can access, process and analyse it remotely by running computers in a cloud hosted next to the data. This model has been exploited in the UK in our own community with JASMIN’s community cloud service, which enables research users to make own their virtual computing resources alongside the CEDA archive held on JASMIN. JASMIN is one of a number of clouds operated by the research community that exploit the same model, such as CLIMB\(^4\) and EMBL-EBI’s Embassy Cloud\(^5\).

Though cloud computing presents many opportunities, it also presents a number of challenges since it brings such a profound change in the way computing resources are procured and accessed. Recognising these challenges, the Cloud Working Group was established in 2015 under the remit of the then Research Councils UK (RCUK), now UKRI (UK Research and Innovation). The group is made up of twelve members with Philip Kershaw, CEDA staff, as chair. Its goal is to support researchers and technical specialists in the application of cloud computing technologies and services for the research community in the UK. In particular, the working group focuses on four key areas for its work: 1) building a community focus around cloud computing for researchers; 2) providing support around the technical integration of cloud with scientific applications; 3) equipping researchers with the skills and finally, 4) exploring the policy changes that may be needed to be addressed around purchasing resources, hosting of data, privacy and security.

Besides the workshops, the group has carried out work in a number of different areas to date: supporting technical pilots to try out different aspects of cloud computing services for research computing scenarios; carrying out analysis on the costs of public cloud and supporting the creation of a training course in ‘ResOps’, equipping researchers with the new skill set they need to deploy applications in the cloud.

---

\(^3\) [http://www.cloud.ac.uk/](http://www.cloud.ac.uk/)
\(^4\) [https://www.climb.ac.uk](https://www.climb.ac.uk)
\(^5\) [http://www.embassycloud.org/](http://www.embassycloud.org/)


1.14, IMPROVING COMMUNICATIONS WITH DATA PROVIDERS

Kate Winfield

CEDA is responsible for curating atmospheric and earth observation data that has been funded by the Natural Environment Research Council (NERC), ensuring long-term preservation and re-use of data. CEDA data scientists liaise with data providers for the lifecycle of their NERC grants. They give advice on archival and metadata standards, set up the data delivery route and capture supporting material needed to create catalogue records about the data. There are currently around 200 active projects in our internal Data Management Plan (DMP) tool, which resulted in many hundreds of emails stored in personal or shared inboxes. This meant communications with data providers could easily get lost when staff left or forgot to put emails in the shared folder. It was also difficult to look back on previous conversations and keep track of emails related to individual projects.

CEDA currently uses a helpdesk system, called Help Scout, where user queries are easily assigned to the relevant member of CEDA staff. However, the CEDA helpdesk runs on a different workflow as it resolves queries on a daily basis, rather than over several years as is common with NERC grant holders. This provided the motivation to develop a separate data management helpdesk using Help Scout; resulting in data providers receiving a quicker response from CEDA as the messages are all stored in one place and can easily be found by all relevant staff.

The first stages were to define our workflow and to see how we could link the DMP tool to the helpdesk. From our workflow, the communications can be split into 4 stages; Initial contact, DMP communication, Progress, Data delivery (figure 17).

In the helpdesk we were able to set searchable custom fields for each query, which means we can easily track each stage of the project. Additionally, in the DMP tool each project is identified with a unique project id number (figure 18). By creating another custom field, we can now link the DMP tool directly to the helpdesk; enabling all previous conversations relating to the project to be easily viewed and stored in one place.

This new system has been hugely beneficial to CEDA data scientists as all email correspondence is now in one place that can be searched easily, the projects can be re-assigned or shared with colleagues effortlessly, and with the added ability to track the progress of each project/grant from the email communications. Ultimately, this means CEDA can provide help and support to the data providers more quickly and efficiently.
Sarah Callaghan, CEDA programme manager, was appointed as Editor-in-Chief of the CODATA Data Science Journal in March 2015, acknowledging her international reputation and expertise in data citation and academic publishing. Taking on this role has given CEDA, via Sarah, valuable insight into the latest developments in data science, policies and practice, as well as influence in international organisations such as CODATA and the Research Data Alliance, raising CEDA’s profile internationally outside the environmental science community.

Over the calendar year, 1 January 2017 to 31st December 2017, 92 submissions to the journal were received (107 in 2016) and 52 papers were published (18 in 2016), with another 61 in the publication pipeline.

