Assessment Information

CoreTrustSeal Requirements 2020–2022

Repository:	Centre for Environmental Data Analysis
Website:	www.ceda.ac.uk
Certification Date:	29 June 2021
This repository is owned by:	**Science and Technology Facilities Council**

Centre for Environmental Data Analysis

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Publication repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

The Centre for Environmental Data Analysis (CEDA) Archive (http://archive.ceda.ac.uk) provides data curation services for the atmospheric, climate change and earth observation research communities as the UK Natural Environment Research Council (NERC: https://nerc.ukri.org/) designated data centre for those research communities. This complements the other 4 data centres provided by NERC, together serving the main environmental research communities within the UK and internationally (see https://nerc.ukri.org/research/sites/data/).

The CEDA Archive services are provided by the Science and Technology Facility Council (STFC: https://stfc.ukri.org/) on behalf of NERC. Both NERC and STFC are now part of UK Research and Innovation (UKRI: https://www.ukri.org/).

The archiving work of CEDA is primarily funded via NERC, though other funding for other activities also comes via specific projects. These additional projects are funded by bodies such as the European Space Agency (ESA) and the UK government's Department for Business, Energy & Industrial Strategy (BEIS) and Department for Environment Food & Rural Affairs (DEFRA).

In addition to CEDA's archive service (CEDA Archive), which this Core Trust Seal application is made in relation to, CEDA also offers a high performance data analysis system, provided via the "JASMIN" infrastructure, supported by the Scientific Computing Department within STFC, co-located at STFC's Rutherford Appleton Laboratory site in Oxfordshire, UK.

CEDA Archive holds data assets from NERC-funded atmospheric, climate change and Earth observation science research grants and programmes alongside other third party data provided as part of a "facilitation" mode of the CEDA Archive service. This facilitation service aids the research community's access to data produced from organisations such as the Met Office which would be otherwise too complex, expensive and/or inaccessible for individual researchers or groups to access, handle or curate themselves.

Though the CEDA Archive also operates the UK's Solar System Data Centre, this application is not made in relation to those data as they remain a largely separate service from the main CEDA Archive. Connections between the CEDA Archive and the UKSSDC are predominantly through shared internal computer and staffing resources and management oversight, rather than end-user services.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

The designated community for the CEDA Archive is principally atmospheric, climate change and earth observation researchers from the UK. More generally we aim to support International Environmental Science.

Users are typically in academia (~74%) and government bodies (~14%), with users also coming from undergraduate users, industry and third sector organisations (e.g. charities). Around 66% of users are UK based with another ~10-15% from Europe, ~15% from elsewhere. (See section 6.1 of the 2019-20 CEDA Annual Report: http://cedadocs.ceda.ac.uk/1489/ for typical figures).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

CEDA’s mission is to deliver long term curation of scientifically important environmental data at the same time as facilitating the use of data by the environmental science community. Data held by the CEDA Archive provides an evidence base for scientific programmes, underpins scientific research and is a source of critically important scientific data for inclusion in future scientific projects, applications, proposals and decision support systems. The CEDA Archive facilitates scientific peer-review processes, open access and re-use of nationally and internationally important datasets as well as the generation of impact for the wider UK industry and businesses.

CEDA was established in 2005, when the activities of two of the Natural Environment Research Council (NERC) designated data centres were merged: the British Atmospheric Data Centre (BADC), and the NERC Earth Observation Data Centre (NEODC). Consolidated annual reports have been produced since 2009. Prior to that time the BADC had been providing long-term archiving since 1994, superseding the Geophysical Data Facility (GDF). The GDF was previously funded by the then Science and Engineering Research Council (SERC) and primarily supported what was then called the "upper atmosphere" remote sensing community (based as it was on studies of the atmosphere between 10 and 400 km). However, a survey of the NERC community (Carruthers and Thornes, 1995: Development of the Atmospheric Science and Technology Implementation Plan), resulted in an increased remit for the newly renamed BADC to support the entire NERC atmospheric science community. Over the years the BADC, and likewise now for CEDA, services are provided to a wider community than just the atmospheric sciences community (e.g. medicine, biology, waste management, marine sciences, ecology).

The NEODC was established in 1998, previously known as the NERC Scientific Services Data Centre (NSSDC), with an initial remit to maintain an archive of all satellite data purchased for the community since 1972 and all Earth Observation (EO) data acquired by the NERC Airborne Research & Survey Facility (ARSF) since 1982, which grew from 2005 to a wider remit as outlined in a NERC Science and Management Audit in 2005.

More recently the CEDA Archive has also assumed the data holdings of the Landmap data service (landmap.ac.uk), continuing to preserve and provide these data for the long-term, and archived satellite imagery previously provided through the NERC Earth Observation Data Acquisition and Analysis Service (NEODAAS) Dundee Satellite Receiving Station, when funding to these services ceased.

The CEDA Archive is also part of the Natural Environment Research Council's (NERC) Environmental Data Service (EDS) alongside its sister data centres: the British Oceanographic Data Centre (BODC, Marine), the Environmental Information Data Centre (EIDC, Terrestrial and freshwater), the National Geoscience Data Centre (NGDC, Geoscience), and the Polar Data Centre (PDC, Polar and cryosphere). The NGDC and EIDC data centres are both holders of CTS certification (https://www.coretrustseal.org/wp-content/uploads/2018/01/National-Geoscience-Data-Centre.pdf and https://www.coretrustseal.org/wp-content/uploads/2019/08/Environmental-Information-Data-Centre.pdf respectively). Through the EDS the data centres provide coordinated services such as the NERC Data Catalogue Service (NERC DCS), which acts as a central catalogue of all NERC data centre assets.

The majority of the datasets held by the CEDA Archive are openly available for future use in accordance with the UK’s Open Government Licence (OGL) or Creative Commons licence with no barriers to re-use by external users or

communities. Other data (mainly third party datasets from organisations such as the UK Met Office, European Centre for Medium-Range Weather Forecasts and ESA) are available under specific licencing conditions. No charge for use of data from the archive is made.

The CEDA Archive has its own website (http://archive.ceda.ac.uk/) which provides users with information on: how to find, use, deposit & cite data; what support is available; and, our guiding principles such as our acquisition and preservation policies.

The CEDA Archive has a range of policies that underpin the functions and services it delivers, including the NERC Data Policy, metadata, collections, data management planning, digital preservation and preferred formats.

Broadly speaking the level of curation offered by CEDA to the research community falls into one of three categories, depending on the requirements of the data providers (primarily NERC funded researchers) and the end-user community. This is defined during the initial Data Management Planning phase of the data lifecycle between the CEDA Archive and the project. The three levels of curation are:

1) Reference Only - data are taken as-is from the data provider, though ensuring that the data either adhere to archive suitable data formats or that suitable documentation encapsulates the data format/structure to ensure they remain future-proof. This is typically used for data where little onward use is envisioned, but a formal archive of the data is required for referencing within the scientific literature.

2) Standard Archiving - CEDA Archive staff provide guidance and assistance to the data providers to ensure that their data is suitably formatted for long-term preservation with detailed metadata (internal file metadata, data catalogue content) captured at all levels following community established practices (e.g. Climate and Forecasting conventions, Dublin Core, GEMINI 2).

3) Enhanced Archiving - In addition to standard archiving levels of curation these data may have additional services built around them to enhance their use. Typical examples are sub-setting tools within the CEDA WPS service or inclusion within dedicated geo-temporal search tools such as the CEDA flight-finder and CEDA Satellite Finder tools. See http://archive.ceda.ac.uk/tools/ for links to available data tools.

For some datasets CEDA Archive staff will help amend provided files to address issues with their formatting and content, for example the CEDA Archive work to support data disseminated through the CEDA Archive as part of the Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC-DDC).

At all levels the data are suitably catalogued in an ISO-19115 compliant data catalogue to ensure data discoverability.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

External back-up copies of the CEDA Archive content are held by the Joint European Torus (JET) facility in the UK for off-site redundancy. This is provided within a reciprocal arrangement between JET and STFC's SCD department.

Additionally, CEDA are a tier-one site within the Earth System Grid Federation (ESGF). Through ESGF CEDA hosts a range of key climate change datasets (e.g. CMIP5 and CMIP6) which are duplicated around the various nodes within the Federation to aid access, redundancy and overall storage of such vast dataset collections.

Within CEDA's host organisation (STFC) system management of the JASMIN infrastructure within which the CEDA Archive resides is provided by the Scientific Computing Department (SCD). SCD also provides tape infrastructure to other STFC departments within which CEDA Archive tape back-ups reside.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

CEDA are also involved in a number of projects with external partners, with varying levels of involvement from the archive side of CEDA activities (e.g. provision of metadata content). See https://www.ceda.ac.uk/projects/ for details of these projects.

