IJDC | *General Article*

Twenty Years of Data Management in the British Atmospheric Data Centre

Sam Pepler
National Centre for Atmospheric Science
and
Centre for Environmental Data,
STFC Rutherford Appleton Laboratory

Sarah Callaghan
National Centre for Atmospheric Science
and
Centre for Environmental Data,
STFC Rutherford Appleton Laboratory

Abstract

The British Atmospheric Data Centre (BADC) has existed in its present form for 20 years, having been formally created in 1994. It evolved from the GDF (Geophysical Data Facility), a SERC (Science and Engineering Research Council) facility, as a result of research council reform where NERC (Natural Environment Research Council) extended its remit to cover atmospheric data below 10km altitude. With that change the BADC took on data from many other atmospheric sources and started interacting with NERC research programmes.

The BADC has now hit early adulthood. Prompted by this milestone, we examine in this paper whether the data centre is creaking at the seams or is looking forward to the prime of its life, gliding effortlessly into the future. Which parts of it are bullet proof and which parts are held together with double-sided sticky tape? Can we expect to see it in its present form in another twenty years' time?

To answer these questions, we examine the interfaces, technology, processes and organisation used in the provision of data centre services by looking at three snapshots in time, 1994, 2004 and 2014, using metrics and reports from the time to compare and contrasts the services using BADC. The repository landscape has changed massively over this period and has moved the focus for technology and development as the broader community followed emerging trends, standards and ways of working. The incorporation of these new ideas has been both a blessing and a curse, providing the data centre staff with plenty of challenges and opportunities.

We also discuss key data centre functions including: data discovery, data access, ingestion, data management planning, preservation plans, agreements/licences and data policy, storage and server technology, organisation and funding, and user management. We conclude that the data centre will probably still exist in some form in 2024 and that it will most likely still be reliant on a file system. However, the technology delivering this service will change and the host organisation and funding routes may vary.

Received 16 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to Sam Pepler, STFC Rutherford Appleton Laboratory. Email: sam.pepler@stfc.ac.uk

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v10i2.379
DOI: 10.2218/ijdc.v10i2.379

Introduction

The BADC was established in 1995 when it superseded a previous facility: the Geophysical Data Facility (GDF). The GDF was previously funded by the then Science and Engineering Research Council and primarily supported what was then called the "upper atmosphere" remote sensing community studying the atmosphere between 10-400 km. However, a survey of the Natural Environment Research Council (NERC) community resulted in an increased remit for the newly renamed BADC to support the entire NERC atmospheric science community. Since then the user community has evolved to encompass other science areas and uses other than pure research, but the principle role of the BADC has not changed: to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects.

The BADC is funded via a Service Level Agreement (SLA) between NERC and the Science and Technology Research Council (STFC). Within STFC, the Centre for Environmental Data Archival (CEDA) delivers the BADC data centre functions, along with other data centres and data related projects. The other data centres delivered by CEDA include the NERC Earth Observation Data Centre (NEODC) and the UK Solar System Data Centre (UKSSDC). Collaboration with these data centres has allowed us to take advantage of economies of scale and share expertise and experience.

Methodology

Annual reports and other internal documents were reviewed for references to key data centre functions, noting the evolution and trends. We focus on three financial years in particular: 1995, 2004 and 2014. Statements and figures from these years are generally from the annual reports Allan and Gray (1995), Pepler, (2005) and Callaghan (2014), respectively. We examine some key data centre functions including: data discovery, data access, ingestion, data management planning, preservation plans, agreements/licences and data policy, storage and server technology, organisation and funding, and user management.

Evolution of Key Data Centre Functions

Data Discovery

All data repositories need some form of catalogue to present users with their wares. In 1995 the catalogue was simply a collection of static web pages describing each dataset. The new web technology made this a reportable advance in functionality:

> 'We have developed a WWW interface to the BADC catalogue allowing basic searches from a Web client' (Allan and Gray, 1995).

In the mid 2000s the NERC Data Grid project aimed to harmonise a number of data services across NERC. One of its results was a standards based metadata model which, amongst other objectives, supported discovery (the Metadata Objects for Linking Environmental Systems, MOLES, discussed in Parton et al., 2015). Several versions of the MOLES schema were implemented up to the present incarnation of the catalogue.

The general trend is towards a richer, more complex catalogue supporting multiple functions.

In contrast to the increasing complexity of the catalogue, the headline list of datasets tends to grow more slowly. The number of datasets in the catalogue progressed from 13 explicitly mentioned in the 1995 annual report to 253 in the 2013 annual report.

