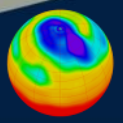
Science & Technology Facilities Council
Rutherford Appleton Laboratory

Twenty years of data management in the British Atmospheric Data Centre

Sam Pepler, Sarah Callaghan
National Centre for Atmospheric Science (NCAS),
Centre for Environmental Data Archival (CEDA)
STFC Rutherford Appleton Laboratory



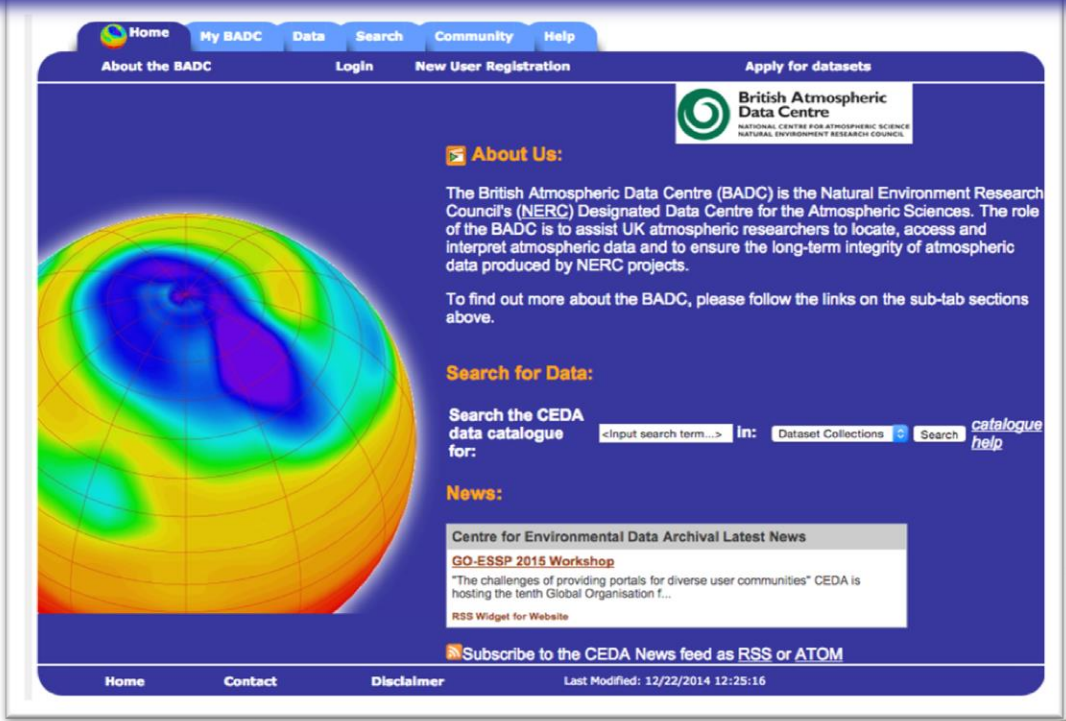
Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

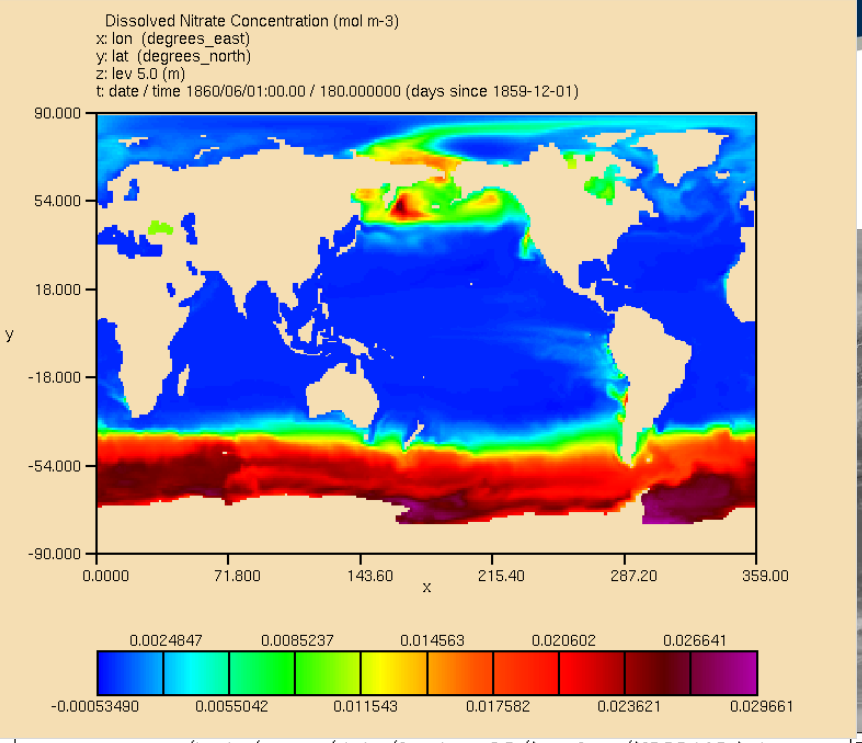
Outline

- What is the BADC?
- Methodology
- Data Centre functions over time
  - Data Discovery
  - Data access
  - Ingestion
  - Agreements, Licences and data policy
  - Data management and preservation planning
  - Storage and server technology
  - Staff, Organisation and funding
  - User management systems
- Conclusions

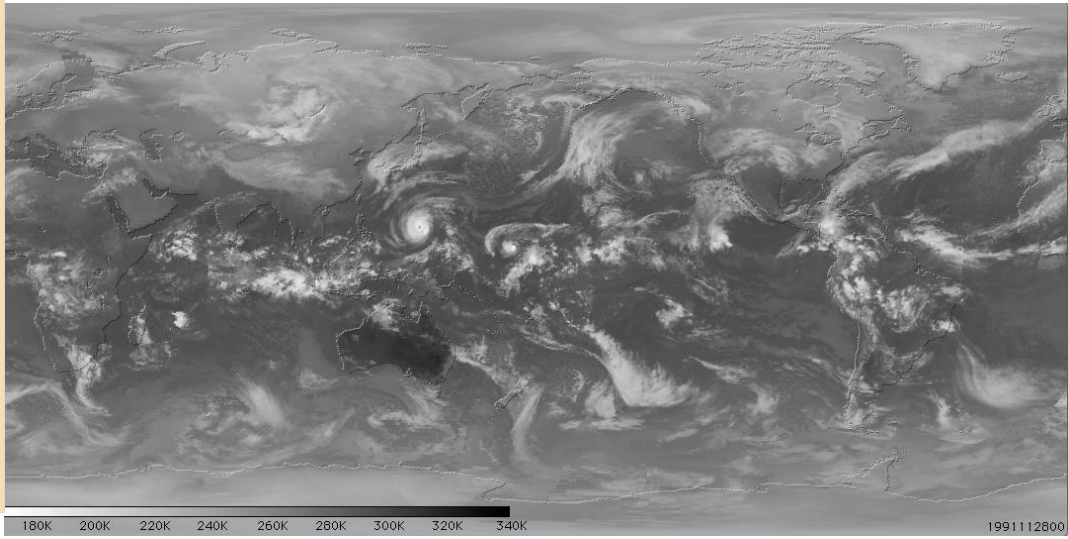
What is the BADC?

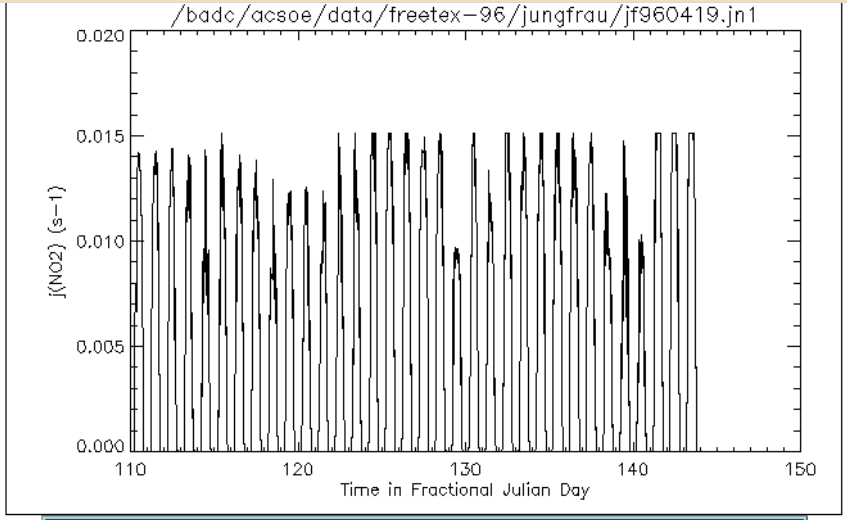
The British Atmospheric Data Centre (BADC) is the Natural Environment Research Council's (NERC) Designated Data Centre for the Atmospheric Sciences. The role of the BADC is to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects.



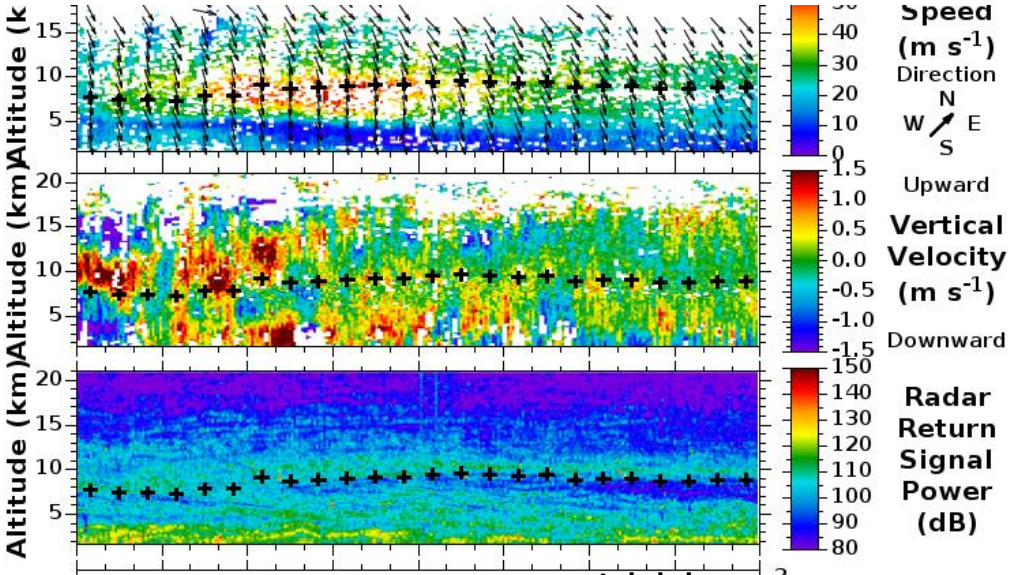


BADC data



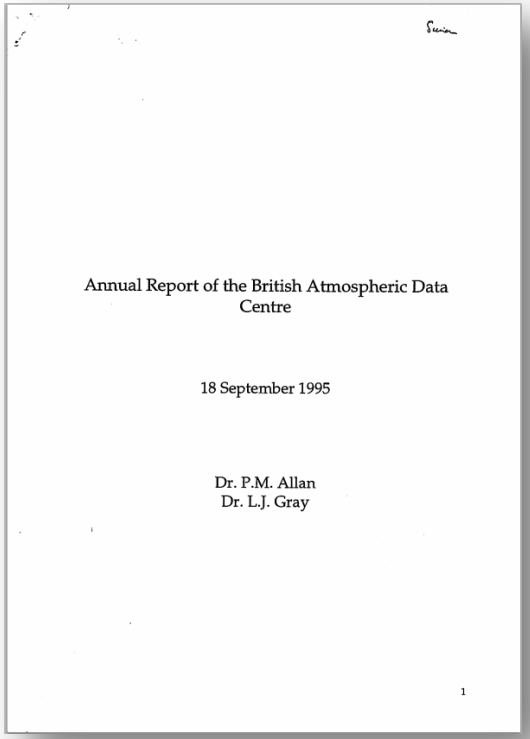


Plot:	j(NO2) (s-1)	against	Time in Fractional Julian Day	Go			
Omit points:	Where: MONTH	=					
Scale	Ymin:	Ymax:	Xmin:	Xmax:	Reset		
Plot Symbol	Plus sign ○	Asterisk ○	Period ●	Diamond ○	X ○	**Connect points?** Yes ● No ○	



Methodology

- Annual reports and other internal documents were reviewed looking for references to key data centre functions.
- We focus on three financial years in particular 1995, 2004 and 2014.

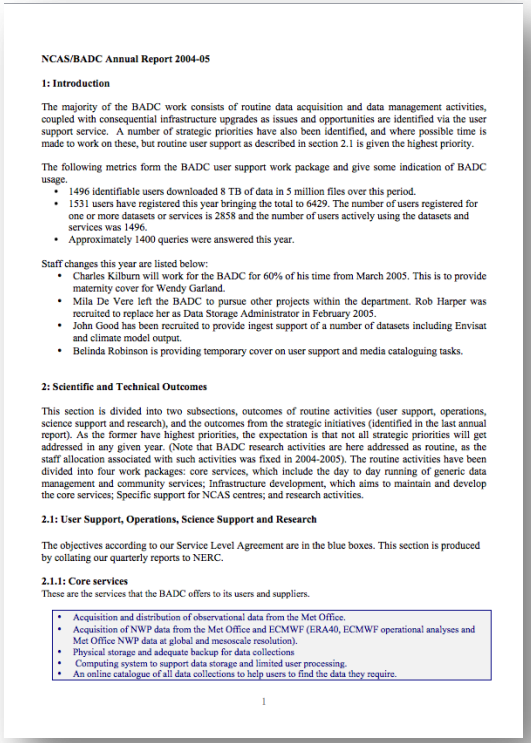


Annual Report of the British Atmospheric Data Centre

18 September 1995

Dr. P.M. Allan
Dr. L.J. Gray

1



NCAS/BADC Annual Report 2004-05

1: Introduction

The majority of the BADC work consists of routine data acquisition and data management activities, coupled with consequential infrastructure upgrades as issues and opportunities are identified via the user support service. A number of strategic priorities have also been identified, and where possible time is made to work on these, but routine user support as described in section 2.1 is given the highest priority.

The following metrics form the BADC user support work package and give some indication of BADC usage.

- 1496 identifiable users downloaded 8 TB of data in 5 million files over this period.
- 1531 users have registered this year bringing the total to 6429. The number of users registered for one or more datasets or services is 2858 and the number of users actively using the datasets and services was 1496.
- Approximately 1400 queries were answered this year.

Staff changes this year are listed below:

- Charles Kilburn will work for the BADC for 60% of his time from March 2005. This is to provide maternity cover for Wendy Garland.
- Mila De Vere left the BADC to pursue other projects within the department. Rob Harper was recruited to replace her as Data Storage Administrator in February 2005.
- John Good has been recruited to provide ingest support of a number of datasets including Envisat and climate model output.
- Belinda Robinson is providing temporary cover on user support and media cataloguing tasks.

2: Scientific and Technical Outcomes

This section is divided into two subsections, outcomes of routine activities (user support, operations, science support and research), and the outcomes from the strategic initiatives (identified in the last annual report). As the former have highest priorities, the expectation is that not all strategic priorities will get addressed in any given year. (Note that BADC research activities are here addressed as routine, as the staff allocation associated with such activities was fixed in 2004-2005). The routine activities have been divided into four work packages: core services, which include the day to day running of generic data management and community services; Infrastructure development, which aims to maintain and develop the core services; Specific support for NCAS centres; and research activities.

2.1: User Support, Operations, Science Support and Research

The objectives according to our Service Level Agreement are in the blue boxes. This section is produced by collating our quarterly reports to NERC.

2.1.1: Core services

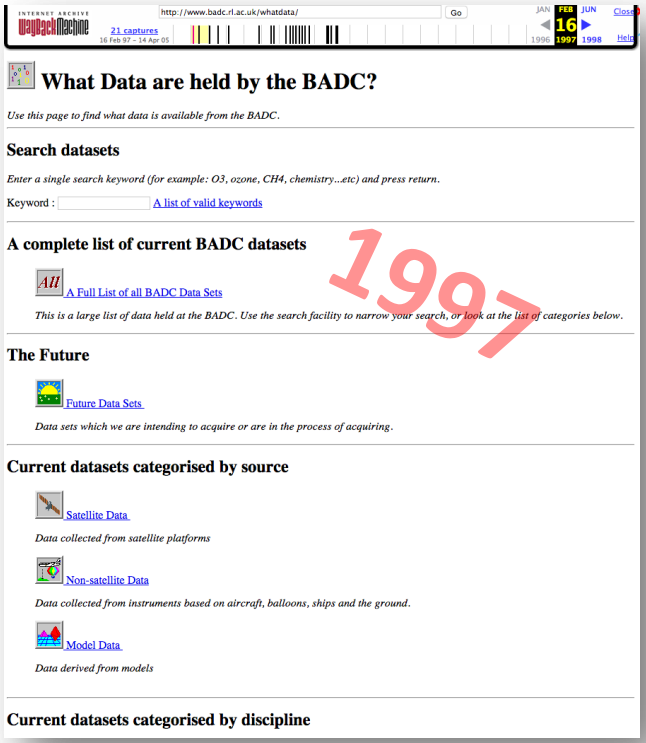
These are the services that the BADC offers to its users and suppliers.

- Acquisition and distribution of observational data from the Met Office.
- Acquisition of NWP data from the Met Office and ECMWF (ERA40, ECMWF operational analyses and Met Office NWP data at global and mesoscale resolution).
- Physical storage and adequate backup for data collections
- Computing system to support data storage and limited user processing.
- An online catalogue of all data collections to help users to find the data they require.

1

Data Discovery

- Trends:
  - More complex catalogues
  - More standards
  - More aggregating portals
  - More records (but not as many as you might think)



Data Access

Year	1995	2004	2014
File Downloaded	Not reported	5 million	9 million
Downloaded Volume	Not reported	8 TB	93 TB
Identifiable distinct users downloading	99	1496	3905
Access methods	FTP	FTP, HTTP, Some web based Sub-setting tools	FTP, HTTP, more sub-setting tools, Direct file system access

Trend: users to take the processing to the data, rather than downloading the data and running the processing at their local institutes.

Ingestion

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

- High volume datasets tend to be climate model output, numerical weather predictions and satellite data - 1PB is the CMIP5 dataset. Homogenous big stuff.
- High effort for ingestion is reported from the smaller scale, heterogeneous datasets. These data require more support to 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.

Agreements, Licences, data policy

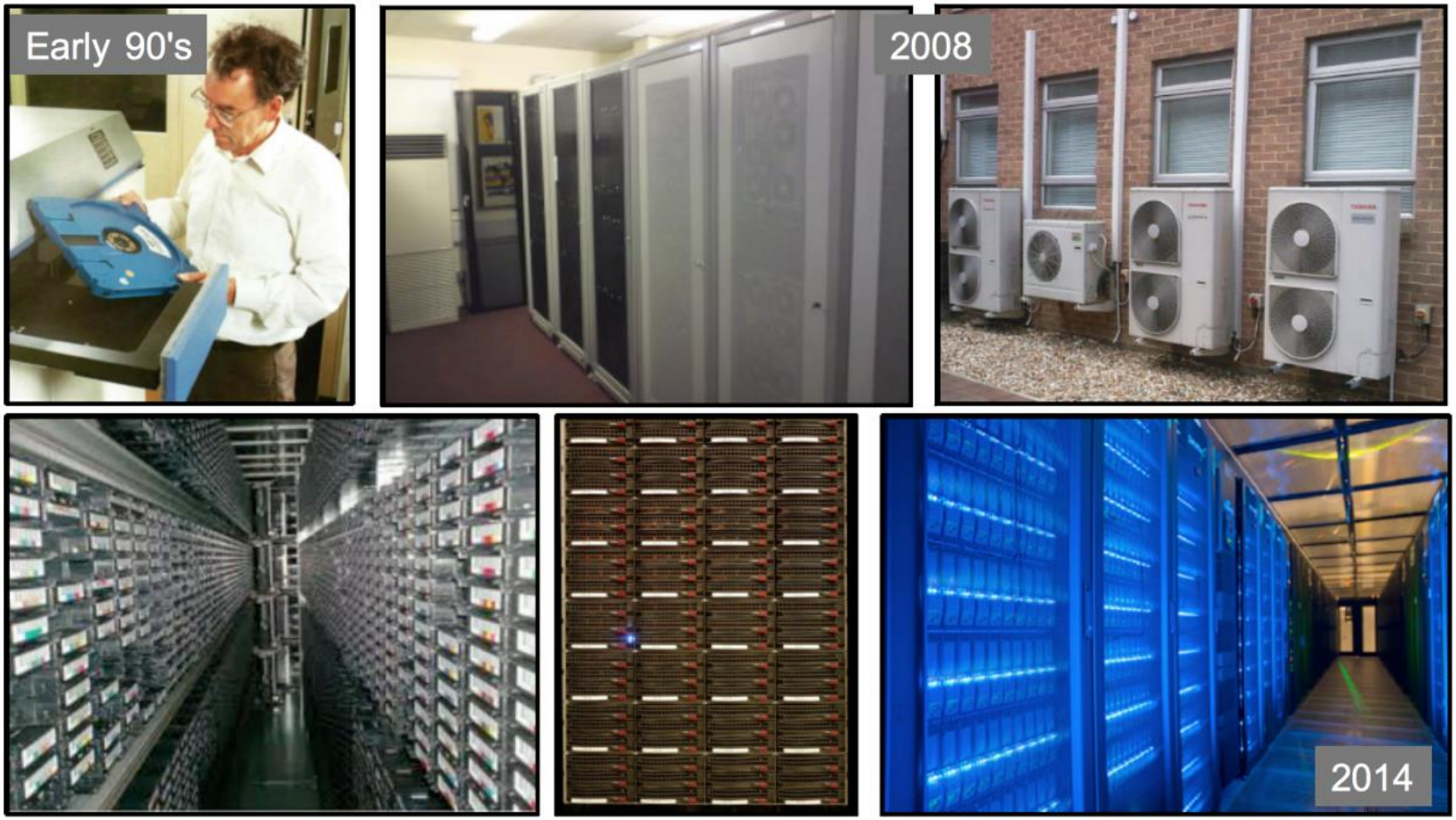
Year	1995	2004	2014
Data shared informally	Met Office University groups NERC facilities	Met Office University groups	Met Office University groups
Data shared via BADC with restrictive license	NERC facilities	Met Office University groups NERC facilities	Met Office
Data share via BADC with open license	NASA	NASA	Met Office University groups NERC facilities NASA

- NERC data policy is now more explicitly open. Data of long-term value (created by NERC funded research) should be curated and must be useable for any purpose after a 2-year embargo.
- Not all data within the BADC is open. Restricted access for non-NERC or NERC data within the embargo period.
- Trend: More open data.

Data management and preservation planning

Year	1995	2004	2014
Is the BADC a Long term data repository?	Maybe	Yes	Yes
Who is picking the data to archive?	DC staff/steering committee	DC staff	DC staff/PI
DMP?	None	Large NERC programmes	All NERC grants
User requirements gathering	Direct contact (small numbers)	User surveys, science meetings	User surveys, science meetings

Storage and server technology



- Migration to new technologies is a continuing task which requires thought and effort.
- 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

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				

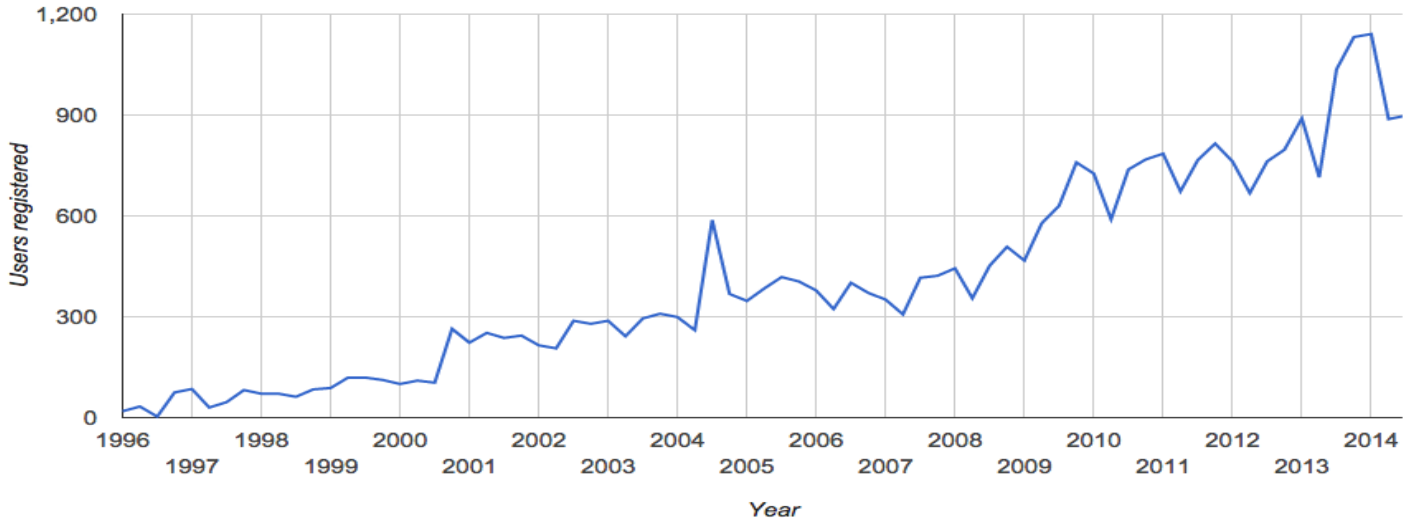
3 BADC staff in group of 5

9 BADC staff in group of 25

User management systems

Year	1995	2004	2014
Annual queries	Not reported	1400	4060
Users from universities	Not reported	75%	70%
Users from UK	Not reported	72%	61%
Total cumulative user registrations	293	6,429	30,000
Query management system	email	Footprints (Proprietary)	Footprints (Proprietary)
User registration system	Manually setup system accounts	Home grown user management system	Home grown user management system

User management systems



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.

Conclusions

- The file system is a great base for an archive.
  - It scales to the volumes needed and allows users to access data in an intuitive form.
- Data outlives researchers (or their roles) and, in most cases, organisations.
  - It is better to have a data management plan in the beginning, and adhere to it.
- The tools of the trade change with time.
  - 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).

The Future

- Will we exist?
- What will the BADC look like in 10 years time?
  - 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. (Due for a change?)
  - 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.