CEDA staff are also involved with various standards groups, for example aiding the management of the Climate and Forecasting (CF) standards names tables.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The Centre for Environmental Data Archive is a NERC Data Centre provided as part of NERC's National Capability Commissioning of the NERC Environmental Data Service (EDS: see https://nerc.ukri.org/innovation/activities/environmentaldata/digitalsolutions/news/townhall/environmental-data-service/ for further details), the delivery of which is through CEDA's host organisation, the Science and Technology Facilities Council (STFC). CEDA manages nationally-important datasets concerned with meteorology, atmospheric composition, climate change and earth observation. The CEDA Archive is required to hold scientific programme outputs from NERC funded research in related research domains and our mission is stated on our website ("To provide data and information services for environmental science") as delivered via out objectives and strategies, see: http://www.ceda.ac.uk/about/our-mission. The CEDA Mission statement, strategy, objectives etc. were approved by the CEDA-JASMIN board in 2017.

We recognise the value of scientific data and the importance attached to the long-term professional management and preservation of data assets. Data are vitally important both as an evidence base for existing scientific projects and for future re-use. This belief is encapsulated in our Acquisition Policy (https://help.ceda.ac.uk/article/4857-acquisition-policy) along with the NERC strategy (http://www.nerc.ac.uk/about/whatwedo/strategy) and Data Policy (https://nerc.ukri.org/research/sites/data/policy/).

Data management plans (DMPs) covering data to be deposited (with the exception of 'third party data') are agreed

between the CEDA Archive and those responsible for the provision of data to the archive (typically the PI of the providing project or equivalent). Each DMP is individual to the project delivering data, but follows standard CEDA Archive DMP templates (examples available on request). DMPs are classed as ‘living documents’ and may be amended as the providing project needs evolve, but set out basic agreements and responsibilities. For example, they will agree on any embargo period for the data (in line with NERC data policy should that apply), end user licencing to use and the Deposit Agreement under which the data will be provided to the archive. The depositors agreement is a standard CEDA Archive agreement (http://artefacts.ceda.ac.uk/licences/depositors_agreement) that each data provider has to agree to for each deposit. A data provider is typically someone working for PIs within a project as opposed to the PIs themselves.

Most data ingested are eligible for digital object identifier (DOI) assignment if they conform to the NERC DOI guidelines (see https://nerc.ukri.org/research/sites/data/doi/), though not all data providers request this. This demonstrates our commitment to retaining and making the data available in perpetuity (subject to any agreed embargo periods, agreed retention periods in the Data Management Plans as informed by our data policies). Not all datasets are eligible for DOI minting as they do not meet the NERC DOI guidelines. Reasons why DOIs may not be given to a dataset include the inability to obtain permission from the dataset owner for the dataset (especially for older datasets) or that they do not meet the requirements laid out in the NERC data citation guidelines: https://nerc.ukri.org/research/sites/environmental-data-service-eds/doi/data-citation-guidelines/ which are used to ensure the long-term viability of the dataset itself. A typical example where CEDA would not mint a DOI for a dataset is where the data are obtained under a 'facilitation mode' - i.e. where the data centre is obtaining a copy under licence from a third party provider (e.g. a national meteorological agency) to aid academic access to such data and the licence may contain a clause which could result in future redaction of the dataset in question (though a tombstone landing page in the data catalogue would remain).

Third party data may also be obtained and archived by the CEDA Archive as part of it's 'facilitation mode' to aid access to data for researchers/wider community that would otherwise be unavailable, e.g. due to costs for individual projects/users or other overheads (e.g. provider unable to support any significant volume of requests for access). Typical examples include observational and modelling datasets from the UK's Meteorological Office.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Licensing information, terms of use and any required citations/acknowledgements are included in the metadata record describing each dataset (e.g. http://dx.doi.org/10.5285/1aa2df5a-798b-46c7-b74a-421f9ca0aa82 within the ‘details’ tab). Licence information is also indicated within catalogue and licence ‘signposting’ files at the top of each dataset within the archive as an aide-memoire for users accessing the archive directly without first referring to catalogue records.

A variety of different licences are in use across the archive, though open data licences are encouraged to be used by data providers where possible. For example, data produced from public funding in the UK should be available under the terms of the UK Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/), whilst non-UK funded data are encouraged to use Creative Commons licences where possible.

Open data licences permit onward sharing of the data which may be undesirable for some data providers and thus two generic ‘closed’ data licence are also available which honour all other aspects of open data, but prohibit onward sharing. These are the Closed-Use General Licence (http://artefacts.ceda.ac.uk/licences/cugl) and Closed-Use Non-Commercial General Licence (http://artefacts.ceda.ac.uk/licences/cuncgl).

Where one of the generic licences referred to above can not be used for a particular dataset a specific licence will be used instead. Often when specific licences are in place access to those data will also be restricted requiring a user to apply for access. Such applications are then vetted either by the data provider themselves or by someone they have delegated this responsibility to. For some cases this role has been delegated to CEDA. Of particular note are Met Office products made available to the research community via the CEDA archives under the NERC-Met Office agreement: http://licences.ceda.ac.uk/image/data_access_condition/ukmo_agreement.pdf. Where access needs to be applied for a specific, time-limited licence will be granted to the successful applicant. CEDA makes use of an in-house tool to link the application details and the licence terms and conditions together to form the specific licence which includes the date-time at which the licence (and access) will expire. (Access may be terminated before this in cases where the user's eligibility ceases or the licencing for the resource is replaced by a more permissive licence - e.g. as data passes from an embargo period to being open access).

A full list of current and historic data licences can be found here:
http://artefacts.ceda.ac.uk/licences/

Where specific access constraints are not required data access is either by the data being ‘publically’ accessible or available to all registered users. Any user (over 18) can self-register for a free CEDA user account. ‘Public’ access means that access is available to both registered and non-registered users. Registered-user access is used primarily to provide useful, anonymised data download statistics and to aid user support.

All access, regardless of access control type, is in agreement to the CEDA Conditions of Use for CEDA services: https://help.ceda.ac.uk/article/4640-ceda-archive-terms-and-conditions. All data usage is required to be inline with the dataset’s licence.

Non-compliance with licence conditions cannot be actively checked once a user downloads data from the archive. However, when breaches of licence or access conditions have been identified the offending user is contacted to raise the issue. Usually this resolves the issue, but CEDA reserves the right to terminate access where this is appropriate, for example where access has been granted in line with ‘academic use only’ and a user has been identified as having changed to a non-academic institute. Such status changes may come to light thanks to the CEDA User database allowing CEDA staff to be aware of changes of user account details. In exceptional circumstances licence breaches will be notified to the data owners for them to pursue further.

No charges are made for access to data held in the CEDA Archives.

Data providers are required to agree to deposit conditions before submitting data to the CEDA Archive: https://arrivals.ceda.ac.uk/agreement/

It should be noted that the Freedom of Information Act 2000 (FOIA), Environmental Information Regulations 2004 (EIR), Public Records Act 1958/67 (PRA) and the General Data Protection Regulations 2016 (GDPR) may override any licensing agreement.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
Accept.

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept.

Response:

The CEDA Archive is part of the NERC Environmental Data Service (EDS), which has funding on a 5 year time horizon. If one of the component discipline based data centres were to fail, due to any number of reasons from reorganisation to catastrophic failure of the physical infrastructure, then the other data centres in the EDS would be the first line of defence in providing continuity of access. The data are catalogued and stored in discrete packages allowing the data to be passed to another data centre.

A formal continuity plan for the CEDA Archive is under development and will be implemented in the near future, probably jointly with the other NERC data centres.

The CEDA Archive is implemented as a tenant within the CEDA managed ‘JASMIN’ high performance data analysis environment hosted at the STFC Rutherford Appleton Laboratory (RAL) this is also funded with a 5-year time horizon by NERC. This enables the CEDA Archive to utilise the enterprise-level data storage, servers, systems and local/wide area networks including high-speed access to the UK joint Academic Network (JANET). This is provided by the Scientific Computing Department (SCD) at RAL.

As detailed in R10, each data holding has an associated Data Management Plan (DMP) which establishes the responsibilities of the data provider and the archive. Furthermore these DMPs record the retention period for each data asset that they cover which the archive.

Reviewer Entry

Reviewer 1

Comments:

A formal continuity plan is indeed required for Compliance Level 4 and should be ready by the time of the next certification.

Thank you for clarifying the final section.

Reviewer 2

Comments:
A formal continuity plan will be needed for level 4 compliance.

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA Archive complies with the STFC Information Security Policy and Data Protection Policy to safeguard and protect the information and assets in its care (see https://help.ceda.ac.uk/article/4639-privacy-and-cookies). All staff working within CEDA are required to undergo training on GDPR and Information Security.