In 2004 we were feeding records from our catalogue into the NERC Metadata Gateway and the Global Change Master Directory (GCMD). The NERC Metadata Gateway attempted to aggregate the data records across all seven NERC data centres, first using the Z39.50 protocol, then later OAI-PMH. In the current incarnation, the NERC data catalogue service uses a Web Catalogue Service (WCS) interface to harvest records from our MOLES catalogue. The idea of a NERC portal has remained constant while the implementation and technology has changed with trends in standards.

A recurring theme across the 20 years of the BADC is that changes in technology are painful, requiring time and effort to implement.

Data Access

Users' primary access pattern has been to select a subset of files from a dataset and then download them. In 1995 the GDF MicroVAX menu interface was being phased out in favour of a web based service to do the selection, but the data was transferred via FTP. A decade on in 2004 and data can be downloaded via a web interface as well as FTP. The FTP route has remained popular, as it is easy to script for regular transfers. Also, by 2004 other specialist services for some datasets are available for subsetting.

The current data access services include all the previous FTP and web interfaces, with new and upgraded dataset specific interfaces. An example of this is the Earth System Grid Federation tools that are used to distribute data from the Climate Model Inter-comparison Project (CMIP5).

Table 1 shows the increase in download volume over time. Note that the users now far exceed the total number of who could possibly be labelled as atmospheric science researchers in the UK, our primary designated community. While the absolute volume of download is increasing, the proportion of the archive downloaded is falling, probably as a result of more advanced, selective services. Parton (2013), estimates that the anonymous access adds an additional ~45% more users to these figures.

Table 1. Download metrics from annual reports. These figures include the vanilla FTP and web downloads only.

Year	Files downloaded	Downloaded volume	Identifiable distinct users downloading
1995	Not reported	Not reported	99
2004	5 Million	8 TB	1,496
2014	9 Million	93 TB	3,905

Another innovation is the JASMIN Linux login service, which offers selected users direct file system access to the data. This allows users to take the processing to the data, rather than downloading the data and running the processing at their local institutes. We anticipate this mode of working will dominate in the future as people move towards a more cloud-based paradigm.

Ingestion

The total volume of the archive gives some indication on the ingest rates over time. The high volume datasets tend to be climate model output, numerical weather predictions and satellite data. However, a lot of the effort for ingestion is reported from the smaller scale, heterogeneous datasets. These data require more support too, as they involve more people and are less consistent when following file formatting guidelines. Volume and number of files are not a useful metric for the ingest effort needed.

Table 2. Archive volumes and sizes from annual reports.

Year	Volume	Number of Files
1995	60 GB	Not reported
2004	17 TB	Not reported
2014	2 PB	89 Million

Agreements, Licences and Data Policy

The sharing culture in science has changed radically over the period reported on in this paper. In 1995 the BADC was primarily a facilitator for large-scale data producers to transfer data to the research community. It was negotiating an agreement with the Met Office to redistribute its data and stockpiling data from NASA satellite missions. Data from university groups was still shared via informal agreements and did not come to the data centre.

By 2004 NERC had a data policy handbook that encouraged data exploitation. The BADC was ingesting data from the larger NERC programmes in order to preserve it, but the primary role was to enable data exchange within those programmes. To make this explicit a data sharing protocol document was signed by programme participants before access or ingest were permitted.

NERC data policy is now more explicitly open[1]. Data of long-term value (created by NERC funded research) should be curated and must be useable for any purpose after a two year embargo. The Open Government Licence[2] is the default licence for NERC data. This does not mean that all data within the BADC is open, as we continue to restrict access for non-NERC, third party data or NERC data within the embargo period. However, the proportion of data with restricted access is falling, and is anticipated to keep doing so in the future.

Data Management and Preservation Planning

1995: Data is collected to facilitate research rather than curate the data. Neither the data producers nor the data centre created any kind of data management plan. Choices about which data to collect are made by data centre staff with advice from the BADC steering committee. As the user base is small we have close contact with them and get direct feedback on dataset, metadata and formats.

[1] NERC data policy: http://www.nerc.ac.uk/research/sites/data/policy/

[2] Open Government Licence: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

2004: The larger NERC research programmes are now supported with data management plans, which encourage sharing within the programmes by depositing data in the BADC. The plans introduce the idea of data release for academic use at the end of the programme. There is an implicit assumption that the data centre will keep the data indefinitely, but the primary aim of the data management plan is to promote scientific reuse of the data. Contact with supplier community is stimulated by attendance at science meetings and basing a member of staff permanently at the Met office. The broader user base requires a shift in the way we gather user requirements, supplementing direct contact with the first user survey.

