Eggleton, F and Winfield, K. 2020. Open Data Challenges in Climate Science. *Data Science Journal,* 19: 52, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2020-052

PRACTICE PAPER

Open Data Challenges in Climate Science

Francesca Eggleton and Kate Winfield

Centre for Environmental Data Analysis, RAL Space, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot, UK

Corresponding author: Francesca Eggleton (francesca.eggleton@stfc.ac.uk)

The purpose of this paper is to explore challenges in open climate data experienced by data scientists at the Centre for Environmental Data Analysis (CEDA). This paper explores two of the five V's of Big Data, Volume and Variety. These challenges are explored using the Sentinel satellite data and Climate Modelling Intercomparison Project phase six (CMIP6) data held in the CEDA Archive. To address the Big Data Volume challenge, this paper describes the approach developed by CEDA to manage large volumes of data through the allocation of storage as filesets. These filesets allow CEDA to plan and track dataset storage volumes, a flexible approach which could be adopted by any data centre. CEDA utilise the implementation of the Climate and Forecast (CF) conventions and standard names within archived data wherever possible to overcome the challenge of Variety. Collaboration from the international science community through contributions to the moderation of CF standard names ensures these data then adhere to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles. Utilising data standards such as the CF standard names is recommended because it promotes data exchange and allows data from different sources to be compared. Addressing these Open Data challenges is crucial to ensure valuable climate data are made available to the scientific community to facilitate research that addresses one of society's most pressing issues – climate change.

Keywords: data; climate; science; FAIR principles; volume; variety

Introduction

The United Nations (2020) describes Climate Change as the defining issue of our time. The UK's top 10 warmest years on record have all occurred since 2002, and July 2019 saw the UK's hottest ever recorded temperature (38.7 °C) (McGrath, 2019). The Centre for Environmental Data Analysis[1] (CEDA) is the UK's atmospheric and earth observation data centre, holding over 15PBs (Petabyte) of data, predominantly open access, in over 250 million files. CEDA provides archive services for climate, atmospheric composition, in-situ and surface observations and model data as well as various Earth Observation datasets, including airborne and satellite data and imagery. Data scientists at CEDA provide data management support covering various parts of the data life cycle from data management planning, data storage, preservation, access and reuse, following the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles (Wilkinson *et al*, 2016).

The data held in the CEDA Archive are used in a wide variety of scientific research to address global climate change, from highlighting rising temperatures across the globe, to producing regional climate change projections and impacts using climate models. CEDA supports the exchange of data, as well as the traceability of results through long-term data management and curation. CEDA curates multiple high value datasets and records that are used in scientific research informing international governmental decision making (e.g. Intergovernmental Panel on Climate Change (IPCC) reports). With such a large and diverse archive there are many challenges that CEDA encounter. Some of which are described by the five Vs of Big Data (Volume, Variety, Veracity, Velocity and Value) (Van Genderen *et al*, 2020). This paper will address the challenges in archive storage (Volume) and data standards and formats (Variety) for open source data.

[1] CEDA website: www.ceda.ac.uk/.

Open Data Challenges

Volume: Archive storage limitations

The CEDA Archive is ever expanding with over 10TB (Terabyte) of new data arriving every day. With research publications in climate science increasing in importance and storing data in a long-term archive becoming a necessity for many researchers, climate data volumes are at an all time high and are predicted to continue rising rapidly (see **Figure 1**). The environmental data stored at CEDA are rarely removed, as the identical environment from which it was collected cannot be repeated. The CEDA Archive is currently at its largest volume and faces challenges with storage limitations due to cost, access, space and software. CEDA attempts to solve these issues by using efficient storage methods across the three storage types it uses: disk, tape and object store. Disk storage is used as a directory file system that hosts the live primary archive. This storage type allows easy access but is not the most cost efficient. Tape is used for large volume datasets that are not accessed frequently and acts as a secondary backup, as this storage type is cost effective. Object store manages and stores data as distinct objects instead of a directory based structure, allowing the data to be associated with labels.