The CEDA Archive also complies with NERC Policies in regards to ethical working (see relevant items under https://nerc.ukri.org/about/policy/policies/). This provides the guiding principles applied to all aspects of the operations of NERC and its component research and data centres. It includes guidance on procedures for staff who have concerns about research procedures or identify breaches in the ethical policy. Serious concerns are referred to the NERC Ethics Board who will consider the issue and have the power to take any necessary action. The Board is accountable to the Chairman of NERC.

NERC also publishes a Research Grants and Fellowship Handbook (https://nerc.ukri.org/funding/application/howtoapply/forms/grantshandbook/) that includes guidance on research ethics. Researchers are required to comply with the UKRI Policy and Guidelines on Governance of Good Research Conduct (https://www.ukri.org/files/legacy/reviews/grc/rcuk-grp-policy-and-guidelines-updated-apr-17-2-pdf/), which should be read in conjunction with the UK Concordat to Support Research Integrity (https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2019/the-concordat-to-support-research-integrity.pdf) and the guidance on Good Research Conduct and Research Integrity (http://www.nerc.ac.uk/about/policy/policies/research-integrity/). These policies and guidelines apply equally to researchers, support staff, research administrators, Research Council staff and all individuals contributing to the Research Councils' peer review process.

The CEDA Archive complies with the Freedom of Information Act 2000 (FOIA), Environmental Information Regulations 2004 (EIR) and the General Data Protection Regulations 2016 (GDPR). These legislative requirements are included in the NERC Records Management Policy (https://nerc.ukri.org/about/policy/foi/records-management-policy/) and NERC Data Policy (https://nerc.ukri.org/research/sites/data/policy/data-policy/). In cases of non-compliance with these conditions, the UKRI can invoke its disciplinary policy to ensure that the highest standards of behaviour and conduct in research are met (https://www.ukri.org/files/termsconditions/rcukukriterms/disciplinary-pdf/).

The CEDA Archive has a rigorous and comprehensive ingestion process in place (https://help.ceda.ac.uk/article/4660-depositing-data-at-ceda-a-step-by-step-guide). Depositors are required to agree to a Depositor Agreement ahead of any deposition (https://arrivals.ceda.ac.uk/agreement/) and to undertake data management practices in accordance with the associated Data Management Plan agreed with the data centre.

The CEDA Archive does not accept data where there is a disclosure risk and ask that any data containing sensitive information be anonymised or be of a sufficiently broad scale that sensitive location information is not made publicly available (https://help.ceda.ac.uk/article/4857-acquisition-policy). CEDA staff are trained in General Data Protection Regulation (GDPR) requirements are will be aware of such considerations when communicating with data providers about the data they wish to deposit with the CEDA Archive. However, given the nature of CEDA's domain (i.e. the physical environment) the likelihood of sensitive information being presented for archive is very low, but where it does occur the assigned CEDA staff member will highlight the situation to the data provider and work with them to ensure that data are sufficiently anonymised ahead of presentation for archiving.

In the unlikely event that data are made available that contains sensitive information, the CEDA Archive will follow actions in accordance with the withdrawal policy (https://help.ceda.ac.uk/article/4730-withdrawal-policy) and a new version of the dataset without the sensitive information would be made available if appropriate.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA archive is managed by staff of the CEDA Division in the RAL Space department of STFC, based on the Harwell Campus in Oxfordshire. The CEDA Division consists of 30 staff (29 FTE), all but 1 permanent, with the CEDA Archive activities accounting for around 12 FTE. CEDA also pays for additional effort in other divisions/departments within the host site to support system and hardware administration and tape backup as required.

The CEDA Archive is, and always has been, dependant on funding from NERC to carry out its activities. The CEDA Archive is currently funded via the NERC Data Centre Commissioning project which began in April 2018 and runs for five years as part of the Environmental Data Service (EDS - see https://nerc.ukri.org/research/sites/environmental-data-service-eds/ for more details). The EDS includes the provision of a data archiving services for a range of environmental domains of which the CEDA Archive has responsibility for Earth Observation, atmospheric and climate change domains. CEDA reports annually to NERC on the number and quality of the agreed functions and services.

The commissioning process for the subsequent funding of NERC's data activities is presently being initiated with an initial 'NERC Environmental Data Service – Future Capability and Service Requirement' Town Hall meeting taking place in March 2021 (see https://eidc.ac.uk/townhall). This demonstrates NERC's commitment to the future of data service

provision for the foreseeable future.

As part of the process to ensure continued funding of the NERC data centres, a commissioning process was initiated in 2016 that included a stakeholder survey to evaluate the services each of the data centres, including the CEDA Archive, must provide to its designated community for the future (see: https://webarchive.nationalarchives.gov.uk/20180103180938/http://www.nerc.ac.uk/about/whatwedo/engage/engagement/datacentres-survey/). The results of this survey have been used to guide the priorities and services delivered to users by the CEDA Archive. The outcomes of this process also form the basis for planning future stakeholder engagement activities that will include user surveys and mechanisms to provide feedback on services delivered by the CEDA Archive e.g. web-based feedback forms.

NERC provides the long term National Capability (NC) funds providing the CEDA Archive in support of all NERC earth observation, climate change and atmospheric science via data management and data dissemination services. CEDA as a whole (i.e both via the CEDA Archive service and the JASMIN data analysis service) maintain close links with NERC's National Centres for Atmospheric Science and Earth Observation (NCAS and NCEO respectively) which enables experienced data specialists to work closely with NCAS and NCEO scientists and other HEI research partners. This has enabled a strong understanding of the needs of researchers across these science disciplines and experience of co-delivery of data-focussed projects. As a result, CEDA constitutes a highly efficient and effective critical mass of expertise in data management and data science.

CEDA staff are experienced environmental data specialists (data managers, data scientists, web developers and software engineers) with in-depth knowledge of the science they support. Most staff members have undertaken relevant undergraduate or post-graduate studies either in the research domains supported and/or related to the technical services provided. CEDA staff continually review and improve the efficiency of technologies and procedures from ingestion to delivery. These staff have expertise in a range of skills including digital preservation, scientific data accession, active data management planning, information architecture, international and regional data standards and web services. CEDA utilises differing proportions of effort from a range of these staff in the course of delivering its functions and services.

CEDA staff have access to a comprehensive learning and development programme provided to all employees of both CEDA's parent organisation (STFC) and principal funding organisation (NERC), ensuring they are up to date with new developments in IT and data management techniques through relevant training. STFC also holds the UK Investors in People accreditation that embodies appropriate professional development strategies.

The CEDA Archive's governance arrangements set the strategic direction, allocating resources, reviewing progress and making decisions. The Head of the STFC Centre for Environmental Data Analysis reports annually to NERC on the delivery of the agreed services and functions identified in the NERC Data Centre Commissioning Proposal.

In addition to the governance associated with the NERC Commissioning Process, CEDA also gets independent advice from a CEDA (& JASMIN) steering board (members represent NCAS, NCEO, NERC, STFC, RAL Space, Scientific Computing Department and the UK Space Agency (UKSA)) which meets at least annually to review and advise on strategy and implementation.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Staff that manage the CEDA Archive are employees of UK Research and Innovation (UKRI) - Science and Technology Facilities Council (STFC). STFC, holders of the Investors in People Gold award, encourages and supports on-going employee training and relevant accreditation to ensure that all staff have appropriate and up-to-date knowledge and skills. Staff are required to set at least one new relevant training and development objective annually and an appraisal is held at least bi-annually to ensure training and development requirements are being met.

CEDA staff work closely with staff from two NERC Research Centres, the National Centre for Atmospheric Science (NCAS) and the National Centre for Earth Observation (NCEO). These two research centres conduct a broad range of scientific programmes and oversee the scientific scope and priorities of CEDA archive services.

CEDA staff proactively engage with a number of initiatives and organisations that provide expertise on a range of relevant topics to ensure that the services provided by the data centre are in line with the current best industry practice. This

includes bodies such as the Research Data Alliance (RDA), and metadata standards groups (e.g. CF standards names list, CMIP).