> ‘A questionnaire was sent out to gather feedback for the coming science and management audit of NCAS. The questionnaire including ten questions and was sent to 3,618 BADC mailing list members on 15th April 2004. The response rate was 8%. The results are encouraging with an impressive 93% of users assess the BADC as providing very good/excellent services overall’ (Pepler, 2005).

2013: NERC policy now stipulates that all research grant proposals must contain an outline data management plan. The data centres are charged with overseeing the creation of full data management plans for funded projects. The data management plans themselves are relatively lightweight documents to encourage researcher engagement. They have sections on roles and responsibilities, data generation activities, in-project data management approaches, metadata and documentation, and data quality[3].

Storage and Server Technology

The system architecture in 1995 and following decade used a single large server to provide all the data centre services. The storage was served over NFS from a second server.

> ‘During the year, we have purchased two Unix services. The larger is a dual processor Digital AXP 2100 which will provide the main Unix service to the users of the BADC for the next few years… Most of the data held by the Geophysical Data Facility (the predecessor of the BADC) was stored on an optical jukebox system. Although ‘high tech’ in its day, it is now obsolete and we have almost completed the migration of data to a Digital StorageWorks system that has a capacity of 120 GB’ (Allan and Gray, 1995).

By 2004 the storage and the servers were being decentralised using Network Attached Storage (NAS) and many Linux servers so that services did not go down together.

> ‘Decommissioning of the ES40 server has started. It is expected to take around a year before final switch off. Most services are already migrated to the new Linux servers. Two new 9TB NAS boxes have been purchased’ (Pepler, 2005).

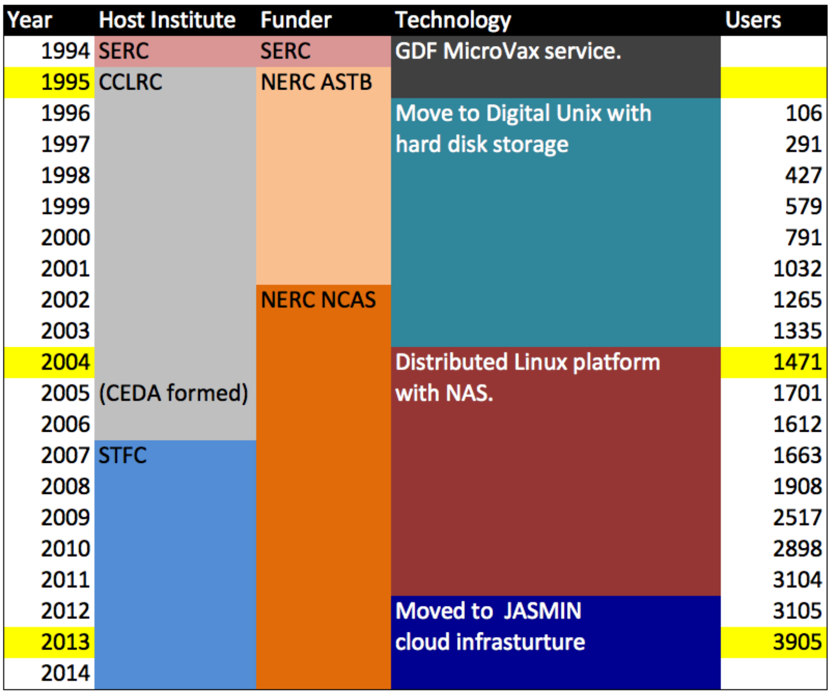
[3] NERC Data Management Plan Template: http://www.nerc.ac.uk/research/sites/data/dmp/

In 2011, a large capital grant allowed the procurement of the JASMIN infrastructure. By 2014 the data is migrated to JASMIN storage and many of the services are moved to virtual machines (Lawrence et al., 2013).

It is clear that migration to new technologies is a continuing task that requires thought and effort. One constant has been the use of a simple POSIX file system as our data store. This is flexible, easy to migrate and copes with large volumes, many files and is understood by users.

