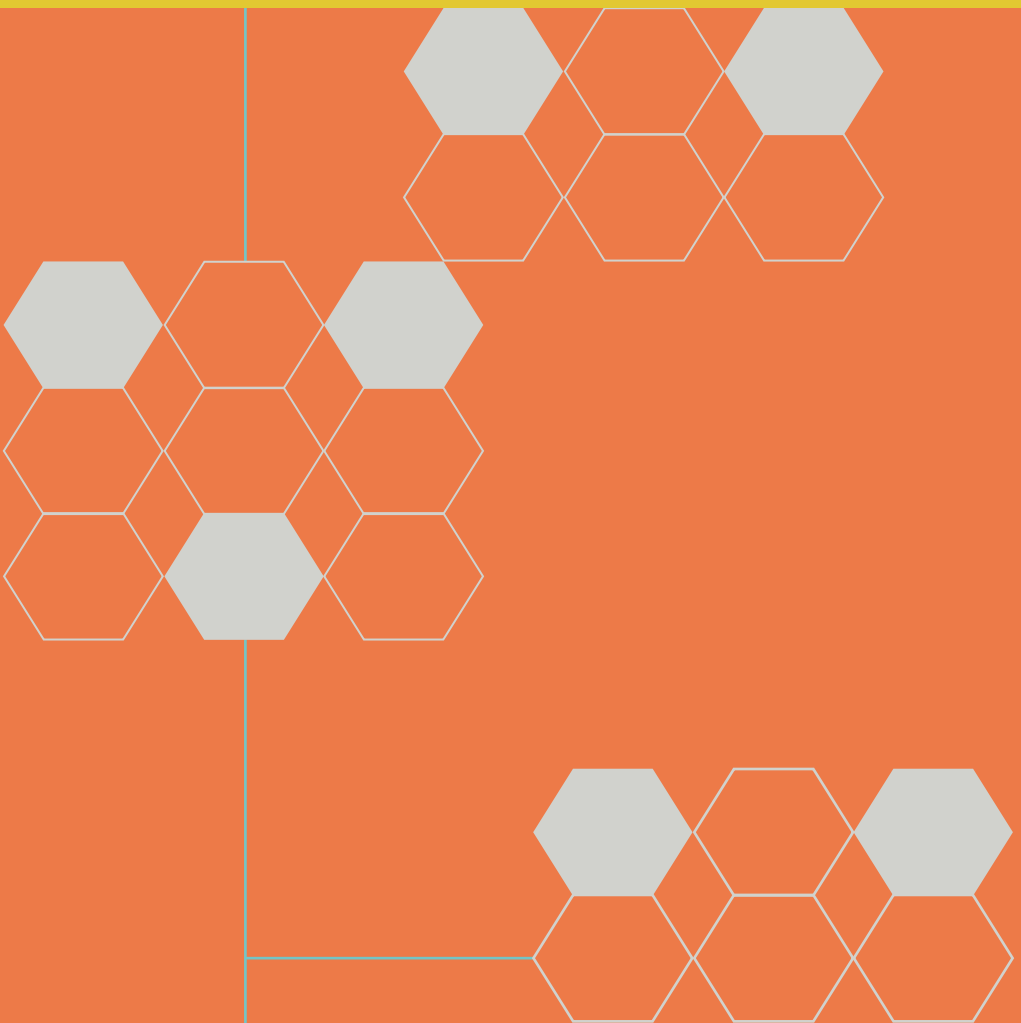
research information network

JISC

Data centres: their use, value and impact

A Research Information Network report

September 2011



Acknowledgements

The research upon which this report is based was undertaken by a team from Technopolis Group. The Research Information Network and JISC would like to thank Paul Simmonds, James Stroyan, Neil Brown and Lark Parker-Rhodes for their work.

It is available to download at www.rin.ac.uk/data-centres

Keep up to date with our work, follow us on Twitter @research_inform

Table of contents

List of abbreviations

The following tables show abbreviations used in this report to refer to UK data centres and research funders included within the study.

Table A: UK data centres

Abbreviation	Data centre
ADS	Archaeology Data Service
BADC	British Atmospheric Data Centre
CDS	Chemical Database Service
EBI	European Bioinformatics Institute
ESDS	Economic and Social Data Service
NCDR	National Cancer Data Repository
NGDC	National Geoscience Data Centre
UKSSDC	UK Solar System Data Centre

Table B: UK research funders

Abbreviation	Research funder
AHRC	Arts and Humanities Research Council
BBSRC	Biotechnology and Biological Sciences Research Council
CR-UK	Cancer Research UK
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
MRC	Medical Research Council
NERC	Natural Environment Research Council
STFC	Science and Technology Facilities Council
Wellcome	Wellcome Trust

Executive summary

Background

In recent years, the value of data as a primary research output has begun to be increasingly recognised. New technology has made it possible to create, store and reuse datasets, either for new analysis or for combination with other data in order to answer different questions. In the UK, academic researchers, funders and institutions have responded to these possibilities by supporting a number of data centres – organisations with responsibility for supplying research data to the academic community, and in some cases for collecting, storing and curating such data as well.

Although researchers may choose to store and share their data in a number of ways, data centres often appear to offer the best way of ensuring that data are preserved and presented in a high-quality way, and made available to the largest number of people. As dedicated, central locations for datasets, data centres are able to ensure deposited collections are highly visible; in many cases, they also help ensure that data are ready for reuse by helping researchers prepare them for deposit.

This study sought to understand usage of UK data centres among researchers, and to examine the impact of such use upon their work. We undertook a series of initial interviews with research funders to understand the role and importance of data and data centres within various academic fields, followed by a survey of the users of five data centres. Finally, through the interviews and surveys, a set of case studies was identified where the data centre had benefited a researcher's work, and in some cases that work had gone on to have an impact in wider society.

Findings

Overall, usage of data centres is high, with most centres supporting thousands of researchers and millions of downloads each year. Academics are the most significant data centre users, and also the most important final audience for research based on data centre assets. However, each centre in the study had a distinctive group of users and audiences, which reflected its holdings and the funding streams within its research field. Data from every centre is used in several ways: for original research, for combination with other data, and for reference. In some centres, it is also used as a basis for further data collection. Users of every data centre overwhelmingly rated the data as important to their research.

Most researchers felt that data centres had improved the culture of data sharing within their research field, either to a small or (in most cases) a large extent. However, not all researchers submitted their own, new, data to data centres, with clear differences between data centres observable in this area. More positively, a majority of researchers cite the centre, the data set or the original creator when reusing content in their own work.

The most widely-agreed benefit of data centres is research efficiency. Data centres make research quicker, easier and cheaper, and ensure that work is not repeated unnecessarily. Research quality was another important benefit, although not rated quite as highly as efficiency. There was mixed evidence about the importance of data centres in stimulating new research questions, with noticeable differences between the data centres. Benefits to researcher training also varied significantly between centres.

The study also showed how data centres enable the benefits mentioned above to be realised. Data centres provide services and support which are highly valued by researchers, including: user support, both online and in person; access to otherwise-unavailable datasets via reciprocal sharing arrangements; and curation, preservation and long-term access for datasets, both for their own research and for datasets created by others. These features are important precursors of the research efficiency, quality and novelty outcomes, and for the most part are a result of the data centre's status as a central and sizeable hub within its field.

It proved more difficult to identify areas where research based upon data centre resources had gone on to have significant social, economic or environmental impacts. This is in part because it is notoriously difficult for researchers to keep track of how their work is used following publication. Many researchers suggested that there was potential for their work to be used in the public, private or third sectors, but that they were not aware of whether this had actually happened. Nonetheless, a few cases at the end of this report illustrate how research based on data centre resources has had a positive impact upon wider society and the economy through the development of new tools and methodologies, new policies and regulatory controls, and new products or services.

Conclusions and strategic implications

The study concludes that data centres play an important part in the modern research infrastructure, in a number of academic fields. They offer many benefits to researchers and their work, and in some cases this work itself offers benefits to wider society and the economy. Researchers believe that many of these benefits emerge because data centres are large, centralised, and offer a range of services beyond the provision of access to data.

The findings suggest that:

1. Data centres are a success story for their users, and funders and policy-makers should continue to support and promote existing national data centres.
2. Data centres are important both for reference purposes, and for novel research. Both these uses should be maintained and encouraged.
3. Data centre staff manipulate, interpret and support use of data sets, and this is highly valued by researchers. The role of data centre staff should be supported, and perhaps investigated further to support advocacy for data centre services.
4. Data centres should continue to collect information about users and usage for planning and advocacy purposes.
5. Although deposit levels are promising, researchers need more encouragement to deposit data. National and international initiatives in this area should be monitored and factored into any consideration of how to improve deposit rates.
6. If data centres are to support the grand challenges of modern research, they need to do more to facilitate interdisciplinary working. Improving facilities for data discovery across data centres may help.
7. The national data centres are just one part of a broader landscape for data curation and storage. Further work needs to be done to investigate how they can work most effectively with local, national and international services.

1. Introduction

In recent years, increasing attention has been paid to data as a primary research output. The advent of powerful computers and monitoring equipment has offered new opportunities to generate and store large amounts of data. Such data then becomes ripe for reuse, either by the research team who first collected it, or by colleagues in – or even beyond – their field.

Such reuse may take two forms. The first can be loosely defined as 'reference'. This includes instances where researchers use existing results as vital context for the data that they themselves have collected or generated, either to aid analysis or to assure the quality of their own work. In other cases, researchers may reanalyse existing data to test the replicability of certain results – a keystone of quality assurance in research. In the second form of data reuse, researchers may perform new analyses on existing data to undertake their own, original research. Data can be aggregated to create new, enormous datasets, which can then be analysed for novel insights using computational techniques via a federated and distributed network – the process known as 'e-Science'(Hey and Trefethen, 2003). Alternatively, new research techniques can be brought to bear upon existing datasets, including the products of so-called 'small science', generating new findings without going through the (often expensive) process of collecting new data (National Science Board, 2005). Both the new research and reference uses of data are important.

For such benefits to be realised, however, data must first be discoverable and reusable. This is a challenge which has been widely recognised by researchers and policymakers (Lievesley & Jones, 1998; Research Information Network, 2008a). Many researchers do not retain data from observations or experiments in a format which could readily be shared with other researchers. Of those who do, relatively few have the time or funding to invest in creating good quality metadata, which is essential for reuse by other researchers, who need to know how the data were collected and processed in order to produce valid conclusions.

Even once metadata are in place, there is no guarantee that the datasets can be found by other researchers, unless they know what they are looking for and are able to approach the original researcher directly. Many researchers also feel concerned that sharing their data before they have fully exploited it could lead to them being 'scooped' by colleagues (RIN, 2008b).

Research councils and other funders recognise both the value of data reuse, and the problems in attempting to make such reuse possible. Figure 1 presents an overview of the data policies of the research councils and two other major funders of data-intensive research (Cancer Research UK and the Wellcome Trust).

As Figure 1 shows, the major UK research funders recognise the importance of data as a research output, with eight of the nine already having a data policy, and the ninth planning to publish one shortly. There is a growing recognition that important data must be preserved for future reuse, with every data policy requiring researchers to deposit their data either centrally or locally. This must be done within a specified time, although for most research funders the details are negotiated project-by-project. Consideration of data sharing and reuse is encouraged, with seven of the nine funders requiring a data management plan from bidders. Furthermore, six research funders are prepared to fund data sharing activities within a project bid, although only three will accept bids for developing tools to aid data sharing. Five funders expect secondary users of data to acknowledge their sources, showing that they recognise the importance of attribution and academic credit, lack of which prevents many researchers from sharing their data.

Figure 1: Data policies of research funders

	AHRC	BBSRC	EPSRC	ESRC	MRC	NERC	STFC	CR-UK	Wellcome
Has the organisation published a data policy?	Y (funding guide)	Y	Y (research guide)	Y	Y	Y	-	Y	Y
When was this policy document published?	2009 (funding guide)	2007	2006 (research guide)	2000	2007	2002	n/a	2009	2007
Are there plans to revise this policy?	-	-	-	Y (published 2010)	Y (ongoing)	-	Y (future)	-	Y (reviewing)
Based on the current data sharing policy…									
Does the organisation support / fund dedicated data centres (how many)?	Y (1)	Y (1)	Y (1)	Y (1)	Y (2)	Y (6+)	Y (1)	Y (2)	Y (2)
Are researchers required to offer to deposit data in a central dedicated data centre (where available)?	Y	Y	-	Y	Y	Y	-	-	Y
Are researchers required to deposit data locally and make it available to others (where data centres are not applicable / available)?	Y	Y	Y	-	Y	-	-	Y	Y
Are researchers required to include a 'data review' (assessing whether data is already available) when applying for funding?	-	-	-	Y	-	-	-	-	-
Are researchers required to include a 'data sharing plan' when applying for funding?	Y	Y	-	Y	Y	Y	-	Y	Y
Does the policy state that research applications can include requests for funding for data sharing activities?	Y	Y	-	Y	Y	-	Y	-	Y
Does the policy state that proposed data sharing activities will be considered within final assessments / funding decisions for projects?	-	Y	-	Y	-	-	-	Y	Y
Are maximum time limits specified for depositing data post-project?	Y (3 months)	Y (~3 years)	Y (variable)	Y (3 months)	Y (variable)	Y (variable)	-	Y (on publ.)	Y (on publ.)
Are minimum time limits specified for retaining data post-project?	Y (3 years)	Y (10 years)	Y (variable)	-	-	-	-	Y (5 years)	-
Does the policy state that secondary users of data are expected to acknowledge their sources?	-	Y	-	Y	Y	-	-	Y	Y
Does the policy state that applications for funding can be made for the development of tools to aid data sharing?	-	Y	-	Y	-	-	-	-	Y
Does the policy state that the organisation will consider purchasing existing data from other sources, or commissioning new data?	-	-	-	Y	-	Y	-	-	-

Source: Published policies and Technopolis interviews with research councils November 2009