CEDA have provided support and training to the environmental science community in a variety of webinars and workshops with material made available online via the CEDA website (https://www.ceda.ac.uk/events/ under Past Events). Other ways that CEDA staff have engaged with the community include:

- Data science skills - as well as the workshops and webinars highlighted above, CEDA staff have also presented at a number of summer schools about data management practices. These include the annual NCAS Arran summer school for new PhD students and the Pratt Summer School for data management Masters students from the US visiting the UK.
- Data management planning, guidance & support - are provided as ongoing support for data providers covering data management planning, preparation and upload for ingestion (central NERC guidelines are available at: https://nerc.ukri.org/research/sites/data/dmp/ and CEDA specific guidelines on the CEDA Archive help site: help.ceda.ac.uk). CEDA are also a key member of the NCAS Data Project, an end-to-end approach to standardise data capture, data production, data delivery and data archiving, centred on data formatting and metadata standards to aid future re-use (more details available under : https://sites.google.com/ncas.ac.uk/ncasobservations/home/background/history).
- Staff at CEDA are also actively engaged with organisations collecting data citation metrics (e.g. Google and Elsevier) and various publishers of data papers e.g. Nature (Scientific data), Earth System Science Data and Elsevier etc.

CEDA staff regularly interact with staff from other NERC data centres (National Geoscience Data Centre at BGS, Polar Data Centre at BAS, British Oceanographic Data Center at NOC and Environmental Informatics Data Centre at CEH) both on an informal basis and also through a formal NERC Data Operations Group (DOG) that coordinates and advises on various aspects of data management policy across the entire research council. In addition, the Science Information Strategy is led by the NERC Information Strategy Group which comprises Directors of Institutes, Heads of NERC Data Centres and external advisors; they are currently developing the new Science Information Strategy for the next five years.

CEDA Archive staff also attend key community meetings requiring data scientist input (for example, observational facility user group meetings). This ensures that such data producing communities are regularly engaged with to ensure data requirements, guidance and information flows between the facilities, end-users and archive.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

All data deposited with the CEDA Archive follows standard CEDA Archive data uploading tools (see section R12 for further details) and requires the depositor to have a CEDA user account. Deposition does not directly lead to archiving of data and data are only ingested into the archive by a designated CEDA data scientist. As part of the ingestion process the data scientist will be in communication with the data provider to discuss the deposit and will undertake checks on the data to ensure that they are prepared for archiving. The exceptions to this workflow will be those cases where CEDA staff pull in data from external third party data sources as part of the facilitation mode of the archive.

During data acquisition the dedicated data scientist ensures supporting metadata are obtained for the accompanying ISO 19115 compliant data catalogue metadata record (https://catalogue.ceda.ac.uk) and to capture any required supporting documentation. The dataset catalogue metadata includes a data lineage field which conveys the origins of the data in the CEDA archive and also records other important changes to the data (e.g. if the data are removed at a later stage then both the date-time and reason for removal are recorded). Additionally, the catalogue record will link to other supporting material where possible about the background of the data (e.g. calibration information, instrument details). Some of this supporting material will be external documents either stored in a remote service or within CEDA's document repository to ensure long term availability or may be other dedicated records within the CEDA metadata catalogue, such as Instrument, Platform and Computation, capturing how the data were produced and Project records capturing why the data were produced. Through connections to these additional, reusable records, inclusion within listings within Dataset Collection records and direct 'related dataset' links users of the data catalogue are able to find other related datasets within the CEDA data catalogue.

The catalogue dataset record also records the completeness of the dataset and any associated update frequency, as well as its publication state. All these fields conveying the completeness of the dataset use standard code lists and are

exported within the ISO19115 version of the catalogue record.

Changes to the metadata entries are done through an administrator tool within the Django web framework used for the catalogue and so are logged automatically. Regular backups of the catalogue ensure that it is possible to roll back the catalogue to earlier states should this be necessary.

Data ingestion to the archive uses standard CEDA tools which default to prevent changes to archive content where the dataset has been marked as complete in the data catalogue. Similarly, for ongoing dataset (e.g. for long-term measurement observation facilities) files already in the archive can not be overwritten unless this option has been actively selected. However, not all the data in the CEDA Archive are immutable. While the majority of datasets are expected to be complete and unchanging, some data have good reason to change. For example a dataset may be added to to extend its coverage e.g. time series. Where the data are too large for limited storage resources, then a rolling archive may be created e.g. satellite data where just the last 2 years of data are retained. To differentiate between these classes of data, the data status is recorded in the catalogue and is used to block changes to immutable datasets.

Data providers may provide checksums with their deposits to ensure data veracity, though this is not typical for small scale, one-off data providers. Where these are provided the CEDA data scientist will use these checksums at various points along the ingestion workflow to confirm that data were delivered and ingested intact. No changes to data are made without the data provider's involvement with the exception of potential migration to alternative formats in the future as agreed within the relevant data management plan. Internal checksums are calculated on all data files following ingestion and used for internal CEDA fixity checks.

For internal fixity checks, CEDA Archive operates an automated audit tool that routinely tests all the checksums for data held by the CEDA Archive and produces a regular report providing warnings if changes to the data are detected. The reports will then trigger any required action to investigate the cause of the changes and, where required, to recover copies from backups to replace corrupted files. Some details and examples of this tool can be seen in Pepler, Sam (2018) Fixity checking a large climate data archive. Available from: http://cedadocs.ceda.ac.uk/1451/.

Supporting documentation may be changed after deposit of the data. Data Centre staff may improve the supporting documentation as part of a continual process of improvement, sometimes resulting from feedback from users or at the request of the depositor themselves. Supporting documentation is either kept in the CEDA Document Repository (http://cedadocs.ceda.ac.uk/) for fixed assets, in the CEDA 'artefacts' service (http://artefacts.ceda.ac.uk/) for information that may change over time (but is held in a SVN repository to capture any changes) or, where material is suitably archived online (e.g. journal articles), content can be directly connected to from data catalogue entries.

The CEDA Archive has a publicly available DOI policy (https://help.ceda.ac.uk/article/146-data-citations-and-dois, but in the process of being updated) which makes clear the limitations of changes that can be made to data once a DOI has been issued. These include handling of growing datasets which may be appended to in limited circumstances, but original data may not be overwritten. If a dataset is later found to contain errors or the depositor wants to change it in otherways a new ingestion must take place with a new metadata record and a new DOI. The dataset being replaced still remains

available for users to access, however, the catalogue entry will clearly indicate that it has been superseded and link to the latest version available. This mechanism is used for non-DOIed datasets too.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Ahead of accepting any data for ingestion into the CEDA Archive initial contact with the providing project will gather basic information about the data to be considered. For NERC research council funded projects the initial list of data products may be provided via an Outline Data Management Plan (ODMPs) the funding grant holder will have submitted with their grant application. Within NERC all ODMPs for successful grant applications are then reviewed to allocate to a relevant NERC data centre, determined by the data centre supporting the research domain the project is most aligned to. A CEDA Data Scientist will then be allocated to co-ordinate the data management for the project who will then undertake correspondence with the project.

Non-NERC funded research projects may also approach the CEDA Archive for data archiving services. Such enquiries initially seek to establish the scope of the data to be archived and, if required, available funding to support any required

work.

Following initial contact with the research project team the CEDA Archive will then assess the data archive request as follows:

- Are the data within the scope of the CEDA Archive and would meet the requirements set out by our acquisition policy (see https://help.ceda.ac.uk/article/4857-acquisition-policy). If the subject domain is determined to be beyond the scope of the CEDA Archive the applicant will be suitably signposted either to an alternative NERC Data centre or another external service.
- Are the data suitable for long term archiving? The assessment is carried out using the NERC Data Value Checklist (see https://nerc.ukri.org/research/sites/data/policy/data-value-checklist/)

If the data are determined to be within the scope of the CEDA Archive and of long-term value, with resources to be able to support the acquisition, the next stage is to establish a full Data Management Plan (DMP) with the project team. This DMP establishes the responsibilities and expectations of both the project team and the archive. It also establishes the data deposit agreement that will be used, any embargo requirements, data retention periods and end-user data licences. It also formally records the anticipated data to be archived and establishes any third-party data that are of interest to the project and where they are to be sourced from (typically, these are products already obtained by CEDA which are reused by many projects).

The DMP remains a living document recognising that it is established (ideally) at the outset of the research project and thus needs to evolve as the project's final data products are realised.

The data management support for the research project is tracked internally with the CEDA Data Management Planning tool and correspondence is carried out through a linked dedicated helpdesk. This system automatically generates dated actions in relation to the project as well as ad-hoc tasks raised by the allocated data scientist to be undertaken as needed.

Once data are ready for archiving an allocated data scientist will work with the data provider to prepare their data in a suitable format and follows a suitable file-naming convention. A series of help articles have been written to guide the data provider alongside support from their supporting CEDA data scientist (e.g. see https://help.ceda.ac.uk/article/4661-depositing-data-faqs). The CEDA Archive encourages the use of a select few non-proprietary formats where possible (see https://help.ceda.ac.uk/article/104-file-formats), though other formats can be accepted. Proposed formats undergo an internal review by CEDA, which may accept the new format if sufficiently supported in terms of documentation, tools and user community or recommending an alternative, already accepted format. Internal file-metadata are also encouraged to follow best-practice (see https://help.ceda.ac.uk/article/4428-metadata-basics) to ensure the long-term usability of the data. However, the level to which data required to follow best-practice is set by the preservation level of the data (see https://help.ceda.ac.uk/article/4691-levels-of-data-curation).