Staff, Organisation and Funding

The CCLRC (Council for the Central Laboratory of the Research Councils) was created in 1995. At this time the responsibility for funding the data centre was moved to NERC and the data centre was renamed the BADC. A service level agreement was agreed between NERC and the CCLRC describing the service. The funding for the BADC was overseen by the NERC Atmospheric Science and Technology Board. A major restructuring of NERC in 2002 created a new virtual research centre: the NERC Centre for Atmospheric Science (NCAS). Oversight of BADC funding now moved into this organisation. In 2007, the CCLRC became the STFC as it merged with the Particle Physics and Astronomy Research Council. None of these organisational changes resulted in significant changes at the data centre level, as the requirement and the service level mechanism were constant throughout. The timeline in Figure 1 shows some of the major organisational changes in the history of the BADC.



Year	Host Institute	Funder	Technology	Users
1994	SERC	SERC	GDF MicroVax service.	
1995	CCLRC	NERC ASTB		
1996			Move to Digital Unix with hard disk storage	106
1997				291
1998				427
1999				579
2000				791
2001				1032
2002		NERC NCAS		1265
2003				1335
2004			Distributed Linux platform with NAS.	1471
2005	(CEDA formed)			1701
2006				1612
2007	STFC			1663
2008				1908
2009				2517
2010				2898
2011				3104
2012			Moved to JASMIN cloud infrasturture	3105
2013				3905
2014				

Figure 1. Development timeline for the BADC. The three years examined are highlighted in yellow. The number of users given is the number of identifiable individuals downloading data, and does not include anonymous access. Acronyms are expanded in the paper text.

In 1995 there were six staff mentioned in a section on training in the annual report. They were not all working full time on the BADC, but around three full time equivalents was the level at which the BADC was funded. Staff levels have since increased to around 9 FTE working on BADC data centre functions. However, the CEDA group, which formed around the BADC in 2005, is staffed at around 23 FTE to support the many data related projects and other data centres that have been added to the portfolio of work. While the BADC remains NERC funded, the other projects supported by the group are funded via EU grants, Government agencies (e.g. the Met Office) and others. This diversification of funding seems to be a growing trend.

User Management Systems

Dealing with many users requires the appropriate tools. In 1995 queries were handled via personal email accounts and user registration involved setting up a system account on the primary server. Obviously, this method of working does not scale to thousands of users. By 2004 we had invested in a commercial query handling system and automated the registration system so that no manual intervention was needed and the need for insure system accounts was removed. These systems have been upgraded, but not changed dramatically since.

Users need to register with the BADC to access many of the datasets. Even datasets that have no access restrictions often need user numbers to report back to funding bodies. Registration enabled us to profile our user base. Table 3 shows how the makeup on the BADC user base has changed. There is a gradual drift towards a more international, and more diverse audience. The long-term trend in user numbers is still increasing, as illustrated by Figure 2. The tools to control register users, apply for data set access and control access have evolved considerably over the 20 years.

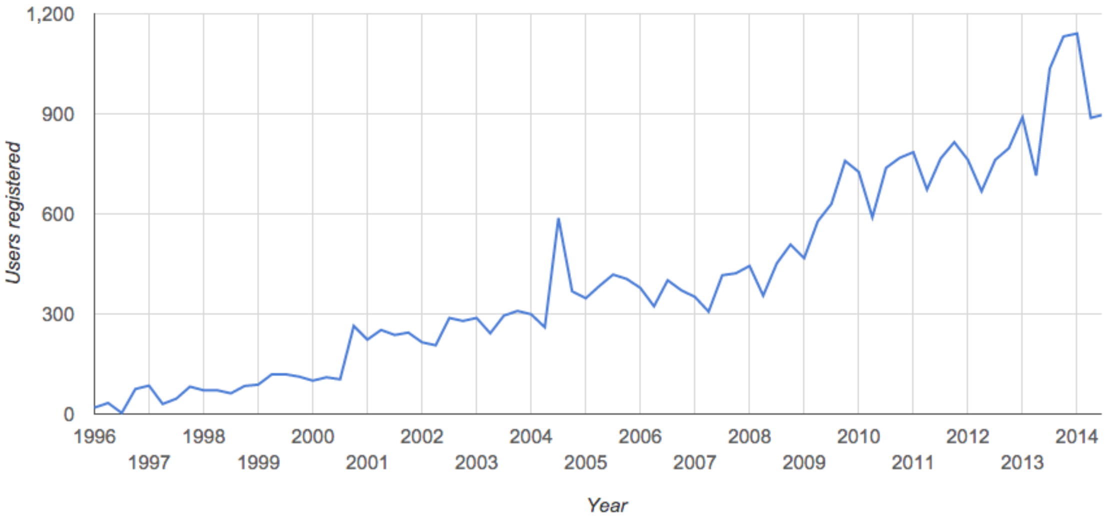


Figure 2. Quarterly user registrations with CEDA. Other services (e.g. the NERC Earth Observation Data Centre) use the same registration system, but the bulk of the user base is from the BADC.

Table 3. User profiles.

Year	Annual queries	Users from universities	Users from UK	Total cumulative user registrations
1995	Not reported	Not reported	Not reported	293
2004	1,400	75%	72%	6,429
2014	9,000,000	70%	61%	30,000

Conclusions

The BADC has been in operation now for over 20 years, and as such we have learned that some changes are inevitable and should be planned for. Significant lessons learned include:

- **The file system is a great base for an archive:** We have found the use of a file system scales to the volumes needed and allows uses for download and use the data in an intuitive form.
- **Data outlives researchers (or their roles) and, in most cases, organisations:** Returning to legacy data many years after its ingestion to add new metadata or clarify terms and conditions is often more awkward and time consuming than dealing with the situation at the beginning. It is better to have a data management plan in the beginning, and adhere to it.
- **The tools of the trade change with time:** Changes in numbers of users, volumes of data, available metadata, external standards, and both producer and user requirements can all prompt changes in the query handling software, storage systems, catalogues and services. These changes are generally disruptive, and mostly necessary. Start planning to retire/replace/extend systems as soon as possible and try and predict changes in user requirements.
- **Openness make running the data centre easier,** but data managers must be aware of licensing and restriction issues, including confidentiality. Not all data should be made open (though that should be the default unless there is a good reason otherwise).

What will the BADC look like in ten years time? Clearly a ten-year time horizon is difficult. There is no guarantee the BADC will even exist, it has already lost considerable independence by incorporation into the Centre for Environmental Data Archival, CEDA. It seems likely that the data will remain important, and that someone will be responsible their management and serving users. Who pays, and where the people and bit and bytes will be, are more interesting questions. Whether or not NERC remains as an independent entity funding environmental data management, it is likely that any significant UK environmental science programme will need, and contribute funding for, a centre facilitating the management and use of data. Where the people will sit, and where the bits and bytes will be kept, are both less clear – although we are aware of no plans to move either people or data from the STFC (and moving the data would be an expensive and complicated logistical activity).

Wherever the activity is carried out, we would assume the following services and systems:

- A catalogue with more detailed inter-related records.
- Access to the data in a direct manner. The file system is still the way the users want the data.
- Proportionally less downloads. Processing goes to the data.
- More data in the archive – o(200)PB.
- More open data.
- Better sub-setting tools.
- A broader user base, becoming more multidisciplinary, international and non-academic.

In delivering these services and systems, we assume the use of whatever technology is available, suitable and affordable, with an increasing interest in and use of the Cloud, will lead to further disruptive – but advantageous – change. The BADC, whatever it is called, and whoever pays for it, will continue to serve its users and provide a valued service, for the next ten years and beyond!

Acknowledgements

The BADC is funded by a range of organisations, with the primary funding from NERC coming via the National Centre for Atmospheric Sciences. Clearly none of the activities carried out here would have been possible without the work of many staff over the last two decades.

References

Allan, P. M., & Gray, L. J. (1995). *Annual report of the British Atmospheric Data Centre*. Retrieved from Centre for Environmental Data Archival Document Repository: http://cedadocs.badc.rl.ac.uk/1122/

Callaghan, S. (2014). *STFC Centre for Environmental Data Archival (CEDA) annual report 2014.* Harwell, UK: Centre for Environmental Data Archival.

Lawrence, B. N., Bennett,V. L., Churchill, J., Juckes, M., Kershaw, P., Pascoe, S., … Stephens, A. (2013). Storing and manipulating environmental big data with JASMIN. In *Proceedings of the 2013 IEEE International Conference on Big Data* (pp. 68-75). doi:10.1109/BigData.2013.6691556

Pepler, S. (2005). *NCAS/BADC annual report 2004-05*. Retrieved from Centre for Environmental Data Archival Document Repository: http://cedadocs.badc.rl.ac.uk/1123/

Parton, G.A. (2013). *BADC user statistics report 2013.* Retrieved from Centre for Environmental Data Archival Document Repository: http://cedadocs.badc.rl.ac.uk/947/

Parton, G. A., Donegan, S., Pascoe, S., Stephens, A., Ventouras, S., & Lawrence, B. N. (2015). MOLES3: Implementing an ISO standards driven data catalogue. *International Journal of Digital Curation, 10*(1), 249–259. doi:10.2218/ijdc.v10i1.365