DCC Data Centres Synthesis Study

August 2007-February 2008

Manjula Patel, Graham Pryor
Digital Curation Centre

Executive Summary

Continuous improvement in instrumentation, measurement and storage techniques mean that scientists, researchers and scholars across the UK are now generating an increasingly vast amount of digital data. This is further supplemented by investment in the digitisation of analogue resources and the purchase of digital content and information. Yet the scientific record created in digital form, and the consequent documentary and cultural heritage, are at risk from a range of threats. These include technology obsolescence, the inherent fragility of digital media and the dearth of basic good practice, such as the provision of adequate explanatory documentation. In this context, and working with other practitioners, the Digital Curation Centre (DCC) [1] has committed to support UK institutions that store, manage and preserve these research outputs to help ensure their enhancement and their continuing long-term use.

This study was undertaken as part of the wider Community Development Programme, under Phase 2 of the DCC, which aims to address the curation and preservation needs and requirements of UK data centres. Its goal was to investigate the perceived roles and responsibilities of data centres, how they operate in practice, what policies and standards they apply, and the strategic and operational challenges they are currently encountering. The aim was to identify where best practice can be shared and which areas would benefit from collaboration, at the same time highlighting issues that are of common interest as potential threats or opportunities to the data centre community. The results of the analysis and the insights gained from the survey will contribute to the planning of national fora sponsored by the DCC, as well as other Community Development activities.

The study comprised a series of visits to UK data centres and an online questionnaire. It was undertaken over the period August - December 2007; the writing up and synthesis of the results taking place between January-February 2008. UK data centres that have substantial data holdings were targeted, with a particular emphasis on those that deal with large research datasets. The short timescale available to the study has meant that we received only six responses to our survey:

Medical Research Council (MRC)
UK Data Archive (UKDA)
EDINA
Archaeology Data Service (ADS)
Science and Technologies Facilities Council (STFC-RAL)
NCAS British Atmospheric Data Centre (BADC)

However, the six data centres that responded are well-established organisations with missions underwritten and funded by national agencies, which defines them as legitimate and important strategic instruments of research data management in the UK. Nevertheless, it should be borne in mind that this survey does not provide a representative sample of the entire UK data centre community, since it reflects the views and opinions of only those individuals who responded to it.

All of the data centres taking part in the survey have either national or international standing. All have been funded by a UK Research Council and all have been

operational for at least 10 years, in some cases much longer. It should be noted that the MRC is in the process of setting up a Data Support Service which is likely, initially, to encompass a centralised facility for data collection.

The data centres are all concerned with managing huge amounts of data, ranging from Terabytes to Petabytes. Most serve a specific discipline or community and collect a diverse range of data from sources such as observations, experiments, surveys, simulation and performance. All of the data centres deal with raw, derived and processed data; four also manage normalised, time series and episodic data.

Although the curation and preservation of data was not listed as an explicit responsibility, it is clear that all of the data centres see themselves as having some sort of stewardship role (i.e. providing long-term access, meeting standards for good practice, protecting rights of data contributors, provision of tools for data reuse, training for deposit etc.). The major data contributors are UK researchers in HE/FE, in particular those funded by the UK Research Councils, but also UK government departments and agencies, Ordinance Survey, Meteorological agencies, local government and some commercial sectors. The majority of users appear to come from the same groups as the depositors, but also include the general public and amenity groups. All the data centres collaborate with a large number of organisations, but only with other specific data centres if they are affiliated to them in some way. Their relationship with the evolving institutional repository community is unclear at the moment.

The UKDA and the ADS have written policies that are available on their websites. A new version of the UKDA's policy is due out in January 2008. The MRC and STFC-RAL follow broader data sharing and eScience strategies.

Between them, the data centres have a diverse range of acquisition strategies and policies. Each of the data centres also has its own procedures and processes for deposit and ingest, which varies with the type of data and organisations involved. Deposit agreements are usually presented in the form of licenses which tend to differ between the data centres. A large variety of media and file formats are accepted for data deposit and the workflows involved in the ingest process vary in complexity. Most of the data centres provide tools to aid both depositors and users of their data.

A wide range of curatorial and maintenance functions are undertaken by the data centres. Appraisal and selection for long-term preservation (following deposit) do not generally feature in the remit of the data centres since they have established robust acquisition processes. The centres tend to perform internal audits, and audit instruments such as TRAC and DRAMBORA are being considered for use by several.

In general, the data held by the data centres is made available for third party use free of charge, although some materials are subject to license. The authorisation and authentication procedures for data users vary from data centre to data centre and include ATHENS and UKAMF (UK Access Management Federation) as well as their own internally-generated procedures. Formal data use or reuse agreements include End-User Licences as well as click-through agreements to online *Terms & Conditions*. All of the data centres allow search and/or browse access to their collections and in addition provide some form of machine access, largely in the form of OAI, Z39.50 and web services. However, it appears that data centres do not generally allow their collections to be harvested or transferred to other services, although some data collections are made available through a sister or related service.

All of the operational data centres require metadata to be deposited alongside data. Most of the data centres make use of metadata standards such as Dublin Core and all encode metadata in XML. Domain based subject access vocabularies such as HASSET (social science) and JACS (eLearning) are also in use.

Standards such as those for information security (ISO 17799), records management (ISO 15489), quality management (ISO 9000) and the OAIS (ISO 14721:2003) tend to be used for guidance rather than strict conformance by the data centres. However, some, such as the UKDA, do have to meet certain standards as a matter of legal obligation.

Several of the data centres are involved in formulating and influencing the development of best practices in the areas of metadata, standards and file formats within their own communities. There is general agreement that "open access" is good, but qualified with a requirement for greater awareness of data protection legislation.

Amongst current issues faced by the data centres are

- the need for standards and tools to enable the transfer and exchange of data
- a requirement for increased storage reliability
- options for the management of large and complex datasets
- processes for archiving of open software where it is integral to the deposited resource
- the achievement of interoperability with other domain and cross-domain datasets

Survey respondents were in agreement that one of the areas requiring urgent discussion and clarification is that of the roles, interactions and relationships between newly created institutional repositories and established data centres. In addition, data centre managers also described their concern to deliver and sustain a pool of appropriately qualified staff, as well as to equip data producers and users with the tools necessary for effective data deposit, access and reuse.

Although best practice differs between data centres, there is scope for

- sharing and collaboration in the areas of organisational structures
- the management of relationships with both depositors and users
- identifying the relevance of services to user communities
- the implementation of standards, metadata and open systems
- provisions made for achieving interoperability
- the value of a lifecycle approach to data curation
- the pursuit and delivery of added value

It is evident that the data centre landscape reflects that of the relevant funding bodies, the UK Research Councils. Each data centre has emerged from specific research disciplines and has been developed to cater for the requirements of those particular communities. Consequently, the policies and strategies adopted by each data centre are heavily influenced by those of its corresponding funding council and the community that it purports to serve.

Notwithstanding the formal relationships with their funding agencies, the data centres are nonetheless independent entities, and at an operational level they have developed

their own policies, procedures and processes for data deposit and access. It is at this level that the provision of effective and coherent services, particularly in terms of enabling access, may be at risk. Evolving data centre practices will inevitably prove idiosyncratic given the complexity and heterogeneity of the data being collected. The achievement of interoperability across data sets and disciplines will depend to a large extent on the adoption of standards as well as open file formats and software.

The DCC's recent creation of a national forum for data centres and repositories [20] is an initiative designed to facilitate greater coherence, at an operational level, between these individual organisations, through the exchange of experience and best practice in the processing and management of data of importance to science, technology and society.

It is interesting to speculate whether the current data centre environment is robust enough to meet the challenges and terms which will be identified in the HEFCE-funded data services feasibility study [17], which has embarked on an examination of the management of multi-discipline research data by institutional repositories.

Document Change History

Date	Contributor	Modification
7/1/2008	M Patel	Initial outline of report
8/1/2008	M Patel	Text for introduction and descriptions of participants
14/1/2008	M Patel	Addition of section on methodology
15/1/2008	M Patel	Checking of websites and refinement of descriptions of participants
16/1/2008	M Patel	Addition of section on survey scope and topics; continued checking of websites and refinement of descriptions of participants
17/1/2008	M Patel	Started section on Analysis and Results
18/1/2008	M Patel	Continued section on Analysis and Results
18/1/2008	G Pryor	Provision of text for section on skills and training
28/1/2008	M Patel	Continued section on Analysis and Results
30/1/2008	M Patel	Continued section on Analysis and Results
31/1/2008	M Patel	Continued section on Analysis and Results
1/2/2008	M Patel	Refinement of text and addition of references
4/2/2008	G Pryor	Addition of comments on the report
5/2/2008	M Patel	Refinement of text based on comments from GP
6/2/2008	G Pryor	Provision of text for discussion section
6/2/2008	M Patel	Refinement of text
7/2/2008	G Pryor	Provision of text for section on roles and responsibilities
7/2/2008	M Patel	Refinement of section on questions, results and analysis
8/2/2008	M Patel	Addition of section on best practices; refinement of text
11/2/2008	M Patel	Addition of conclusions; refinement of text
12/2/2008	M Patel	Addition of executive summary; refinement of text
15/2/2008	G Pryor	Comments and refinement of text
15/2/2008	M Patel	Further refinement of text
6/3/2008	M Patel	Incorporation of feedback from survey participants

Contents

1. Introduction

The Digital Curation Centre (DCC) [1], a JISC-funded project,

> provides a national focus for research and development into curation issues and promotes expertise and good practice, both in the UK and internationally, for the management of all research outputs in digital format. The second phase of the project (March 2007-February 2010) aims to provide strategic leadership in digital curation and preservation for the UK research community, with particular emphasis on science data."

The practice of digital curation is described on the DCC website:

> Digital curation is maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific materials.

Scientists, researchers and scholars across the UK are generating an increasingly vast amount of new digital data. This is further supplemented by investment in the digitisation of analogue resources and the purchase of digital content and information. Yet the scientific record thus created in digital form, and the consequent documentary and cultural heritage, are at risk from a range of threats, including technology obsolescence, the inherent fragility of digital media and the dearth of basic good practice, such as the provision of adequate explanatory documentation. In this context, and working with other practitioners, the DCC has committed to support UK institutions that store, manage and preserve these research outputs to help ensure their enhancement and their continuing long-term use.

This study is being undertaken as part of the wider Community Development Programme, under Phase 2 of the DCC, which aims to address the curation and preservation needs and requirements of UK data centres. Its goal is to investigate the perceived roles and responsibilities of data centres, how they operate in practice, what policies and standards they apply, and the strategic and operational challenges they are currently encountering. It will deliver a comparative study that identifies where best practice can be shared and which areas would benefit from collaboration, at the same time highlighting issues that are of common interest as potential threats or opportunities to the data centres community. It is expected that the results of this analysis and the insights gained from the survey will contribute to the planning of national fora sponsored by the DCC as well as other Community Development activities.

2. Methodology

The study, which comprised a series of visits to UK data centres and an online questionnaire, was undertaken over the period August - December 2007; the writing up and synthesis of the results being undertaken between January-February 2008.

Time Period	Activity
August-September 2007	Scoping of survey and development of questionnaire
October 2007	Peer review and refinement of questionnaire
November-December 2007	Collection of responses to questionnaire
January-February 2008	Analysis of results and writing of report

UK data centres that have substantial data holdings were targeted, with a particular emphasis on those that deal with large research datasets. Individuals at the following organisations were initially contacted by email and invited to take part in the study:

Medical Research Council (MRC)
Natural Environment Research Council (NERC)
UK Data Archive (UKDA)
EDINA
Archaeology Data Service (ADS)
Science and Technologies Facilities Council (STFC)
NCAS British Atmospheric Data Centre (BADC)
European Bioinformatics Institute (EBI)

A web-based platform (SurveyMonkey [2]) was used to implement and administer the online questionnaire. In addition, respondents were asked if they would be prepared to take part in follow-up interviews.

With regard to data protection, the survey platform complies with EU Safe Harbor provisions [3] and the questionnaire was made available via a secure URL. Survey respondents were advised that

> No personal details will be used without permission, although individual input to the questionnaire may be quoted anonymously to validate the analysis.

It should be borne in mind that this survey does not provide a representative sample of the entire UK data centre community, since it reflects the views and opinions of only those individuals who responded to it.

3. Respondents

In order to provide some context and an indication of the type and nature of the organisations taking part in the survey, we provide a brief overview and some related background information for each. From this overview it is evident that all of these data centres perform a stewardship role with respect to preservation and curation, in addition to their core activity of making data accessible for the purpose of research and reuse. Further, each data centre is a discrete entity that has developed its own policies, procedures and processes for data deposit and access. Several data centres also provide best practice guidelines, as well as training courses and workshops, covering the management and use of the data encountered within their particular domains.

3.1 UK Data Archive (UKDA)

The UK Data Archive (UKDA) [4]

is a centre of expertise in data acquisition, preservation, dissemination and promotion and is curator of the largest collection of digital data in the social sciences and humanities in the UK. It is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and the University of Essex. Founded in 1967, it now houses several thousand datasets of interest to researchers in all sectors and from many different disciplines.

The UKDA provides resource discovery and support for the secondary use of quantitative and qualitative data in research, teaching and learning. As a lead partner of the Economic and Social Data Service (ESDS) [5], the UKDA is responsible for: overall integration and management of the ESDS access and preservation programme, focusing on the central activities of data acquisition, processing, preservation and dissemination; ESDS Qualidata (a specialist service for a range of qualitative datasets); and ESDS Longitudinal, undertaken jointly with the Institute for Social and Economic Research (ISER).

The UKDA also provides preservation services for other data organisations, also supporting the National Centre for e-Social Science (NCeSS) and facilitating international data exchange through agreements with other national archives. The UKDA hosts AHDS History, one of the five Centres of the Arts and Humanities Data Service, and Census.ac.uk, enabling access to the census data resources for UK higher and further education.

The UKDA's Data Catalogue provides access to over 5000 computer-readable datasets for use in research and teaching, within a range of different disciplines. Comprehensive online information and links to enable access to data are provided for each dataset. Searching and browsing of information about the data is free and does not require registration, but registration *is* a prerequisite to accessing any raw data. Types of data administered by the UKDA include:

quantitative

- micro data are the coded numerical responses to surveys, with a separate record for each individual respondent
- macro data are aggregate figures, for example country-level economic indicators

Data formats include SPSS, Stata and tab delimited formats

qualitative

- data include in-depth interviews, diaries, anthropological field notes and complete answers to survey questions

Data formats include Excel, Word and RTF

multimedia

- a small number of datasets include image files, such as photographs, and audio clips

non-digital material

- paper media includes photographs, reports, questionnaires and transcriptions
- analogue audio or audio-visual recordings

Over 200 new datasets are added each year and come from different sources. These include official agencies (mainly central government); international statistical

agencies; individual academics with (ESRC) research grants; market research agencies; historical sources; as well as other data archives worldwide.

3.2 EDINA

EDINA [6] is the JISC national academic data centre based at the University of Edinburgh. Its mission and purpose is to 'enhance the productivity of research, learning and teaching' across all universities, research institutes and colleges in the UK. Access to most services involves licence or subscription by universities and colleges, and requires some form of authentication. Current online services include:
Film and Sound: delivers downloadable film, video and audio from 17 collections. These high-quality films cover a broad range of subjects from medicine to 20th-century history and include one classical audio music collection.
Education Image Gallery: offers access to 50,000 downloadable photographic images licensed from Getty Images. A further 10,000 images will be selected over the new agreement period from August 2007 to July 2010. The photos cover a diverse range of subject areas such as sport, fashion, major events, buildings, politics, social history, key personalities, transport, industry, work, leisure and music. The number of subscribing institutions has grown steadily to 133 by end of July 2007. The images are copyright-cleared and can be downloaded at screen resolution for use in learning, teaching and research.
UKBORDERS: an integral part of the ESRC Census Programme, for which EDINA acts as the Geography Data Unit, UKBORDERS offers access to more than 50 digital boundary datasets for past and present geographic areas. A recent additional service offers the ability to download postcode directories.
Digimap Collections: provides online access to current national mapping from the Ordnance Survey (OS). Since April 2005, this has extended to earlier raster maps from Landmark Historic, and since January 2007 users have had access to mapping from the British Geological Survey (BGS), as Geology Digimap. There are plans to add hydrographical map data (Marine Digimap) in 2008. In addition, Historic Digimap provides access to historical maps of Great Britain for the period 1843 to 1996 and Geology Digimap delivers geological maps and data from the British Geological Survey (BGS).
Land, Life & Leisure: covers current practice and developments in temperate agriculture and all rural topics, conservation, estate management, forestry, horticulture, organic husbandry, rural planning, recreation and tourism, and environmental issues.
Index to the Times: comprises Palmer's Index to The Times (covering the years 1790 to 1905), and the Official Index to The Times (covering the years 1906–1980).
Statistical Accounts of Scotland: JISC funding assisted the set-up of an online access service for the Statistical Accounts of Scotland, probably the best source of contemporary comment on Britain's experience of the agricultural and industrial revolutions (covering the period 1790 to 1845). This provides both a free service and a value-added service accessible by institutional subscription.
CAB Abstracts: a bibliographic database compiled by CAB International and offered by EDINA since 1999. It covers the significant research and development literature in the fields of agriculture, forestry, aspects of human and animal health, conservation and leisure & tourism. It now contains over five million records from 1973 to date. 150,000 records are added each year from over 10,000 serial titles, books, monographs, technical reports, proceedings, patents and published theses. In addition

to the contemporary service an archive is offered by separate subscription covering the period 1910-1972.

For further details and information on emerging services see the EDINA website [6] and the EDINA Community Report, 2007 [7].

3.3 Archaeology Data Service (ADS)

The Archaeology Data Service (ADS) [8] is part of the Arts & Humanities Data Service (AHDS) [9] and hosts AHDS Archaeology. The ADS is primarily funded by the Arts and Humanities Research Council and the Joint Information Systems Committee (JISC).

The aim of the ADS is to collect, describe, catalogue, preserve and provide user support for digital resources that are created as a product of archaeological research. The ADS also has a responsibility for promoting standards and guidelines for best practice in the creation, description, preservation and use of spatial information across the AHDS as a whole. A mission statement is available on the ADS website [8]:

> The Archaeology Data Service (ADS) supports research, learning and teaching with high quality and dependable digital resources. It does this by preserving digital data in the long term, and by promoting and disseminating a broad range of data in archaeology. The ADS promotes good practice in the use of digital data in archaeology, it provides technical advice to the research community, and supports the deployment of digital technologies.

Archaeology occupies a special position in that much of the creation of its data results from the destruction of primary evidence, making access to data all the more critical in order to test, assess and subsequently reanalyse and reinterpret both data and the hypotheses arising from them. Over the years, archaeologists have amassed a vast collection of fieldwork data archives, a significant proportion of which remains unpublished.

The ADS is working with national and local archaeological agencies and those research councils involved in the funding of archaeological research, to negotiate the deposit of project data. This includes data derived from fieldwork as well as desk-based studies. The types of data involved include text reports, databases (related to excavated contexts or artefacts, for example), images (including aerial photographs, remote sensing imagery, photographs of sites, features and artefacts), digitised maps and plans, numerical datasets related to topographical and sub-surface surveys and other location data, as well as reconstruction drawings.

The following agencies either require or recommend that datasets produced by their grant holders should be offered to the ADS or one of its sister services in the Arts and Humanities sector:

- Arts and Humanities Research Council
- British Academy
- Carnegie Trust
- Council for British Archaeology
- Economic and Social Research Council

- Leverhulme Trust
- Natural Environment Research Council
- Wellcome Trust's History of Medicine Programme

For further information regarding details of collections and background information, see the ADS website [8].

3.4 The Medical Research Council (MRC)

The Medical Research Council (MRC) is a publicly-funded organisation dedicated to improving human health. It supports research across the entire spectrum of medical sciences, in universities and hospitals, in its own units and institutes in the UK, and in its units in Africa. The following explanation of its mission is available on the MRC website [10]:

> The heart of our mission is to improve human health through world-class medical research. To achieve this, we support research across the biomedical spectrum, from fundamental lab-based science to clinical trials, and in all major disease areas. We work closely with the NHS and the UK Health Departments to deliver our mission, and give a high priority to research that is likely to make a real difference to clinical practice and the health of the population.

The terms and conditions of MRC grants [11] include the following statement relating to data sharing:

> All proposals for MRC funding submitted after 1st January 2006 must include (costed) plans for preparing and documenting research data for preservation for sharing in line with MRC data sharing policy, access principles, and guidance. As part of the end of grant reporting process, MRC funded researchers will also be expected to report on data management and sharing activities relating to these plans.

The MRC is currently in the process of implementing a Data Support Service that will have stewardship responsibilities for medical research data. This development has emerged from the broader strategic context of the MRC's data sharing policy, which promotes new and extended use of data beyond the originating research teams. MRC data access principles and guidance set a community-wide framework.

3.5 NCAS British Atmospheric Data Centre (BADC)

The BADC is one of seven Natural Environment Research Council (NERC) [12] data centres established to administer and support NERC data policy. It provides services to the National Centre for Atmospheric Sciences (NCAS), which is responsible for the core research programme in atmospheric science funded by the NERC. The NERC data policy is outlined in the NERC Data Policy Handbook [13], which describes the responsibilities of both NERC-funded researchers and the NERC designated data centres. The mission statement of the BADC is available on its website [14]:

> The role of the NCAS British Atmospheric Data Centre (BADC) is to assist UK researchers to locate, access and interpret atmospheric data and to ensure

> the long-term integrity of atmospheric data produced by Natural Environment Research Council (NERC) projects.

The BADC has substantial data holdings of its own and also provides information and links to data held by other data centres. Two types of dataset are held by the BADC: those produced by NERC-funded projects (treated as high priority since the BADC may be the only long-term archive of the data) and third party datasets, which are required by a large section of the UK atmospheric research community (such as those produced by the Meteorological Office and the European Centre for Medium Range Weather Forecasting).

There is also considerable interest in the BADC's data holdings from the international research community, in particular the Meteorological Office data. The BADC is also required to take a proactive role in the data management practices of NERC thematic programmes.

All BADC data are available online through a World Wide Web interface or via an ftp service. At present the data index page lists 148 datasets as being available. Software is provided to assist in the manipulation of the data and extensive information is provided on the data collection procedures, formats, data quality, contact names and references to journal papers. Other specialist services include the development of value added data products such as averaged and girded data, and the generation of video sequencing of data to facilitate viewing of large datasets.

Some datasets held at the BADC are restricted to academic use only and require registration as a condition of access. Only selected project members are authorised to submit data to the BADC. The BADC website [14] provides additional background and other community related information including guidance on preferred file formats, and access and deposit procedures.

3.6 STFC-RAL Data Cluster

Within the Science and Technology Facilities Council (STFC), the RAL (Rutherford Appleton Laboratory) Data Cluster [15] is a part of the National Grid Service (NGS) which

> …aims to provide coherent electronic access for UK researchers to all computational and data based resources and facilities required to carry out their research, independent of resource or researcher location.

The Data Cluster was originally funded by PPARC, but is now largely funded by users through Service Level Agreements or grants with contributions from CCLRC and the national e-Science Centre at RAL.

The very large scale data storage and management production services at STFC are aimed primarily to meet the needs of the high priority long term projects identified in the STFC delivery plan: Large Hadron Collider (LHC), ISIS-Target Station 2 and the

Diamond Light Source (DLS), but also, where appropriate, to help meet the requirements of UK science, as directed by (part of) the STFC mission statement[1].

The capacity of the existing archive systems is currently 5 Petabytes, with an additional 2 Petabytes of disc based storage, with a plan to move to new technology over the next 3 years that will double this capacity. Much of the investment in long term, high performance, and large scale data storage services is driven by the needs of the LHC experiments, which are expected to continue to generate data for at least the coming decade, and for which the STFC-RAL Data Cluster has archival responsibility for the UK. Plans are under development for future massive storage capability through an existing collaboration with CERN. The development of this service has become a central focus, and is laying the foundation for future very large-scale services to be offered to non LHC users.

Non LHC users of the data storage services are also growing. Deployment of long term data storage services to all institutes of the BBSRC is continuing. This service is now being extended to similar institutes in Scotland. In addition, data archive and storage services are provided to many scientific groups including the British Atmospheric Data Centre, the NEODC, the Arts and Humanities Data Service, National Crystallography Service, Southampton University, to EISCAT, WASP, and VIRGO Consortium, as well as to others and the growing STFC facilities. In addition, discussions are underway with many new users to provide long term data storage and archive services across many different scientific communities, as the business of managing and storing the ever faster growing volumes of science data continues to cause serious problems for the science community.

4. Survey Scope and Topics

This survey examined data centre curation and preservation strategies and policies as well as operational practices, with the aim of understanding both current practice and future needs and plans. A major aim of the survey was to gain an understanding of both the diversity and synergies in current data centre policies, practices and services. Broadly speaking the areas under investigation included:

- Data holdings
- Responsibilities and relationships
- Curation and preservation strategies and policies
- Principal operating practices and services
  (Acquisition; Deposit and ingest; Data management; Access and dissemination; Metadata; Standards)
- Open issues and future development

5. The Survey: Questions, Results and Analysis

The short timescale available to this study has meant that we received only six responses to our survey, despite several reminders and extensions to the survey

[1] "promote and support high-quality scientific and engineering research by developing and providing, by any means, facilities and technical expertise in support of basic strategic and applied research programmes funded by persons established in the United Kingdom and elsewhere."

deadline. Regrettably, promises to participate failed to materialise in a number of cases. NERC in particular was targeted in the hope of acquiring responses from each of its seven data centres, but we received a response from only one (BADC).

5.1 General Information

All of the data centres taking part in the survey have either a national or an international standing. All are or have been funded by a UK Research Council and all have been operational for at least 10 years, in some cases much longer. It should be noted that the MRC is in the process of setting up a Data Support Service which is likely initially to encompass a centralised facility for data collection.

All of the data centres play a stewardship role incorporating measures to enhance access to data, the management of digital storage, facilitating the re-use of data, long-term preservation, the promotion of standards and best practice, and the development of community collaboration tools.

1. *Please provide some details about yourself and the Data Centre (these will be used to notify you when the survey report is published)*

   The respondents included:
   Head of Digital Preservation and Systems
   Director (of Centre)
   User Services Manager
   Research Programme Manager
   Curation Manager
   Deputy Division Head, Data Services Division

2. *What are your specific duties in relation to your role or position?*

   The responsibilities of the respondents varied considerably:
   - My main task is to develop, implement, and maintain a comprehensive digital preservation policy framework
   - Leadership and overall responsibility
   - External liaison, resource promotion, research and management
   - Lead on data sharing, eHealth and research governance
   - Oversee operation of the data centre. Direct curation practice. Look into emerging curation issues
   - Manage teams who run the data storage and archive services for a wide user base

3. *What are the primary purposes of the Data Centre? (please choose all that apply)*

	UKDA	EDINA	ADS	MRC	BADC	STFC-RAL
Enhancing access to data	√	√	√	NA	√	√
Long-term preservation	√		√	NA	√	√

Facilitating re-use of data	√	√	√	NA	√	
Managing digital storage	√	√	√	NA	√	√

Additional responses included:

- We do not have a direct remit for long term preservation but we are considering the implications for taking on that role
- Not applicable yet. MRC is in the process of establishing a Data Support Service which in the first two years will focus on facilitating data re-use and enhancing long-term preservation for research access. It will involve establishing a centralised archive/repository in the early years
- Community collaboration tools. Research into atmospheric science
- Promoting standards and best practice

4. *How and when was the Data Centre established?*

   All the data centres have been in operation for at least 10 years and several for considerably longer, e.g. the UKDA recently celebrated its 40th anniversary.

5. *How is the Data Centre funded?*

   All of the data centres are or have been funded by their respective UK Research Council. The STFC-RAL Data Cluster, which was funded by PPARC, is now principally funded by users through Service Level Agreements or grants with contributions from CCLRC and the national e-Science Centre at RAL.

6. *What is the status of the Data Centre?*

   Two of the data centres have an international standing, whilst the other two are UK national and serve only the UK community. The international data centres accept data largely from UK organisations or researchers, but allow use of the data to be made outside of the UK.

5.2 Data Holdings

The data centres are all concerned with managing huge amounts of data, ranging from Terabytes to Petabytes. Most serve a specific discipline or community. A diverse range of data is collected, including from sources such as: observations, experiments, surveys, simulation and performance. All of the data centres deal with raw, derived and processed data; four also manage normalised, times series and episodic data.

7. *What is the approximate size of the Data Centre's data holdings (e.g. number of data sets or collections)?*

   The data centres hold considerable amounts of information. They reported:

- …as of August [2007] 5094 studies comprised of 1,450,436 files. A further single collection is made up of 394,687 files. Total file size approx 1.5 Tb
- About 20 main services, with variable number of datasets/collections, sometimes numbering hundreds per service
- 150 data sets, 100TB, 100,000,000 files
- C.500 Collections/Resources

8. *Which discipline(s) does the Data Centre cater for?*

   A number of data centres have been established to serve a specific community (e.g. social science, humanities, archaeology, biomedicine, atmospheric science etc.), nonetheless, together with the other Centres, they do offer services to a broader base of users and multiple disciplines (e.g. STFC-RAL).

9. *What types of data are collected? (please select all that apply)*

   Observational data is collected by all the data centres, followed by computational and survey (5) and then experimental and simulation data (4). Three of the data centres also collect performance data.

	UKDA	EDINA	ADS	MRC	BADC	STFC-RAL
Experimental			√	√	√	√
Observational	√	√	√	√	√	√
Computational		√	√	√	√	√
Survey	√	√	√	√		√
Simulation			√	√	√	√
Performance			√	√		√
Other	historical	words, numbers, pictures & sounds	All digital archaeological data sets	All from MRC-funded research		

10. *What are the characteristics of the data collections? (please select all that apply)*

    All of the data centres deal with raw, derived and processed data; four also manage normalised, times series and episodic data.

	UKDA	EDINA	ADS	MRC	BADC	STFC-RAL
Raw	√	√	√	√	√	√
Derived	√	√	√	√	√	√
Processed	√	√	√	√	√	√
Normalised		√	√	√	√	
Time Series	√		√	√	√	
Episodic		√	√	√	√	
Other			All digital archaeological data sets	All from MRC-funded research	Climatology	

5.3 Responsibilities and Relationships

Although, curation and preservation of data was not listed as an explicit responsibility, it is clear that all of the data centres see themselves as having some sort of stewardship role (providing long-term access, meeting standards for good practice, protecting rights of data contributors, provision of tools for data reuse, training for deposit etc.). The major data contributors are UK researchers in HE/FE, in particular those funded by UK Research Councils, but also: UK government departments and agencies; Ordinance Survey; Meteorological agencies; local government and some commercial sectors. The major users appear to come from the same groups as the depositors, but also include the general public and amenity groups. All the data centres collaborate with a large number of organisations, but only with other specific data centres if they are affiliated to them in some way. Their relationship with Institutional Repositories is unclear at the moment.

11. What are the major responsibilities of the Data Centre? (please select all that apply)

Meeting standards for good practice was indicated as a key responsibility by all respondents; whilst five also indicated managing data for the long-term, protecting the rights of data contributors and promoting the repository service; four are also responsible for the provision of tools for the reuse of data and two for providing training for deposit. Other responsibilities specified included:

- Brokering access to existing long-term preservation services in the first instance; not centralised archiving in first two years
- Data acquisition
- Develop technology to meet the future needs of digital scientific data archive and storage systems

12. Who are the major data contributors?

The major data contributors are UK researchers in HE/FE, in particular those funded by UK Research Councils, but also: UK government departments and agencies; Ordinance Survey, Meteorological agencies; local government and some commercial sectors.

13. Who are the major users of the data collections?

The major users appear to come from the same groups as the depositors, but also the general public and amenity groups.

14. Does the Data Centre collaborate with any other bodies (e.g. another Data Centre, a publisher or a learned society)? If so, in what way?

All the data centres collaborate with a large number of other organisations; however, responses were qualified:

- Yes, but probably not in the way in which this question is framed

- Yes, this is important. In addition to working with vendors/licensors, EDINA works v. closely with Mimas (esp. for Jorum) and others in the JISC community
- Yes, the MRC DSS will need to co-ordinate with a wide range of different stakeholders too numerous to mention in detail here. Stakeholders will be looking for different things from the DSS
- Other NERC data centres. Development projects include collaborators in industry, university groups, Meteorological Office, institutional repositories and the Royal Meteorological Society
- Yes - endless. I'd need a month to complete this. We collaborate with most major research councils, with JISC, many universities, with CERN, JET, the DCC, and with other international bodies
- Numerous collaborations inc. deep storage, cross-searching, hosting of Learned Society journals

15. *Do you foresee a relationship between the Data Centre and Institutional Repositories which are currently being set up?*

    Of the respondents, five definitely envisage a relationship with Institutional Repositories (IR), one of which is already providing storage services to IRs. The remaining data centre is currently unclear on this issue.

5.4 Curation and Preservation Strategies and Policies

The UKDA and the ADS have written policies which are available on their websites. A new version of the UKDA's policy is due out in January 2008. The MRC and STFC-RAL Data Cluster follow broader data sharing and eScience strategies. The Research Data Principles and Guidelines [16], published by the Research Information Networks (RIN) is beginning to influence strategies and policies in some data centres.

16. *Does the Data Centre have a written curation and preservation strategy and/or policy? If so, what does it state and who is responsible for maintaining it?*

    The UKDA and the ADS have written policies which are available on their websites. A new version of the UKDA's policy is due out in January 2008. The MRC and STFC-RAL follow broader data sharing and eScience strategies.

17. *Have external factors, such as the RIN's Research Data Principles and Guidelines, affected strategy and/or policy? If, so in what way?*

    One respondent was unaware of the RIN (Research Information Networks) and its Research Data Principles and Guidelines [16]. Two respondents said that the guidelines had influenced their strategies, while the remaining three indicated that it had not had any influence as yet.

5.5 Principal Operating Practices and Services

This section is partitioned into: acquisition; deposit and ingest; data management; access and dissemination; metadata and standards.

5.5.1 Acquisitions

Between them, the data centres have a diverse range of acquisition strategies and policies, which tend to vary according to the type of data and community they are managing or serving.

18. Please describe briefly the Data Centre's current acquisitions strategy and/or policy

Each of the data centres has its own acquisitions strategy/policy. They vary in coverage and style according to the type of data involved:

- We have two: http://www.esds.ac.uk/andp/create/policy.asp and a non-public 'collection themes' policy for AHDS History (to become History Data Service). Essentially AHDS History prioritises digital resources created by historians with AHRC/JISC funding or of value to historians in UK HE/FE
- This is largely determined by the requirements put upon EDINA by JISC, and within that by our Strategy Plan and the responses made by EDINA business development staff to various ITTs. Identify academic scientific users who require our services. Talk to them. Find out what they need. Deliver a service for an initial period. Agree and sign-off an SLA. Varies with different communities
- It's dataset dependent. NERC programmes are obliged to give us the data.
- Met agencies we go and get the data. Others case by case
- All archaeological data in digital format - cost model varies by nature of depositor

5.5.2 Deposit and ingest

Each of the data centres has its own procedures and processes for deposit and ingest, which further varies with the type of data and people or organisations involved. Deposit agreements are usually presented in the form of licenses which vary considerably between the data centres. A large variety of media and file formats are accepted for data deposit and the workflows involved in the ingest process vary in complexity. Most of the data centres provide tools to aid both depositors and users of their data.

19. Please describe the authorisation/authentication process for depositors, if one exists

- the Depot & Jorum use Athens (and will be UK AMF); most other ingest is from single source 'vendors'/'publishers', but also including transfer from projects
- This is very complex for us. Look at http://badc.nerc.ac.uk/data/rules.html for summary
- Different for different user communities
- Dialogue with collections development team followed by deposit validation

20. Please describe any formal deposit agreements that are in place

- License agreement for ESDS is at: http://www.esds.ac.uk/aandp/create/licence.asp . License agreement for AHDS is similar
- Try depot.edina.ac.uk and jorum.ac.uk
- Many and various
- Numerous and varied (all are covered by the standard ADS deposit licence, available on line)

21. What are the most common types of media used for data deposit?

The most common forms of media used for data deposit are network file transfer, and CDROM, followed by others including: magnetic tape, Linear Tape-Open (LTO), both Digital Linear Tape (DLT) and Super Digital Linear Tape (SDLT), Digital Audio Tape (DAT) and other data cartridges, Compact Discs (CD-ROM, CD-R, CD-RW), DVD (all types), Zip Disk as well as satellite links.

22. What are the most common types of file format used for data deposit?

All the operational data centres accept a large variety of file formats for data deposit, varying from international standards to domain standards to *ad hoc* formats.

23. Are any of the file formats above prescribed by the Data Centre? If so, which?

Three responses indicated that the respective data centre does prescribe file formats for deposit; the others provided no response to this question. However, one respondent seems to have misread "prescribed" for "preserve". The BADC and the ADS both provide online information on prescribed or preferred formats.

24. Once data has been deposited with the Data Centre, who has ownership of the data?

According to two data centres, it is the data owner that retains ownership of the data. However, most indicated that it depends on license agreements and the use to which the data will be put.

25. During the deposit process, are contributors required to provide any metadata? If so, is this a prescribed set of information?

All five of the currently operational data centres indicated that data depositors are required to provide metadata during deposition. However, the conditions under which depositors are required to conform to a set of prescribed information tend to vary.

26. Following deposit are any integrity, authenticity or provenance related checks performed by the Data Centre? If so, please provide details

All of the data centres perform checks on the data, but the types of checks vary with the types of data being deposited:

- Yes, depends on dataset
- Integrity checking
- Yes, full validation to file level inc. MD5 - authenticity and provenance covered by deposit licence (i.e. warranted by depositor).

27. Does the Data Centre perform any normalisation and/or file format conversion on the data?

Of the operational data centres, four process deposited data, whilst one does not.

28. What tools are provided by the Data Centre to aid depositors, if any?

Both the UKDA and the ADS provide online advice, guidance, forms and templates for depositors. In the case of the BADC, a web-based up-loader and checker is provided for data submission by authorised science programme project participants. For the STFC-RAL Data Cluster, the tools vary according to the community from which the data is being supplied:

- Many and various - often bespoke for the community concerned. Sometimes developed and shared by the user community itself.

29. Are depositors provided with a method for submitting contextual and/or semantic information alongside the data?

All of the operational data centres provide a method for the submission of contextual and/or semantic information with the data deposited.

30. Please describe any workflow associated with the deposit/ingest process

The ESDS provides an overview of the multistage deposit/ingest workflow at http://www.esds.ac.uk/aandp/create/Stagethree.asp. It comprises the following stages: acquisitions review; processing the data; quantitative data processing; qualitative data processing; creating the catalogue record; Service Level Definition; completion of processing; and preservation of the data.
The response from EDINA was

- varies. suggest that you try Depot & Jorum, or download the documentation from the website Jorum: this has a workflow that is controlled by the depositor, but metadata is added/edited by Intute. the Depot: user has control over workflow, that includes check with publisher policy, and option for closed access

As the response from BADC indicates, the workflow can vary with the datasets:

- This is way too small a box... Dataset dependant

The ADS follows the workflow of the OAIS Model.

5.5.3 Data Management

A wide range of curatorial and maintenance functions are undertaken by the data centres collectively. Appraisal and selection for long-term preservation (after deposit) do not generally feature in the remit of the data centres since they have established robust acquisition processes. The centres tend to perform internal audits, and audit

instruments such as TRAC and DRAMBORA are being considered for use by several data centres. As far as active long-term preservation of data is concerned; the UKDA has a policy on its website which is currently under review; BADC's main strategy is to adopt open file format standards; whilst the ADS follows the OAIS Model.

31. What type of curatorial and maintenance functions are performed by the Data Centre?

Here the answers varied as follows:

- migration and refreshment of preserved data
- value add and keep-safe
- fixity check, backup, metadata review, media migration, storage and metadata standards adaptation
- Long term integrity checking - Bit level migration; technology migration. Growing requirements for metadata
- See OAIS Ref. Model

32. Please provide details of any techniques used in the appraisal and selection of deposited data for long-term preservation. For example, are data-sets appraised after a certain period of time or are particular data-sets marked as being of significance to the domain or user community?

- Essentially all datasets ingested are considered to be of permanent value. Our policy is:
  - the data collections selected must be suitable for scholarly re-use;
  - datasets accessioned should contain information which has permanent or continuing value
- we do not have such a procedure; as indicated we provide continuing access, until we are asked to stop ....
- NERC data overseen by programme committee. Data scientist from BADC assesses data
- User communities drive this - not the data centre
- The ADS follows the processes and procedures in the OAIS Model.

33. Are audits performed on the Data Centre?

Four of the operational data centres responded positively.

- Internal. Varied scope. Varied time scales. Many are *ad hoc*, but in the process of implementation
- NERC science management audit. Test OAIS audit
- Both - every few years
- Internal e.g. DRAMBORA (TRAC also done internally)

34. Does the Data Centre undertake active long-term preservation of its data collections?

The UKDA has a policy on its website, 2005 (currently under review): http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf . BADC's main strategy is to adopt open file format standards, whilst the ADS follow the OAIS Model.

35. How does the Data Centre manage identification of data sets?

BADC uses URLs for datasets, while ADS uses "Internal/External classification systems". EDINA does not manage dataset identification.

36. How does the Data Centre handle versioning of data?

- If this means what I think it does, we create DIPs directly from processed AIPs
- Badly. Depends on dataset
- In house Collections Management System, developed by ADS in use by all AHDS subject centres

37. Are there any pertinent technical issues? If so, please provide details

No responses were provided.

5.5.4 Access and Dissemination

In general, the data held by the data centres is made available for third party use free of charge, although some materials are subject to license. The authorisation and authentication procedures for data users vary from data centre to data centre and include: ATHENS, UKAMF (UK Access Management Federation) and their own procedures. Formal data use or reuse agreements include End-User Licences as well as click-through agreements to online Terms & Conditions. All of the data centres allow search and/or browse access to their collections and in addition provide some form of machine access, largely in the form of OAI, Z39.50 and web services. However, it appears that data centres do not generally allow their collections to be harvested or transferred to other services, although some data collections are made available through a sister or related service.

38. Please provide details of any data collections that are available for third party use

All the data centres make all of their collections available, free of charge, for third party use (mostly UK HE/FE) although some materials are subject to license. Some data sets are also made available for commercial use.

39. Describe the authorisation/authentication process for data users, if one exists

UKDA makes use of ATHENS and a special license. For details see http://www.esds.ac.uk/aandp/access/login.asp .
EDINA uses UKAMF [UK Access Management Federation], ATHENS and its own authentication.
BADC allows anonymous access to some datasets, as well as licenses for restricted access data sets; the rules are available at http://badc.nerc.ac.uk/data/rules.html.
ADS enables access to all its data via a click-through agreement to the Terms & Conditions published on its website at http://ads.ahds.ac.uk/catalogue/ .

40. Please describe any formal use/re-use agreements or licenses that are in place

UKDA uses End-User Licenses (EUL) both as part of the registration process as well as for particular EULs for special services or data.
EDINA uses many types of agreements or licenses as appropriate.
BADC rules for access vary according to the data involved:
http://badc.nerc.ac.uk/data/rules.html.
ADS require agreement with the Terms & Conditions published on its website.

41. Does the Data Centre provide access to metadata and/or contextual information to aid third parties in the use of data sets?

All of the operational data centres provide third party access to metadata and/or contextual information.

42. Does the Data Centre provide search and/or browse access to its data collections?

All of the operational data centres provide search and/or browse access to data collections.

43. Does the Data Centre provide machine access to its data collections?

All the operational data centres provide some form of machine access to their data collections, mostly OAI, Z39.50 and web services.

44. Describe any additional tools or information provided by the Data Centre to help third parties make use of data collections?

This question was only answered by one participant describing the ESDS Humanities and Social Science Electronic Thesaurus (HASSET):
http://www.esds.ac.uk/search/hassetAbout.asp .

45. Does the Data Centre allow harvesting or transfer of data collections to other Data Centres, repositories or archives?

Two respondents answered "yes" and two answered "no". However, it appears that in general, data centres do not allow their collections to be harvested or transferred to other services, although some data collections are made available through a sister or related service.

46. Are any data collections indexed by external search engines, database services, aggregators or similar services?

Data collections are indexed by a variety of services including: FSOL and the Depot (EDINA); web search engines, such as Google, goGeo and GCMD (BADC); and AHDS Cross search engine and EH Heritage Gateway (ADS).

5.5.5 Metadata

UKDA, EDINA, BADC and ADS all require metadata to be deposited alongside data. UKDA provides a deposit form for contributors to fill in whilst ADS provides online guidelines for depositors. Metadata is created and maintained by depositors, technical staff, software and administrative staff as well as through outsourcing (e.g. to Intute). Most of the data centres make use of metadata standards such as the Dublin Core and all encode metadata in XML. Domain based subject access vocabularies such as HASSET (social science) and JACS (eLearning) are also in use.

47. Please provide details of any metadata required during the deposit/ingest procedure?

UKDA, EDINA, BADC and ADS all require metadata to be deposited. UKDA provides a deposit form for contributors to fill in, see http://www.esds.ac.uk/aandp/create/depform.asp.
ADS also provides guidelines for depositors at http://ads.ahds.ac.uk/project/userinfo/deposit.html.

48. How is metadata created and maintained? (please select all that apply)

Metadata is created and maintained by depositors, technical staff, software and administrative staff as well as through outsourcing (e.g. to Intute)

49. Please provide details of any standards, schema or profiles used in recording metadata

The predominant metadata standard used by UKDA is DDI (Data Documentation Initiative). BADC uses: DIF (ISO19115), MOLES, NERC data grid schema CSML, CDML and Dublin Core. The ADS' metadata is based on Dublin Core.

50. Is the metadata encoded in XML?

All the operational data centres make use of XML for recording metadata.

51. Please provide details of any subject access vocabularies such as keywords, thesauri, classifications or ontology in use

UKDA: HASSET, http://www.esds.ac.uk/search/hassetAbout.asp
EDINA: jacs, learndirect, unesco, and lots of geo-spatial, plus proprietary
BADC: CF names aimed at GCMD discovery terms
ADS: Numerous, based on MIDAS, EH Thesaurus of Monument, INSCRIPTION - also engaged in ontology development.

52. Please provide details of any named entity control e.g. domain category names, institutional names, chemical names etc. applied or enforced by the Data Centre

Responses indicated that either the participants did not have this information at hand, or did not see the difference with Q51.

53. *If encoding schemes are used for metadata values e.g. dates, molecules, formulae, keywords etc., please give details*

Only one response indicating the use of Dublin Core

54. *Are identifiers used for metadata elements e.g. URIs, handles, DOIs, internal identifiers etc.?*

Responses varied between URI, DOI, handles and internal identifiers

55. *Does the Data Centre provide tools or any other aids in the creation of metadata?*

There were two "yes" replies and two "no" replies; one indicated help in the form of documentation and templates.

5.5.6 Standards

Standards such as those for information security (ISO 17799), records management (ISO 15489), quality management (ISO 9000) and the OAIS (ISO 14721:2003) tend to be used for guidance rather than strict conformance by the data centres. However, some, such as the UKDA, do have to meet certain standards as a matter of legal obligation.

56. *Does the Data Centre make use of any information security standards, such as ISO 17799?*

Out of four replies, two asserted "yes" and two "no" (mainly used as a guideline).

57. *Does the Data Centre make use of any records management standards such as ISO 15489?*

Three out of four replies were "yes"; one indicated that ISO 15489 is used as a guideline

58. *Does the Data Centre make use of any quality management standards such as 1SO 9000?*

Three out of four replies indicated "no"; BADC uses it as a guideline

59. *Does the Data Centre make use of the Reference Model for an Open Archival Information System (OAIS, ISO 14721:2003)?*

Three out of four replies were affirmative: UKDA has further information at http://www.data-archive.ac.uk/news/publications/oaismets.pdf; for BADC, OAIS is the organisational framework.

60. Does the Data Centre make use of audit and risk management instruments such as the TRAC (Trustworthy Repositories Audit and Certification Checklist) or DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)?

Two data centres are starting to make use of these instruments

61. Are there any other standards critical to the operation of the Data Centre?

One affirmative response from UKDA:

- Again, best to see the Preservation Policy, for details. These have been updated for the 2008 version. They fall into categories: storage devices, security, fire and safety, information security and information management. (That's the ISO/BS covered). There is also a raft of legislative requirements which UKDA faces as the UKDA is a registered Place of Deposit for Public Records

62. Are there any domain standards that the Data Centre uses?

DDI (Data Documentation Initiative) Standard is important to the UKDA.

5.5 Open Issues and Future Development

Several of the data centres are involved in formulating and influencing the development of best practices in the areas of metadata, standards and file formats within their own communities. There is general agreement that "open access" is good, but qualified with a requirement for greater awareness of data protection legislation.

Amongst current pertinent issues for the data centres, there is a need for: standards and tools for the transfer and exchange of data; increasing storage reliability; management of large and complex datasets; archiving of open software where it is integral to the deposited resource; and interoperability with other domain and cross domain datasets.

63. How does the Data Centre follow domain best practices in the areas of metadata, standards, file formats etc.?

- technology watch and professional standards
- Sets file format standards and drives community towards them. These file formats contain structured metadata
- The ADS are participant members of the Forum for Information Standards in Heritage (FISH), setting and deploying best practice

64. Do you see any aspects covered in the survey changing in the near future (2-3 years)? If so, how?

- UKDA is moving from a policy which has been predicated on usability to one which has a stronger emphasis on reliability, integrity and authenticity
- greater formalisation and 'compliance'
- more controlled vocabs

- Certain file formats require further research in order to comply with OAIS practices such as AV and VR formats. Very large data sets and their management are a current subject of research at the ADS

65. What specific comments or views do you have in relation to open access?

- Good, when it's appropriate. Many datasets may have restrictions which prevent this. Both researchers and "open data" promoters need to have greater understanding of research ethics AND data protection legislation
- This is do-able, but harder than first supposed; Generating the deposit-habit is the key
- Open access is something to be striven for, but it will not be totally open for data. For example, programme data is only used by the programme participants to start with, i.e. embargoed
- ADS requires open access, subject to terms and conditions, for all its deposits

66. What specific comments or views do you have in relation to preservation?

EDINA
- Important

BADC
- NERC see this as a good thing, but to get the community totally on board data citation is needed. This means they get academic rewards for their data.

67. Please list 3 issues which are currently pertinent to the Data Centre?

UKDA
- 1) Belief (by others) that institutional repositories manage and preserve complex datasets with special rights issues 2) Open data initiatives 3) Standards and tools for the transfer and exchange of data

EDINA
- 1. Whether to offer a (general) preservation service to complement the continuing access services 2. How to invest in appropriate process 3. What partnerships to forge

BADC
- Parameter names from CF list. Increasing storage reliability. Metadata population

ADS
- Very large data sets e.g. LIDAR, Bathemetry - Archiving of open s/w where it is integral to the deposited resource. - Interoperability with other domain and cross domain datasets via SOA

68. Are there any other comments you would like to make?

EDINA: request for a copy of the questionnaire and own answers.

6. Discussion

The motivation for this study was not merely to record the characteristics of UK data centres but to produce a synthesis of those characteristics that would expose both similarities and divergences. Where researchers are producing ever-increasing volumes of research data in digital form, and with research funders casting an ever more critical eye upon options for securing a return on their investment through data sharing and re-use, opportunities to exploit synergies and standard platforms, as well as close attention to the data curation lifecycle, have become paramount.

In this discussion we have identified those areas where data centre policy and practice differ, as well as instances where they exhibit complementarity of execution and in the challenges still to be met.

6.1 Roles and Responsibilities

The six data centres that responded to our survey are well-established organisations with missions underwritten and funded by national agencies, which defines them as legitimate strategic instruments of research data management in the UK. Whilst there are differences in breadth of service profile, with perhaps the most expansive being the UKDA's claim to support researchers in all sectors and disciplines, none of them as data providers are strictly limited to meeting the needs of the disciplines they were created to serve. That said, responsibilities for addressing the needs of those disciplines are made clear in the policy statements of the funding agencies with responsibility for specific sectors of UK research and scholarship, including those for the MRC and its emergent Data Support Service.

The role of EDINA is somewhat unique amongst this group, in that its service is exclusive to institutions of further and higher education and research, which are normally required to subscribe and to be authenticated at point of use.

Policy and principal funding agent	Arts and Humanities Research Council (AHRC)	Joint Information Systems Committee (JISC)	Medical Research Council (MRC)	Natural Environment Research Council (NERC)	Science and Technologies Facilities Council (STFC)	Economic and Social Research Council (ESRC)
Data Centre	**Archaeology Data Service**	**EDINA**	**MRC Data Support Service**	**British Atmospheric Data Centre**	**RAL Data Cluster**	**UK Data Archive**
Council/Committee research data policy	RCUK position statement on open access	Strategic plan: to promote the use of IT in support of research and learning	MRC policy on data sharing and preservation, 2007	NERC Data Policy Handbook, December 2002	In fEC research grants handbook, section 8	ESRC data policy, April 2000
Council/Committee engagement with data sharing/preservation	No data policy in place but recommends grant holders offer datasets to the ADS	National service portfolio includes statements for each JISC-funded service	Data management plan a prerequisite to funding. Data support service approved	Expectation that grant holders will deposit with NERC data repository	Requires identification of exploitable intellectual property	Deposit of data and appropriate documentation required of grant holders

Data Centre mission	Web page: specific to archaeological research and on behalf of eight funding units within the Arts & Humanities Data Service	Web page: a contracted level of service to enhance productivity in UK research, learning and teaching	ITT: support service designed to improve effectiveness of world-class medical research	Web page: to enable access and ensure long-term integrity of NERC-funded research	Web page: as part of National Grid Service aims to provide coherent electronic access to STFC-funded output	Web page: emphasises position as centre of expertise and curator acting on behalf of the ESRC, JISC, and other HE organisations

In their mission statements, all express an ambition to enhance the productivity of research, and to support good practice through the application of standards in the management of digital resources. This identifies them as more than a community of data archives, since not only are they committed to the preservation of data but each exercises discrimination in the deposit of data, adds value through the provision of additional services (including guidance and training), and applies curation techniques to ensure the long-term usability and audit of the data they hold. This more complex suite of operations is demanding in terms of capital investment and particularly for the provision of expert staff (see section 6.3). It is a characteristic that currently illustrates a difference between the data centres and most institutional data repositories, which generally receive very limited funding, and clearly represents a key challenge for policy makers who are examining future options for the curation of digital research output [17].

Whilst the UK data centres receive funding and a statement of responsibilities directly from national agencies, institutional repositories have yet to gain a comparable status; for the moment they lack a national mandate and the development or recurrent funding necessary to the creation of a sustained and trusted resource. The survey respondents were in agreement that one of the areas that need urgent discussion and clarification is that of the roles, interactions and relationships between newly created institutional repositories and established data centres.

6.2 Best Practice

Whilst at a generic level best practices can be observed to involve global methodologies and an adherence to standards, open formats and software, it is inevitable that to be effective the data centres need to take into account the specific practices of the disciplines and communities that serve. Consequently, each of the participating data centres has developed (and evolved) best practices in the management of the particular and sometimes very specialised data that they manage, based on many years of experience and expertise. Furthermore, any guidance and advice that is provided by a particular data centre is tailored to the specific community within which it operates and serves. In addition, many of the data centres view their role as being over and above that of merely managing and making available data, to one of adding value, again within the context of the community that they serve. As a result, unsurprisingly, cross-domain interoperability and the provision of services for inter-disciplinary use of data have not featured very highly on the data centres' agendas.

However, there is scope for sharing and collaboration with regard to best practice in the areas of

- managing organisational structures
- the conduct of relationships with data depositors and users
- the relevance of services provided to user communities
- implementation of standards, metadata and open systems
- provision for interoperability
- the value of a lifecycle approach to data curation
- the pursuit and delivery of added value.

6.3 Skills & Training

During the survey, data centre managers described their concern to deliver and sustain a pool of appropriately qualified staff, as well as to equip data producers and users with the tools necessary for effective data deposit, access and reuse.

At the UKDA, for example, the data centre team is comprised of individuals from a range of discipline backgrounds relevant to the social sciences domain, complemented by expertise and qualifications in information, data or systems management; but typically, no structured career path was apparent.

The limited supply of an appropriately qualified workforce, which is generally acknowledged to require both discipline knowledge and data handling skills, has been identified as a significant risk. In the context of an escalating volume of digital research output, if 'there is a dearth of skilled practitioners' [18] there will follow an inevitable inadequacy in the maintenance and enhancement of scientific knowledge. At a meeting to plan the forward programme for a data centres Forum, a consistent theme raised by senior staff from both the established UK data centres and from institutional data repositories was the need to address 'such critical organisational issues as the provision of an appropriate skills base/career path for informaticians' [19].

With respect to the skilling of data users, formal programmes already feature as a component of some data centre services. As an illustration, training in the management of data is provided to UKDA clients through the provision of data creation workshops, based on the rationale that it is important to 'get to' researchers early in the data collection process.

The NERC and ESRC are currently considering how this approach should be developed for a broader community of researchers and, during March 2008, as part of the ESRC Festival of Social Science, the Economic and Social Data Service (ESDS) is holding two workshops on managing and sharing research data. Guidance will be based on advice and support provided to ESRC award holders as part of their contractual obligations to share data, and to the cross Research Council Rural Economy and Land Use (RELU) Programme which requires formalised data management plans. These hands-on workshops will cover key issues on data management and sharing as applied to socio-economic research involving people as participants (interviews, questionnaires, focus groups, observation etc.). The workshops focus on dealing with confidential research information and personal data; developing consent agreements for obtaining informed consent from participants;

appropriate anonymisation techniques to enable the use and sharing of research data. These workshops will provide both theoretical information and practical exercises, and are designed to prepare researchers for a practical approach to managing their research data needs.

7. Conclusions

From our survey it can be deduced that the data centre landscape reflects that of their funding organisations, in the majority of cases the UK Research Councils. Each data centre has emerged from specific research disciplines, and has been expected to cater for the requirements of those particular communities. Consequently, the policies and strategies adopted by each data centre are heavily influenced by those of its corresponding funding council and the community that it purports to serve.

Notwithstanding the formal relationships with their funding agencies, the data centres are nonetheless independent entities, and at an operational level they have developed their own policies, procedures and processes governing data deposit and access. It is at this level that the provision of effective and coherent services, particularly in terms of enabling access, may be at risk. Evolving data centre practices will inevitably prove idiosyncratic given the complexity and heterogeneity of the data being collected. The achievement of interoperability across data sets and disciplines will depend largely on the adoption of standards as well as open file formats and software.

The DCC's recent creation of a national forum for data centres and repositories [20] is an initiative designed to facilitate greater coherence, at an operational level, between these individual organisations, through the exchange of experience and best practice in the processing and management of data of importance to science, technology and society.

It is interesting to speculate whether the current data centre environment is robust enough to meet the challenges and terms which will be identified in the HEFCE-funded data services feasibility study [17], which has embarked on an examination of the management of multi-discipline research data by institutional repositories.

Acknowledgements

The DCC is funded by the Joint Information Systems Committee (JISC), an independent advisory body that works with further and higher education establishments, and the e-Science core programme.

We are indebted to all the participants in this study who gave generously of their valuable time in undertaking both the online questionnaire and in face-to-face meetings.

References

[1] The Digital Curation Centre, http://www.dcc.ac.uk/

[2] SurveyMonkey, http://www.SurveyMonkey.com/

[3] Safe Harbor, http://www.export.gov/safeharbor/sh_overview.html

[4] The UK Data Archive (UKDA), http://www.data-archive.ac.uk/

[5] Economic and Social Data Service (ESDS), http://www.esds.ac.uk/

[6] EDINA, http://edina.ac.uk/

[7] Burnhill, P., *What EDINA does: A Community Report,* Sept. 2007 http://edina.ac.uk/about/annrep/communityreport_sept07.pdf

[8] The Archaeology Data Service (ADS), http://ads.ahds.ac.uk/

[9] The Arts & Humanities Data Service (AHDS), http://ahds.ac.uk/

[10] The Medical Research Council (MRC), http://www.mrc.ac.uk/

[11] MRC Terms and Conditions -5th Oct 2006, http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC001898

[12] Natural Environmental Research Council (NERC), http://www.nerc.ac.uk/

[13] NERC Data Policy, http://www.nerc.ac.uk/research/sites/data/policy.asp

[14] NCAS British Atmospheric Data Centre (BADC), http://badc.nerc.ac.uk/

[15] The Science and Technology Facilities Council (STFC) RAL (Rutherford Appleton Laboratory) Data Cluster, http://www.ngs.ac.uk/sites/ral/

[16] *Research Funders' Policies for the management of information outputs*, A report commissioned by the Research Information Network, Jan 2007, http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20%20Final%20Report.pdf

[17] *A shared research data service for the UK: Feasibility Study*, Invitation To Tender, 15th Nov 2007, RUGIT/CURL: Shared Services Working Group, http://www.curl.ac.uk/ukrds/links.htm

[18] Lyon L., *Dealing With Data: Roles, Rights, Responsibilities and Relationships*, JISC Consultancy Report, June 2007, section 6.8, p54 – Training and Skills http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

[19] Thorley M.R., Minute of the Data Forum Round Table Planning Meeting, London, 19th November 2007

[20] Pryor G. H. S., Research Data Management Forum Terms of Reference, December 2007