Data delivery to CEDA may be via various routes as appropriate to the nature of the data to be supplied. For one-off, small

scale deposits a dedicated web based data deposit service (https://arrivals.ceda.ac.uk) is available. This requires the depositor to also submit a metadata file alongside the deposit which gathers key discovery level metadata for use with the CEDA data catalogue (see R7 for further details). Other routes, more suited to large-scale or continuous deposits, include FTP and RSYNC upload and CEDA pulling data from dedicated local workspaces or external resources. In all cases the data are reviewed by the data scientist to ensure the file format, file-names, overall structure and internal metadata are sufficient for archiving. Where further action is required by the data provider (e.g. amending of internal metadata) this is communicated via the dedicated helpdesk. Data checks carried out on the data include use of format checkers and metadata checkers, e.g. the CF-netCDF and BADC-CSV format checkers.

All data discovery level metadata feed into the CEDA data discovery catalogue which follows the EU INSPIRE/UK government (UK GEMINI) Standard. Additional, CEDA specific, metadata are also sought to complete the cataloging content to ensure full background context to the data are also curated. The data catalogue produces ISO standard 19115:2003 for geographic information metadata https://www.iso.org/standard/26020.html quality metadata records. Note, as per the CEDA Archive withdrawals policy (https://help.ceda.ac.uk/article/4730-withdrawal-policy) deaccessioned datasets (for whatever reason) will retain a their catalogue entry as a landing page for citations. Such 'tombstone' pages retain the original information about the deaccessioned asset along with additional, date-time stamped details regarding the deaccession. This additional information is appended to the data lineage statement for external visibility both on the CEDA data catalogue page and though the associated marked-up representations (e.g. schema.org or ISO record XML) for external harvesting. See https://catalogue.ceda.ac.uk/uuid/7a5fce32050a4d2fac92ac9b6c2d56d0 for an example of a deaccessioned dataset's landing page.

The metadata catalogue records associated with the data go through an internal review process ahead of the dataset's formal publication. The dataset records are also exported to the NERC Data Catalogue Service (NERC DCS, data-search.nerc.ac.uk). The NERC DCS provides a central catalogue for all of NERC's data centres the content of which undergoes reviews on key discovery fields (title, abstract, data lineage, etc). These reviews have provided a structured way to track the progress of NERCs data catalogue holdings as each NERC data centre has worked to improve their completeness, accuracy and quality over time. The outcome of these reviews has been a series of internal reports to the NERC Data Operations Group. The next round of reviews is due to take place in 2021.

Reviewer Entry

Reviewer 1

Comments:
The DMP approach you describe, with explicit and shared responsibilities, is a great example of how DMPs can benefit repositories.

Reviewer 2

Comments:

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA Archive carries out data storage according to documented processes and procedures. The documentation is maintained within a private documentation site within CEDA's cloud-based helpdesk system provided by Help Scout. This documentation service is version controlled and is additionally backed up to a local copy on a regular basis for added redundancy as detailed below.

Ingestion of data is carried out following a series of processes documented within an internal, version controlled documentation system. Any alterations to processes are first discussed internally within the data scientists team as necessary and changes agreed with the archive manager. Enacted changes to systems and processes lead to updated/new documentation. All internal documentation is stored within our help desk system which provides version control and therefore both the ability to track changes and to roll back if required. Significant changes to systems undergo pre-release acceptance testing.

Secure access for CEDA Archive staff to data ingestion systems is provided via a ssh-key login and restricted to only IP addresses with our host organisation's firewall. Secure login (via https) is also required for access to Data Management Planning tools, CEDA Data catalogue administration interface and User details/access control systems which are all locally hosted web-based applications.

Overall the repository follows the best practice guidance for archive procedures which seek to adhere to ISO14721 (OAIS Model) for storage as well as the other preservation functions. Briefly, files are backed up on tape in another building on CEDA's host campus within a fire-suppressed storage facility. Additionally, as part of the CEDA's disaster recovery plan, copies of the backup tapes are stored off-site in a secure facility. As such, if any corruption of deposited data files is discovered, backed-up copies can be retrieved initially locally or from the remote back-up where necessary.

Once the original data (Submission Information Package (SIP)) has been obtained (either pushed to the CEDA arrivals area by web-upload, FTP or RSYNC or pulled in by CEDA staff) the data are stored securely awaiting ingestion. In the main, depositors are required to take responsibility for uploading the data they are depositing, though data centre staff may pull in data to aid large and complex data transitions, especially those created within ‘group work spaces’ co-located in the JASMIN data analysis system within which the CEDA Archive and services reside. Once data, metadata information and supporting documentation have been received it is checked against what was previously agreed in the Data Management Plans (file names, format, size etc). Metadata details are entered into the CEDA Data Catalogue service by a CEDA data scientist. This retains an audit trail of changes that are made to these records and is regularly backed up. Static supporting documentation is stored in the CEDA Document Repository (http://cedadocs.ceda.ac.uk). This dedicated EPrints repository allows a full audit trail of these curated documents. In some cases supporting documentation may alternatively be added to the CEDA Artefacts Service, mainly where these may be a changeable resource over time. This is built on a SVN repository, allowing full audit trail to be established for each item and permits rolling back content if required.

All ingestion of data to the archive is undertaken through a dedicated deposit client-server system. Data providers have no write access to the archive. As of October 2019 direct write access to the CEDA archive from the ingestion servers was removed for data staff ensuring that all ingestion took place through this dedicated tool. As part of this deposit system an ingestion log is generated which is publically available at: http://data.ceda.ac.uk/badc/ARCHIVE_INFO . This system also permits archive management functions such as renaming, moving, symlinking and deletion. All processes are logged by the deposit server. Additionally, changes to the archive are flagged to CEDA’s ‘File Based Index’ system. The FBI holds an entire index of all archive content in an Elasticsearch index, storing details for all files, including basic information such as archive location, size, file name. Further file metadata may also be held where these can be obtained, including: file format and parameters. (There are some complementary indexes which also store richer metadata such as geo-temporal information, but these are not for all data types).

Authorisation of access to the archive storage holding the data is limited via permissions to the minimum viable number of systems support staff responsible for the physical infrastructure. Access to the network systems is also strictly limited. CEDA Archive systems are hosted on Virtual Machine and some physical servers. Administration (root) access to these systems is limited to the systems administration team only. The building housing the repository is equipped with a security-protected card access system within a campus with gated access control.

CEDA's host organisation, STFC, regularly updates and maintains a full organisation wide Risk Register documenting significant corporate risks using a scoring system, and outlines appropriate mitigation scenarios, including relevant risks to CEDA services.

Disaster recovery procedures are in place that include data recovery provisions involving restoring data from tapes and controlled restoration of CEDA Archive services e.g. catalogue, website etc.

CEDA's operation manual is compiled within the cloud-based Help Scout helpdesk solution used by CEDA for user-support purposes. Though the Help Scout system is a well managed cloud-based solution utilising Amazon Web Services, the CEDA Operations manual is backed up daily from the Help Scout system to local systems. Furthermore, an additional copy is made weekly to a memory stick for an offline version in an alternative building to CEDAs infrastructure to provide further redundancy.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA Archive's role as the national data centre for atmospheric, climate change and earth observation research assumes due care of the data in its care with appropriate retention periods. The present UKRI data policy states that all research data should be retained for 10 years after it was last used. However, the nature of the data held by the CEDA Archive often, but not always, warrants the data being kept in perpetuity.

The CEDA Archive has a digital preservation policy (https://help.ceda.ac.uk/article/4860-data-preservation-policy) which

documents the model under which the data centre operates, funding and the legal and regulatory frameworks the data centre is bound by.

The ‘preservation level’ for each data resource is determined within a Data Management Plan, agreed with the data provider and in accordance with the NERC data policy. The DMP details:
- The responsibilities of all parties involved in the production and archiving of the data outputs.
- Details the data products to be archived, embargo periods, deposit agreements, end-user licence(s) and retention periods for the data.

The data are not accepted by the data centre unless the DMP has been agreed by both parties. Once a DMP has been agreed, an anticipated data delivery date is set, though this may be revised as the providing project progresses. Likewise, the resulting datasets to be archived may be revised as part of the DMP remaining an active document to respond to changes in circumstances. Depositing of data within the CEDA Archive is in accordance with the agreed Depositor Agreement (see https://artefacts.ceda.ac.uk/licences/depositors_agreement). This confirms data ownership is retained by the appropriate party, that CEDA has the right to provide access to the data and translate the data to any medium or format for the purpose of future preservation and accessibility.