The CEDA Archive is held within JASMIN, a storage and cloud computing facility, which provides infrastructure for data analysis. To some extent, every data centre is limited by the capacity of its storage platform which presents limitations on how the data are organised. JASMIN allows flexibility in the type of storage architecture (e.g. file system, object store, tape storage). The way this storage is organised and divided has an influence on how effective it is. CEDA introduced the use of filesets in the archive, a method of grouping together data in an efficient way for storage management. The filesets are defined as an allocation of disk space for a particular dataset/project or portion of a dataset. The fileset is manifested in the disk system as a logical path and a directory on a storage (disk) partition. The allocation given to a fileset is an estimate of the final storage volume. This enables data scientists to plan and track such volumes. This is important as the organisation of these filesets can have an effect on the performance of the storage, due to the size and

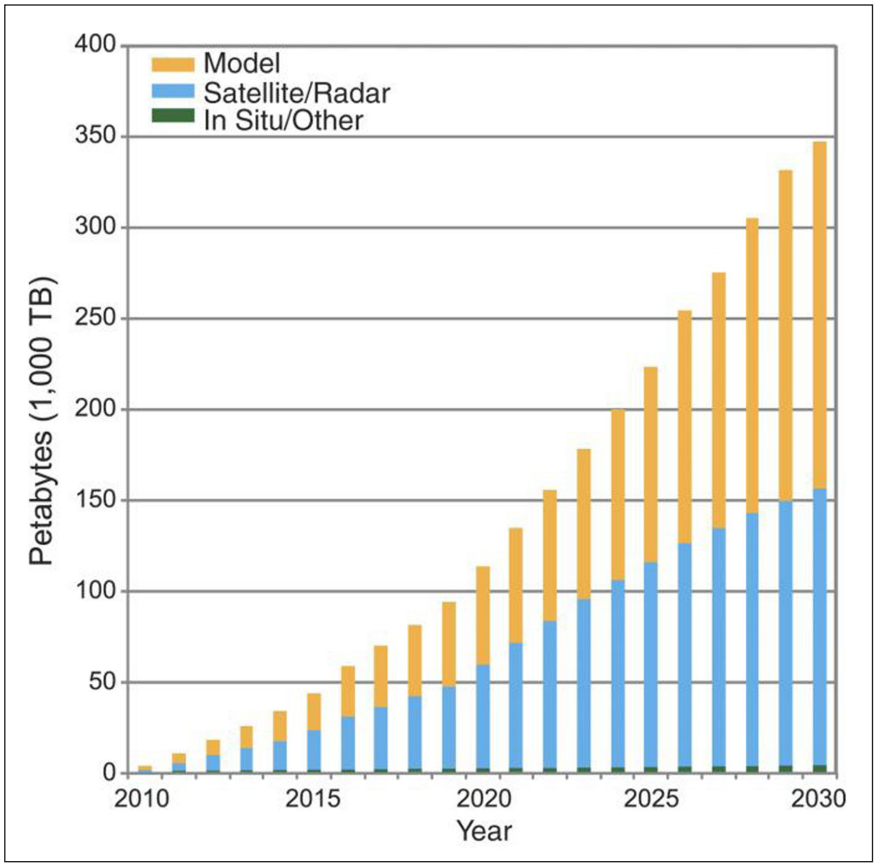

Figure 1: Global climate data volumes and projections into 2030 for climate models, remotely sensed data and in situ instrumental/proxy data. From Overpeck, J T et al. 2011 Climate data challenges in the 21st century, *Science*, 331(6018): 700–702. doi: 10.1126/science.1197869. Reprinted with permission from American Association for the Advancement of Science (AAAS).

number of files within them. For example, lots of small files within a fileset could impact the speed of download for a user, making open data less available. This technique has been advantageous for archive storage at CEDA, and could be adopted by other data centres or organisations working with large volumes of data.

At CEDA, storing data in specific filesets is useful for large open datasets such as Sentinel satellite data, as it can be archived per observation type and per day. Sentinel satellite data are part of the Copernicus Programme, which comprises seven Sentinel missions, some of which have up to four satellites collecting various measurements and images for land, ocean, atmospheric and climate monitoring (Copernicus, 2020). CEDA currently holds over 8PB of global data from most Sentinel missions to support the UK environmental research community. Creating filesets per observation type/per day allows specific filesets to be moved to tape and retrieved in an efficient manner, freeing disk space. However, this wouldn't necessarily be efficient if a user wanted to retrieve data for a specific region, as data may be in several tape locations and on different storage types, highlighting the need to consider the organisation of the filesets carefully.