2017 also saw the launch of two special collections of papers:
1. 20 Years of Persistent Identifiers: Applications and Future Directions
2. SciDataCon
(A third special collection “Open Data and Africa” was launched in January 2018).

In comparison with 2016, there was an increase in web sessions of 51.46%, in users of 46.72% and in page views of 47.29%. This growth in usage of the journal, along with the increasing number of publications shows that the journal is developing as a source of impactful information on data management, data science, and the policy and practices of data use and reuse.

These increases would not have been possible without the expertise and enthusiasm of a dedicated editorial team from all around the world, along with the support of the Ubiquity Press staff.

The CODATA Data Science Journal is a peer-reviewed, open access, electronic journal, publishing papers on the management, dissemination, use and reuse of research data and databases across all research domains, including science, technology, the humanities and the arts. The scope of the journal includes descriptions of data systems, their implementations and their publication, applications, infrastructures, software, legal, reproducibility and transparency issues, the availability and usability of complex datasets, and with a particular focus on the principles, policies and practices for open data. All data is in scope, whether born digital or converted from other sources.

The increase in usage of the Data Science Journal website, and in the number of published publications and special collections demonstrates the increasing importance of data science and data management to the academic and research communities. Sarah’s work as Editor-in-Chief keeps CEDA up to date with latest developments in the field, while also promoting CEDA’s influence and position as a leading exponent and developer of data curation practices.
CEDA GRADUATES COLLABORATE WITH PARTNER ORGANISATIONS

William Tucker and Ed Williamson

As part of the STFC graduate scheme, all graduates are encouraged to conduct a placement of up to 3-months in another department within STFC or external organisation. Two CEDA graduates, William and Ed, undertook their placements during 2017.

Placement at ISIS Facilities Business Applications - William Tucker

William, a software engineer at CEDA, did his placement with ISIS’s Facilities Business Applications ("Bus Apps"), a software focused group who develop and maintain web applications and services for teams at the ISIS Neutron and Muon Source. This was a rare opportunity to share experience and best-practice with another STFC-based developer group.

His assigned task was to engineer a web application which could generate QR-code labels for lab equipment at ISIS. These would allow engineers to quickly and easily identify items – which might need replacing or repairing – using a QR reader.

ISIS Bus Apps employ many “Agile” practices as part of their software development lifecycle. For instance, regular code-reviews allow members of the group to comment on significant code changes. A high level of customer involvement is also desirable; in this case, a final review of the project by the ISIS Instruments division was held at the end of the placement. These sorts of practices are gradually being adopted by CEDA’s development group and it was very useful to have one of our developers gain this experience.

Graduate placement at NERC-Airborne Research Facility (ARF) Data Analysis Node (DAN) - Ed Williamson

Ed, a data scientist at CEDA, spent his placement at NERC-ARF-DAN in Plymouth. This team process all of the raw data from the flights by the NERC-ARF aircraft (formally Airborne Research and Survey Facility (ARSF)). CEDA hold data produced from these flights, which dates back to 1981. These data are on the archive for users to access via the web or on JASMIN. CEDA also run the EUFAR Flight Finder, which enables users to search flights from a variety of aircraft data that CEDA hold.

The aim of Ed’s placement was to produce a PostGIS database for all the flight data that NERC-ARF-DAN hold; this would allow data to be added to the flight finder tool. The data produced on a flight is typically hyperspectral, thermal and RGB imagery and LiDAR data. Ed harvested metadata for the PostGIS database, this allowed production of GeoJSON files (the correct format for the flight finder). This means that older flights can now be searched geographically at CEDA. Future flights will have this work integrated into the initial processing so that the GeoJSON is automatically produced and metadata added to the PostGIS database. Ed’s placement allowed for improvements to the NERC ARF data, allowing for the data to be added to the CEDA flight finder; ultimately meaning these data are now more easily searchable.
PART 2: CEDA AND JASMIN SUMMARY INFORMATION 2017-2018

The Centre for Environmental Data Analysis (CEDA) exists 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. CEDA enhances this role through the development and maintenance of tools and services to aid data preservation, curation, discovery and visualisation; all of which add value for the world-wide user community.

The JASMIN data analysis environment is increasingly providing data-compute capabilities for the UK and wider environmental research communities. This section of the annual report presents summaries of CEDA and JASMIN usage.

2. USAGE OF CEDA DATA

CEDA delivers Data Archive services for the National Centre for Atmospheric Science (NCAS) and the National Centre for Earth Observation (NCEO). In addition, CEDA delivers the NERC/STFC funded UK Solar System Data Centre (UKSSDC) and the IPCC Data Distribution Centre for the Intergovernmental Panel on Climate Change (IPCC).