All DMP related activities are tracked through an internal DMP management tool and associated ‘helpdesk’ for correspondence with the data providers. Routine DMP tasks are automatically generated for each data project supported within the DMP tool. These relate to specific touch points with the project during the project's lifecycle. Additionally, specific tracking is made in relation to initial contact, agreement of a DMP and the project’s end date and specific actions are raised and tracked within the tool. The archive manager and a dedicated staff member regularly review the DMP tasks which are reviewed both within a data management group as a whole on a monthly basis and periodic individual catch-ups to ensure tasks are progressing. One of the aims of this approach is to ensure that data providers meet their data provision requirements as stated in the DMP and, for NERC funded activities, the CEDA Archive reports back to the funding body on the status of projects for which data management is being supplied. This may, if required, also lead to sanctions on the data provider from the funder should they fail to meet their data management responsibilities as covered in the NERC Data Policy.

Deposits of both data and supporting metadata and documentation are checked for archive compliance by the data centre staff before being ingested into the archive, added to the data catalogue or the data centre’s documentation services. Guidance is available to depositors on the submission standards required by the data centre. Data catalogue records undergo an internal review process before publication. Lifting of data embargo periods for data access (and changes to associated licencing) is automatic within CEDA’s data access control system once the appropriate rules and expiry dates have been set up to ensure these are enacted in a timely manner.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

NERC expects the scientific quality of the data generated by its funded grants and programmes to be assured by the scientific staff delivering such programmes. Whilst it is not feasible for the data centre to check the scientific quality of the diverse data it receives, the data centre does review the technical quality of the data submitted for ingestion through several 'sanity' checks upon receipt of data. Such checks include: data files being in the correct format; data files open and are readable in an appropriate application; the content of the file(s) is as described in the metadata; data types are appropriate and consistent; the data contains no information of a personal nature (e.g. names & email addresses) beyond those relevant to the contributors to the data. All data centre staff have backgrounds in environmental data science, and will use their experience to judge whether the content of a data file is sensible, though as stated previously, they are unable to make an assessment of scientific quality of the data.

Within some communities (most notably those of the climate modelling community) there have been steps towards reviewing data and creating data quality reports. Meanwhile other providers (most notably those from the observational community) provide details of the measurement errors and instrument calibrations where these are available. Where possible such data quality information is sourced and linked to on the appropriate catalogue record.

The CEDA Archive requires that all deposits are accompanied by comprehensive discovery metadata. Discovery

metadata are assessed against a set of quality rules incorporating the UK GEMINI metadata standard (https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini) and the NERC metadata quality guidelines. The data catalogue entries undergo an internal review process to ensure conformance to these guidelines. The metadata are also subject to a regular appraisal conducted by external reviewers organised by the NERC Data Operations Group. The CEDA data catalogue contents are harvested through a CSW service which also validates the content against the UK GEMINI standard for European spatial datasets.

Discovery metadata records are publicly available via a searchable catalogue at https://catalogue.ceda.ac.uk. Records are available both in HTML (marked up with schema.org) and GEMINI xml. They are also harvested and made available through a number of other systems including the NERC Data Catalogue service (http://data-search.nerc.ac.uk/), Find Open Data (https://data.gov.uk/) and Google Dataset Search (https://toolbox.google.com/datasetsearch/).

Discovery metadata records all contain details of how to cite the resource, but may also contain links to other related data resources and/or documents e.g. academic papers which have used the resource, grey literature related to the resource, etc. and through to other records in the catalogue that describe related details such as the related project and the tools user for data production (e.g. instruments, platforms, models etc). However, these are not mandated as the data centre is primarily concerned with the data resource itself. Related datasets are given specific links within the data catalogue, such as either being within a common Dataset Collection or some other direct dataset to dataset linking (e.g. to show supersedence/versioning).

These extra, descriptive metadata are important resources to help future users of the data to interpret and reuse them. In short, this supporting metadata should be sufficiently detailed as to enable re-use of the resource by a member of the community without further recourse to the authors of the data or the data centre. Such documentation must include an account of the data structure with a description of all variables including data types and units of measurement. It is also likely to include details of: the experimental design/sampling regime; data collection methods; data transformation methods; fieldwork/laboratory instrumentation used; analytical methods and any quality control measures employed. The descriptive metadata in the CEDA data catalogue and connected resources are available for download separately from the data resource itself in order to permit users to make an informed decision regarding the data prior to agreeing any licence and/or initiating an order/download of the data. CEDA reserves the right to amend or enhance the supporting metadata documents at any time after the deposit is complete.

Data users have the option to comment on data and/or metadata supplied by the CEDA Archive via by contacting us using our embedded help ‘beacon’ on all CEDA Archive services. Contact in this manner has led to several useful amendments or corrections to resources held by CEDA, including catalogue records, supplementary documentation, etc. On occasion data issues have been highlighted back to the data providers for their investigation and have led to new versions being issued as a result.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

All ingestion of data to the CEDA archive is undertaken according to defined processes. These processes are documented in a secure internal area of the CEDA help documents by the CEDA team, and any changes are approved by data centre management before becoming part of the working practices. All Data Centre Operatives (referred to below as the Data Scientist) conducting work ingesting resources to the CEDA archive adhere closely to the documented processes. Deviations from the processes are identified and referred to the data centre management team for approval. Communications with depositors of data are made using a dedicated official CEDA Data Management helpdesk, distinct from the general user-support helpdesk, using standard templates which are amended with information unique to the relevant ingestion.

The details and progress of all data management activities are recorded at all stages in a custom project tracking database tool. This primarily covers all data management activities relating to successfully funded grant applications by CEDA's main funder, the Natural Environment Research Council (NERC), within the scope of the data centre operated by CEDA (other science domains are covered by other NERC data centres). Other activities, such as acquisition of 3rd party datasets on behalf of the research community, are recorded within this tool too.

There are 3 categories of data ingestion into the CEDA archive based on the ingestion workflows needed:

1) New, discrete, small-scale datasets such as those from short term projects uploaded using web-tools by the data provider
2) Large (in terms of volume and/or number of files) or ongoing datasets, often spanning years where the data provider uses tools such as FTP and RSYNC to push data to CEDA
3) Large, ongoing or third-party datasets where data are pulled from external sites by CEDA staff

For the first two cases the data deposition workflows begin through initial conversations between the data providing project or individual wishing to deposit data and an assigned Data Scientist at CEDA. In most cases these initial conversations follow on from successful research grant applications with the Natural Environment Research Council (NERC) for which an Outline Data Management Plan (ODMP) will have been prepared. These ODMPs inform, where possible, initial conversations with the project/individual (typically a project PI) to establish a Data Management Plan which records data to be archived as identified using the NERC Data Value Checklist (see https://nerc.ukri.org/research/sites/data/policy/data-value-checklist/). This checklist assesses data against criteria including any statutory requirements, long-term value, reusability, uniqueness, etc. If a resource is identified as not suitable for ingestion by the CEDA archive, the individual requesting deposit will be notified via email and given reasons for the decision. If possible, another suitable data repository may be recommended to the depositor (e.g. if not within the scope of the CEDA Archive).

Assuming suitable resources have been identified and recorded within the DMP, as the data producing project progresses discussions will then move to those responsible for data provision (hereafter the Depositor). Note, the actual Depositor may be someone other than the person with whom the DMP has been agreed (e.g. a researcher working for the project PI). Ingestion of resources then follows a clearly defined process from start to finish, details of which are available to data depositors on the CEDA help documentation webpages: (https://help.ceda.ac.uk/article/4660-depositing-data-at-ceda-a-step-by-step-guide - note the variations highlighted for use of FTP and RSYNC used for large and/or ongoing data deposits).

The DMP remains a ‘living document’ and will evolve as the project or long-term data provisioning activity continues to evolve over time. The DMP will also cover matters such as end-user licencing, responsibilities for those involved with the data lifecycle and data retention periods.

‘Third-party’ data are periodically identified as a valuable resource for the CEDA Archive to obtain on behalf of the research community to aid access (e.g. due to cost, licencing and/or access constraints to source material). In such cases ad-hoc workflows will be required to set up the required mechanisms to ‘pull’ the data into the CEDA system, either as ongoing data feeds or one-off deposits, for onward ingestion. Alternatively, where possible, 3rd party data suppliers are instructed how to use standard CEDA delivery workflows (typically FTP or RSYNC). Beyond data delivery, i.e. for ingestion and onward cataloguing and documentation, all other standard workflows are followed by CEDA staff.