The increasing data volumes of large open access datasets (such as Sentinel data) has driven the development of CEDA's Near Line Archive System (NLA). This is used to move less frequently accessed data onto tape storage. Space can then be freed up for more environmental data to be made available to the scientific community. Usually, this would mean compromising availability and access. However, the NLA system allows users to request the data they want from the tape storage (up to 10TB). This request means the data are available on disk for every user during the predefined time limit.

The large data volumes produced by the Sentinel missions will not diminish, more satellites are due to be launched soon e.g. Sentinel Six B which is due to launch sometime in 2025, following Sentinel Six A which provides sea surface topography data, including sea level height (EUMETSAT, 2020). It is important for the CEDA Archive to hold as much of these data as possible as it is crucial for the scientific community in addressing climate change. By having the data in the CEDA Archive this allows users to utilise the data on the JASMIN analysis platform, bringing together data, services and expertise - underpinning academic environmental science. Sentinel data have been used to address this global issue in a number of ways, including: monitoring ice sheet changes in Antarctica; providing data on forest fire scar mapping; monitoring land use change; and, recording sea surface temperature data (Amos, 2019; ESA, 2020). Storage limitations is one issue that will continue far into the future, hence CEDA will continue to innovate and apply new technologies (like the NLA) to allow users to easily find and use environmental data.

Variety: Data standards and formats

The FAIR data principles concept provides guidance for scientific data management, promoting the maximum use and reuse of research data through interoperability. Data standards provide common formats that facilitate the creation of tools for data access, curation, ingestion, and visualization (Bermudez, 2017). There are many different ways to format data, but as hardware and software technologies change, certain formats may not be accessible in the future. CEDA encounters many challenges when managing heterogeneous datasets, including ensuring data are interoperable. Therefore, using suitable standard data formats maximises the potential for re-use and long-term usability. For this reason, CEDA encourages the use of standard formats such as NetCDF[2] (Network Common Data Format), NASA Ames (Gaines and Hipskind, 2006) or BADC-CSV (Pepler and Parton, 2010) to ensure long term readability and access. These formats contain metadata inside the data files which allows CEDA to extract information to create catalogue records with detailed information, such as parameters and file format information.

When using NetCDF, CEDA requires that files adhere to the CF (Climate and Forecast) conventions. The CF conventions for NetCDF metadata are designed to promote the processing and sharing of files (CF conventions, 2020). These conventions ensure a definitive description are given of what the data values in the file represent (CF conventions, 2020). This information enables users to decide which variables are comparable when looking at data from different sources, and promotes the creation of software to extract, regrid, and display the data (CF conventions, 2020). For example, the Python language tools cf-python and cf-plot as well as the command line tools Climate Data Operators (CDO) and NetCDF Operators (NCO) (NCAS CMS, 2018; Max-Planck-Institut für Meteorologie, 2019; NCO, 2020). CEDA encourages data users to utilise these tools for NetCDF data handling.

Implementing these standards poses challenges to CEDA as there are many versions of the standard formats, and unique datasets come with subtle differences from various scientific areas. To overcome this

[2] For more information on the NetCDF format: www.unidata.ucar.edu/software/netcdf/.

challenge CEDA has developed checking tools that can be accessed by typing a command on JASMIN or, alternatively, through a web interface to allow providers to check their data before they are submitted to the CEDA Archive. Additionally, using file naming conventions can allow insight into the content and version of a file, without needing to open the file. However, there is no standard file naming convention endorsed by the climate community. CEDA has developed its own file naming convention, enabling quick access to pertinent metadata and avoiding the need to open and read the file in order to assess its contents. This allows CEDA to improve the searchability of the data holdings in the archive and maintain an index of information.