<table>
<thead>
<tr>
<th>Annual CEDA Usage: April 2017 to March 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of users</td>
</tr>
<tr>
<td>Total data downloaded</td>
</tr>
<tr>
<td>Total number of accesses</td>
</tr>
<tr>
<td>Total days activity</td>
</tr>
</tbody>
</table>

Table 2.1: 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 fairly static, there is a slight trend upwards in the amount of data downloaded (table 2.2 and figure 2.1)

<table>
<thead>
<tr>
<th>Date</th>
<th>Users</th>
<th>Methods</th>
<th>Datasets</th>
<th>Number of accesses</th>
<th>Size</th>
<th>Activity days</th>
</tr>
</thead>
<tbody>
<tr>
<td>201704</td>
<td>2504</td>
<td>12</td>
<td>2300</td>
<td>964207</td>
<td>39.1 TB</td>
<td>41711</td>
</tr>
<tr>
<td>201705</td>
<td>2362</td>
<td>10</td>
<td>2314</td>
<td>1322628</td>
<td>56.5 TB</td>
<td>16573</td>
</tr>
<tr>
<td>201706</td>
<td>2982</td>
<td>10</td>
<td>2239</td>
<td>921575</td>
<td>55.6 TB</td>
<td>20196</td>
</tr>
<tr>
<td>201707</td>
<td>1804</td>
<td>11</td>
<td>1077</td>
<td>1699459</td>
<td>76.0 TB</td>
<td>8597</td>
</tr>
<tr>
<td>201708</td>
<td>1763</td>
<td>11</td>
<td>1010</td>
<td>971330</td>
<td>73.9 TB</td>
<td>9002</td>
</tr>
<tr>
<td>201709</td>
<td>2056</td>
<td>11</td>
<td>1722</td>
<td>529870</td>
<td>69.7 TB</td>
<td>9188</td>
</tr>
<tr>
<td>201710</td>
<td>2609</td>
<td>9</td>
<td>2044</td>
<td>1266255</td>
<td>78.5 TB</td>
<td>13407</td>
</tr>
<tr>
<td>201711</td>
<td>2409</td>
<td>9</td>
<td>1538</td>
<td>768742</td>
<td>52.0 TB</td>
<td>10354</td>
</tr>
<tr>
<td>201712</td>
<td>2025</td>
<td>7</td>
<td>988</td>
<td>658517</td>
<td>84.5 TB</td>
<td>8038</td>
</tr>
<tr>
<td>201801</td>
<td>2526</td>
<td>7</td>
<td>1352</td>
<td>1055389</td>
<td>31.4 TB</td>
<td>10391</td>
</tr>
<tr>
<td>201802</td>
<td>2151</td>
<td>7</td>
<td>1309</td>
<td>1710856</td>
<td>111.9 TB</td>
<td>9374</td>
</tr>
<tr>
<td>201803</td>
<td>2460</td>
<td>8</td>
<td>1091</td>
<td>1218142</td>
<td>65.4 TB</td>
<td>10439</td>
</tr>
<tr>
<td>Total</td>
<td>27651</td>
<td>112</td>
<td>18984</td>
<td>13086970</td>
<td>794.5 TB</td>
<td>167270</td>
</tr>
</tbody>
</table>

Table 2.2: Monthly summary figures for usage by CEDA consumers during the reporting year
Note that additional metrics also appear in the data centre metrics section, chapter 7.

Figure 2.1: Breakdown of CEDA usage by month
JASMIN is a globally-unique data analysis environment, designed, integrated and operated by STFC providing four main functions: storage, batch computing, managed infrastructure compute and a private cloud to provide Infrastructure-as-a-Service (the JASMIN “unmanaged”, “community” or “external” cloud).

During this period, JASMIN received an £8M capital investment by The Natural Environment Research Council (NERC) to provide JASMIN’s Phase 4 upgrade. In parallel with procuring and integrating new hardware and services, JASMIN continued to provide a cutting-edge research environment for NERC-related environmental science.

All Phase 4 procurements were completed successfully during this reporting year. Significant further work has since been underway to bring the items procured into operation as functioning components within JASMIN services, and to decommission deprecated hardware which can no longer be supported.

### 3.1. STORAGE

The JASMIN Phase 4 investment enabled us to decommission the oldest (Phase 1) hardware, and still double the available storage by the end of 2018 and begin the transition to new types of storage more suited to future needs.

Data storage capacity was procured as a mix of main storage types:

- Panasas Parallel File System (PFS), of the same type which dominated earlier JASMIN phases
- PURE Storage Solid State Disk (SSD) for home directories and scratch storage, to help performance with operations handling small files.
- QuoByte Scale Out Filesystem (SOF): scalable software-defined storage which is interchangeable in function between a traditional file system and an Object Store on the same hardware.
- Caringo dedicated Object Store

In total, 38.9 PB new storage was purchased. Work has taken place to migrate from old hardware to newer systems, and deploy the new storage.

Object Storage is considerably cheaper than traditional (POSIX) file system storage, and offers the possibility of significantly improved data access into and out of JASMIN cloud tenancies, so will feature prominently in JASMIN’s storage mix in the future. For now, most analysis workflows rely on file systems, so a transition to Object Storage at volume is not yet feasible. However, a significant deployment of Object Storage, procured as part of Phase 3.5 and now deployed as a proof-of-concept enables us to start work towards that goal, initially by providing an S3 interface to object storage from tenancies within the JASMIN unmanaged cloud. Once the proof-of-concept project on this QuoByte Object Store has completed, the storage will be re-tasked as the first major (1.2 PB) deployment of QuoByte SOF storage.

At the start of this reporting year there were 154 Group Workspaces with allocations totalling 8.8 PB, serving a total community of 1445 users. By the end of the period this had increased to 164 group workspaces in use with allocations of 9.5 PB respectively, serving a total community of 1815 users.

The Elastic Tape service, a near-line tape storage facility available to JASMIN 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.

### 3.2. MANAGED COMPUTE INFRASTRUCTURE

22
The Phase 4 procurement provided significant enhancements to JASMIN’s compute fabric in the form of 210 24-core servers for use either as hypervisors or LOTUS compute nodes, to be deployed in the following reporting period.

The previous Phase 3.5 upgrade had stabilized LOTUS utilization from regularly topping 70% beforehand to around 70% (even with that new capacity) and the number of active LOTUS users per month was around 150 but increasing steadily. It is expected that the Phase 4 deployment, when complete, should add around 40% capacity to LOTUS to help cope with the increasing demand.

A new VMWare vCentre system was deployed to manage JASMIN’s virtualized compute estate (i.e. for managed VMs not in cloud tenancies). This replaced older, separate JASMIN and CEMS systems.

JASMIN’s “Science DMZ” or “Data Transfer Zone” continued to provide an efficient route for high-performance data transfers for cases where large volumes of data need to be moved efficiently. In addition to providing services for individual JASMIN users, work continued in collaboration with international partners in the ESGF infrastructure, helping to prepare CEDA for participation in large-scale institutional data transfers as part of the anticipated CMIP6 data replication activity. The relay hub for ESA Sentinel data also operates in this dedicated network zone and was used throughout the period to help ingest and disseminate increasing volumes of Sentinel data, with data flows of 10TB/day expected once all Sentinel satellites are operational.

The “JASMIN Analysis Platform” software stack continued to evolve with a several new versions (1.1-29 to 1.1-33) released and deployed across the managed compute infrastructure during this period.

A service to provide OPeNDAP for Group Workspaces was introduced, with the capability for Group Workspaces to autonomously expose their GWS data via a common web interface.

A testbed for high performance OPeNDAP was also deployed, for use with the CEDA Archive.

3.3. NETWORK AND SUPPORTING INFRASTRUCTURE

During this period, access to the UK JANET network was improved via an upgraded RAL site network link (increasing from 40 to 100 Gbit/sec, with JASMIN a significant contributor to the demand for the upgrade).

The Phase 4 upgrade procured a new “super-spine” internal network and future expansion capability, refreshing the internal network to ensure maximal I/O performance. Additional machine room hardware including new racks, power distribution units, cabling, environment monitoring equipment and management network components were also added to refresh essential infrastructure.

3.4. JASMIN CLOUD

The JASMIN cloud portal continued to provide tenants of JASMIN's unmanaged cloud with a web-based interface enabling them to deploy and manage virtual machines at will within their tenancies. During the period, the portal’s capabilities were expanded, and VM templates created so that managed-cloud tenancies could be managed through the same system.

A strategic decision was taken to move away from VMWare to OpenStack a) because license costs associated with the current VMWare software would not scale, going forwards and b) because OpenStack offers a set of features more compatible with key technological aspirations for the future, such as “cluster-as-a-service” and improved data access. Initially, the VMWare VIO OpenStack product was selected, but migration to OpenStack opens the possibility to open source alternatives in the future.

A production OpenStack instance was deployed and progress made towards integrating this with existing infrastructure components such as networking and LDAP.
A new version of the JASMIN Cloud Portal, developed for use with VIO OpenStack, was tested successfully towards the end of the year.

Other items on the development roadmap at this stage included:

- Cluster-as-a-Service testbed deployment
- Containerised Jupyter Notebook service deployment proof-of-concept

### 3.5. USER SUPPORT AND OUTREACH

June 2017 saw the second annual JASMIN user conference, hosted over two days by RAL Space. The event aimed to bring together members of JASMIN’s virtual community in a physical meeting, with opportunities to learn about recent developments of the infrastructure and service, to share experiences of using JASMIN with other users and with the JASMIN team, and to gain training in key aspects of the JASMIN system. As with the previous event, tours of the JASMIN infrastructure laid on by the Scientific Computing Department proved very popular with attendees.

A variety of science projects supported by JASMIN presented details of their work including the challenges they had faced. Many identified aspects in which JASMIN had provided a “turnkey” scale or performance benefit. A novel “pitch your poster” slot provided light-hearted rapid-fire introductions to a poster session held in the informal setting of STFC’s visitor centre. This was also used as the venue for technical training sessions, including topics such as “Getting started with JASMIN”, “Batch processing using LOTUS”, and “High-performance data transfer using JASMIN’s Globus Online endpoint”.

The event has attracted 50-70 attendees each year so far, and as a result of feedback and discussions with users, is now planned as an annual event, but as part of a wider programme of user engagement activities.

Other engagement events held this year include:

- JASMIN training as part of CEH workshop at CEH Lancaster (February 2018)
- JASMIN Webinar (March 2018)

The webinar event trialled remote delivery of training to JASMIN users (or potential users) in their home institutions. Based on its success, it has been decided that further events will follow on specific topics, to be guided by user community demand.

A significant amount of work has been done this year to improve the documentation about JASMIN and its component services, to help users understand how to use JASMIN resources efficiently as part of a multi-user environment. The introduction of “HelpScout” help desk software at the beginning of 2017 (jointly with CEDA) provides those involved with user support with integrated tools to manage user queries and documentation. Recently, JASMIN documentation has moved to its own dedicated site, with further content and enhancements in progress. Training materials from sessions at the JASMIN conference and other events have also been made available on JASMIN’s own YouTube channel as another method for addressing user education.

A number of VIP tours of the JASMIN infrastructure were conducted throughout the year by the SCD team, to a variety of visitors including:

- STFC CEO
- DEFRA Permanent Secretary
- UKSA CEO
- British Geological Survey
- Representatives from NCI (Australia)
- Representative from DKRZ (Germany)
4. 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). In Europe, we collaborate with partners in the ENES (European Network for Earth System Modelling) grouping, to efficiently distribute high-volume climate model simulation data.

Other major international collaborations include the European Space Agency’s Climate Change Initiative Programme.

CEDA works closely with STFC’s Scientific Computing Department to deliver the JASMIN infrastructure.

4.1. MAJOR COLLABORATIONS

In 2017/2018, significant national and international collaborations have continued. 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 (ESGF) 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. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.
4. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) Open Data Portal
5. 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. CEDA works with ECMWF to provide EO scientists with the high resolution atmospheric analyses they need to process satellite observations
7. With partners in Germany and the USA, CEDA provides data services on behalf of the IPCC (Intergovernmental Panel on Climate Change) through the Data Distribution Centre
8. Supporting the Climate and Forecast Metadata (CF) Conventions with partners in Uni. Reading, UKMO, and multiple US research institutions.
9. With 20+ partners in the European Network for Earth System Modelling, CEDA is working to develop software and services for climate model data archives;
10. Working with academic partners in the UK Research and Innovation UKRI Cloud Working Group to share best practice, knowledge and strategy for use of cloud computing in the research domain

5. FUNDING AND GOVERNANCE

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), EC Copernicus Climate Change Service, BEIS, DEFRA and others, as well as participating and coordinating major European projects.

5.1. ANNUAL TOTAL FUNDING

<table>
<thead>
<tr>
<th>Financial Year</th>
<th>10-11</th>
<th>11-12</th>
<th>12-13</th>
<th>13-14</th>
<th>14-15</th>
<th>15-16</th>
<th>16-17</th>
<th>17-18</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCAS income</td>
<td>906</td>
<td>883</td>
<td>935</td>
<td>829</td>
<td>829</td>
<td>808</td>
<td>808</td>
<td>808</td>
</tr>
<tr>
<td>NCEO income</td>
<td>450</td>
<td>419</td>
<td>445</td>
<td>392</td>
<td>390</td>
<td>393</td>
<td>393</td>
<td>393</td>
</tr>
<tr>
<td>Other NERC</td>
<td>341</td>
<td>527</td>
<td>287</td>
<td>272</td>
<td>600</td>
<td>621</td>
<td>825</td>
<td>816</td>
</tr>
<tr>
<td>Other income</td>
<td>1144</td>
<td>1099</td>
<td>1283</td>
<td>1486</td>
<td>1394</td>
<td>1505</td>
<td>1092</td>
<td>1280</td>
</tr>
<tr>
<td>Total income</td>
<td>2841</td>
<td>2928</td>
<td>2950</td>
<td>2979</td>
<td>3213</td>
<td>3327</td>
<td>3118</td>
<td>3297</td>
</tr>
</tbody>
</table>

Table 5.1: Overall funding for CEDA for financial years 2010 — 2011 to 2017 — 2018 (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.

5.2. EXTERNALLY FUNDED PROJECTS FOR THE YEAR 2017-2018

The table below shows CEDA’s externally funded projects which were active during the reporting year.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Funder</th>
<th>Start date</th>
<th>End date</th>
<th>Value (£k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA CCI Data Portal</td>
<td>Archive and data services for ESA CCI programme</td>
<td>ESA</td>
<td>01/05/2015</td>
<td>31/03/2018</td>
<td>195.3</td>
</tr>
<tr>
<td>EUFAR-2</td>
<td>The European Facility for Airborne Research (EU-FAR), coordinating the operations of the European fleet of aircraft in the field of environmental research</td>
<td>FP7</td>
<td>01/02/2014</td>
<td>31/01/2018</td>
<td>174.0</td>
</tr>
<tr>
<td>EUSTACE</td>
<td>EU Surface Temperature for All Corners of Earth</td>
<td>H2020</td>
<td>01/01/2015</td>
<td>30/06/2018</td>
<td>142.9</td>
</tr>
<tr>
<td>FIDUCEO</td>
<td>JASMIN and data support for Fidelity and uncertainty in climate data records from Earth Observations</td>
<td>H2020</td>
<td>01/02/2015</td>
<td>21/02/2019</td>
<td>102.5</td>
</tr>
<tr>
<td>Project</td>
<td>Description</td>
<td>Lead Agency</td>
<td>Start Date</td>
<td>End Date</td>
<td>Value</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>--------</td>
</tr>
<tr>
<td>BACI</td>
<td>JASMIN support for BACI (Biosphere Atmosphere Change Index)</td>
<td>H2020</td>
<td>01/04/2015</td>
<td>31/10/2018</td>
<td>25.7</td>
</tr>
<tr>
<td>Forestry TEP</td>
<td>ESA Thematic Exploitation Platform</td>
<td>ESA</td>
<td>01/01/15</td>
<td>31/07/17</td>
<td>125.2</td>
</tr>
<tr>
<td>PRIMAVERA</td>
<td>JASMIN and archive support for new generation of global climate models</td>
<td>H2020</td>
<td>1/1/2015</td>
<td>31/10/2019</td>
<td>141.0</td>
</tr>
<tr>
<td>ESA Sentinel Data Hub Relay</td>
<td>STFC as part of network of ESA relay hubs for Sentinel data</td>
<td>ESA</td>
<td>1/5/2015</td>
<td>30/10/2017</td>
<td>311.2</td>
</tr>
<tr>
<td>UKCP09 User Interface 15-18</td>
<td>UK Climate Projections platform, user interface development</td>
<td>EA</td>
<td>01/04/2015</td>
<td>31/03/2018</td>
<td>86.5</td>
</tr>
<tr>
<td>C3S CDS Infrastructure</td>
<td>Consultancy role for C3S Climate Data Store project</td>
<td>C3S</td>
<td>01/07/2016</td>
<td>30/06/2017</td>
<td>14.7</td>
</tr>
<tr>
<td>CDS Zone</td>
<td>Support for ECV processing</td>
<td>UKSA/NCEO</td>
<td>01/08/2016</td>
<td>15/03/2019</td>
<td>155.0</td>
</tr>
<tr>
<td>MEDIN Metadata Services</td>
<td>Metadata services for MEDIN</td>
<td>MEDIN</td>
<td>1/4/2015</td>
<td>28/02/2018</td>
<td>30.7</td>
</tr>
<tr>
<td>C3S ESGF Data Node (CP4CDS)</td>
<td>Operational ESGF data node for C3S</td>
<td>C3S</td>
<td>1/9/2016</td>
<td>31/8/2017</td>
<td>309.0</td>
</tr>
<tr>
<td>Pest Risk Modelling in Africa (PRISE)</td>
<td>JASMIN support for UKSA IPP project</td>
<td>UKSA</td>
<td>1/12/2016</td>
<td>31/03/2019</td>
<td>97.7</td>
</tr>
<tr>
<td>C3S ECVs production SST</td>
<td>JASMIN and archive support for C3S SST ECV processing</td>
<td>C3S</td>
<td>01/02/2017</td>
<td>31/10/2018</td>
<td>63.9</td>
</tr>
<tr>
<td>BEIS DDC</td>
<td>UK component of IPCC Data Distribution Centre</td>
<td>BEIS</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>83.0</td>
</tr>
<tr>
<td>C3S CORDEX4CDS</td>
<td>Regional Climate Projection data for C3S</td>
<td>C3S</td>
<td>01/05/2017</td>
<td>30/04/2021</td>
<td>160.0</td>
</tr>
<tr>
<td>MOHC Data Pipeline</td>
<td>Supporting CMIP climate model data movement from the Met Office to the CEDA archive and ESGF</td>
<td>BEIS</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>70.0</td>
</tr>
<tr>
<td>Project Description</td>
<td>Funding</td>
<td>Start Date</td>
<td>End Date</td>
<td>Cost</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------</td>
<td>---------</td>
<td>------------</td>
<td>-----------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>To archive the next generation of climate projections for the UK (2018)</td>
<td>Met Office</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>75.0</td>
<td></td>
</tr>
<tr>
<td>To provide data services to support access to the next generation of climate projections for the UK (2018).</td>
<td>Met Office/Defra</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>191.0</td>
<td></td>
</tr>
<tr>
<td>CEDA support to multi-model climate ensemble archives</td>
<td>Met Office</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>45.0</td>
<td></td>
</tr>
<tr>
<td>In-situ observations for Copernicus</td>
<td>C3S</td>
<td>01/07/2017</td>
<td>31/03/2019</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>Funding Esther Conway to attend ESA LTDP WG and organise PV conference in UK</td>
<td>UKSA</td>
<td>01/04/2017</td>
<td>31/03/2018</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td>JASMIN resources to support CSSP China project</td>
<td>Met Office</td>
<td>01/01/2018</td>
<td>31/03/2018</td>
<td>35.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 5f.2: Externally funded projects for 2017-2018 (non-core NERC)

5.3. GOVERNANCE

The CEDA/JASMIN board reflects our funding arrangements: CEDA and JASMIN are NERC funded facilities based at STFC Rutherford Appleton Laboratory. NERC funding includes both central provision and support from NCAS and NCEO. Thus, all three, as well as the host organisation, are represented on the board.

Membership of the CEDA/JASMIN board (March 2018) is:
- Bryan Lawrence, NCAS (Chair)
- Sarah Callaghan, CEDA, STFC (Secretary)
- Victoria Bennett, Division Head CEDA, STFC
- Chris Mutlow, Director RAL Space
- Stephen Mobbs, Director NCAS
- John Remedios, Director NCEO
- Martin Wooster, Divisional Director EOIF, NCEO
- Amber Vater, NERC
- Beth Greenaway, Head of EO, UKSA
- Tony Hey, SCD, STFC

The board aims to meet at least annually, and up to semi-annually when necessary. This year, the CEDA board met on the 19th of March 2018; the next board is planned in June 2018.
CEDA is required to provide metrics quarterly in a number of categories. Some additional metrics to those provided in Chapter 1 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.

### 6.1. ACCESS RELATED METRICS

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

<table>
<thead>
<tr>
<th>Area</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>2459</td>
<td>60.6%</td>
<td>2422</td>
<td>60.8%</td>
</tr>
<tr>
<td>Europe</td>
<td>483</td>
<td>11.9%</td>
<td>465</td>
<td>11.7%</td>
</tr>
<tr>
<td>Rest of the world</td>
<td>1043</td>
<td>25.7%</td>
<td>1021</td>
<td>25.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>73</td>
<td>1.8%</td>
<td>74</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

Table 6.1: Users by area

<table>
<thead>
<tr>
<th>Institute Type</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>2961</td>
<td>73.0%</td>
<td>2908</td>
<td>73.0%</td>
</tr>
<tr>
<td>Government</td>
<td>610</td>
<td>15.0%</td>
<td>606</td>
<td>15.2%</td>
</tr>
<tr>
<td>NERC</td>
<td>158</td>
<td>3.9%</td>
<td>155</td>
<td>3.9%</td>
</tr>
<tr>
<td>Other</td>
<td>233</td>
<td>5.7%</td>
<td>228</td>
<td>5.7%</td>
</tr>
<tr>
<td>Commercial</td>
<td>51</td>
<td>1.3%</td>
<td>45</td>
<td>1.1%</td>
</tr>
<tr>
<td>School</td>
<td>43</td>
<td>1.1%</td>
<td>38</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Table 6.2: Users by Institute type

### 6.2. DATA HOLDINGS

<table>
<thead>
<tr>
<th>Data Centre</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEDA</td>
<td>4421</td>
<td>4444</td>
<td>4456</td>
<td>4459</td>
</tr>
</tbody>
</table>

Table 6.3: Number of dataset discovery records held in the NERC data catalogue service.
Table 6.4: Number of dataset collections and datasets identified by CEDA and displayed via CEDA catalogue.

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>datasets</td>
<td>5152</td>
<td>5075</td>
<td>5089</td>
<td>5333</td>
</tr>
<tr>
<td>collections</td>
<td>538</td>
<td>476</td>
<td>481</td>
<td>573</td>
</tr>
</tbody>
</table>

6.3. HELP DESK RESPONSIVENESS

Table 6.5: 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.

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received</td>
<td>566</td>
<td>553</td>
<td>629</td>
<td>685</td>
</tr>
<tr>
<td>Closed</td>
<td>543</td>
<td>562</td>
<td>586</td>
<td>702</td>
</tr>
</tbody>
</table>

Table 6.6: Help desk queries specifically about access authorisation for restricted CEDA datasets and services received and closed by quarter, including the three-day closure rates.

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received</td>
<td>472</td>
<td>383</td>
<td>476</td>
<td>461</td>
</tr>
<tr>
<td>Closed</td>
<td>476</td>
<td>382</td>
<td>456</td>
<td>496</td>
</tr>
</tbody>
</table>
7. PUBLICATIONS AND PRESENTATIONS

Bennett, ‘JASMIN - Supporting Diverse User Communities on a Private Cloud’ Platform, Presentation, UKTO Collaboration Meeting, Coseners House, 14-16 March 2018

Callaghan, S. ‘The Devil’s in the Details-Lessons in Operationalising Data Sharing and Credit in the Geosciences’, AGU Fall Meeting Abstracts, 2017


Conway, E., ‘PV2018 (Preservation and Value Add Conference 2018), Poster at NCEO/CEOI Conference, Bath, June 2017


Donegan, S., ‘EO Data and Services at CEDA’, Poster at NCEO/CEOI Conference, Bath, June 2017


Kershaw P. J., Identity, Entitlement and Access Management Working Team, Presentation, 7th Annual ESGF Face-to-Face Meeting, 2017 San Francisco, 4-8th December 2017

Kershaw, P. J., J. Churchill, A. Iwi, S. Pepler, N. Massey, M. Pritchard, M. Pryor, A. Stephens, Managing growth and complexity – technologies to meet the challenges of operating data, services and infrastructure at scale, Poster, 7th Annual ESGF Face-to-Face Meeting, 2017 San Francisco, 4-8th Dec 2017


Pamment, J.A.; ‘CF Controlled Vocabularies’, Oral presentation for EarthCube Advancing CF-NetCDF Workshop, Unidata, Boulder, Colorado 6-8 September 2017


Townsend, P. ‘Weighing up big data…’, Poster, NCAS staff meeting, February 2018.


Williamson, E. and CEDA Team, CEDA Services and EO Data, NCEO Researchers Forum. 5-6 February 2018

Williamson, E. ‘Sentinel Data at CEDA’, Poster at NCEO/CEOI conference, Bath, June 2017