Once data delivery is set to commence the assigned Data Scientist will undertake the following actions:
1) Request the required Archive space, set-up the required directory structure and establish a backup schedule

2) Add an entry to CEDA's 'Security database' to set access control for download tools. This may include expiry dates to automatically open up access for embargoed data once the embargo period expires
3) Basic checks are performed on the incoming data including file format checks which confirm that the files can be read, are correctly structured and required metadata fields are present. Filenames are also checked to ensure they conform to agreed file naming conventions to be used for the dataset (e.g. for observational data that follow the CEDA FIle naming convention). Where checksums have been provided by the Depositor these are used to ensure that the data were delivered successfully with no corruptions. Note - the data themselves are not checked by the data centre.
4) Assuming all basic checks are passed, the data are then ingested to the archive using CEDA's standard deposit tools that:
4a) copy files to the archive, ensuring that these are not overwriting existing files or into already completed datasets
4b) check the files have been deposited correctly
4c) logging the successful deposits
4d) trigger scanning of files to collect file metadata (archive location, name, etc) and internal file metadata where possible (e.g. parameter details) for CEDA's File Based Index (FBI). This ensures that the files are then available through CEDA's web download tool.
5) Writing of metadata catalogue records, prepared using metadata provided by data providers for the dataset itself and also to detail related information on relevant background information such as Instrument or Project details. The principal record within the catalogue is the Dataset record relating to the ingested dataset within that catalogue. This stage will also initiate automatic harvesting of some metadata such as parameter details from the FBI tool or accompanying manual metadata store to ensure completeness and timeliness of information, especially important when dealing with large datasets. Supplementary records types are also created where required and connected to the Dataset record. These include:
5a) Dataset Collections - an umbrella record which acts to link related datasets under a common theme (e.g. all datasets used for a given project; produced from a given long-term observatory) and may also include additional third-party datasets to use alongside the principal datasets curated in the Collection. These curated collections are themselves citable objects and may also be assigned DOIs, though the requirements for such minting are more constrained than for Dataset records
5b) Instrument and Platform records - detailing the data production tool and where these are located or the vehicle on which they are mounted as part of observational data production
5c) Computation records - covering data production via algorithms and modelling
5d) Project records - capture the background reason why the data were produced
6) The catalogue records are then checked for conformity to CEDA's catalogue metadata standards (an extension of the NERC metadata guidelines) by a different Data Scientist. This is important to ensure that the catalogue record correctly describes the data archived, is understandable by future users (including those beyond the initial target audience for the data) and to ensure discoverability.

Only once all stages have been successfully completed and the catalogue record has passed quality control will the dataset be formally published. To ensure that these are complete the final publication is undertaken by the reviewing Data Scientist in step 6 above.

Following formal publication of the dataset via the data catalogue a DOI may be requested. Such a DOI request originates with the data provider/PI for the project and needs to have agreement from all dataset authors to proceed. Additionally, the dataset must also meet a series of requirements as listed on https://help.ceda.ac.uk/article/146-data-citations-and-dois. If a DOI is possible for the dataset then the assigned Data Scientist will submit a request internally where another Data Scientist will review the record to ensure that it is suitable for DOI assignment and that all required DOI metadata fields are correctly completed. The actual DOI minting is then initiated by the DOI reviewing Data Scientist which automatically uploads DOI metadata to the DataCite metadata store.

Once all publication/DOI stages have been completed the Data Scientist will inform the Data Provider and, if suitable, may also arrange for a news item to be issued by the CEDA Archive service to raise community awareness of the published asset. This is typically done for third party datasets or those principal datasets which are anticipated to have widespread importance or usefulness.

With the dataset ingested into the CEDA Archive the archive continues to manage the asset to ensure that it remains available to users for the preservation timescales as indicated in the data management plan. Typically this will be indefinite for observational data. Such onward management covers aspects such as continual archive audits to identify (and resolve) data corruptions; catalogue content reviews as part of the wider NERC Data Catalogue Service review process which covers content from all NERC data centres to monitor quality as well as ad-hoc improvements as and when identified needing to take place; migration of content to new storage and/or moving to 'near-line archive' storage where required as datasets are superseded or storage requires. Where datasets are superseded catalogue records will be adjusted to ensure that superseded datasets are flagged as such and linked to the latest versions. In the rare circumstance where data are deaccessioned the Remove Data Procedure (internal documentation, available on request) ensures that a record is kept of the removed content to ensure that an audit trail remains of such assets and that dataset citations resolving to the dataset's landing page contain details of the deaccession.

Where improvements or errors within the ingestion workflow are identified, these are raised with the data centre management, detailing the improvement or problem and the work needed to effect the required change. Subsequent actions are then taken and, where necessary, workflows are adjusted to accommodate the required changes.

Reporting of data centre outputs to stakeholders includes determining the number of data resources published in the metadata catalogue on a quarterly and annual basis, and the number of new, completed, and currently active projects in our custom project tracking database tool. The progress of the data management for larger projects is reported quarterly to our main funding bodies as per our service level agreements with them.

Reviewer Entry

Reviewer 1

Comments:
The workflow described is very good. I would highly recommend that it is published on your website for transparency and inspiration.

Reviewer 2

Comments:

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA Archive provides access to the data resources it holds via a bespoke discovery metadata catalogue (https://catalogue.ceda.ac.uk/). Data resources are described to the INSPIRE/UK government UK GEMINI metadata standard (https://guidance.data.gov.uk/publish_and_manage_data/harvest_or_add_data/harvest_data/gemini/), with the exception of a small number of data resources with no geographic element e.g. purely lab-based studies. Catalogue records are also harvested by metadata aggregator sites such as the NERC Data catalogue Service and data.gov.uk. Additionally, a CSW endpoint (https://csw.ceda.ac.uk/geonetwork/srv/eng/csw) and OAI-PMH end point (https://csw.ceda.ac.uk/geonetwork/srv/eng/oaipmh?verb=ListRecords&metadataPrefix=oai_dc)allow external users to programatically interrogating the CEDA Catalogue using either the Catalogue Service for the Web (CSW) service API or the OAI-PMH API.

Catalogue records have schema.org markup supporting harvesting by services such as Google's Dataset Search service (https://datasetsearch.research.google.com/).

The CEDA Archive is listed in the Registry of Research Data Repositories (http://doi.org/10.17616/R3CC7N), further facilitating exposure for the data resources held by the CEDA Archive.

The CEDA Archive catalogue allows users to search for data resources of interest via keywords searches matching title, abstract and keyword content. Search results can be further filtered based on the record type. Enhancements to the search functionality are presently under development. A CEDA Archive dataset catalogue record displays information about the data resource (description, accessibility, temporal and spatial extent etc.) and also displays a 'Citable as:' section which details the recommended citation for the data. Other record types in the data catalogue are supporting records providing related information detailing aspects such as the project for which the data were collected and how the data were produced/collected. These are connected in such a way as to enhance the user-experience by showing related content in the catalogue to aid further discoverability of related datasets (e.g. datasets from a common instrument or project).

The CEDA Archive is an issuing agent for DataCite DOIs (https://www.datacite.org/). The data centre has a documented process for assigning DOIs. Data providers are encouraged to request a DOI, but this is not done for all datasets as some are third party datasets where the data centre has obtained these as copies of a resource already available elsewhere to aid community access. A DOI is a unique identifier for a data resource, which can be dereferenced in a web-browser to direct the user to a ‘landing page’ containing metadata about the data resource. The CEDA Archive uses the catalogue record for the dataset and, occasionally, collection records covering multiple datasets as the 'landing page' for the DOIs it issues. In order for the CEDA Archive to assign a DOI to a data resource it must be fully ingested into the data centre to ensure it is of the required quality, and has all of the necessary supporting information to enable re-use.

It should also be noted that the present catalogue is the third iteration of a data catalogue from CEDA or its predecessors (the British Atmospheric Data Centre and NERC Earth Observation Data Centre, BADC and NEODC respectively) where citation strings including URLS have been provided to aid data citation. Care has been exercised with each migration to the next iteration to ensure that catalogue content is maintained. URLs used for records in previous catalogues have been retained on their equivalent records in the latest CEDA data catalogue with a redirection service utilising a dedicated catalogue API endpoint ensuring that users of old URLs found in the literature, for example, will be redirected as required.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Discovery metadata complying with INSPIRE 19115/19139, which in the UK also complies with the UK GEMINI v2.2 schema is required when data are deposited to the CEDA Archive. This discovery metadata ensures datasets are described in sufficient detail to be found using search parameters that include geographical coordinates or location, free text against title, abstracts, keywords etc. This also enables datasets to be exposed through appropriate external gateways e.g. data.gov.uk and the NERC data catalogue service (data-search.nerc.ac.uk). Provision of detailed discovery metadata enables users searching for data to make an initial, high-level assessment of whether the data resource being described is suitable for their requirements.