It is important for environmental data to be open and easily accessible so analysis can be done around the world in support of climate science and climate change policy. The Coupled Model Intercomparison Project Phase Six (CMIP6) is an example of a complex collection of climate datasets held in the CEDA Archive which has developed standards to improve consistency and ease of use. The World Climate Research Programme Working Group on Coupled Modelling oversees CMIP6, which will inform the IPCC Sixth Assessment Report[3]. CMIP6 involves more than 100 models from more than 40 institutes across 27 countries (Balaji *et al*, 2018). This global collaboration project combines climate modelling experiments and produces large quantities of data archived around the world by members of the Earth System Grid Federation (ESGF)[4] (Williams *et al*, 2011). The CEDA Archive currently holds 1.8PB (as of June 2020) out of the 13.5PB available through ESGF (ESGF, 2020). This is a prioritised subset of data, chosen for its importance to the community, determined primarily by a prioritised list of data for Working Group One (WG1 assesses the physical science of climate change) of the IPCC. ESGF allows data access by scientists globally, promoting collaboration which is needed for comparison of the model output. These data are held across multiple data centres around the world, CEDA being the representative for the UK (Stockhause and Lautenschlager, 2017). As such it is crucial to have standards across the data formats, file naming conventions and metadata. Without these standards, projects such as CMIP6 would not work.

Due to the highly specific and vast number of variable names, standardisation of these is crucial to the CMIP6 project, therefore they are formally identified by their CF standard name. The CF standard name table is an international community lead vocabulary which defines climate and forecast variables and provides unambiguous identification (CF conventions, 2020). Hundreds of new names were added to the standard name table due to the increased complexity of climate models creating new output variables in the CMIP6 project. Data scientists at CEDA play a significant role in maintaining and updating the CF standard name table. The CF standard names promote data exchange and allow data from different sources to be compared, hence its relevance in CMIP6. In CMIP6, each CMIP Endorsed Model Intercomparison Project (MIP) produces its own variables, therefore systems which list them rely heavily on the CF conventions for standardisation (Juckes et al, 2020).

CMIP is set to continue its journey onto a seventh phase and beyond, therefore continuing to adapt and improve standards across institutes is important, especially for data centres such as CEDA. The raw output from the CMIP6 models results in very heterogeneous data. The standards introduced have meant these data are restricted to the agreed formats and institutes are required to conform to these standards which greatly improves the homogeneity of the data. This is important to facilitate the intercomparison of these data. Climate models continue to improve and increase in complexity, as does the output and the amount of variables. CMIP6 is an example of where standards have provided a solution to a complex collection of datasets that needs to be openly available and accessible by researchers across the world. This could be applied to other climate datasets that experience challenges in the variety of data to ensure findability, reusability and interoperability. CEDA continues to introduce and encourage the use of standards across data formats, metadata and filenames in order to make open data comply with FAIR principles.

Conclusions

Environmental data will only continue to grow. Projected demand is for 300PB of data needing to be stored within the next few years (Quobyte, 2019). Data Volume and Variety are just two of the five V's (Volume, Variety, Veracity, Velocity and Value) of Big Data that CEDA faces (Van Genderen *et al*, 2020). In this paper, these two challenges and CEDA's approach to them, have been explored. It has been recognised that obstacles such as these require new developments in computing, data management, information extraction, and knowledge discovery (Van Genderen *et al*, 2020). At CEDA, increasing data volumes of large open access

[3] For more information on the IPCC AR6: https://www.ipcc.ch/assessment-report/ar6/.
[4] For more information see ESGF homepage: https://esgf.llnl.gov/.

datasets, such as Sentinel, have driven the development of CEDA's filesets and the NLA. This enables data scientists to plan and track dataset storage volumes and allow movement of data between different storage types. Archive space can then be made available for more environmental data of importance to the scientific community. Such techniques could be adopted by other data centres to ensure efficient use of their storage platforms.

Within the CEDA Archive there are many heterogeneous datasets. Data standards provide common formats that facilitate the creation of tools for data access, curation, ingestion, and visualization (Bermudez, 2017). Using suitable standard data formats and metadata conventions maximises the potential for re-use and data comparison. This helps to ensure the data adheres to the FAIR data principles – Interoperability and Reusability. The CF standard names are used in the CMIP6 datasets to improve the homogeneity of the data by utilising a community lead standard. Working with the community can be beneficial to data centres by ensuring standards are communally agreed upon before data publication. However, ensuring these standards are implemented across the scientific community can be difficult to enforce. For CEDA, addressing these open data challenges is crucial to ensure valuable climate data are made available to the scientific community to facilitate research that addresses society's most pressing issue - climate change.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Both authors contributed equally.