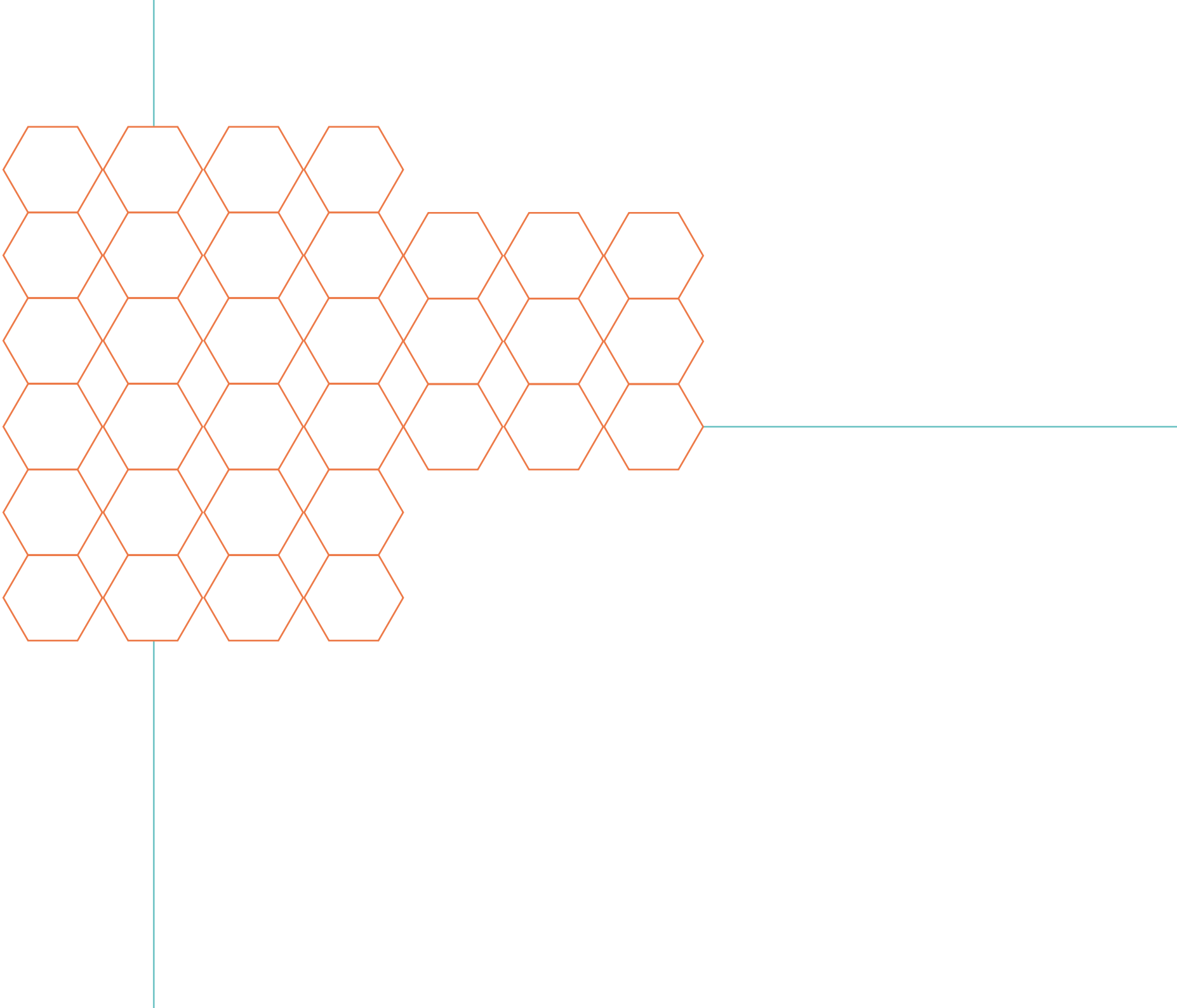
Each of the major UK research funders supports at least one national research data centre. Such centres represent an important – although not the only – way to ensure effective data sharing and reuse. This study does not set out to explore the relative merits of various ways of storing data, and focuses on the impact and benefits of national data centres. However, it is important to recognise that researchers are using other methods to store and share their data. Other solutions include self-archiving and subsequent promotion via local computer systems or a personal website, or collation at the institutional level via institutional repositories. Data centres, however, possess two important characteristics which help researchers to find and access quality data within their field.

The first characteristic is their function as a central, national service where researchers can both deposit data they have created (in the hope that it will be useful to other researchers) and also find data they can reuse within their own work. For many data centres, however, comprehensive coverage remains an aspiration rather than an achievement: Beagrie et al. found that only 18% of researchers deposit their work in a data centre, although a more encouraging 43% use centres to access data (2009). Even where centres have accumulated loose aggregations of data, many datasets are subject to restrictions on access or reuse, for reasons of commercial sensitivity or confidentiality, or because the original researchers want to get as much use out of it for publications before making it available to colleagues (RIN, 2008b). Nonetheless, data centres represent an important step in the right direction towards data sharing and reuse, as they can offer guidance and support to researchers operating within these limitations to ensure that data is at least preserved and then made available when that is considered permissible.

The second important role that data centres can play is helping researchers to prepare their data for wider presentation and reuse – in particular, the creation of appropriate metadata.

Researchers working with a dataset habitually develop their own private code and shortcuts, which can work well while a project is underway. However, once data is offered for reuse by others, idiosyncratic ways of describing data become a barrier to, rather than an enabler of, effective research. Most researchers are not used to preparing their data for reuse by others, and few have the time to do so, or to learn to do so (Research Information Network, 2008b). This was one of the important barriers identified by the Australian National Data Service Technical Working Group when considering the possibility of fully open research data (2007). Data centres can make this process of reconfiguration easier for researchers by offering advice, guidance, standards and structures to ensure that data is ready for reuse.

Of course, in order to do this, data centres need to understand what their users (both depositors and those who want access) are doing, particularly in the context of a rapidly changing research data environment. Furthermore, in straitened economic times it is important that data centres should be able to demonstrate that the funding they receive from Research Councils and other public sources is put to good use (Research Information Network, 2008a). Recently, the notion of 'impact' in academic research has become an important criterion for perceived success (Grant et al., 2009). The precise definition and nature of 'impact' of academic work has spawned a large and complicated literature of its own. This study, therefore, set out to understand the benefits of data centres to research activities, and then to trace (wherever possible) any impacts that such activities had upon wider society, businesses and the economy. This could be through their translation into new policies, or through the development of new technologies, products or services. The impact of the data centre is thus twofold: first, upon the researcher's work and then second, and via that work, upon wider society. All this is set within a context of identifying who uses data centres, and how, to understand what might be done to improve their services to researchers.



2. Data centres included within this study

Eight centres were included in this study. They were selected to represent a cross-section of the types of centre which operate in the UK, supporting a variety of disciplines and offering a range of data types and services. Particularly important is the inclusion of centres that operate as portals to a collection of externally-hosted databases as well as centres which curate and maintain their own collections: some centres in the study provide both kinds of service.

The centres also vary considerably in size, both of their holdings and of their budgets. The holdings are examined in greater detail later in this report. Figure 2 presents the annual budgets of each data centre for the financial year 2008/09, ordered from highest to lowest.

A final, important, distinction between the data centres is the terms and conditions under which they offer access to their holdings. In many cases, this is not something the data centre can control: where content is held by an external supplier, they are bound by their licence agreements. Thus the CDS, which acts as a gateway to a range of externally-held datasets, states explicitly that no commercial use may be made of its services. Where the centres hold original data, they have more control and can, if they choose, offer wider access. Some, such as the BADC, view their data as a public good, and allow reuse by anybody for any purpose. Others, such as the ADS, allow commercial use only where the final output will be placed in the public domain.

Figure 2: Data centre budgets 2008/09

Centre	Full name	Annual budget
EBI	European Bioinformatics Institute	c.£4m
ESDS	Economic and Social Data Service	c.£2m
BADC	British Atmospheric Data Centre	c.£1m
ADS	Archaeology Data Service	c.£640k
NGDC	National Geoscience Data Centre	c.£350k
CDS	Chemical Database Service	c.£250k
NCDR	National Cancer Data Repository	c.£200k
UKSSDC	UK Solar System Data Centre	c.£150k

Archaeology Data Service (ADS)

The ADS was established in 1997 as a central archive for archaeology data, and is based at the University of York. It aims to collect, describe, catalogue, preserve and provide user support for digital resources that are created as a product of archaeological research. It is also responsible for promoting standards and guidelines for best practice in the creation, description, preservation and use of archaeological data and for collaborating with existing bodies in order to promote greater use of their services.

The ADS holds 300 collections, comprising a total of 1 million records. Data includes text reports, databases, images, digitised maps and plans, numerical datasets and reconstruction drawings.

The ADS is funded primarily by the Arts and Humanities Research Council (AHRC), although it generates additional income by supporting projects or providing services to other organisations. Its core user groups include academics, teachers, professional archaeologists and the general public.

British Atmospheric Data Centre (BADC)

The BADC was originally established in 1984 as the Geophysical Data Facility (GDF), and is based at the STFC Rutherford Appleton Laboratories. It aims to assist UK researchers in locating, accessing and interpreting atmospheric data, and to ensure the long-term integrity of atmospheric data produced by NERC projects.

The BADC holds data produced by NERC-funded projects and programmes alongside other data sets which are required by the UK atmospheric research community. This last category includes ground-based observations, numerical model outputs and satellite data. The BADC provides access to 228 data sets, some of which contain ongoing observations and therefore grow over time, and some of which are project outputs and therefore static.

Core funding for the BADC comes from NERC, although the centre has secured additional funding from development contracts with third parties, such as the Department for Environment, Food and Rural Affairs (DEFRA) Climate Impacts Project. Its core user group is academics.

Chemical Database Service (CDS)

Unlike many of the other data centres in this study, the CDS does not curate and archive data itself. Rather, it provides enhanced access to existing chemical databases offered by third party providers. The service was established following a 1992 EPSRC committee's recommendation that there would be a number of advantages to a dedicated central resource for chemical database provision. The CDS's primary aim, therefore, is to ensure that information from global chemical research is accessible to UK researchers.

The CDS provides access to ten core databases in four areas: structures, spectroscopy, thermophyiscal data and organic chemistry. These databases exist independently of the service, but the majority are not currently made available at institutional, regional or national level other than via the CDS.

The CDS is financed by the EPSRC, and usage is restricted to UK 'academics', defined as anybody employed by, or working at, a UK university.

European Bioinformatics Institute (EBI)

The EBI is one of several satellite research organisations that form part of the European Molecular Biology Laboratory (EBML). It was established in 1994, and is based at the Wellcome Trust Genome Campus in Cambridge. It maintains the world's most comprehensive range of molecular databases and is the European node for globally coordinated efforts to collect and disseminate biological data.

The EBI provides access to 64 datasets, all of which are available for download from its website. These cover areas including DNA and RNA sequences, genomes, microarray data, protein sequences, macromolecular structures, protein-protein interactions and pathways. It also provides a scientific literature portal.

The majority of the EBI's funding comes from the governments of EMBL's 20 member states. However, other major funders include the European Commission, Wellcome Trust, US National Institue of Health, UK Research Councils, industry partners and UK government. The EBI's main audience is academics, although it is freely available to all users.

Economic and Social Data Service (ESDS)

The ESDS is the primary data provision service within the UK Data Archive (UKDA), and is based upon a collaboration between the UKDA and the Institute for Social and Economic Research (ISER) at the University of Essex, and the Cathie Marsh Centre for Census and Survey Research (CCSR) and Mimas at the University of Manchester. It aims to acquire, curate and provide access to the UK's largest collection of social and economic data.

The ESDS main catalogue has approximately 5000 'collections', each of which may contain a number of datasets ranging from one to several thousand. It holds data from, among others, large-scale government surveys, multi-national aggregate databanks and survey data, major UK surveys following individuals over time, and cohort studies.

The ESDS is primarily funded by the Economic and Social Research Council (ESRC), with the Joint Information Systems Committee (JISC) making a significant contribution (£390,000 from a total of around £2 million). Its main audiences are academics, practitioners, policy-makers and the general public.

National Geoscience Data Centre (NGDC)

The NGDC was established in 1983, and is located at the headquarters of the British Geological Survey (BGS). It aims to maintain, link and provide access to data generated by the BGS, in addition to data provided by external organisations.

The NGDC holds over 9 million items, some of which date back over 200 years. Data include earth science datasets, physical collections, records and other information gathered or created by the BGS, and similar information from organisations including oil companies and the Coal Authority.

The NGDC is financed by the BGS, which is itself funded by the Natural Environment Research Council (NERC), and its main audiences are academics, policy-makers, businesses and the general public.

National Cancer Data Repository (NCDR)

The NCDR was established in 2008 as a part of the National Cancer Intelligence Network (NCIN), itself a project within the National Cancer Research Institute (NCRI). The NCDR aims to provide a centralised and co-ordinated approach to the management and reuse of cancer data, and to enable access to that data by research groups.

The NCDR combines the English cancer registries' data into a national dataset, and links this with data from other sources, including the General Practice Research Database. This linkage provides vital information on primary care that is not included in the other available datasets, thereby making possible new kinds of research.

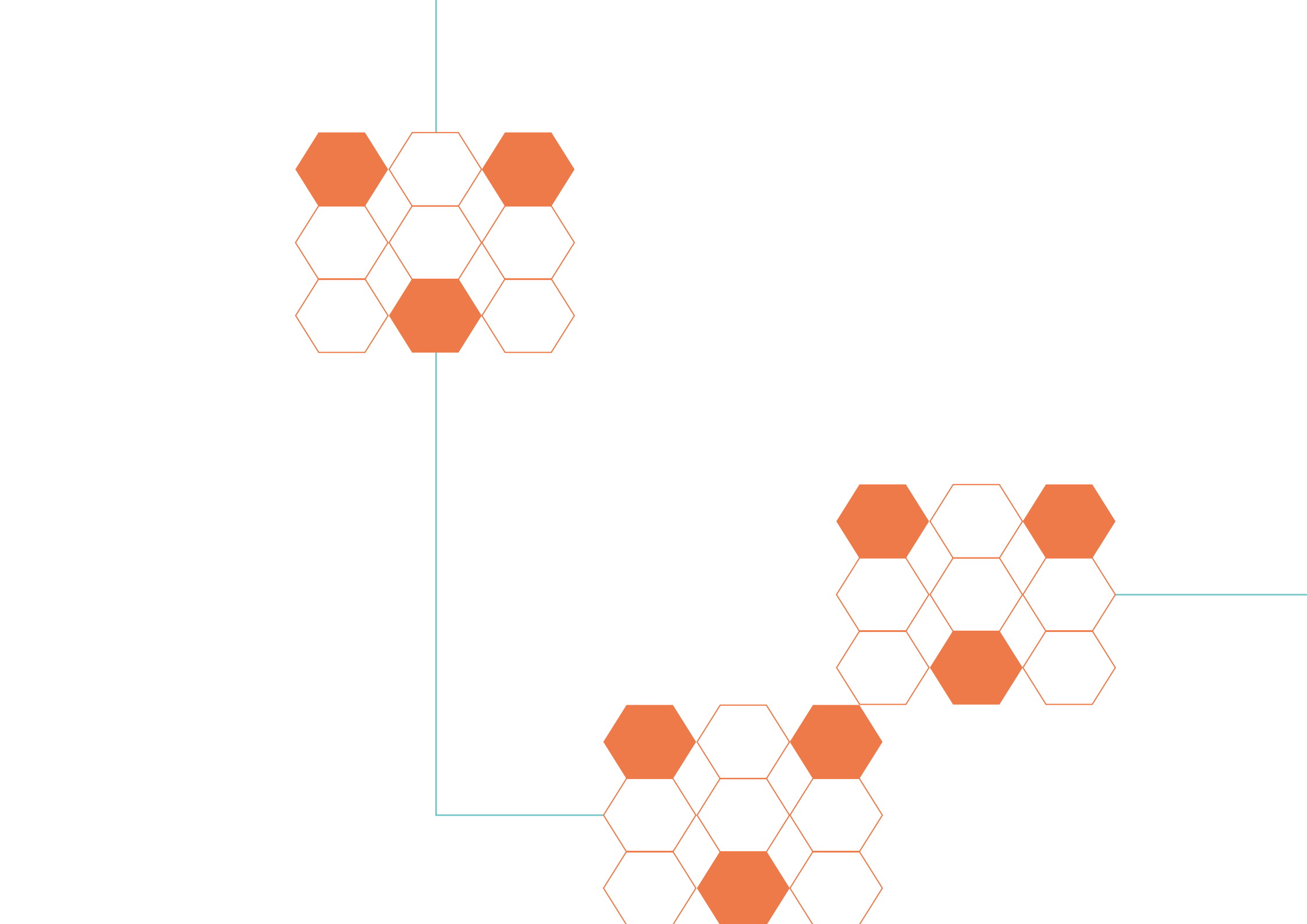
The NCDR is funded primarily through the NCRI, a partnership of 21 funders of cancer research in the UK including Cancer Research UK, Wellcome, the Medical Research Council, the EPSRC and the Department of Health. Its primary audiences are academics, practioners and policy-makers.

UK Solar System Data Centre (UKSSDC)

The UKSSDC has been in operation for over 50 years, albeit on a relatively small scale. It aims to provide a portal to, and well-maintained archives of, international solar system data, for use by the UK solar system community.

The UKSSDC provides access to important international data holdings, as well as archiving around a dozen of its own collections. It has a particular strength in time series data, and provides access to real-time data derived from multiple international sources.

The UKSSDC's main funder is the Science and Technology Facilities Council (STFC), and its main audience is academics, although data is also used by industry, particularly those in the communications sector.



3. Methodology

The study on which this report is based employed a range of methods to gather evidence on how researchers use data centres, and the benefits that such use brings to their work. A series of initial interviews with research funders helped to establish the role and importance of data and data centres within various academic fields. These interviews were followed by a survey of users of the ADS, the BADC, the CDS, the ESDS and the NGDC. Practical difficulties meant that it was not possible to survey users of the EBI, the NCDR and the UKSSDC.

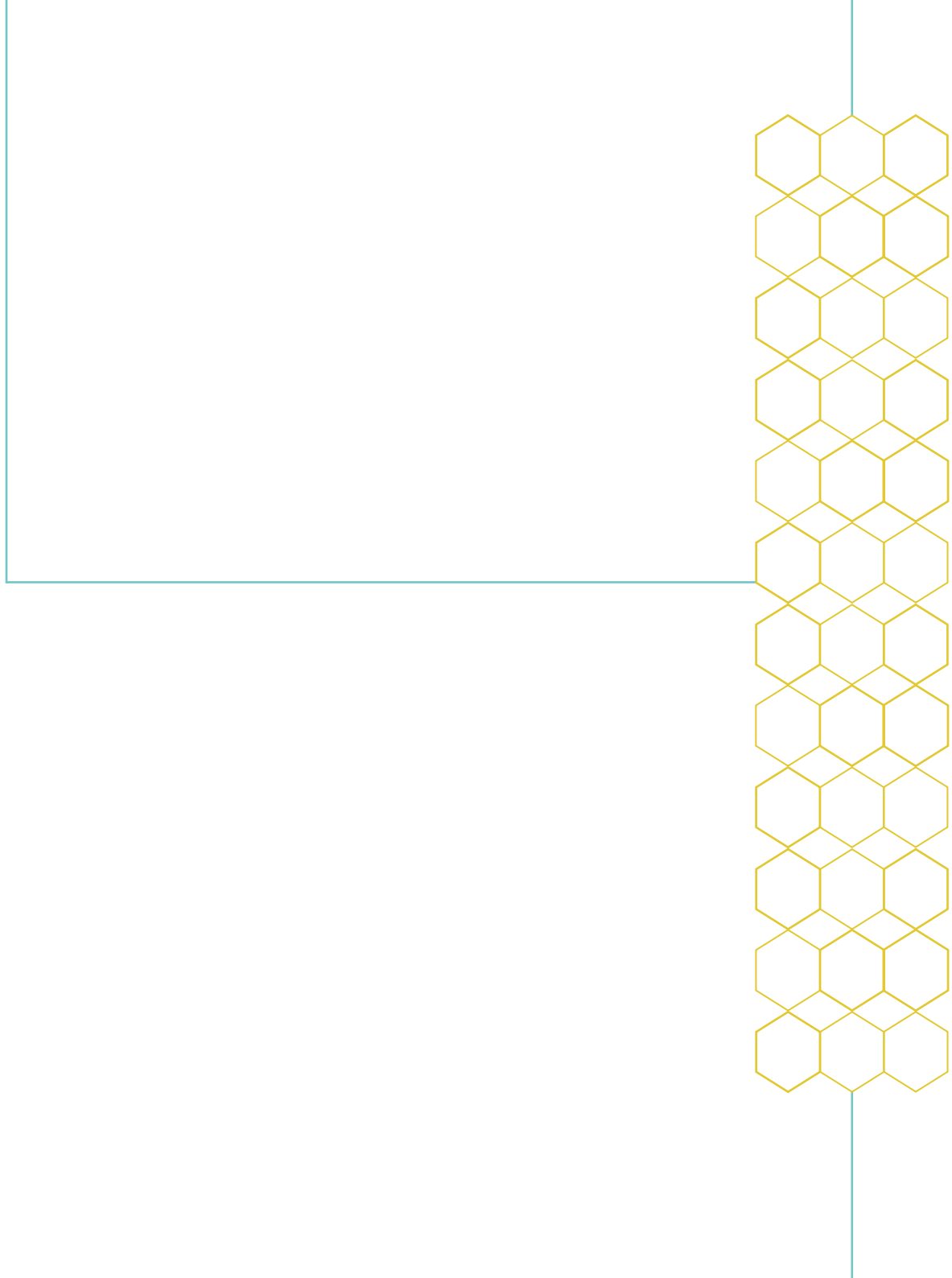
The survey used a range of questions to investigate researcher use of data centres, and their perceptions of the benefits of this use. The different ways in which the surveys were publicised to the various user communities means that the data centres achieved varied levels of responses: for this reason, we have not presented aggregate totals for all five data centres in the report.

Throughout the report, we have used 'heat maps' to represent the relative importance of each response within a question. Figure 3 shows how these heat maps should be interpreted each time they are used.

Figure 3: Interpretation of tables in report

Shade	Key
	80%-100% of respondents agreed with this statement
	60%-79% of respondents agreed with this statement
	40%-59% of respondents agreed with this statement
	20%-39% of respondents agreed with this statement
	0%-19% of respondents agreed with this statement

Finally, through both the interviews with the research funders and the survey, a set of case studies was identified where the data centre had benefited the researcher's work, and in some cases that work had gone on to have an impact within wider society. The researchers involved in these case studies were contacted and interviewed, in order to understand how the work made possible by data centres can bring benefits to society, the economy or business.



4. Users and usage

Users

The centres' own monitoring statistics show encouragingly high levels of content and usage. Most centres support thousands of individual researchers, and process millions of downloads each year, as shown in Figure 4.

Figure 4: Overview of basic statistics on data holdings, users and usage

Data centre	Data	Users	Usage
ADS	300	n/a	12 million pages / year
BADC	130 (2007)	2,000 active	14 million downloads / year
CDS	9	5,000 registered users	5,000 visits a month
EBI	64	340,000 enquirers a month	3 million data requests / day
ESDS	5000 data collections	40,000 registered users	5,200 downloads / month
NCDR	1 linked database	30 projects / organisations	1,500 logins / month
NGDC	400	11 million visits a month	11 million visits a month
UKSSDC	36 databases / archives	5,400 registered users	1.5 million data requests / year

Source: Technopolis compilation of material provided by data centres, November 2009

However, the centres' statistics are not always directly comparable. For example, the ADS has no user statistics because it does not require registration to access datasets, while other centres count either registered users, active users or enquirers – all of which mean different things. Similarly, the distinctions between downloads, visits and data requests within the usage statistics are important, and mean that the numbers cannot be compared directly.

In order to get a clearer idea of who uses the centres, and how, we included some profiling questions within the online user survey. There is no way of knowing whether the groups that we reached are representative of the entire user population, since the surveys were voluntary and broadcast in ways that may have reached certain groups more than others – indeed, we know that this is the case with the NGDS survey, which attracted most responses from BGS employees. However, the results provide useful context for other questions about usage, the answers to which will be analysed in more detail later in this report.

Figure 5 shows the background of respondents to each survey. Most were based in academic institutions or research institutes, which is in line, for the most part, with the internal statistics collected by some data centres. Four of the five data centres received more than 80% of their survey responses from these types of user.

The NGDC's high response rate from public research organisations is almost certainly due to the large proportion of respondents who are BGS staff. The BADC produced a smaller, but still significant, number of responses from the research institute sector, which probably reflects NERC's funding strategy, which supports both universities and institutes. The ADS has the strongest representation from non-academic respondents, possibly because expert practitioners in archaeology are located in institutions such as museums and galleries, county archaeological services and commercial organisations rather than public research organisations (as is the case for other disciplines). The ADS also attracted a high number of users from government, business and charitable organisations, possibly reflecting the wide-ranging presence of archaeology in urban planning, education and so on.

20

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

The other surveys do not show such a high response rate from sectors other than academia or public research organisations. This may be because the survey did not, for whatever reason, reach researchers outside these sectors, or because the data centre's services are not relevant or accessible to researchers outside those sectors; as with, for example, the CDS, which is only available to UK academics. The latter case would imply a lower level of engagement with non-academic sectors, which would in turn reduce the likely impact of the data centre on non-academic research.

Figure 5: Profile of data centre users responding to the questionnaire

	ADS	BADC	CDS	ESDS	NGDC
Academic	51%	67%	95%	78%	4%
Public research organisation	0%	21%	4%	4%	86%
Private research organisation	8%	2%	1%	4%	0%
Private / independent researcher	8%	2%	1%	2%	2%
Central / local government	10%	5%	0%	7%	2%
Business	5%	1%	0%	1%	0%
Community / charity organisation	11%	1%	0%	2%	0%
Other (please specify)	8%	3%	1%	3%	6%
N=	83	759	200	292	51

Source: Technopolis survey of data centre users, January 2010

Types of data sought

The BADC, CDS and NGDC surveys invited respondents to indicate what kind of research data they were most likely to use. Respondents could tick all categories that applied. Figure 5 shows the responses, which clearly reflect the types of data provided by the three centres. The BADC and NGDC provide access to observational data, for the most part, whereas the CDS is primarily a portal to databases containing experimental results which describe the properties and reactions of various compounds.

Figure 6: Types of research data accessed, by centre

	BADC	CDS	NGDC
Observational data	81%	13%	91%
Experimental data	13%	74%	19%
Simulated data	22%	14%	5%
A mixture	10%	23%	0%
Other	2%	3%	23%
N=	723	188	43

Source: Technopolis survey of data centre users, January 2010

The online surveys also invited respondents to indicate how they used data from the centre within their own research, ticking all categories that applied. Results are shown in Figure 7. There is something of a divide here, with ESDS and BADC users more likely to use data for original research, whether as a standalone dataset or in combination with other data. ADS and CDS users, on the other hand, are more likely to use data for reference. Using data centre resources as a basis for further data collection is the least common response, although it seems to be linked more to the 'reference' type of usage than the 'original research' kind of work.

The divide between use for original research and reference probably relates, at least in part, to the types of data held and norms within the disciplinary field. The CDS primarily provides access to experimental data which researchers reference routinely and frequently. The ESDS, on the other hand, has a heterogeneous collection of datasets which, in the social sciences, are much more likely to form the basis of novel research.

That said, it is clear that usage is relatively high across the board, and that many researchers use data in more than one way. This is true of all centres, but particularly of the NGDC. As mentioned, the respondents to this survey were primarily researchers at the BGS, and the NGDC is central to most research work there, which may explain the diverse uses to which data from the centre is put.

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

Figure 7: Use of research data, by centre

	ADS	BADC	CDS	ESDS	NGDC
For research	51%	75%	48%	88%	72%
For combining with other data	46%	46%	39%	34%	81%
For reference	79%	22%	82%	21%	72%
As a basis for further data collection	51%	7%	24%	17%	58%
Other	14%	2%	7%	6%	16%
N=	70	713	190	289	43

Source: Technopolis survey of data centre users, January 2010

Users were asked to rate the importance of the data that they accessed to their work. As Figure 8 shows, the majority of users for every data centre consider the data they access from data centres to be 'very important' to their research or teaching, with most of the remainder considering it 'quite important'. No more than 8% of users in any single data centre said that the data was 'not very important' or 'not important' for their research.

Figure 8: Importance of data for users' research

	ADS	BADC	CDS	ESDS	NGDC
Very important	60%	61%	64%	68%	95%
Quite important	34%	32%	28%	25%	5%
Not very important	6%	7%	7%	5%	0%
Not important	0%	1%	1%	2%	0%
N=	65	700	189	282	40

Source: Technopolis survey of data centre users, January 2010

Duration of use

Figure 9 shows how long survey respondents had been users of each centre's data holdings and services. Overall, a significant minority of users were relatively new, having first used the service less than 12 months ago, while another significant minority were long-established, having benefited from the service since its launch, in some cases for more than 20 years.

The NGDC has a noticeably higher mean and median number of years of usage than any other data centre. This may be due to the bias towards BGS employees in the survey response: such a sample is likely to be more uniform than a group of users drawn from a wide range of institutions. Furthermore, as we have noted, most BGS staff rely on the data centre for their day to day work, and so are unlikely to have come to it later in their career as part of a specific and time-limited project.

The BADC and the ESDS show relatively low median responses for the number of years of use. This could relate to the age of the users, and since both data centres offer pedagogical materials and simulations as well as core data sets it seems plausible that they are particularly attractive to PhD students and younger researchers. However, the low numbers may also be due to the relative importance of BADC and ESDS data for original research (noted in Figure 5, above). This would explain a shorter period of use, timed to fit with a specific project based on the data set, rather than the long-term and established patterns of use that we would expect to see where the data set is used routinely for reference, as is the case with the ADS and CDS.

Figure 9: Number of years using a centre's data

	ADS	BADC	CDS	ESDS	NGDC
Mean	7	4	9	4	17
Min	1	0	0	0	1
Max	14	20	35	30	39
Median	6	3	10	2	14
N=	83	757	200	165	51

Source: Technopolis survey of data centre users, January 2010

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

Frequency and extent of use

Figure 10 presents the survey results for number of uses for each of the five centres, in absolute terms and adjusted to a number per year, to aid comparison. These responses should be treated with some caution, and as broad brush indicators rather than precise figures, since it may have been difficult for long-term users to estimate the total number of productive visits. More useful, perhaps, is the data presented in Figure 11, which shows how often researchers currently make use of the data centre.

Looking at these two tables together, the NGDC is, once again, a clear outlier, with a mean and median number of uses that is significantly higher than the other data centres. The frequency of use was also much higher than other data centres, with 45% of respondents using the NGDC daily; for comparison, between 0-4% of researchers at other centres reported daily use. Among the other four centres, the ESDS and BADC researchers once again show quite a different profile from those using the ADS and CDS. As before, this is probably related to the way in which researchers use the data that they access from the centre. ESDS and BADC researchers tend to use data for their own original research projects, and so it seems plausible that they would visit the data centre less frequently: once they have downloaded the relevant data sets, there is no pressing need to return to the site. ADS and CDS users, on the other hand, reference data on a regular basis as part of their work, with more frequent visits to 'check' certain data sets for citation in a specific publication or as background to a new project.

It is also worth noting that on the supply side there are differences in research infrastructure: so, for example, the NGDC databases are accessed online (in part for reasons to do with the size of the database), while the ESDS is set up to provide users with downloads of data sets for them to work on offline. This clearly affects the number of visits that users have to make to the centre in order to access the data that they need.

Figure 10: Number of times users had accessed data from the data centre

	ADS	BADC	CDS	ESDS	NGDC
Mean	270	115	445	19	12070
Min	2	0	0	0	0
Max	5000	2000	20000	300	250000
Median	50	6	100	6	1000
Mean / year	49	28	41	5	731
Median / year	10	2	8	3	74
N=	83	757	200	165	51

Figure 11: Frequency of use

	ADS	BADC	CDS	ESDS	NGDC
≤ monthly (more frequently)	69%	31%	79%	43%	88%
≥ monthly (less frequently)	31%	69%	21%	57%	12%
N=	83	757	200	165	51

Source: Technopolis survey of data centre users, January 2010

5. Trends in users and usage

Overall, with respect to data holdings and usage, trends were broadly consistent across all eight centres, with each data centre's internal statistics showing growth in one or more of each of four fundamental dimensions:

- Number of collections;
- Size of individual holdings;
- Number of users; and
- Intensity of use.

Figure 12 picks out selected examples of trends on each of the four dimensions. The table presents a truncated list of centres, due to the limitations in available time series data. We were unable to obtain relevant time series for the NCDR, the NGDS or the UKSSDC. Nevertheless, the statistics provided by the other five centres indicate a significant increase in the numbers of data collections being curated, an expansion in the size of individual data holdings and an increase in users and usage. All this suggests that data centres have value for researchers.

Figure 12: Examples of trends in holdings and use of research data centres

Dimension	Example
Increasing number of collections	ADS: 300 databases and collections in 2009, up from 100 in 2005 BADC: 20 data sets / 20 terabytes (2002); 130 / 60 terabytes (2007) ESDS: adding 250 – 300 data collections annually
Increasing size of individual holdings	EBI storage: 20 terabytes in 2006, rising to 4.5 petabytes in 2009 EBI nucleotide database: 170 million records; doubling every 2-3 years
Increasing number of users	CDS: increase in registered users, from 2,000 in 2000 to more than 5,000 in 2009
Increasing intensity of use	ADS: 12 million pages downloaded in 2009, up from 4 million pages in 2005 ESDS: 5-fold increase in monthly downloads, rising from 1,000 a month in 2003 to around 5,000 a month in 2008

Source: Technopolis compilation of material provided by data centres, November 2009

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

The usage statistics reported by almost all the centres show reasonably strong growth over time. As Figure 11 shows, the online surveys of users reveal a more mixed picture, with three of the five centres seeing relatively stable levels of usage by individual respondents. However, since this data is not necessarily representative of all users, it does not contradict the data recorded by the centres themselves.

Figure 13: Trends in individual use

	ADS	BADC	CDS	ESDS	NGDC
Decreased over time	13%	26%	16%	21%	12%
Same / fluctuated	39%	62%	54%	60%	33%
Increased over time	48%	12%	30%	20%	55%
N=	71	730	200	256	51

Source: Technopolis survey of data centre users, January 2010

The most widely cited reasons provided by users for increasing their frequency of use of data were as follows, in descending order:

- New research questions have driven the need to acquire more data (28%)
- Improvements to the range and quality of data available (26%)
- Change in role / position leading to an increased need for data (21%)
- Increased familiarity with the databases (11%)
- Improved ease of access to the data (8%)

This list hints at a virtuous circle of extending the range of data sets and new research questions, leading to increased demand for and use of the centres and their holdings.

The main reasons provided by users for decreasing their frequency of use of data were as follows, in descending order:

- Research questions have been addressed / emphasis has shifted (42%)
- Change in role / position leading to a decreased need for data (33%)
- Switch in use to alternative sources of data (14%)
- Lack of availability, accessibility, etc, (5%)

The two most widely cited factors, by a significant margin, reveal the importance of changed circumstances on the demand side: the completion of a particular piece of work or the onward movement of the individual concerned, through promotion, changed employer or retirement. Supply-side challenges were much less commonly cited, and where they were, the issue was most often the emergence or discovery of an alternative source, which was either better or more convenient.

Outputs and end-users

The surveys asked about the types of outputs respondents produce, making use of information from the data centres. Figure 14 presents the results (respondents were able to select as many answers as applied).

Overall, research papers were the most common output of work that used data from a centre, reported by over 60% of respondents from four of the five centres. The BADC, CDS and ESDS respondents were strongly skewed towards these outputs. For these centres, internal reports were the only other significant output. The NGDC and ADS had a greater range of outputs, with quite a strong showing for contract research outputs (the most frequent response under the 'other' category). This may reflect the relative importance of commercial and regulatory clients for BGS staff using NGDC services, and the importance of public agencies, such as English Heritage, for the ADS.

Figure 14: Types of outputs produced using data from the centres

	ADS	BADC	CDS	ESDS	NGDC
Research papers	46%	73%	75%	62%	64%
Internal reports	43%	28%	30%	28%	77%
Enhanced data	39%	14%	15%	17%	61%
No formal outputs	26%	14%	13%	17%	9%
Other	30%	6%	5%	19%	30%
N=	70	712	190	289	44

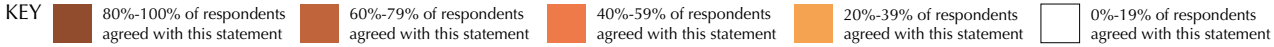
Source: Technopolis survey of data centre users, January 2010

The surveys went on to ask respondents about the intended end-users of their outputs. There is of course no guarantee that these are the people who are actually using the research outputs, although it seems unlikely that researchers would persist in producing work for such audiences if they are not interested. Figure 15 presents the results.

Figure 15: Types of intended end users

	ADS	BADC	CDS	ESDS	NGDC
Academics	69%	84%	90%	79%	68%
Individuals in your organisation	43%	21%	26%	26%	82%
Policy-makers	16%	28%	2%	47%	66%
Businesses	19%	4%	7%	7%	75%
Own use only	31%	13%	26%	16%	16%
Unknown	4%	3%	2%	1%	7%
Other	24%	6%	8%	10%	25%
N=	70	710	189	287	44

Source: Technopolis survey of data centre users, January 2010



Academics are important end-users for all centres. The relative importance accorded to other end-users is probably linked to the fields in which researchers are working. For example, business users are rated quite highly by the NGDC and the ADS; this probably reflects the commercial consultancy role of the BGS, and the importance of archaeology within the planning process for property developers and utility providers (among others). Similarly, ESDS and BADC users are likely to be working in areas of interest to public policy-makers, explaining their high rating for this particular audience.

The ADS and CDS had a relatively high number of respondents saying that the research outputs were for their own use only. This may be because, as noted in Figure 7, ADS and CDS data tends to be used for reference. Many researchers presumably download information in order to check their own findings and conclusions, but do not necessarily need to share this more widely.

Some respondents gave further details of the kinds of user that fell within the 'other' category. In many cases, these were students, both undergraduate and postgraduate, or the general public. It would have been interesting to see how much more recognition these audiences gained had they been offered as options within the predetermined responses.

Finally, it is worth noting the low numbers of respondents who do not have a specific end-user in mind for their research outputs. This applies across all data centres, and suggests that researchers are, broadly speaking, aware of the types of impact that they would like their work to have.

The culture of data sharing

Data centre users were asked about the impacts of the centres in improving the culture of data sharing and re-use within their research communities. As Figure 16 shows, most respondents agreed that the centre had had an impact on data sharing culture 'to a large extent'. The ADS users were the strongest supporters of the centre as an agent of cultural change. ESDS users, on the other hand, were more likely to see the change as existing only 'to a small extent' or, indeed, 'not at all'. That may be because the UK Data Archive has been in operation for more than 40 years and had an online presence for more than 15 years, and any behavioural impacts are thus already well advanced.

Figure 16: Impact on culture of data sharing, by data centre

	ADS	BADC	CDS	ESDS	NGDC
To a large extent	84%	69%	72%	54%	68%
To a small extent	16%	29%	27%	40%	30%
Not at all	0%	2%	1%	7%	3%
N=	61	601	164	244	37

Source: Technopolis ranking based on survey of data centre users, January 2010

We asked respondents whether they ever added value to data, or created new data of their own which might be of interest to the wider research community. We did not define what we meant by 'adding value' and researchers could therefore have interpreted this in a variety of ways. Figures 17 and 18 show their responses, and indicate that for most data centres, the proportion of researchers who add value to existing data is lower than the proportion who are creating novel data through their own research projects. The only exception is the ESDS, where a larger proportion of researchers add value to data, either 'sometimes' or 'always', than those who create new data of their own. This is probably due to the cost of collecting new data in the social sciences.

It is also worth noting that the ADS, ESDS and NGDC have a particularly high proportion of researchers who add value to data, and that the ADS, CDS and NGDC have, among the respondents, a relatively high proportion of data creators.

Figure 17: Researchers 'adding value' to data, by data centre

	ADS	BADC	CDS	ESDS	NGDC
No, never	54%	84%	69%	51%	55%
Yes, sometimes	43%	14%	25%	35%	34%
Yes, always	4%	2%	6%	14%	12%
N=	54	610	160	243	77

Source: Technopolis survey of data centre users, January 2010

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

Figure 18: Researchers creating new data, by data centre

	ADS	BADC	CDS	ESDS	NGDC
No	20%	59%	30%	65%	30%
Yes	80%	41%	70%	35%	70%
N=	59	608	162	241	37

Source: Technopolis survey of data centre users, January 2010

Responses to the questions of whether researchers ever submitted their new data, or resubmitted data to which they had added value, to a data centre are presented in Figures 19 and 20. Only respondents who said that they created new data, or added value to downloaded data (either 'always' or 'sometimes') have been included in these figures: a subset of the main survey. BADC users were asked a slightly different question – whether the data was submitted to BADC or to another data centre – and so the results are shown separately.

Figure 19 shows that for all data centres except the NGDC, most respondents did not resubmit data. This may be because they do not feel that the value they added was sufficient to justify resubmission. The relatively high level of resubmission in the NGDC is probably because respondents, as employees of the BGS, were required to observe the institution's data sharing policies.

Figure 19: Re-submitting data to data centres

	ADS	CDS	ESDS	NGDC		BADC
No, never	58%	64%	75%	13%	No, never	74%
Yes, sometimes	29%	26%	22%	42%	Yes, to BADC/NEODC	18%
Yes, always	13%	10%	3%	45%	Yes, to other data centres	8%
N=	24	50	116	31	Total	90

Figure 20, showing the proportion of respondents who submit new data, tells a slightly more encouraging story. Most researchers using ADS, CDS, BADC and NGDC do submit new data to a centre, although for all centres except NGDC it is more common for them to do this 'sometimes' than 'always'. The very low level of ESDS users submitting new data may again reflect the nature of research in the social sciences, where new data sets may be hard-won, and mined for new publications for several years. There is a clear disincentive to share the work if such generosity could result in being scooped by another researcher. The ESRC has one of the strongest policies regarding the submission of new datasets, and these results suggest that the policy is not being complied with across the social science research community as a whole (where much research, of course, is supported by other funders).

Figure 20: Submitting new data to research data centres

	ADS	CDS	ESDS	NGDC		BADC
No, never	36%	32%	70%	8%	No, never	49%
Yes, sometimes	52%	41%	20%	27%	Yes, to BADC/NEODC	28%
Yes, always	11%	28%	11%	65%	Yes, to other data centres	23%
N=	44	111	82	26	N=	241

Source: Technopolis survey of data centre users, January 2010

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

The culture of data sharing

Finally, we asked users to explain their citation behaviour when reporting results that made use of 'third party' data obtained from one of the centres. Figures 21 and 22 show the results. They suggest that citation of data centres or sets is relatively common, particularly among users of the ADS, BADC and ESDS, most of whom always cite the data centre or set. Citation rates for data creators or gatherers are slightly lower, but remain impressive in view of the ongoing debate about the difficulty of citing datasets.

Figure 21: Citation of data centre

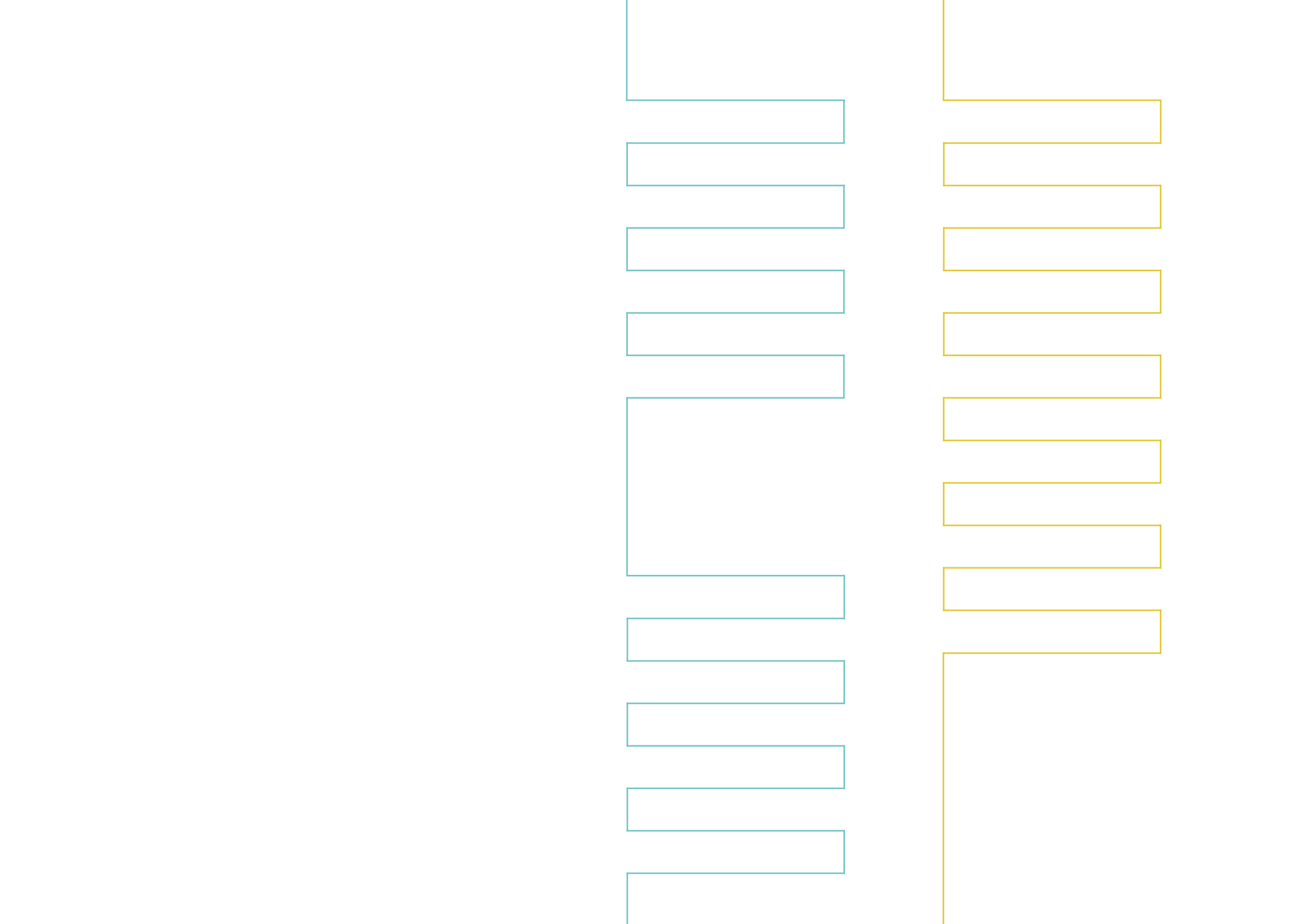
		ADS	BADC	ESDS	NGDC			CDS
Do you cite the data centre / set?	No, never	4%	11%	6%	12%	Do you cite the CDS / dataset?	No, never	24%
	Yes, sometimes	32%	22%	16%	56%		Yes, sometimes	42%
	Yes, always	63%	67%	78%	33%		Yes, always	34%
	N=	68	691	277	43		N=	184

Source: Technopolis survey of data centre users, January 2010

Figure 22: Citation of original data creator

		ADS	BADC	ESDS	NGDC			CDS
Do you cite the data creator / gatherer?	No, never	3%	17%	9%	22%	Do you cite the original developer / provider of the database?	No, never	41%
	Yes, sometimes	37%	27%	21%	59%		Yes, sometimes	29%
	Yes, always	60%	55%	70%	20%		Yes, always	30%
	N=	67	528	249	41		N=	163

Source: Technopolis survey of data centre users, January 2010



6. Impact on research

Research benefits

The survey collected information about the benefits of data centres in two ways. Respondents indicated their level of agreement with a set of statements about perceived benefits, and then expressed any further thoughts or experiences in a free-text response. Researchers were also asked about the implications for their work if the data centre did not exist.

The statements can be arranged into four broad groups: research efficiency, research quality, research novelty and researcher training. Overall, the level of agreement was very high for each group of benefits, although the most widely-agreed statements relate to research efficiency. Respondents' comments also highlighted a number of benefits which do not fall directly into any of the categories used for the offered statements.

Research efficiency

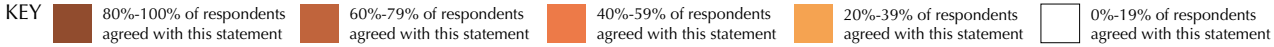
One central justification for data centres is that they help researchers to work more efficiently. They save time and money, by eliminating unnecessary re-creation of data, and by making it easier for researchers to find what they need. This is reflected in the responses, with researchers from all five centres ranking research efficiency benefits very highly. Figure 23 shows the percentage who agreed 'to a large extent' with the various statements about research efficiency.

Overall, time appears to be the most important efficiency benefit. The statements about general efficiency and the financial cost of using data garnered a fairly consistent level of agreement across all data centres. However, views on the duplication of effort were more divided, with this being a relatively important benefit for CDS and ESDS users, but less widely-supported by users of the ADS and BADC. NGDC users were particularly positive across the board, as can be seen in their responses to the statements about duplication of effort and greater quantities of research, the last of which was not so strongly supported by users of other centres.

Figure 23: Research efficiency benefits, by data centre

Benefit	ADS	BADC	CDS	ESDS	NGDC
It has reduced the time required for data acquisition / processing	79%	68%	76%	80%	92%
N=	67	618	174	262	36
It has improved the efficiency of research	79%	62%	75%	67%	89%
N=	67	622	179	261	36
It has reduced the financial cost of data acquisition / processing	65%	62%	61%	73%	78%
N=	66	612	175	262	36
It has reduced duplication of effort (i.e. unnecessary recreation of data)	57%	57%	68%	62%	81%
N=	65	609	176	258	36
It has enabled me to undertake a greater quantity of research	52%	42%	50%	54%	77%
N=	63	614	176	257	35

Source: Technopolis survey of data centre users, January 2010



The free text responses to questions about the benefits of the centre, and the implications should it no longer exist, broadly confirm this picture; but they also allow us to delve more deeply into the specific nature of the time efficiencies that centre users have experienced.

Most respondents who cited time efficiencies mentioned that the centres save them time in finding and accessing the information they need:

> Having worked outside the UK (without access to the CDS), I can fully appreciate the enormous benefit of having easy access to this resource, which enables UK academics to access high quality data with minimal effort, and thus to concentrate their efforts on pushing their research forwards at a greater rate than would otherwise be possible.

> [If the centre no longer existed] time would be wasted trying to track down data and organise access to restricted datasets.

The other main benefit under time saving was the ability readily to access contextual data. As seen in the previous section, many researchers require data from earlier experiments or long-term observations to support their own work; the centre makes it easy to find and use this data, thereby speeding up their new research.

> Calculations using crystal structures from the CDS are used to predict and interpret experimental data. Without the CDS making such structural information readily accessible, the predictive side of my work would either be much slower, or non-existent. This in turn would make the experimental and interpretive side of the work slower and more challenging.

Several respondents mentioned the reduced possibility of duplication of effort:

> [If the centre no longer existed] I would have no idea if my data was new, or if it had already been reported in some form.

> Unless we could access data from previous research we would expensively reinvent the wheel. Much research is cumulative.

Researchers also mentioned financial efficiencies that resulted from using the centre, especially when considering the implications if it ceased to exist. ‘It would cost more money’ featured in a number of responses. Others went into more detail about exactly how money has been saved. In some instances this related back to time: not having a data centre ‘would waste a lot of my (costly) time’. In other cases, respondents focused on the savings that can be made, both by individual institutions and by the research system as a whole:

> BADC provides a way for all the data (which is expensive to produce in the first place) to be collected and managed, which may be too costly and difficult for individuals/institutions to do.

Case study – research efficiency

Understanding how lightning and radiowaves interact

Dr Chris Davis uses images of the sun to study the mass projections emitted from it and the effect these have on the Earth's atmosphere. The sun is highly variable and volatile, and the storms it creates in space weather can affect the ionosphere, which has knock-on consequences on issues ranging from radio transmission to lightning. Dr Davis argues that we need to understand the ionosphere to have good models of our global environment.[1]

1 This may suggest a wider impact for the research.

Dr Davis's work in this area has made extensive use of the ionospheric data held at the UK Solar System Data Centre (UKSSDC) in the World Data Centre for Solar-Terrestrial Physics data subset. Dr Davis explains that he approached the UK WDC because it has a much larger holding of ionospheric data[2] than the other centres.

2 Having all the data in one place is important to researchers.

Is the current basic state pension sustainable?

Dr Martin Weale was director of the National Institute of Economic and Social Research (NIESR) from 1995 to 2010. NIESR was commissioned by the Department of Work and Pensions (DWP) to investigate the effect of means-testing on pension eligibility and the effect this has on saving behaviour. In 2003, the way that pensions are calculated was changed, so that the state pension was reduced by 40% of any savings income received, rather than 100% as was previously the case. The aim was to remove the disincentive to save.

Dr Weale used data from the ESDS to show that individual responses to the policy change would vary, depending upon how wealthy people are. For the poorest third of households, the improved rates of return would motivate an increase in savings of 17-23%. The middle third of households, however, could be expected to reduce their savings by 27-29%. The policy change would also affect the decision to retire, with the poorest households waiting an extra 0.4-0.5 years, and middle-income households bringing their retirement forward by 0.3 years. Overall, government expenditure[3] could be expected to rise by a small amount.

To make these predictions, Dr Weale used the Family Resources Survey, the Expenditure and Food Survey and the British House Panel Survey. Dr Weale explained that he turned to the ESDS rather than the original producers of the data because he wouldn't think to look anywhere else for it.[4] He agreed that the data came in a useful and accessible format[5] and that there were clear time efficiencies in having all the information he needed in one place.[6]

3 We might expect to see an impact on policy making in the longer term.

4 Data centres become a first resource for researchers seeking information.

5 Data centre protocols help ensure that data is produced in a format that is valuable to researchers, which may not be the case if data were accessed directly from creators.

6 Time efficiencies are related to the size and coverage of the data centre's collections.

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

Research quality

This set of questions was designed to test the hypothesis that having ready access to a range of data through data centres helps researchers to produce better research. This may be related to increased availability of data or better quality data. As Figure 24 shows, the overall level of agreement with statements about research quality, although high, was much lower than for statements about research efficiency. However, the NGDC respondents remain noticeably more enthusiastic than other groups.

Figure 24: Research quality benefits, by data centre

	ADS	BADC	CDS	ESDS	NGDC
It has increased the use of data in my research	48%	40%	38%	46%	75%
N=	63	614	175	254	36
It has improved the quality of the data I use within my research	55%	47%	48%	51%	72%
N=	62	613	174	258	36
It has improved the evidence base of my research	58%	46%	60%	56%	77%
N=	65	624	173	259	35
It has helped to improve the quality of my research outputs	56%	47%	58%	56%	69%
N=	66	620	177	259	36

Source: Technopolis survey of data centre users, January 2010

Relatively few free text statements on the benefits of data centres referred to their impact on research quality. However, more responses were given to the question about the implications if a data centre were no longer available. These give us a more detailed view of exactly how the data centre supports quality in research.

Many of the responses about research quality were variations on a theme stated succinctly by one respondent: 'papers of lower quality'. Concerns about getting published weigh heavily in the minds of researchers because of the important role that published output plays within the academic system. Others, however, concentrated more on the impact on their results and conclusions, and the concomitant effect on the scientific process.

> Research findings would be incomplete and not as full as they should be.

> Poorer quality analysis and science would be carried out.

A couple of respondents mentioned that data from centres can be used to test the quality of previous work. This could be an important benefit of centres, in a context where quality assurance is becoming an increasingly large challenge within the scholarly communications process. Data centres can help researchers to check the validity of conclusions drawn by their colleagues.

> We have the potential to revisit data, do more research with the data and check past results - in my mind this is only done feasibly with a central data centre.

They can also help researchers to check the quality of their own work and analyses:

> [If the data centre no longer existed] I might have less data on similar compounds which I could use to compare my complexes with.

Other responses focused on the quality of the data provided by the centre, and the fact that it offers a 'quality stamp', 'quality controlled data', or some kind of guarantee of 'reliability' or 'accuracy'. Perhaps more realistic was the respondent who said:

> Although I don't expect BADC to ensure the data is perfect I have confidence that it is what it says it is.

Others, however, commented upon the quality, not of individual datasets, but of a centre's entire collection:

> [If the centre did not exist] the depth of data from the longitudinal datasets would never be replicable, so arguably despite the increased costs [of recollecting data] the data would not be of the same quality.

The curatorial role of the centre thus affects two important elements of data quality: first, ensuring that individual datasets are academically 'good' (as much as it can) and second, ensuring that it creates and preserves collections which can be a useful starting point for new research.

Case study – research quality

Chemical methods in the conservation of cultural heritage

Mark Dowsett is a Professor at the Department of Physics at the University of Warwick. He is studying the corrosion processes of metals such as lead, copper, bronze and silver by analysing the reactions between them and chemical compounds. He is using this to develop methods to improve the conservation of irreplaceable historical artefacts ranging from jewellery to armour to musical instruments.

Prof. Dowsett relies on CDS to provide reference spectrum data that helps him to identify corrosion products. He can also compare his results with the CDS database of known compounds to check when a new product has been discovered. One of the features that Dowsett finds useful is the way that CDS closely ties the spectrum data to scientific literature providing information about components that he would not otherwise have come across during his research.[7]

The electrochemical and synchrotron techniques that Prof. Dowsett and his team have developed have extensive potential for creating preservation treatments for cultural heritage, and he is working on methods that will be adapted by cultural organisations in several countries.[8] Industry could benefit[9] from the compounds found and coating methods developed by Dowsett and his team, as they will have a general application.

7 The quality of the research would have been lessened without this additional service (beyond mere access to the data sets) offered by the CDS.

8 Evidence of wider benefit of research carried out using data from a data centre.

9 There is also potential for further benefit in the commercial sphere.

KEY 80%-100% of respondents agreed with this statement | 60%-79% of respondents agreed with this statement | 40%-59% of respondents agreed with this statement | 20%-39% of respondents agreed with this statement | 0%-19% of respondents agreed with this statement

Research novelty

The notion of data as the 'fourth paradigm' of scientific discovery is well-rehearsed. Our survey tested the idea that the large collections of datasets held by data centres offer new intellectual opportunities, or permit new types of research to go ahead.

The responses about research novelty show bigger differences between the five data centres than those to the previous two groups of statements.. The large number of BGS employees who responded to the NGDC survey probably accounts for the strong agreement with all these statements.

Figure 25: Research novelty benefits, by data centre

	ADS	BADC	CDS	ESDS	NGDC
It has created new intellectual opportunities (e.g. merging of several data sets to answer new questions)	51%	33%	27%	40%	69%
N=	63	600	171	256	36
It has enabled research to go ahead that otherwise might not have done	62%	48%	41%	60%	89%
N=	66	636	180	263	35
It has permitted more novel research questions to be answered / tackled	46%	38%	44%	51%	69%
N=	63	616	175	261	36
It has enabled new types of research to be carried out	56%	34%	33%	44%	77%
N=	63	604	175	259	35

Source: Technopolis survey of data centre users, January 2010

Some trends are discernable across data centres. For example, the most widely-agreed statement seems to be about enabling research to go ahead that otherwise would not have done. Conversely, the statements about new intellectual opportunities and new types of research were not so widely supported. The free text responses shed some light on why this might be.

There were relatively few concrete examples where the availability of data had changed the way that researchers approached their field, although some hoped that this would become possible in the future:

> The information in the ICSD helps new research ideas to form.

> As ADS and other centres develop we will become more fluent in using large quantities of information rather than just using the two or three bits we happen to have, and that ought to create a revolution in the interpretation of archaeology.

It was much easier for respondents to identify situations where data centres make it possible to answer questions which have probably existed in the minds of researchers for some time, but could not be addressed without large-scale, openly-accessible collections of data:

> You have to work with data which are available. If certain data are not available we may not be able to answer specific research questions.

> [If the centre no longer existed] research in this country would return to much narrower surveys (based on objects, regions, periods, etc). It is the broad scope of ADS resources, and the detail of some of the collections that have enabled me to work on larger datasets, addressing broader themes, across larger areas, and across disciplines.

In some instances, such questions arose because research methods have moved on since the data was collected:

> These data are very useful for applying new data analysis methods and this gives a chance for obtaining completely new results.

Collaborative research is evidently an important subset of the kinds of research made possible by data centres:

> It is an excellent medium for collaborative work with colleagues overseas.

> [If the data centre no longer existed] collaboration with other organisations would not be possible.

Some respondents also mentioned that having free access to the database enabled them to work across a wider stretch of their own discipline:

> It enables research on fringes of main interest of department for which in-house versions of databases could not be afforded, and would not be cost effective for small number of users.

Case study – research novelty

Landscapes of Anglo-Saxon governance

Stuart Brookes' research maps out the emergence of English kingdoms during the early medieval period (AD400-1100) by tracing military, social, political, and judiciary processes. He has shown that powerful families maintained a grip on their territories through legal proceedings and fora for settling disputes, which were at least as important as their battlefield victories. Open-air assembly sites were important at all levels of medieval society, and examination of these sites has revealed pre-Roman, Roman, Anglo-Saxon, Viking and other origins. Many of the geographical, governmental and judiciary configurations of modern England can be traced back to the early medieval period, and Dr Brookes' findings challenge many of the tacit assumptions[10] made about English culture and society. By re-examining the sites and institutions of medieval England, his research not only leads to a greater appreciation of the history of medieval landscapes, but also identifies a diversity of cultural values and identities that can open up a dialogue about contemporary English political ideology and contribute toward a more consistent view of the origins of 'English' identity. Another modern myth that Brookes' work challenges is the widely applied benchmark of Greek and Roman social models by giving us a framework for understanding the social complexity of post-Classical societies. By doing so, his work portrays alternative versions of European democracy.[11]

One of Dr Brookes' techniques is the creation of a detailed map of the contemporary landscape of war and governance. This requires a thorough understanding of all the sources available in his area, so he uses the Heritage Highway in ADS to provide evidence about sites, grey data and further sources. All these resources are essential to elaborate the mapping and to progress new discoveries:[12] 'ADS is the best source to map different periods.'

10 Research undertaken using data centre resources has challenged assumptions within the discipline…

11 …and may have broader significance in areas of study beyond archaeology.

12 The data centre is essential to novel research.

Training

Only one question was asked about researcher training, and it appears to have generated the most mixed response across data centres. The low response numbers for this question make the quantitive data unreliable, so it has not been made use of here.

Many of the free text responses referred to researcher training or development as a benefit of data centres. For the most part, respondents were talking about PhD students, and the benefits that data centres can have for both teaching and learning and for research.

In teaching and learning, many comments focused upon how data centres could help students to become familiar with the types of data, methods and even theoretical concepts that they would use in their research careers.

> Cambridge database software [is] very user friendly so that students are able to pick up ideas about molecular conformations and supramolecular interactions without realising.

> For students engaging in the formal use of large scale datasets for the first time - the UKDA/ESDS is simple, straightforward, easy to use and not intimidating. Choice of format is also very useful.

> The archive provides an easily accessible source of examples of social research which are useful in teaching research methods. Without these, examples could be found elsewhere but the archive encourages students to be more independent in their exploration of research ideas.

Responses about training were particularly strong among ESDS users. This may relate to the importance of existing data to new research in the social sciences, and the need to train novice researchers in the use of such data. The ESRC has emphasised the need to develop data 'handling' skills (covering data management and analysis); the ESDS and UKDA have been active in developing and delivering training materials.

Several responses also mentioned the importance of data centres in helping postgraduate students to find topics and undertake original work:

> The rainfall archive has spawned a PhD in its own right - and the outcome of analysis here will probably lead to other PhDs in related areas of climate model validation - all made possible by access to archive data.

> Climate change data sets are most exciting for students to analyse, and also for me to supervise, and have lead to many successful MSc projects under my supervision

Almost all of these kinds of comments came from BADC users, suggesting that this dataset is a particularly rich resource for postgraduate researchers in the environmental sciences.

Additional value

The previous sections have looked at how data centres can benefit research projects. However, respondents to the free text questions also mentioned several characteristics of centres which, although they do not directly affect research efficiency, novelty, quality or training, are considered important in achieving these benefits. These responses help to clarify what users value about data centre services.

Most data centres provide support for users on how to access and use the data they hold. This support can be via the website, or from data centre staff, and both are clearly important:

> Data can be accessed quickly; online help with specific questions has been very useful. All my questions have produced responses from BADC staff.

> A particularly strong feature of CDS is the excellent technical support.

> The support from the main office and the help-mail is very efficient.

> BADC staff have expertise of both atmospheric science and informatics (metadata technology, in particular). I think it is the strength of BADC.

In one case, the centre staff were able to help researchers access a collection that they needed for their research, even though it did not hold that collection itself:

> The UK Data Archive was able to assist in obtaining data from a US data archive (ICPSR) which was essential to my research. These reciprocal arrangements are extremely useful!

Indeed, the value of data centres as a national resource was mentioned by several respondents:

> A nice example of effectiveness that should be followed in other European countries (specifically France and Germany).

One ESDS user mentioned that the wide availability of UK-based research outside the UK through the data centre made it more likely that projects in other countries would follow the same template, thereby facilitating comparability of international research:

> The Panel data, together with the easily-accessible supporting information and documentation have recently proved to be superior to similar services in other countries. The availability of the data has also meant that other countries have been able to replicate surveys making cross-country analysis possible, and potentially increases comparability of information if they are using the UK format as an example.

Many researchers also welcomed the curatorial and data preservation work undertaken by the centres. Several respondents felt strongly that this work would not be undertaken, or even be feasible, if it were left to individual data creators, which would lead to loss of valuable data which, in many cases, simply cannot be re-created.

> Secure curation of data is vital - I know of instances where data collected at considerable public expense in the 1980s and 1990s is now sadly no longer available - much to the regret of current researchers.

Through their curation work, data centres also set a high standard for data formatting and presentation. This may be explicitly stated as part of the centre's mission, or it may be a more informal consequence of researchers knowing that their data will be shared:

> [Benefits include] setting quality and documentation standards (that we are still striving to achieve).

> [If the centre no longer existed] the same data would be available from the primary gatherers, but this process would be slower, more dependent on individuals responding to requests and a centralised repository encourages better data formatting, manual pages, transparency and usability.

Finally, many researchers mentioned the benefits of the centre as a repository for their own work:

> [If the centre no longer existed] we would have nowhere suitable (and affordable) to archive our project information in the future, and so the information would be restricted and not publically disseminated. This would be a disaster for a public, community minded archaeology project such as ours.

In some cases, the efficiency of the centre as a platform for data dissemination played an important role in allowing the researcher to undertake original data collection:

> [If the centre no longer existed] as a data owner (and supplier to UKDA) I would have a smaller user base, so harder to justify funding for survey.

These benefits might be considered important precursors to, or enablers of, the research benefits that were evaluated through the quantitative questions, and raise some important points about how data centres enable research benefits. A common thread, running through all the characteristics identified here, is the importance of the centre as a sizeable and central hub for data within a discipline or field. It can thus collect and curate useful collections, offer support to researchers in using the collections and encourage good practice in data usage and storage. If the centre were to be disbanded and responsibility for data handed back to individual researchers it seems likely that many of these benefits would be lost.

Case study – additional value

Correlating radiocarbon dating and collagen degradation in dinosaur bones

Professor Matthew Collins has been trying to understand how dinosaur bones deteriorate, in order to help improve dating techniques at archaeological sites. Bones are composed of two main types of material: mineral and organic. The mineral component is mainly calcium phosphate and the organic component is mainly bone collagen, which is a group of naturally-occurring proteins. When bones interact with their environment, a process of diagenesis occurs which alters their original chemical structure. The speed of deterioration of the collagen depends on environmental factors and tends to increase in warmer climates. Professor Collins hypothesised that protein sequences found in the bones of T. Rex found in warmer climates and much longer ago could not have survived.

To test this hypothesis he applied to ADS staff to collate all the archaeological bone findings from anywhere in the world that had been analysed by UK scientists and were on the ADS record.[13] ADS sent over 1,400 observations. Professor Collins explained that without the support of ADS staff who collected this information, his project would not have been viable. Professor Collins et al. plotted the radiocarbon age of the bone findings against the estimated effective collagen degradation temperature and this produced a correlation that showed the age at which collagen should be expected to disintegrate in various climates.[14] This research indicates that any case in which collagen had not disintegrated after 68 million years represents a substantial outlier.

Uncovering the effect of the sun's influence on the Earth's atmosphere

Mike Lockwood is Professor of Space Environment Physics at the University of Reading and also works at the Rutherford Appleton Laboratory in the Space, Science and Technology department. His research examines the theory that the reason the Earth's climate has been warming up for the last 400 years is because of increasing activity from the sun in the form of solar flares. Professor Lockwood agrees that it is undeniable this has been the case until very recently, but he theorises that there has been sufficient divergence between the sun's activity and the current increase in global warming that the discrepancy can only be explained anthropogenically.

One of Professor Lockwood's earlier articles made use of UK Solar System Datacentre (UKSSDC) data. This article looks at how the increased activity in the geomagnetic field surrounding the globe can be ascribed to the increased activity of the sun's corona

13 This research could not have been conducted without the assistance of the data centre and its records.

14 Through this research, the degradation of collagen in dinosaur bones could become – in theory – a proxy indication of climate.

by affecting the Interplanetary Magnetic Field, which then impacts on the earth's geomagnetic field. To do this Lockwood made use of the geomagnetic aa index available through UKSSDC. The index was originally assembled from the observations of the Greenwich and Melbourne magnetometer stations.

Professor Lockwood is aware that UKSSDC has many images and written records that could still be digitised and put online but that UKSSDC currently lacks the necessary technology and resources. He says that one of the things he has learnt from his work in climate change is that 'every single data point is valuable because you can't predict what use it may be put to in the future[15] and you can't go back and repeat a measurement.'[16]

One of the other practical reasons why Professor Lockwood had made full use of UKSSDC data was simply their service orientation and the fact that they are willing to supply the data he needs in the format he requires. He uses a programme called MATLAB, which requires spaces rather than colons to separate the data. He explains how at one point some of the data he accessed came with minus signs between the dates. But when he contacted UKSSDC they happily converted the data for him.[17] He also explains that although he could himself assemble from original sources the data he has used, it would take an enormous amount of labour.

15 Research practices change, and data may have new uses in the future. Data centres ensure that researchers will still have access to historic data for such new investigations.

16 Data in some fields can never be recreated, and the data centre is crucial in ensuring such research is preserved and accessible.

17 The support of data centre staff adds value to the data that they supply.

7. Wider impact

An original aspiration of this research project was to indicate any areas where there may be broader benefits from research undertaken using data centre resources. This proved rather difficult. We asked survey respondents:

> Can you identify examples of wider impacts (e.g. on society, the economy, policy, etc.) that have resulted from your research, where data accessed from the data centre has played a significant role? (please briefly explain).

Since most respondents were still at an early stage in the research which involved data centre resources, it was hard for them to identify wider impacts. As a result, many answers suggested areas where researchers considered their work might eventually have impact, or simply stated that it was too soon to say anything concrete. Even researchers who had finished their work found it difficult to identify solid examples of impact, because it is hard for them to trace the different ways in which their findings have been used by people with whom they are not directly collaborating.

Nonetheless, some tentative conclusions can be drawn from the information collected via the surveys and case studies about impacts upon wider society and different types of enterprise (public, private or third sector), through various types of output – from new methodologies to products or services. The following sections illustrate some of these impacts – and, indeed, some impacts may legitimately be considered significant.

Impact tends to be focused on fields which relate closely to the subject area of the centre. So, for example, impact identified by users of the BADC tended to relate to environmental issues:

> My research using BADC data has helped direct conservation efforts for NGOs working to protect two different bird species.

> Government have decided not to fund research into Amplitude Modulation Noise from wind farms as this is a relatively minor noise problem within the wider context of noise annoyance. The MIDAS wind data statistics of stations close to a wind farm were very important to make an estimate of the severity of the problem at that specific wind farm.

Researchers using ESDS resources, on the other hand, tended to have an impact on social policy:

> I have carried out research into road casualties, and this has had a major impact on road safety policy in the City of London.

> A body of work from the academic community on gendered impacts of pension system design over many years using archived data has led to major pension reforms which should improve retirement prospects for women. The latter-day focus on gender issues within government has been entirely research led.

Users of the other three data centres made fewer mentions of identifiable impact – partly because of the smaller number of respondents to these surveys. However, CDS researchers tended to notice impact in medical fields, while ADS impact was around cultural policies, and NGDC impact primarily around the environment and building or planning. There was relatively little crossover between data centres, which suggests that interdisciplinary research using data centres is not particularly widespread.

Researchers mentioned impacts which affect the public, private and voluntary sectors. The public sector dominates the small sample of responses generated by the survey, but this is probably because the ESDS and BADC offered significantly more examples of impact than users of other centres and, as we have noted earlier, policy-makers are an important target for users of these two centres. The small number of responses given by CDS and NGDC users seemed to indicate that business could be an important audience for their users. For the CDS this is probably due to the medical applicability of some of their research, while the BGS (whose employees formed the majority of respondents to the NGDC survey) offers consultancy to the commercial sector as part of its service.

Impact can be grouped into three broad categories:

- New tools and methodologies
- New policies and regulatory controls
- New products or services

New tools and methodologies

New tools and methodologies are observed when the data centre has contributed to research which helps other organisations improve the way they plan their work. As distinct from the types of project which have been categorised as 'new policies and regulatory controls', these research projects do not suggest a conclusive way of 'doing things better'. Rather, they give organisations a better way to understand their own data and make specific improvements to the way they prioritise their services. As the following quotes show, this occurs in both public and private sectors.

> It has helped develop models of vegetation productivity used in decision support tools for settling upland grazing regimes.

> I have used the BADC data (in combination with DEFRA sheep density data and CEH land cover data) to assess the risk of sheep scab, a disease in sheep, in the UK. Such models can be used to better target disease management programmes where they are needed most – in the hotspot areas.

> My research on poverty dynamics based on the BHPS (British Household Panel Survey) provided a 'template' for what is now undertaken by DWP in their own published statistics.

> The risk group at Lloyd's of London has shown interest in modelling of landfall for tropical storms.

Case study – new tools and methodologies

Improving cancer care through more robust clinical trials

In 2008, the National Cancer Intelligence Network (NCIN) and the National Cancer Research Network (NCRN) began to explore the potential for using the National Cancer Data Repository (NCDR) to improve the quality of results obtained through clinical trials. This would have a positive impact on the translation of research to practical patient care.

Clinical trials are very costly and time-consuming. Such constraints can mean that trials have to work with quite significant gaps in their evidence base. This research set out to test whether such gaps can be filled by using routinely-collected data and data from other trials, held in the NCDR.

18 Data from the NCDR offers new possibilities for research, with important clinical implications.

For one trial, the NCDR allowed a further comparison (not possible in the original study[18]) of the trial population with the general population and showed some significant differences on several important dimensions. The trial population had small biases on gender (more males), on age (more younger patients) and on the stage of illness (a greater proportion of people with earlier stage cancer). Together, these indicated that the trial was likely to have understated mortality rates by 5-10% after five years, in comparison with the average for the population as a whole.

These preliminary results suggest that the NCDR has potential to inform clinical trials generally, through (i) backfilling missing data – particularly with respect to longer-term, follow-up data; and (ii) through permitting quantitative adjustment of trial results for any selection biases evident in the trial population as compared with the general population. They also suggest that a more general move to using NCDR data to support clinical trials should produce benefits by encouraging more consistent implementation of data collection protocols[19] and analytical treatment.

19 The data centre encourages standardisation of data collection, making it easier to aggregate data for this kind of work.

New policies and regulatory controls

A number of projects described by respondents have supported regulatory systems, policy decisions and in some cases influenced legislation. This has happened at local, regional, national and even international levels.

- My studies on the impact of weather / climate on the dynamics of freshwater lakes have proved of interest to regulators working on a European as well as national level.

- Some of my research on contamination of private water supplies in Scotland used rainfall data from BADC, and this research is now embedded in legislation.

- Influencing the GLA London Plan by giving an insight into population, employment and transport interactions.

One response highlights the challenge faced by anybody attempting to track the wider 'impact' of research activity and the difficulty of establishing direct causal links:

- Not sure how much policy impact my research has had – research carried out with an ESRC grant on childhood disability and household circumstances has attracted interest in the DCSF [Department for Children, Schools and Families, now Department for Education] but unsure if policy has been influenced.

Once the research has been completed, the researcher may have some sense of where it has raised interest, but following this through to evidence of actual impact is much more challenging.

Case studies – new policies and regulatory controls

UK Energy from Renewables

Graham Sinden has researched the future development of wind, wave, tidal and solar energy resources. He also looks at issues surrounding the integration of electricity from these resources into larger electricity networks, including the implications for system security and costs associated with development of large-scale renewable energy sources.

Dr Sinden's research involved simulating energy output from different renewable resources for up to 34 years. He then validated these results by comparing predicted and actual renewable energy generation over a ten-year period and on an hour-by hour basis. The aim was to find a mix of renewable technologies to provide the best match between electricity supply and demand patterns. He used 34 years of hourly data from up to 66 different UK sites, all of which came from the BADC (together with additional data on wave and tidal resources from the BODC). Dr Sinden believes that his research would have not been possible without the added value that came from BADC[20] through their efforts to put multiple inputs of data together and to convert them in a structured and usable body.[21]

As a result of this research, Dr Sinden demonstrated that substantial reductions in stand-by generating capacity requirement can be achieved by implementing renewable energy systems in a planned way. He gave both written and oral evidence to the House of Lords Select Committee on Science and Technology where he emphasised that providing diverse sources of renewable electricity would smooth the rough edges of variable generation. The findings were accentuated in his subsequent "Wind power and the UK wind resource" report for the former Department of Trade and Industry.

His research supported subsequent studies that found that the UK has the largest potential for wind energy in Europe and contributed to a more informed perception about renewable energy which is a key part of the UK Renewable Energy Strategy[22] published by the Department of Energy and Climate Change in July 2009.

Virtual Worlds and legal aid entitlement

Graham Stark is the owner of 'Virtual Worlds', a small business that specialises in microsimulations of the economy; mathematical models of individual behaviour which can be manipulated by economists to understand how different policies would affect outcomes. He recently build a microsimulation of legal aid eligibility to inform changes to the Scottish and Irish legal aid entitlement system.

20 The data centre adds value to the research data, and this can be critical to the success of research.

21 Reframing, restructuring and combining datasets is an important function of the data centre.

22 The research has had a direct influence upon strategy development.

Mr Stark used the Family Resources Survey (FRS) dataset from the ESDS to construct an estimate of how many people would be entitled to legal aid. He could have made an estimate of Scottish and Irish income levels without the FRS by using other datasets, but the accuracy of his results would have suffered considerably [23] and the questions in alternative surveys would not have been as relevant. He then used data from legal aid payments in Scotland and the Northern Irish Crime survey to understand which groups of people are most likely to apply for legal aid, and applied this to the estimates derived from the FRS to develop rules about how we might expect this population to behave.

23 Data centres ensure that researchers can access the data that they need to get the highest quality results, particularly important where policy decisions rest upon the outcomes.

Both the legal aid models were used to inform policy and contributed to a restructuring of the Scottish and Irish legal aid entitlement systems.[24] In Ireland, the system was simplified to make it easier for people to establish their eligibility. In Scotland, a graduated system was introduced to allow for both full and partial support, enabling up to a million more people to have access to legal aid.

24 Research made possible by the data centre has changed policy, and potentially helped a very large number of people.

New products and services

There was less evidence of new products and services, although in some cases the role of the data centre was very important, as in the example of research into ageing shows.

- Ageing research has huge societal impact. One collaborator was able to source a compound (via my search on ACD) that he could not find in usual catalogues, that is now the lead compound in a cure for premature ageing.
- Research feeds into interventions to reduce climate change impacts, e.g. Department of Health heat-health warning system.
- Improved Scottish flood prediction service using Met data for Loch Linnhe area.

The Hidden Structure of Inorganic Azides

Michael Cartwright, recently retired from Cranfield University, was commissioned to look into the toxicity of sodium azides, which are used to inflate airbags. This research led Dr Cartwright to examine the sensitivity of inorganic azides.

Although data on crystal structures are available from other sources, Dr Cartwright maintains that the labour involved in assimilating all the crystal structures he required in one place would have been prohibitive[25] and he would not have undertaken the project without access to CDS. He found the search engine very accessible because all that he needed to do was to specify that he wanted compounds comprising of nitrogen and any metal, and all of the known permutations of these structures were immediately available. He also appreciated the function that displays an illustrated version of the crystal structure when a formula is highlighted.[26]

Inorganic azides form volatile compounds that are easily triggered to decompose and release nitrogen, which is the inert gas used in airbags. Although sodium azides were popularly used in airbags and aeroplane chutes, other inorganic azides are frequently used in detonators and explosives, and are increasingly being used in mining and quarrying.[27] Their sensitivity as explosives is thus an important practical issue for those handling them.

Combining databases to accelerate drug discovery

Recent developments in molecular biology have begun to focus attention on 'cell choreography': the interactions between macromolecules and other biological and chemical molecules, to understand how these interactions activate or block metabolic pathways (for example leading to illness or disease). For historical reasons, there is less information available in the public domain about how macromolecules interact with chemical entities (e.g. drugs) than with their biological equivalents. This has limited opportunities for research.

The ChEMBL team at the EBI aims to plug this gap and provide an indispensable public resource for drug discovery and development. It seeks to identify genes that are affected by drugs, and link them with chemical databases which hold information on those drugs or chemicals. This allows researchers quickly to identify which existing drugs or chemical entities affect genes of interest by searching through a highly annotated database. ChEMBL adds value by extracting information and drawing on multiple resources and

25 The data centre brings data together and allows research to go ahead that would otherwise be impossible.

26 The data centre adds value to the data itself by providing supplementary services and links to other important information sources.

27 Research enabled by the data centre has commercial application.

datasets[28], enabling a 'systems view', providing a published history of all existing modern day drugs that have potential to work on new and novel targets, with links to other relevant data.

ChEMBL enabled recent research into potential drug targets and candidate drugs for the parasite Schistosoma mansoni, which is responsible for the tropical disease schistosomiasis. It affects 210 million people in 76 countries, and leads to chronic illness that can damage internal organs and, in children, impair growth and cognitive development. Other examples include identifying the role of statins (a class of drugs that lower cholesterol) in combating hay fever, and the use of Viagra against Chrohn's disease.[29]

This new approach can short-circuit the need for many years of costly research and development[30] for new drugs. Given its relative efficiency and speed, it may hold out the promise of therapeutic breakthroughs in many other neglected areas. It also provides a practicable strategy for dealing quickly with emerging pathogens by providing insights into existing or new candidate molecules.

28 The data centre combines different sets of information in an intelligent way to support researchers.

29 Research made possible by the data centre has positive medical outcomes.

30 The data centre saves time for researchers, and for the research process as a whole.

53

8. Conclusions

Users and usage

Usage of research data centres is high, and the services provided by data centres are highly valued by researchers in all fields within the study, most of whom consider the data they access to be 'very' or 'quite' important to their research. The trends for increasing holdings and usage are encouraging, and suggest that researchers are using centres not just as sources for new data, but as places to deposit and share the work they themselves create. This is confirmed by the relatively high levels of data deposit among data centre users.

Users are overwhelmingly from the academic community, employed either in higher education institutions or research institutes. However, where the data held in the centre has relevance outside academia – in local government planning, for example, for the ADS – this is reflected in the user profiles. A similar story holds true for the intended audiences of data centre users' research. Most users are writing primarily for a scholarly audience, but where there could be an interested external audience – in public policy, for example, for the ESDS and BADC – this is recognised by researchers.

The types of data accessed, and the uses to which it is put, relate closely to the holdings offered by the data centre. For each centre (except the NGDC) there is a clear dominant type of use, but also evidence of some use within each category we defined. This suggests that data centres are meeting a range of needs, but that they may have certain areas where additional services could be closely targeted, in order to satisfy the biggest group of their users.

Benefits to research and wider impact

There were high levels of agreement across all data centres with most of the statements about research benefits. Benefits to do with research efficiency were the most widely supported, with researchers mentioning ways in which the centres had saved them time, money and effort. Benefits to do with research quality related both to the quality of their own work, and the quality of the data that they access from the centre in order to undertake such work. In both cases, the data centres are perceived to add quality. Researcher training was more important in some centres than others.

Benefits relating to research novelty received a more mixed response. There was some agreement that data centres opened up new types of research question, but the detailed qualitative evidence for this is rather patchy. One possible explanation may be the way that academics understand research questions to develop. Many respondents reported that data centres had enabled them to answer questions which had always existed in their minds, but which would have been too expensive, time-consuming or just impossible to explore without the resources provided by the data centres. In one sense, these questions are not 'new' – they may be a long-pondered facet of a much bigger research problem. However, the data centre makes it possible to engage with them, so in that sense they are encouraging new research.

Clearly, then, researchers believe data centres have improved their ability to undertake high-quality research. The research is better in the sense that it is more rigorous, more thorough, and wider-reaching in its conclusions. This, we would expect, has made it more likely that such research would have greater social, economic and academic impact. The case studies outlined here demonstrate how research made possible by data centres has gone on to have an impact on wider society and the economy, through the development of new tools and methodologies, new policies and regulatory controls and new products or services. Overall, evidence of such benefits was fairly limited, but this is due at least in part to the problems inherent in attempts to capture wider impact.

Characteristics of research data centres

The qualitative evidence showed very clearly that the benefits which flow from use of research data centres are closely related to the characteristics of these centres. It is not just about making research data available – although this is clearly an important and non-trivial precondition to further exploitation. Data centres provide many other services – and encourage many types of behaviour – which are crucial to achieving the research benefits and wider impacts that this study has identified. In some cases, these are the result of conscious policies; in others, they are side-effects of collecting, storing and managing vast quantities of data.

Many researchers mentioned the value of having a large number – and wide range – of datasets in one place. This encouraged them to explore work that might be peripheral to their original interests, but which sometimes turned out to be important. Furthermore, the links that many data centres made between their holdings and a wider literature were highly valued by a number of researchers.

Other benefits relate more to the quality of the data, and many respondents framed their comments on this in contrast to the likely outcomes if researchers were left to manage their own data. The quality of the data was felt to be higher because researchers knew it would be submitted to a data centre. Equally, the data centres were able to ensure good metadata, which enabled researchers to enjoy the efficiency benefits that they identified.

The technical support provided by data centres was cited by many as an important research benefit. This may range from providing ad-hoc help with small glitches to collating and reformatting entire data series in order to meet a researcher's specific needs. Some researchers also mentioned the value of training offered by the data centre, saying that this was important for the next generation of students.

Finally, several researchers mentioned the value of having a national data centre with an international reputation. This enabled them to make connections with data creators in other countries, opening up data sources that they would otherwise have been unable to access. So the data centre provides a way into datasets beyond those that are held in its own archives. Others suggested that having a widely-accessible UK source of data may influence the design of research carried out in other countries, thereby enabling cross-comparability and perhaps giving the UK a competitive advantage when it comes to such activities.

9. Strategic implications of findings

1. Data centres are clearly a success story, as far as their users go. This report has found evidence of benefits for research in a number of ways and evidence of broader impacts most obviously, but not exclusively, in areas relating to social and environmental policy. Funders and policy-makers should continue to support and promote existing national data centres.

2. The reference function of data centres and services is important. Similarly, users look to data centres as a source of high quality trusted data. Both these functions of data centres are highly valued and should be maintained and encouraged.

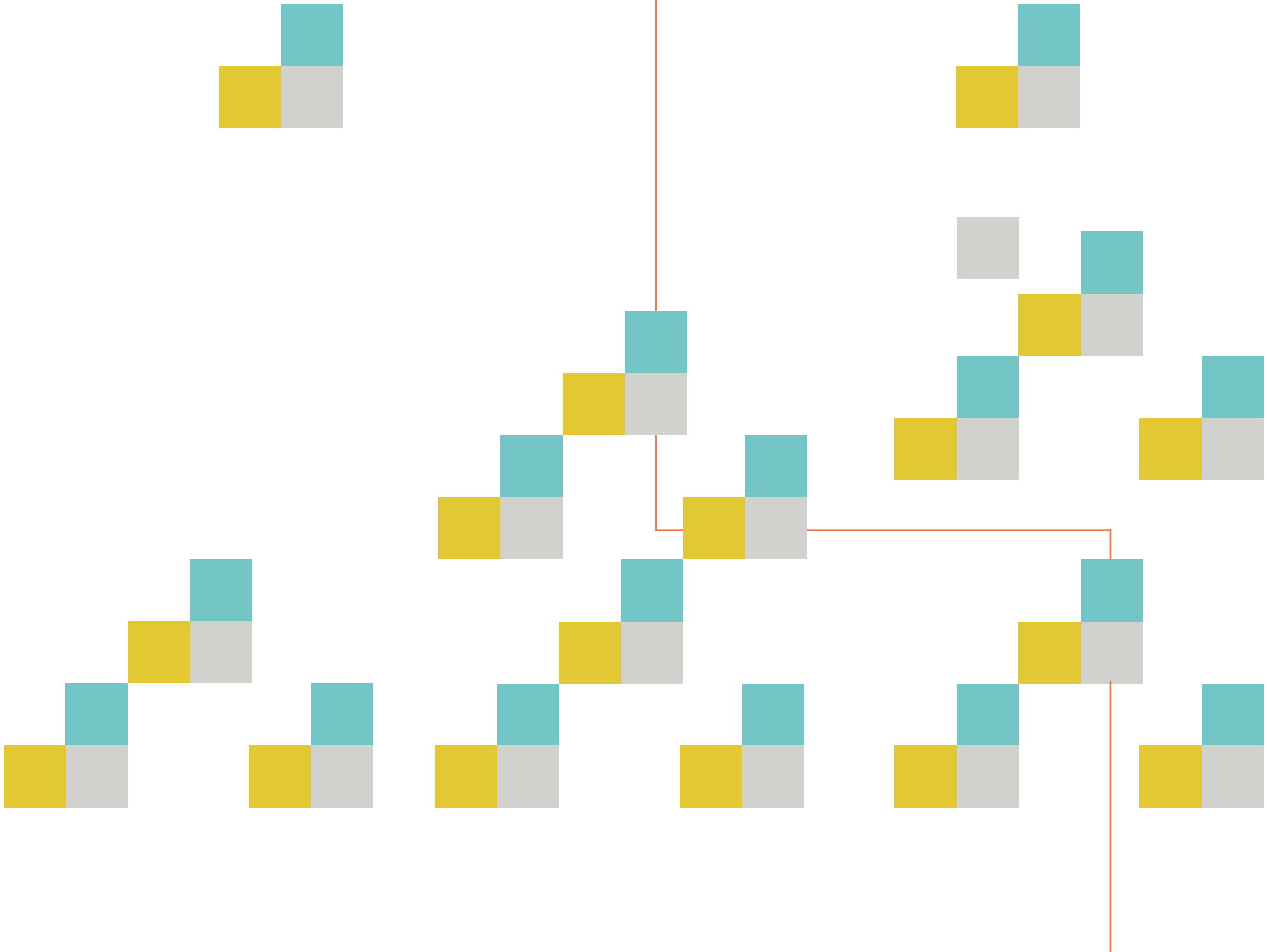
3. There is a clear value in aggregating a large number of data sets by subject or discipline, but many of the benefits that are associated with data centres rely upon the availability of data centre staff to manipulate, interpret or support use of these data sets. This implies a case for resourcing such data centre activities; it would be worthwhile to accumulate further evidence to support this case.

4. It is important for data centres to understand and support their users, but also to promote the work that they are doing and the value that they add to the research process. For this reason, centres should continue to collect information about users and usage for advocacy and planning purposes.

5. While deposit levels are promising, researchers need to be encouraged to deposit data. It may not be enough for Research Councils to apply strict rules about deposit; as this study has shown, even where there are strong funder mandates, this may not bring high levels of compliance across the discipline as a whole. Encouraging researchers to deposit their data raises a number of questions that have not been tackled within this project: most particularly, the capacity and funding which is available to support the curation and storage of an increased volume of deposits. Nationally and internationally, there are a number of initiatives exploring how best to support research data management 'upstream' of deposit to data centres. The outcomes of these initiatives should be tracked closely and factored into any consideration of how to improve deposit rates.

6. If – as is widely asserted – the grand challenges of modern research require interdisciplinary collaboration, then data centres are likely to be required to play an important role. However, the present study found relatively little evidence of interdisciplinary research resulting from the role of data centres. Improving facilities for data discovery across data centres would seem a necessary first step towards supporting greater levels of interdisciplinary research.

7. The national data centres which are the subject of this report cover only a part of the broader research landscape. If disciplinary data centres are valuable – as this study suggests – is there a case for 'backfilling' such services to support other disciplines and sub-disciplines? Resourcing expansions of the data infrastructure will prove challenging. Initiatives are likely to proceed, in response to need, at local, national and international levels. In such circumstances, better understanding and coordination are needed of the relationship between local, national and international provision of data curation and aggregation services.



References

The ANDS Technical Working Group (2007). *Towards the Australian Data Commons: A proposal for an Australian National Data Service.* Canberra: Australian Government Department of Education, Science and Training.

Beagrie, N., Beagrie, R. and Rowlands, I. (2009) 'Research Data Preservation and Access: The Views of Researchers' in *Ariadne* 60.

Grant, J., Brutscher, P.-B., Kirk, S., Butler, L. and Wooding, S. (2009). *Capturing Research Impacts: A review of international practice.* Cambridge: Rand Europe.

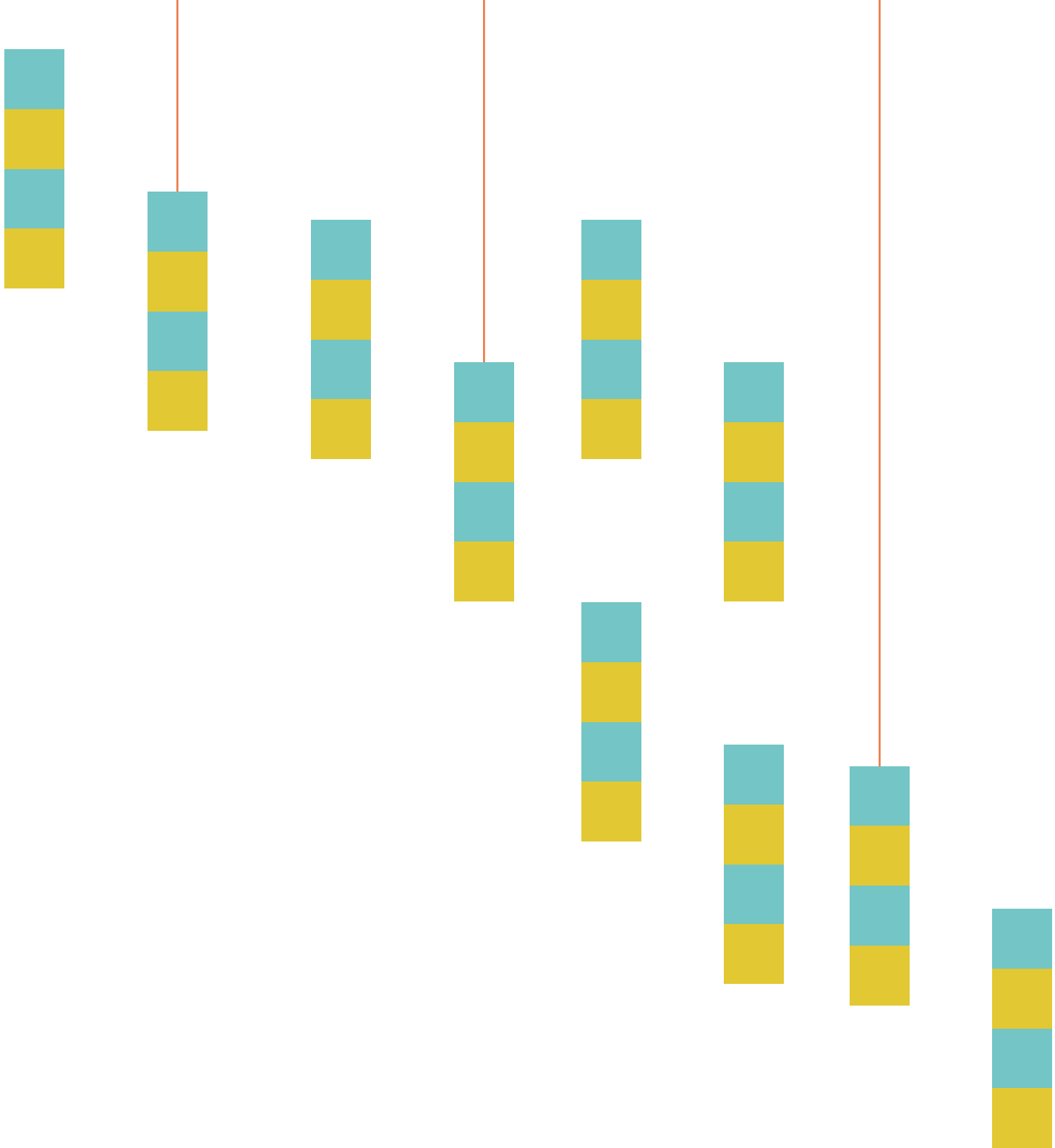
Hey, T. and Trefethen, A. (2003). 'The Data Deluge: An e-Science Perspective'. In Berman, F., Fox, G. and Hey, A., Eds. *Grid Computing – Making the Global Infrastructure a Reality* pp. 809-824. Chichester: John Wiley & Sons.

Lievesley, D. and Jones, S. (1998). *An Investigation into the Digital Preservation Needs of Universities and Research Funders.* Bath: UKOLN.

National Science Board (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Arlington, VA: National Science Board.

Research Information Network (2008a). *Stewardship of digital research data: A framework of principles and guidelines.* London: Research Information Network.

Research Information Network (2008b). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs.* London: Research Information Network.



The Research Information Network has been established by the higher education funding councils, the research councils, and the national libraries in the UK. We investigate how efficient and effective the information services provided for the UK research community are, how they are changing, and how they might be improved for the future. We help to ensure that researchers in the UK benefit from world-leading information services, so that they can sustain their position as among the most successful and productive researchers in the world.

The Research Information Network
96 Euston Road
London NW1 2DB

+44 (0)207 412 7946
contact@rin.ac.uk
www.rin.ac.uk