The CEDA Archive provides guidance on the file formats accepted by the data centre (https://help.ceda.ac.uk/article/104-file-formats) to encourage deposit of data in formats that are at less risk from technology/software obsolescence, or require costly proprietary software to open. The preferred formats are generally non-proprietary, or open, industry-standard formats and can be used by anyone wishing to access the data. For example, a depositor holding their data in an MS Excel spreadsheet would be asked to deposit the data as a BADC-CSV file, a comma-separated ASCII format with a rich metadata header section. The ‘gold-standard’ for data archiving in CEDA is binary netCDF data within metadata adhering to the CF-Conventions. Other formats may be accepted (see link above for those that have been previously sanctioned for ingest). New formats suggested by data providers undergo an internal review process to assess their suitability for long-term archiving. Where they are not deemed to be acceptable for long-term archiving or where an existing suitable sanctioned format already exists these will be recommended to the data provider. Acceptable formats will be those that are shown to be fully supported by documentation to allow future understanding of the format used and, where possible, an active community supporting the format and associated tools.

The CEDA Archive recognises the importance of providing as much contextual information as possible for its data holdings in order to facilitate its re-use. In order to ensure continued understanding of the data, detailed supporting documentation (in addition to discovery metadata) is captured at the data ingestion stage. The CEDA Archive provides guidance to depositors as to the depth of information required in the supporting documentation (https://help.ceda.ac.uk/article/143-supplementary-info). This supporting information may appear on dedicated catalogue

records (e.g. instrument or project details) or linked to from the catalogue page to where it is stored in suitable service for its long-term availability. Such services may be the open literature, another long-term repository or, where no suitable long-term solution is available, within the CEDA Document Repository (cedadocs.ceda.ac.uk) for fixed assets to CEDA's 'artefacts service' (artefacts.ceda.ac.uk) for assets likely to evolve over time.

Upon receipt of the data and supporting documentation, data centre staff check the contents of the supporting documentation to ensure that they are sufficient to support the long-term use of the deposited data. Staff also check that included links/references resolve to suitable end-points.

To ensure proper maintenance of the CEDA Archive storage media are routinely replaced in accordance with best practice. As such data will be migrated over to new storage media. Details of such migrations can be found in section 1.10 Data Migration report of the 2018-19 CEDA Annual report (http://cedadocs.ceda.ac.uk/1466/1/ceda_annual_report_2018-2019.pdf).

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Online storage for the data centre repository is based on the JASMIN research infrastructure. This is currently a scale out filesystem from Quobyte. This is administered by a dedicated in-house IT infrastructure team within the host organisation, STFC. Data are also stored in tape libraries according to policies for different datasets; some datasets are kept on disk; some on disk and tape; and some large datasets are on tape only. The process of making tape copies is continuous. The tape technologies used are Spectra TFinity Tape Library has an initial capacity of 65PB, 48 Drive Bays and a mix of 17 industry open standard LTO-8 (Linear Tape Open) Drives and 16 TS1160 Drives. An older Oracle library and drives are also used. Tape copies are also stored off-site in a secure facility.

The rate of accumulation of stored data on all media is closely monitored so that increases in data storage capacity can be planned in advance in the line with evolving requirements. In turn these requirements are fed into the periodic procurement exercises for the JASMIN infrastructure. As hardware components reach their end-of-life careful migration of data and services is undertaken to replacement storage/hardware. See the 2018-19 CEDA annual report for examples of the latest round of JASMIN infrastructure procurement and details of data migrations: http://cedadocs.ceda.ac.uk/1466/ .

Technical support for the CEDA archive comprises three different teams operating on the various levels of the system. The hardware systems are maintained by a dedicated team within the Scientific Computing Department, whilst virtualised servers supporting the archive services are managed through a combination of internal CEDA team members and local departmental (RAL Space) systems administrators. Finally, service development and deployment is carried out by a developer team within CEDA. At all levels documentation and monitoring systems are employed to ensure a full hardware, software and service inventory is maintained. All servers are linux based servers. (Internal documentation available on request)

Software for data centre services is a mix of in-house written, maintained open-source, and proprietary code. For example, the data centre metadata catalogue has been developed in-house, a standard open source ftp server (ProFTP) is used, and our helpdesk uses the commercial cloud based Help Scout service. All in-house code is held within GitHub repositories and deployed with Ansible playbooks. Before deployment into production, developments are checked via deployment onto a staging server to ensure that changes are successful and that deployment will take place with minimal service interruption. CEDA maintains detailed databases of CEDA services and infrastructure as well as a detailed Operations Manual within the cloud-based Help Scout helpdesk solution. The Help Scout system provides a commercially supported cloud based solution with in built versioning and auto-save on edits. The services database and related operations documentation are regularly reviewed with service operations meetings maintaining a weekly overview. All services are routinely monitored using a range of dedicated tools. All operation manual documentation backed up daily locally and weekly onto an offline copy for further redundancy. (internal documentation of procedures available on request)

The CEDA Archive resides within the STFC Rutherford Appleton Laboratory (RAL) site network, which is currently connected at 4 x 10Gbit/s to the JANET backbone. (CEDA Archive is a tenant within the JASMIN infrastructure - see JASMIN documentation for external connection details : https://help.jasmin.ac.uk/article/221-jasmin-external-connections)

Discovery metadata conforms to the UK GEMINI schema (https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini) and is thus also compliant with INSPIRE metadata requirements.

Delivery of small datasets is performed via a dedicated ‘arrivals’ service which supports HTTP web-uploading and provides access to FTP and RSYNC tools if required. Larger and ongoing data transfers are usually conducted by FTP or RSYNC. The CEDA Archive provides access to data via web-download, FTP and under the DAP protocol. Some data are also available via OGC compliant web services. (See https://help.ceda.ac.uk/article/142-sending-data-to-ceda for upload information, https://help.ceda.ac.uk/category/15-ceda-services for download service information).

The research nature of the infrastructure does not permit 24/7 support for the system to ensure round the clock support for incoming data feeds and support of end-users. However, a wide range of near-real-time data feeds from a number of data providers are routinely delivered and ingested into the CEDA Archive with little disruption.

The CEDA Archive Disaster Recovery Plan (internal documentation, available on request) covers the steps to take in the event of a catastrophic failure of the CEDA archive that shut down all services. In such circumstances CEDA management, in consultation with the NERC EDS management and CEDA, would make a plan dependent on the impact of the catastrophic event.

The priority would be to recover primary archive data rather than restarting services.

The order of events would be:

1. New servers and storage would be purchased or rented as soon as possible.
2. Data from tape backups would be recovered.
3. Recovery of non-primary archive data from external sources would start.
4. Ingest services restart.
5. Access Services restart.
6. Community services restart.
7. Disaster Recovery

For non-catastrophic instance requiring recovery from back-up CEDA have a standard 'recovery from back-up' procedure to follow (internal documentation available on request).

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The CEDA Archive has a comprehensive set of procedures in place to ensure rapid recovery and return to normal operations in the event of a system failure (see also infrastructure documentation in section R15).

Data are stored on a range of systems, in most cases a Scale Out File System from Quobyte with secondary copies on tape. Some larger dataset collections are stored on tape only, as cost of disk storage is unjustifiable. Tape copies are replicated with one copy being kept securely off site. (See section R15 for further details).

Access to data is authorised using a range of services, depending on the access type. An internal security database is used to set up required access control for http, ftp and OpenDAP access via appropriate policy files, whilst local read-only access to archive content is controlled through linux groups. No user has direct write access to any part of the CEDA archive; CEDA staff are only able to perform write operations using a series of dedicated, version-controlled tools via a cluster of deposit-servers. This limits the potential for human error.

Access to all servers is controlled using SSH with public-key authentication.

The CEDA Technical manager is responsible for applications that run as part of CEDA services. These are routinely subjected to security checks by CEDA's host organisation with appropriate action to resolve any identified issues as swiftly as possible. Additionally, software vulnerabilities (e.g. due to out of date imported packages) are reported automatically through the GitHub service used by CEDA for its code repository.

The security of the underlying JASMIN infrastructure is maintained by a dedicated systems support staff member in CEDA's host organisation's Scientific Computing Department. For specific STFC policy under which CEDA operates see : https://stfc.ukri.org/about-us/how-we-are-governed/policies-standards/monitoring/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

Reviewer Entry

Reviewer 1

Comments:
An outstanding element of your Archive's process is the way you apply Data Management Plans. This is not key for certification, but it is a really admirable approach and you might want to share it as a good practice with the research data management community.

Reviewer 2

Comments:

Additional supporting evidence, i.e. links, and textual clarifications have consolidated the application, warranting CTS-certification.