References

Amos, J. 2019. Climate change: Satellite fix safeguards Antarctic data. Available at: https://www.bbc.co.uk/news/science-environment-47461199 [Last accessed 17 November 2020].

Balaji, V, et al. 2018. Requirements for a global data infrastructure in support of cmip6. *Geoscientific Model Development, Copernicus GmbH*, 11(9): 3659–3680. DOI: https://doi.org/10.5194/gmd-11-3659-2018

Bermudez, L. 2017. New frontiers on open standards for geo-spatial science. Geo-Spatial Information Science. *Taylor & Francis*, 20(2): 126–133. DOI: https://doi.org/10.1080/10095020.2017.1325613

CF Conventions. 2020. CF Conventions and Metadata. Available at: http://cfconventions.org/ [Last accessed 22 May 2020].

Copernicus. 2020. Copernicus Homepage. Available at: https://www.copernicus.eu/en [Last accessed: 17 June 2020].

ESA. 2020. Climate Change Overview. Available at: https://sentinel.esa.int/web/sentinel/thematic-areas/climate-change [Last accessed: 19 May 2020].

ESGF. 2020. Published data over the federation. Available at: http://esgf-ui.cmcc.it/esgf-dashboard-ui/federated-view.html [Last accessed: 8 June 2020].

EUMETSAT. 2020. Sentinel-6: Monitoring the global ocean. Available at: https://www.eumetsat.int/website/home/Copernicus/copernicus-sentinel-6/index.html [Last accessed: 19 May 2020].

Gaines, SE and **Hipskind, SR.** 2006. Ames Format Specification (v2.0). NASA. Available at: https://espoarchive.nasa.gov/content/Ames_Format_Specification_v20 [Last accessed: 23 June 2020].

Van Genderen, J, et al. 2020. Digital Earth Challenges and Future Trends. In: Guo, H, Goodchild, MF and Annoni, A (eds.), *Manual of Digital Earth*. Singapore: Springer Singapore, pp. 811–827. DOI: https://doi.org/10.1007/978-981-32-9915-3_26

Juckes, M, et al. 2020 The CMIP6 Data Request (version 01.00.31). *Geoscientific Model Development Discussions*, 1–35. DOI: https://doi.org/10.5194/gmd-2019-219

Max-Planck-Institut für Meteorologie. 2019. CDO. Available at: https://code.mpimet.mpg.de/projects/cdo [Last accessed: 19 June 2020].

McGrath, M. 2019. Climate change: UK's 10 warmest years all occurred since 2002. Available at: https://www.bbc.co.uk/news/science-environment-49167797 [Last accessed: 23 June 2020].

NCAS, CMS. 2018. Tools and Utilities. Available at: http://cms.ncas.ac.uk/wiki/ToolsAndUtilities [Last accessed: 28 May 2020].

NCO. 2020. NetCDF Operator (NCO) site. Available at: http://nco.sourceforge.net/ [Last accessed: 19 June 2020].

Overpeck, JT, et al. 2011. Climate data challenges in the 21st century, *Science*, 331(6018): 700–702. DOI: https://doi.org/10.1126/science.1197869

Pepler, S and **Parton, GA.** 2010. The BADC Text File Guide for users and producers. http://cedadocs.ceda.ac.uk/772/.

Quobyte. 2019. Digital Infrastructure for a World-Leading Environmental Data Facility. Available at: https://www.quobyte.com/case-studies/stfc [Last accessed: 2 June 2020].

Stockhause, M and **Lautenschlager, M.** 2017. CMIP6 data citation of evolving data. *Data Science Journal*, 16: 1–13. DOI: https://doi.org/10.5334/dsj-2017-030

United Nations. 2020. Climate Change. Available at: https://www.un.org/en/sections/issues-depth/climate-change/index.html. [Last accessed: 23 June 2020].

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1–9). DOI: https://doi.org/10.1038/sdata.2016.18

Williams, DN, et al. 2011. The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5. *Proceedings of the Asia-Pacific Advanced Network*, 32(0): 121. DOI: https://doi.org/10.7125/APAN.32.15

How to cite this article: Eggleton, F and Winfield, K. 2020. Open Data Challenges in Climate Science. *Data Science Journal,* 19: 52, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2020-052

Submitted: 20 October 2020 **Accepted:** 25 November 2020 **Published:** 16 December 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS