Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes

Jenny Fry, Suzanne Lockyer and Charles Oppenheim
Department of Information Science
Loughborough University

John Houghton and Bruce Rasmussen
Centre for Strategic Economic Studies
Victoria University, Melbourne

November 2008

CENTRE FOR STRATEGIC ECONOMIC STUDIES

Department of Information Science

Loughborough University

Contents

Summary

It is becoming increasingly clear that effective and efficient management and reuse of research data will be a key component in the UK knowledge economy in years to come, essential for the efficient conduct of research and its dissemination and use. In recognition of this, there have been many calls for access to science data at national and international levels. JISC and other UK funding bodies have developed a number of initiatives concerned with the management and curation of research data. The report by Lyon (2007) was pivotal in delineating the issues that need to be addressed and this project aims to take forward Recommendation 30: *JISC should work in partnership with the research funding bodies and jointly commission a cost-benefit study of data curation and preservation infrastructure.* The project's objectives are to:

- Identify the benefits of curating and sharing research data;
- Identify a methodology by which to estimate the benefits to UK Higher Education and the UK more generally of curating and openly sharing research data produced by researchers in UK HE;
- Use the methodology, as far as possible, to derive an estimate, expressed in financial terms where possible, for the identified benefits;
- Document case studies and examples of data re-use, where that re-use led to tangible benefits.

Potential benefits of the open sharing and re-use of research data include: maximised investment in data collection; broader access where costs would be prohibitive for individual researchers/institutions; potential for new discoveries from existing data, especially where data are aggregated and integrated; reduced duplication of data collection costs and increased transparency of the scientific record; increased research impact and reduced time-lag in realising those impacts; new collaborations and new knowledge-based industries.

Broader indirect benefits might include transparency in research funding, use of data sets in education to enhance data awareness of students, enhanced researchers' skills through access to a broader range of data, tools and standards have potential to increase data quality, and increased visibility and promotion of institutions and researchers.

The project used a mixed method approach, including a literature review and qualitative case studies to inform the development of a model on which to build a business case for data sharing in UK Higher Education. The case studies investigated were the European Bioinformatics Institute (EBI) and Qualidata, which is part of the Economic and Social Data Service. The case-studies were based on semi-structured interviews with service providers and users of the service. The interviews were supported by documentary evidence in order to identify and illustrate the benefits and costs for the different stakeholders.

Benefits may accrue in a variety of ways, including cost savings, efficiency gains, and new opportunities to create value through doing things in new ways and doing new things. These are, successively, more difficult to quantify: not least because they often emerge over time and can only be realised in the future. We present a simple example of cost-benefit analysis applicable to an individual dataset or repository, based on costs and potential cost savings. It describes the data requirements and walks the reader through the process step-by-step. The approach is then extended to explore the more diffuse benefits of data curation and sharing at the institutional and disciplinary levels.

The recommendations of this study address three key areas:

Recommendation 1 – Baseline reporting

A key finding of this research is that there is, as yet, no standardised and consistent system of reporting of the data necessary to make a business case. Therefore, we recommend:

- The development of guidelines for data collection and reporting through consultation with stakeholders, taking account of the need to minimise the reporting burden.
- Further classificatory work to explore how costs and benefits differ according to institutional and disciplinary factors such as intellectual field, objects of research, data types, analytic techniques and approaches.

Recommendation 2 – Model Questionnaire

This project has focused on identifying the data necessary to make a compelling business case for data curation and sharing. In doing so it has provided a foundation for the development of a model data collection framework that could be further developed. We recommend that this is taken forward by:

- The development of a model questionnaire building upon the questions outlined in Section 9.2 information sources. This could then be combined with an extended version of the 'Beagrie model', which would capture repository cost data.
- In order to reduce duplicative effort in building business cases we recommend that JISC host a web based survey/data gathering instrument, and invite repository/data centre management staff and users (from both deposit and withdrawal sides) to use the instrument for reporting purposes.
- Public dissemination of such a survey could be confidential and anonymised by aggregating repositories according to the key institutional and disciplinary factors identified as part of recommended guidelines for baseline reporting.

Recommendation 3 – Developing a community resource

In order to achieve an empirical and scalable evidence base upon which policy makers and funders can evaluate benefits at different levels of granularity, e.g. across types of repository/centre or discipline, a system of consistent recording and reporting needs to be developed. Given the differences in practices and types of re-use across disciplines that this study has highlighted this system would need to be implemented in a culturally sensitive way. We recommend that:

- The centralised collection and collation of data resulting from the development of guidelines for baseline reporting and participation by community members in the model questionnaire be made available as a shared community-level resource and that;
- Such a resource should stipulate what basic core data might be collected and reported annually.
- The model might include collection of the following data in a consistent way: annual acquisitions (data submitted, data accepted), annual usage (downloads, requests), citations, external funds received, annual spend (split across main budget headings).

1 Introduction

It is becoming increasingly clear that effective and efficient management and reuse of research data will be a key component in the UK knowledge economy in years to come, essential for the efficient conduct of research and its dissemination and use. In recognition of this, there have been many calls for access to science data at national and international levels.

There have been a number of development projects and related work concerning the management of research data, for example the JISC's Digital Repositories Programme (JISC 2005) and the scoping of a set of principles for the stewardship of research data (RIN 2008a). Lyon (2007) prepared a review of data management in the UK, which presented a number of recommendations. One recommendation was for a cost-benefit study of data curation and preservation infrastructure. Following this report, the JISC commissioned a suite of projects to take forward the recommendations. This project aims to address Recommendation 30: *JISC should work in partnership with the research funding bodies and jointly commission a cost-benefit study of data curation and preservation infrastructure.*

1.1 Aims and objectives

The aim of the project is to identify the benefits of the curation and open sharing of research data, using quantitative and qualitative methods.[1]

The project's objectives are to:

- Identify the benefits of curating and sharing research data;
- Identify a methodology by which to estimate the benefits to UK Higher Education and the UK more generally of curating and openly sharing research data produced by researchers in UK HE;
- Use the methodology, as far as possible, to derive an estimate, expressed in financial terms where possible, for the identified benefits;
- Document case studies and examples of data re-use, where that re-use led to tangible benefits.

The purpose is not to present a definitive answer or benefit/cost ratio, but rather outline one or more examples of costs-benefits in order to give some guidance to those preparing a 'business case' for institutional and/or disciplinary data curation and preservation.

1.2 Method

The project used a mixed method approach, including a literature review and qualitative case studies to inform the development of a model on which to build a business case for data sharing in UK Higher Education.

[1] By 'research data' is meant the evidence base on which academic researchers build their analytic or other work, where this evidence base is typically gathered, collated and structured according to declared and accepted protocols.

1.2.1 Literature Review

The aim of the literature review was to identify the elements of cost and benefits associated with data sharing in UK Higher Education. As research data are heterogeneous, a key feature of the literature review was to highlight the differences in disciplinary needs and practices with regard to data curation and sharing. The literature review, therefore, provided illustrative examples of reuse and the views of stakeholders in various disciplines as reported in the literature, including informal dissemination channels such as websites and blogs.

1.2.2 Case studies

The case studies investigated further the issues raised in the literature by examining data activities in two contrasting disciplinary areas. The focus of the case studies was established data centres serving:

- The bioinformatics community (European Bioinformatics Institute, EBI), and
- The social science community (ESDS (Economic and Social Data Service) Qualidata).

The criteria for including these data centres was:

- Data are researcher generated;
- Data are held in established, domain specific data centres which have an open access policy, and provide added value services to their research communities;
- These centres provide contrasting examples, and so demonstrate disciplinary differences; and
- The JISC recommended EBI and UK Data Archive (UKDA) for inclusion in the study. It was decided to focus on Qualidata (within UKDA), for the reasons outlined above. Although other data centres were suggested, the timescale of the project restricted the scope.

Each case study used semi-structured interviews with service providers and users of the service, supported by documentary evidence, to identify and illustrate the benefits and costs, for the different stakeholders. Contact was made with the director of each service and their suggestions followed for further provider interviews. For EBI, researchers are located at EBI and so contact was made via service provider interviews. For Qualidata, emails were initially sent to recent ESRC award holders (identified from the ESRC website), but no response was received. To facilitate face to face interviews contact was made with Social Sciences Department and associated research institutes at Loughborough University, which have an international reputation in their areas of research. The Qualidata website was used to identify established users. An interview schedule was designed for each of the three categories of interviewee, to guide but not constrain the interview:

- Service provider
- Data depositor
- Data user (download/reuse)

Copies of the schedules are given in Appendix 2.

1.2.3 Framework for making a business case

Whatever the motivation for preserving and sharing research data it is important to have a solid grasp of the costs involved and commitment required. However, “an appeal to the Newtonian vision of ‘standing on the shoulders of giants’ may fall short of what is needed to make a persuasive case for adding these costs to already-strained budgets” (Beagrie *et al.* 2008, p16). To make such a case it is necessary to examine the costs-benefits.

It is always more difficult to identify and quantify benefits than costs. Benefits may accrue in a variety of ways, including cost savings, efficiency gains, and new opportunities to create value through doing things in new ways and doing new things. These are, successively, more difficult to quantify: not least because they often emerge over time and can only be realised in the future.

An obvious starting point is to begin with the most direct and directly measurable, namely cost savings. This entails extending the coverage of the ‘Beagrie Model’ for data repository costing to more fully cover the costs faced by users, be they ‘depositors’ of data to the repository or ‘withdrawers’ of data from it.

In Section 4 and Appendix 3, we present a simple example of cost-benefit analysis applicable to an individual dataset or repository, based on costs and potential cost savings. It describes the data requirements and walks the reader through the process step-by-step *(Example I)*. The approach is then extended to explore the more diffuse benefits of data curation and sharing at the institutional and disciplinary levels *(Example II).*

2 Benefits

There are many reasons to support the curation and sharing of research data, ranging across a continuum from the altruistic through to the pragmatic. The benefits of curation and open sharing of research data include, *inter alia*:

- Opportunities afforded for wider access than has been typical with less structured and informal channels for data sharing, such as access for researchers outside the core HE and public sector research networks, for researchers in industry, government and non-government organisations, thereby enabling greater cross sectoral collaboration as well as considerable opportunities within education and training.
- Opportunities for use and re-use of data, including reduced cost of collection and duplication, sharing the direct and indirect costs of collection (e.g. avoiding survey fatigue and thereby improving response rates), new uses unforeseen at the time of collection and data mining opportunities.
- Opportunities to create a more complete and transparent record of science, with implications for improved detection of fraud and plagiarism, easier assessment and peer review at the grant application and assessment, publication and research evaluation stages.
- Opportunities to better align research evaluation with what researchers produce, by providing recognition for a wider range of 'outputs' and contributions than is typical in current publication-centric evaluation programmes.
- Opportunities to raise the visibility of researchers, repository host institutions and funders, by linking them to valued resources.
- Opportunities for the emergence of re-use 'industries' in particular areas of research and observation, as has happened with geospatial, meteorological and oceanographic data, etc.
- Opportunities for the emergence of support and service 'industries', focusing on providing value adding products and services that enable easier storage, discovery and access to datasets.

A comprehensive identification of the potential benefits of more open access to scientific and scholarly publications has been presented by Houghton *et al.* (2006; forthcoming), and many apply to the curation and sharing of research data. A number of authors have identified the benefits of data curation and sharing, as well as outlining the barriers to be overcome (for example, Beagrie *et al.* 2008; Ball *et al.* 2004, the Joint Standards Study, 2005; RIN, 2008, Carlson and Anderson, 2007 and David, 2006). A review of the literature, illustrating the costs-benefits from the perspectives of different disciplines is given in Appendix 1.

3 Case studies

3.1 Introduction

The aim of the case study phase of the project was to provide qualitative and quantitative evidence of data sharing in different disciplines. Interviews with service providers and users of established data centres in bioinformatics and social sciences illustrate attitudes to data sharing, as well as examples of reuse. Quantitative material, for example staff and infrastructure costs, was obtained from documentary evidence including Annual Reports and information provided by the data centres directly.

Although limited in scope because of the timescale of the project, the case studies provide a rich picture of data sharing needs and practices in these two contrasting disciplines, illustrating that any cost-benefit analysis must consider the cultural as well as the financial elements.

3.2 ESDS Qualidata

ESDS was established in January 2003 with the aim of developing resources for social science research and learning in UK Higher and Further Education. The service is funded by the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC), initially from 2003 – 2007, with extension of funding to 2012 following a successful mid-term review in 2005.

ESDS is a distributed service based on collaboration between four centres: UK Data Archive (UKDA), Institute for Social and Economic Research (ISER), Manchester Information and Associated Services (MIMAS) and the Cathie Marsh Centre for Census and Survey Research (CCSR). ESDS has two central functions – Management and Accessions and Preservation. In addition it has four specialist divisions providing targeted value added services: ESDS Government, ESDS International, ESDS Longitudinal and ESDS Qualidata. The core activities of data acquisition, processing and preservation, as well as user registration are dealt with centrally. Qualidata provides specialist support for users and depositors of qualitative data or mixed datasets containing qualitative data. Qualidata is hosted by the UKDA at Essex University.

The purpose of the case study is to illustrate how the dynamics of costs/benefits that underlie data sharing play out in the social sciences. The case study used documentary evidence, primarily ESDS Annual Reports and documents provided by the Director of Qualidata, and interviews with:

- The Associate Director and Head of ESDS Qualidata to supplement printed documentation and obtain a personal perspective on the current activities and future plans of Qualidata
- Two researchers who had deposited data to Qualidata, as a requirement of their ESRC grant. Both had deposited one data set, and could be described as novice ‘users’ of the Qualidata service. Their respective areas of research were conversational analysis (using spontaneous conversations requiring highly specialised transcription), and women in science (mixed method – survey, and interview/focus group).
- One user of Qualidata datasets, this respondent had also deposited data and was an experienced user. His area of research was primarily social and cultural geography (use of oral and life history tapes and transcripts, interviews and other primary data collection).

Although the timescale of the project limited the number and scope of interviews, several key issues emerged surrounding the sharing of data in social sciences, not least the wide range of disciplines ESDS Qualidata and its parent centre must serve. The examples obtained from respondents serve to indicate the potential benefits, but also highlight the barriers to be overcome for a data sharing culture to evolve in the Social Sciences and the need for data to be collated that would relate to quantitative metrics. The full case study is given in Appendix 2 *(Section 8.2)*

3.3 EBI

The EBI hosts the major core biomolecular data resources in Europe. These include EMBL-Bank (nucleotide sequence), Ensembl (genomes), ArrayExpress (microarray-based data), UniProtKB (protein sequence and functional information), MSD (macromolecular structures) and InterPro (protein families, motifs and domains). The EBI is part of an International biomolecular information landscape and collaboration plays a significant role in its activities, including contributing to the public record of science and the development of data standards. In particular, it collaborates with the neighbouring Sanger Institute, in the delivery of specific value-added services. Furthermore, the EBI obtains a substantial percentage of its data from the Sanger Institute, which is why it is included in the case study to provide insights into data production costs. Other data centres that are part of the EBI's broader landscape are the National Center for Biotechnology Information (NCBI) hosted by the National Institutes of Health and the National Library of Medicine in the USA and the DNA Data Bank of Japan.

The purpose of this case study is to illustrate data sharing practices in the biosciences and provide examples of the benefits of data re-use. The case study is based on documentary evidence drawn mainly from annual reports and interviews with the following key personnel:

- The Associate Director of the European Bioinformatics Institute (EBI) to provide an overall picture of the role of the EBI in the biosciences and give a historic perspective of the creation of networked databases in the biosciences based on his experience establishing the EMBL data library.
- The team leader for Vertebrate Genomics at the EBI to give the researcher's perspective on working with raw sequence data and also insights into developing value-added services through the Ensembl Project, which is a software system that produces and maintains automatic annotation on selected eukaryotic genomes.
- The group leader for ArrayExpress at the EBI, again to provide both the researcher's perspective of using large-scale networked databases and the perspective of providing value-added services.
- The Head of Outreach and Training at the EBI to get an overview of personnel at the EBI and insights into skills training.
- The group leader for the Human Genome Analysis Group and team leader of The Ensembl project at the Sanger Institute to obtain a data processing perspective and the costs associated with it.
- The Head of IT at the Sanger Institute to obtain a picture of the infrastructure costs and establish what percentage of data production and processing costs are associated with infrastructure costs.

The examples obtained from participants serve to indicate the potential benefits of data sharing, but also highlight some of the unforeseen costs, including indirect costs, and potential impact on researchers' careers of making data available for re-use. In particular, the case-study highlights the benefits/costs at the institutional level relating to data sharing, rather than at the level of individual databases and services. Two specific value-added databases (Ensembl and ArrayExpress) and one research project (the 1000 Genomes Project) are highlighted in order to illustrate the potential benefits to the institution, the research community and individual researchers. A major benefit of a data sharing infrastructure in the biosciences, of which the EBI and the Sanger Institute play a pivotal role, is the centralised data integration function that enables researchers to search across different types of data and multiple data sets produced from thousands of individual investigations. Key factors in having a critical mass of searchable 'open' data are the data sharing policies of funding agencies, participation of journals in a data sharing culture, and the provision of value-added services such as standards, methodologies for processing large-scale data, development of search and analytical tools, and training. These latter value-added costs are borne at the institutional level, and measuring the direct benefits gained by these institutions is problematic not least due to the well-understood problem of time lags between expenditure and impact, but also because of the initial investment in infrastructure costs. The EBI case-studies show that the ratio between benefits/costs is reducing as services mature and usage increases exponentially. IT infrastructure is a significant component of service costs and whilst the cost per unit of computer storage may be reducing system requirements (e.g. storage and compute power) are increasing in order to keep pace with the increased rate of data generation and demands for effective analysis. The full case study is at Appendix 2 *(Section 8.3).*

3.4 Findings from the case studies

3.4.1 Qualidata

The ESRC is a major funder of research in the Social Sciences and requires its award holders to offer data to UKDA (qualitative data to ESDS Qualidata). However, a culture of data reuse is not fully established in the research community. One of the key roles for Qualidata is to develop and encourage a data sharing culture by improving methods of organisation and dissemination of data and by education and training.

Acquisition of data in this discipline is labour intensive, both during the original data collection stages and during data deposit. As a result, researchers feel a strong sense of ownership and a reluctance to share; this is compounded by what they view as extra work to prepare data for deposit. Furthermore, the manual processing required by Qualidata limits the number of datasets that can be disseminated. Other constraints are the confidentiality and consent issues which are a feature of much social science research: in 2007/2008, 30% of data sets offered were rejected because of confidentiality/consent issues. These factors limit the number and range of datasets and reinforce the reluctance to share and reuse data, as it is felt to be 'one way'. However, Qualidata undertake research activities, with additional grants, to develop methods of automating aspects of data processing, including methods of anonymising data. This will result in a greater number of data sets being accepted, increase processing rates, and also provide data in different media. Together with outreach activities, the increasing range and number of data sets available may encourage new users within the research community.

Therefore, this case study demonstrates that investment of both financial and human resources is required to reap any benefits from data sharing, and that such benefits may be delayed.

However, the study has also provided some positive examples of how data can be reused in the social sciences to give a fresh perspective on research questions.

3.4.2 EBI

The EBI's main funders are the European Molecular Biology Laboratory (EMBL; an intergovernmental organisation to which 20 member states and one associate member state contribute funds) and the European Commission. Within the UK, the EBI is mainly funded by the Wellcome Trust, the MRC and the BBSRC. It also receives funding from the US National Institutes of Health. All of the relevant UK funding agencies have open data policies, though the extent to which data sharing is actively supported varies. For example, since 2006 the MRC has stipulated that grant applications must include costed plans for preparing and documenting research data for preservation and sharing. The BBSRC also stipulates that all grant applications should include a data management plan or provide explicit reasons why data sharing is not possible or appropriate. The Wellcome Trust's policy on data sharing is underpinned by a philosophy of 'open science', but as Lyon (2007) points out it is difficult to assess the impact of the Wellcome Trust's data sharing policy on research.

From an individual researcher's perspective, the most effective incentives to share and re-use data are not necessarily those derived from funding policies. For example, popular value-added databases such as ArrayExpress are doubling every 14 months in terms of data acquisition. With over 50% of journals in the 'omics' and bioinformatics fields mandating that the underlying data need to be submitted to a specified data centre and conform to the kinds of data standards being developed by the EBI, it is highly likely that a researcher trying to publish an article in a journal will submit their data to a publicly accessible database.

Despite proactive data sharing policies within the biosciences, there are still uncertainties about building long-term infrastructure. It is recognised by the bioscience community that the infrastructure to support data sharing is difficult to fund and there are many questions around long-term sustainability of such infrastructure.

Overall, growth in data holdings at the EBI is exponential. The databases vary in scale and scope, and are having to cope with the submission of ever-larger datasets. For example, the 1000 Genomes Project[2], an international research consortium that includes scientists from both the Sanger Institute and the EBI, amongst other institutions, is producing

> *"somewhere between enormous and terrifying amounts of data".*
> (Team Leader Vertebrate Genomics, EBI, 2008)

The data management task alone associated with the 1000 Genomes Project is, according to the EBI, beyond the capacity of any individual researcher or research group.

The popularity of a particular database is often linked to its maturity. At the time of the case study, the fastest growing data type, in terms of accessions, was DNA sequence. DNA sequence is collected, curated and made publicly available by an international collaboration between the EBI, the NCBI and the DNA Data Bank of Japan. This joint collection acquires about three new sequences per second, twenty four hours a day, three hundred and sixty five days a year.

[2] Details can be found at http://www.1000genomes.org/page.php. Accessed 9th October 2008.

The data made publicly available through the types of value-added services developed by the EBI, and similar institutions across the world, appear to be at the heart of research within the biosciences. Users are not necessarily interested in the raw unprocessed data, and filtering data is a key aspect of data curation. It is the intermediary data, the traces or summary datasets that enable researchers to re-analyse the data. According to the Associate Director of the EBI, if the data were held in a standalone archive that mainly functioned as a scientific record, the data would be of very little value. It is the data integration task, coordinating data of different types and from different sources, that is important to users. This is because the databases are a composite of thousands of investigations and therefore afford new forms of knowledge to be created that otherwise would not be possible if the data were scattered across thousands of individual investigators.

In terms of the benefits of data sharing for institutions, the experience of the EBI indicates that those scientists or institutes that are linked with major data centres are the most highly visible amongst their scientific communities.

3.4.3 Convergence and divergence across disciplines

The two case studies had very different characteristics, in terms of the types of disciplines they represented, the users they supported, the types of data produced and curated, and the value-added services that are important to the communities they serve and that are feasible to provide.

In certain respects, therefore, these two case studies represent diverse examples. Rather than comparing like-with-like we have chosen two extreme examples to illustrate the points of divergence and convergence in thinking about ways in which benefits might be offset against costs and the disciplinary characteristics that underlie the interaction between costs-benefits.

Table 3.1 shows points of divergence and convergence which emerged from the case studies. These provide examples of the issues to be addressed in developing data sharing, and where barriers need to be overcome and benefits may emerge.

Table 3.1 Points of divergence and convergence for EBI and Qualidata

	EBI	Qualidata
	Divergence	
Data production	Born digital	Manual collection and processing
Data processing and cleaning	Automated and manual	Largely manual
Key value added services	Development of tools and standards	Outreach, training and re-presentation of datasets
Culture	Data sharing embedded in culture	Data sharing culture evolving; debates over value of reuse in some areas.
	Convergence	
Outreach and training	Provides skills for researchers which further establishes data sharing culture Provides skills for data scientists and curators with potential new industries in data management	
Reward and recognition	Requirement for method of ‘tracking’ data set use Production and deposit of data sets to contribute to academic reward and recognition structure Opportunities for increased visibility for researchers and institutions, leading to new opportunities	

4 A framework for making a business case

Whatever the motivation for preserving and sharing research data it is important to have a solid grasp of the costs involved and commitment required. The guidelines suggested by Beagrie *et al.* (2008) provide an excellent foundation for costs, but as Beagrie *et al.* themselves note:

> *"An appeal to the Newtonian vision of "standing on the shoulders of giants" may fall short of what is needed to make a persuasive case for adding these costs to already-strained budgets"* (Beagrie *et al.* 2008, p16)

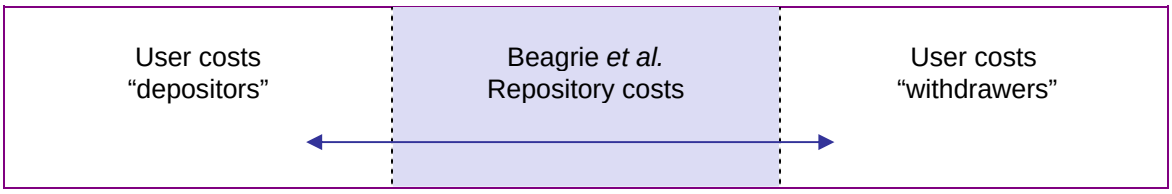
To make such a case it is necessary to examine both costs and benefits, and to that end we outline one possible approach to making a 'business case' for data curation and sharing that takes account of the very real limits to what is practicable *(see Appendix 3 for details)*.

4.1 Costs

Beagrie *et al.* (2008) explored the costs of curation of research data. Their study provided a detailed description of cost elements and an activity costing framework focused on costs relating to staff, equipment, travel, consumables, estate and indirect costs of establishing and operating a research data repository (*i.e.* full economic costing). The major activity elements identified related to three phases, namely:

- **Pre-archive** – a phase primarily relating to research projects in universities creating research data for later transfer to a data archive, in which implications for repository costs are considered and data collection/creation designed and implemented with curation and sharing in mind;
- **Archive** – a phase primarily relating to the acquisition/disposal, ingest, storage and management of data, but also expanding into the provision of access and user support; and
- **Support services** – covering administration, common services and estates.

Figure 4.1 Scope and coverage of costs



Source: Authors' analysis

The focus of the 'Beagrie Model' is on data repository operation, although there are clear interfaces and some overlaps with the activities of repository users, be they researchers who deposit data or those who withdraw it. Nevertheless, for the purposes of exploring benefits it is necessary to extend the 'Beagrie Model' to more fully include these user costs (*Figure 4.1*).

4.2 Benefits

It is always more difficult to identify and quantify benefits than costs. Benefits may accrue in a variety of ways, including cost savings, efficiency gains, and new opportunities to create value through doing things in new ways and doing new things. These are, successively, more difficult

to quantify: not least because they often emerge over time and can only be realised in the future. An obvious starting point is to begin with the most direct and directly measurable, namely cost savings.

Possible cost savings can be reaped by the users of data repositories, be they on the input side ('depositors') or the output side ('withdrawers'). Both will also face costs that must be factored into any cost-benefit analysis. The production-side costs of 'depositors' include the total costs of production of the datasets held in a repository, any additional research costs relating to researchers working-up the data for use by others, the application of standards and metadata above and beyond that necessary for the project itself, additional costs of anonymisation, etc. The user-side costs of 'withdrawers' are also important, with users facing costs in searching for, accessing and assessing data held in repositories, as well as reaping potential savings through reduced data collection demands and realising a range of wider potential benefits.

The costs of production of the data, its preservation and access, can also be used in estimating first order benefits, with one benefit of data repositories being that the data would not have to be re-created – making the data's production costs multiplied by the number of times they might have been re-created if not made available through curation and sharing one path towards estimating potential benefits.

4.3 Calculations and examples

In this section we present some hypothetical examples, which focus on the way that available information might be used to explore benefits/costs and build a 'business case' in the most simple and practicable way possible.

4.3.1 Example I: Cost savings

Here we suggest one approach to exploring the direct and indirect impacts of cost savings using a hypothetical example which is applicable to individual datasets and/or repositories.

Data requirements and sources

Data requirements are modest and relate to the major cost elements involved, including:

- **Depositor costs:** the costs of data collection/creation and preliminary preparation faced by the depositor (sourced through consultation with users who have deposited or are depositing data);
- **Repository costs:** the costs of preparation and storage faced by the repository management, including an allowance for a share of the overall repository costs (sourced from existing repositories and/or consultation with repository managers and based on the 'Beagrie Model' for full economic costing); and
- **Withdrawer costs:** the costs of search, discovery and access faced by the withdrawers and users of the data (sourced from repository users).

These costs can be set against the savings realised from the reduction or elimination of duplication of collection/creation costs achieved through data curation and sharing.

Step 1: Direct cost savings from an individual dataset or repository

Direct cost savings can arise from the use and re-use of data made available for sharing, as the costs of collection/creation are shared across multiple users and the value of the data realised

by each user. A simple calculation is to sum the costs and set them against potential savings, which can be done for individual datasets or a repository.

To take a hypothetical example in which:

- The cost of data collection/creation (faced by the depositor) is £200,000;
- The cost of additional preparation for sharing (faced by the depositor and/or repository) is £10,000;
- The cost of searching for, accessing and assessing the data (faced by the user/withdrawer) is an average of £2,500 per use;
- The annualised cost of storage and access provision for the data concerned (faced by the repository) is £10,000; and
- The data are used/re-used 6 times over the 10 year life-cycle for which they are stored.

Table 4.1 Example I: Use/re-use leading to research cost savings

Direct cost savings from data use/re-use		Value
Data requirements:		
Cost of data collection/creation (faced by the depositor)	C1	£200,000
Cost of any additional preparation for sharing (faced by the depositor/repository)	C2	£10,000
Cost of searching for and accessing the data (faced by the user/withdrawer)	C3	£2,500
Annualised cost of storage of the data concerned (faced by the repository)	C4	£10,000
The life of the data in years	L	10
Number of times the data are used/re-used over the life-cycle	N	6
Direct cost savings (Step 1):		
Direct Costs = C1 + C2 + (C3 * N) + (C4 * L)	DC	£325,000
Direct Benefits = C1 * N	DB	£1,200,000
Direct benefit/cost ratio = (C1 * N) / (C1 + C2 + (C3 * N) + (C4 * L))	DBCR	3.7
Indirect cost savings (Step 2):		
Effective additional R&D spending = DB – DC	ARD	£875,000
Additional returns to R&D from that spending @ 20% = ARD * 0.20	AR	£175,000
Indicative total benefits = DB + AR	TB	£1,375,000
Indicative total benefit/cost ratio = TB / DC	TBCR	4.2

Source: Authors' analysis.

The costs would be:

*£200,000 + £10,000 + (£2,500 * 6) + (£10,000 * 10) = £325,000*

The benefits would be:

*£200,000 * 6 = £1,200,000*

And the benefits would be almost 4 times the costs, calculated as:

£1,200,000 / £325,000 = 3.7

While hypothetical, this example shows that given the relative costs of research and curation, relatively low levels of use/re-use can justify the activity in terms of direct cost savings alone.

Step 2: Indirect cost savings from an individual dataset or repository

For the purposes of estimation it is reasonable to assume that the research costs saved would be spent on additional research, thereby effectively increasing R&D spending by the amount of the saving (i.e. doing more research for the same expenditure).

So in the example outlined above, R&D spending would, effectively, increase by:

£1,200,000 - £325,000 = £875,000

Returns to R&D vary considerably between fields of research, but if we assume that the R&D expenditure savings went back into the overall pot of public research funding, then at the conservative estimate of 20%[3] social returns would be worth an additional £175,000[4] calculated as:

*£875,000 * 20% = £175,000*

In our hypothetical example, this would increase the attributable benefits to almost £1.4 million over the life-cycle, calculated as:

£1,200,000 + £175,000 = £1,375,000

Thereby, lifting the benefits to more than 4 times the costs, calculated as:

£1,375,000 / £325,000 = 4.2

Due to the lag between research expenditure and impacts these indirect impacts are no more than indicative *(see footnotes),* and it is important to note that the benefits accrue over time while many of the costs accrue up-front, and most costs accrue to producers and repository operators while many of the benefits accrue to external or third-party users.

4.3.2 Example II: Potential benefits

In addition to these cost savings, we note a number of potential benefits from data curation and sharing relating to collaboration and enhanced outcomes, better education and research training, new opportunities and uses, a more complete and transparent record of 'science', and greater visibility and reward. All of these impact the quality and efficiency of research over time.

However, impacts will vary from case to case and are by their nature diffuse and uncertain.[5] Nevertheless, in order to get some sense of the possible value of these potential benefits we explore a number of scenarios based on percentage point increases in returns to R&D expenditure at the institutional, disciplinary or sectoral levels *(see Appendix 3 for details).* These impacts will be in addition to those outlines in Example 1.

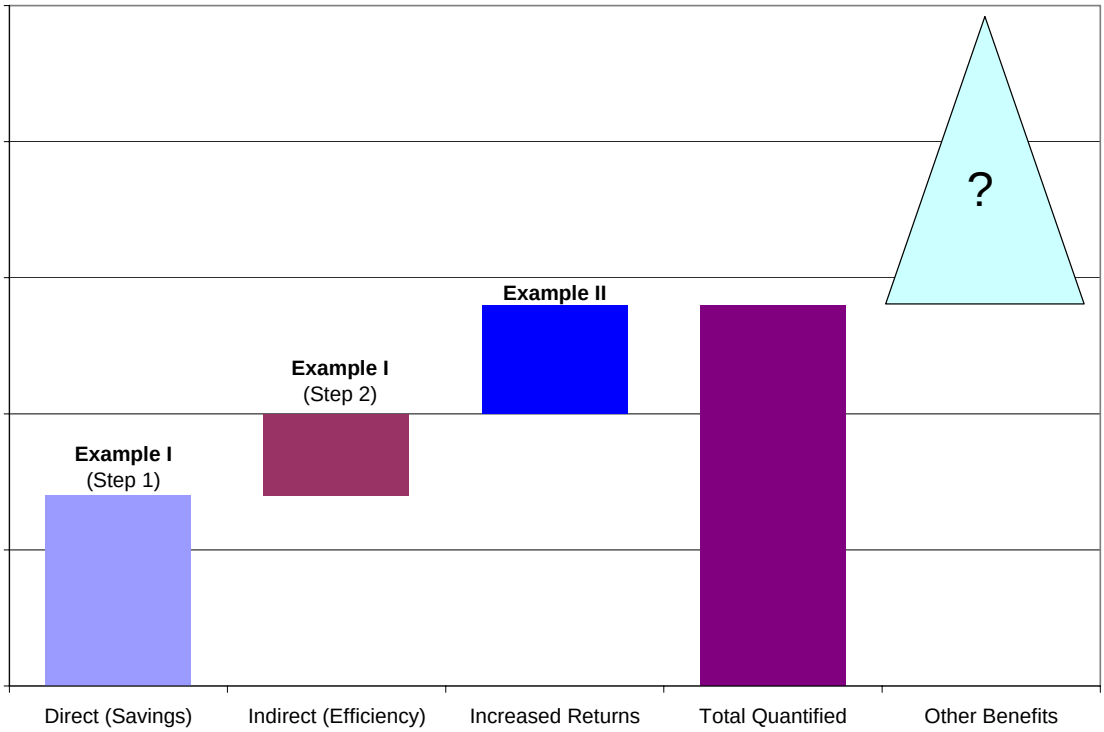
---

[3] In one of the most thorough summaries of the literature, Martin and Tang (2007) noted that studies spanning more than 30 years have found a rate of return to public R&D of between 20% and 50%. Therefore, 20% is a very conservative estimate of average social returns to publicly funded research.

[4] These are recurring gains, albeit lagged to account for the time between the conduct of research and its impacts. Such returns can be expressed in Net Present Value (NPV), lagged and recurring over the useful life of the knowledge. However, NPV calculations are sensitive to the discount rate applied and for the sake of simplicity and transparency we simply take the original value as indicative of the value of the returns *(see Appendix 3 for details).*

[5] Those exploring costs and benefits in order to develop a business case for a data repository may wish to go no further than the cost savings scenario outlined in Example I (above).

Figure 4.2 Types of benefits explored in examples



Note: Not to scale

Source: Authors' analysis

4.3.3 Limitations and caveats

There a number of issues to consider in the interpretation of these hypothetical examples, relating to the difficulty of attributing particular impacts to specific activities or expenditures, the variation in returns to R&D spending between fields and disciplines, and difficulties encountered when trying to account for inflation and the lagged and recurring nature of returns to R&D. These issues are discussed in Appendix 3.

5 Conclusions

A limitation of this study is that it was based on small-scale case studies that were limited to two diverse disciplines. It is difficult therefore to extrapolate on how typical the benefits/costs identified would be across a broader base of data repositories/centres and disciplines. It is clear, however, that there are some common issues that need addressing. Further comparative research could be based on like-with-like estimations e.g. based on scale and scope, discipline, data type, types of value-added services[6].

It is well understood by policy makers and funding agencies that measuring the impact of infrastructure resources, such as data repositories and centres, is a complex and challenging process. Units of measurement need to move away from enumerating outputs to more dynamic metrics that capture the qualitative enhancements that investment in these resources bring to science and, ultimately, society. Yet, as this study has found, systematic practices and systems for recording and reporting the type of data upon which robust and culturally sensitive metrics might be developed are not yet in place.

Current uncertainty around sustainability of research infrastructure such as data repositories and centres, for example the withdrawal of funding from the Arts and Humanities Data Service and rationalisation of STFC infrastructure services, means that the role of business cases in policy making and governance models for data services is likely to become more prominent.

We found there to be significant differences in practices relating to data curation and sharing. There was also variation in the scale and scope of services offered and levels of re-use between the two disciplinary case studies. Beyond these two case studies we anecdotally found there to be a lack of consistent recording and reporting both year-on-year within services and across services. This lacuna is a barrier to measuring the benefits of data sharing and makes comparative analysis across service types and disciplines problematic.

Despite these constraints the evidence supports an argument in favour of enhancing and promoting data sharing. The benefits are not just short or medium term, such as economies of scale, time saved, saved duplication and faster access to a broader range of data, but also longer term such as new discoveries, new collaborations, and efficient and rapid knowledge transfer.

[6] Raivo Ruusalepp of the Estonian Business Archives Consultancy has done some preliminary work in this area based on broad categories such as sources of funding (used as an indicator for national/international remit and disciplinary grouping), Higher Education Institutions and individual projects.

6 Recommendations

6.1 Recommendation 1 – Baseline reporting

A key finding of this research is that there is as yet no standardised and consistent system of reporting of the data necessary to make a business case for data curation and sharing. Therefore, we recommend:

- The development of guidelines for data collection and reporting through consultation with stakeholders, taking account of the need to minimise the reporting burden.
- Further classificatory work to explore how costs and benefits are differentiated according to institutional and disciplinary factors such as intellectual field, objects of research, data types, analytic techniques and approaches (the types of value-added services required by constituent user communities), and any other key dimensions that make a material difference to data curation.

Intended outcome

- Consequently, Higher Education Institutions and other organisations evaluating possible approaches to research data curation could:
- Use the framework proposed by Beagrie et al (2008), or a subset thereof, to estimate their costs, and then draw on the worked examples from the recommended classificatory study.
- Based on these existing worked examples and worked scenarios service providers could scope their user communities, e.g. the field(s) of research and data types that they might expect, and easily identify the kinds of impacts and benefits that might be expected from relevant examples.

6.2 Recommendation 2 – Model questionnaire

This project has focused on identifying the data necessary to make a compelling business case for data curation and sharing. In doing so it has provided a foundation for the development of a model data collection framework that could be further developed. We recommend that this is taken forward by:

- The development of a model questionnaire building upon the questions outlined in Section 9.2 information sources. This could then be combined with an extended version of the 'Beagrie model', which would capture repository cost data.
- In order to reduce duplicative effort in building business cases we recommend that JISC host a web based survey/data gathering instrument, and invite repository/data centre management staff and users (from both deposit and withdrawal sides) to use the instrument for reporting purposes.
- Public dissemination of such a survey could be confidential and anonymised by aggregating repositories according to the key institutional and disciplinary factors identified as part of recommended guidelines for baseline reporting.

Intended outcome

- The survey could be developed as a community resource for building business cases by providing a spreadsheet model and linked website where dynamic context specific scenarios and examples could be generated.
- It might contribute to a broader suite of resources relating to data and repository management and would complement existing tools, such as the Data Audit Framework currently funded by the Digital Curation Centre and the JISC.
- Furthermore, It could be a source of information about key drivers and barriers of interest to JISC internally and others, to guide activities.

6.3 Recommendation 3 – Developing a community resource

In order to achieve an empirical and scalable evidence base upon which policy makers and funders can evaluate benefits at different levels of granularity, e.g. across types of repository/centre or discipline, a system of consistent recording and reporting needs to be developed. Given the differences in practices and types of re-use across disciplines that this study has highlighted this system would need to be implemented in a culturally sensitive way. We recommend that:

- The centralised collection and collation of data resulting from the development of guidelines for baseline reporting and participation by community members in the model questionnaire be made available as a shared community-level resource and that;
- Such a resource should stipulate what basic core data might be collected and reported annually.
- The model might include collection of the following data in a consistent way: annual acquisitions (data submitted, data accepted), annual usage (downloads, requests), citations, external funds received, annual spend (split across main budget headings).

Intended outcome

- Reduction of the burden on individual data repositories and centres to devote resources to building business cases from scratch.
- Consistent reporting across data repositories and centres and estimates of benefits on the basis of ‘like-with-like’ aggregated data.

7 Appendix 1: Literature Review

7.1 Introduction

Networked digital resources are increasingly central to knowledge creation practices. The increasing centrality of data is not restricted to 'big science' disciplines alone, such as high energy physics, genomics, and bioinformatics. In some areas of the social sciences, such as geography, researchers are carving out the Internet as a virtual social science laboratory (Batty, 2005) and in the humanities classics research has been revolutionized by combining the three-dimensional virtual rendering of classical sculptures with powerful image-based search tools that allow comparison of digital objects across entire archives. Such shifts are revolutionising disciplinary landscapes.

The collection, organisation, analysis, management and dissemination of data is an essential part of the research process and the role of data sharing and re-use has been the subject of debate in recent years, both within disciplinary communities themselves and amongst policy makers. For example, a report for the science research councils and the DTI (Department of Trade and Industry) in 2005 investigated the issues surrounding data sharing, based on several case studies of data sharing. The report recommended high-level actions to overcome the current barriers to data-sharing, noting that difficulties in sharing have a considerable cost. The problems which emerged from the case studies included absence of standards and tools, confidentiality and consent issues and lack of support for researchers. Recommendations required co-ordination between funders, researchers and data repositories to ensure data management plans became a part of the funding application, development of standards and tools, clear goals for data centres and addressing career structures for researchers.

More recent reports have focussed on specific aspects of data sharing such as preservation costs across different types of repository or data centre (Beagrie *et al*, 2008), the roles and responsibilities of different institutions (Lyon, 2007) and principles and guidelines for the governance of data sharing (RIN, 2008a). The costs and barriers to data sharing are generally agreed, as are the potential benefits, not only from reports, but also from formal and informal literature in several disciplines and these are discussed below. Somewhat underplayed are the disciplinary differences in attitudes to data sharing. The importance of recognising and understanding these differences is reflected in a number of studies recently funded by the Research Information Network (RIN) and the JISC. For example, a study of data management practices in several diverse disciplines noted the importance of acknowledging these differences (RIN, 2008). An ongoing RIN project aims to investigate researchers' use of information resources, including analysis and dissemination of data.[7] The DDC (Digital Curation Centre) SCARP project is investigating disciplinary attitudes to data sharing, preservation in 10 case studies.[8]

7.2 Role of data (disciplinary differences and similarities)

Knowledge is not a homogenous whole; rather it is constituted of heterogeneous disciplinary communities. Different intellectual and social concerns shape the role of data and the practices associated with them. In discussing the factors that may influence the costs and potential

[7] RIN Disciplinary Case Studies, expected completion May 2009 – see http://www.rin.ac.uk/case-studies

[8] DCC SCARP project – see http://www.dcc.ac.uk/scarp/

benefits of sharing data in an 'open' way[9], the fundamental nature of data in research should not be overlooked.

Lyon (2007) provides an interpretation of the research life cycle adapted to e-research. This clarifies the various stages, but also highlights the fact that the early stages of research – when ideas are formulated and tested – are quite independent of recent developments and discussions on open access and/or e-science. Researchers have *always :*

- formulated ideas;
- tested by observation/experiment;
- preserved data for validation;
- communicated results, leading in turn to new ideas and hypotheses.

Data are therefore essential and this is not new. However, there are considerable disciplinary differences in the types of data, scale of collection, and methods of analysis, which have led to different attitudes to sharing. The RIN carried out an extensive survey within eight disciplines/sub disciplines to investigate attitudes to, and awareness of, various aspects of data sharing (RIN, 2008). The report found a strong tradition of sharing in astronomy, genomics and classics, but a weak sharing culture in social sciences, and concluded that disciplinary differences should not be ignored in any initiatives to promote data sharing.

Although researchers in some social science fields, such as poverty studies, use large-scale datasets collected nationally for quantitative research, the deposit and re-use of qualitative data is in an earlier stage of development (Moore 2007). In the social science literature, the usefulness of secondary analysis is often brought into question (Moore, 2007; Parry and Mauthner, 2005). One explanation posited for the reluctance to share is that researchers do not believe other researchers would be interested in their data given the specificity of the research context in which it was probably created (RIN, 200*8).* This perspective contrasts with the Classics, where there is a 'grass roots' motivation to share data given the scarcity and fragility of the object of research. The value of data as a teaching resource is also recognised in the Classics (RIN, 2008). In fact, the re-use of data for research often goes in hand with teaching and this has been reflected in the Cyberinfrastructure programme in the U.S. (Borgman, 2006). The urgency for addressing data management issues and affording data sharing are most obvious in disciplines where advances in technology have led to data being created on an industrial scale – what Hey and Trefethen (2003) have coined as the 'data deluge'. These technology centric disciplines, such as high-energy physics and genomics, are producing data at an increasing rate due mainly to global collaborations.

CERN (European Organization for Nuclear Research) was founded in 1954 and is one of the world's largest scientific laboratories. Experiments involve the collaboration of researchers worldwide, for example the ATLAS experiment involves 1700 collaborators in 144 institutes in 33 countries. (Wouters and Reddy 2003).

The scale of this experiment and the Large Hadron Collider was described by Peter Murray-Rust:

> *"This is a "wow" experience – although I "knew" it was big, I hadn't realised how big. I felt like Arthur Dent watching the planet-building in the* Hitchhiker's Guide to the Galaxy*. It is*

[9] Fry, Schroeder and Den Besten (forthcoming) observed a stratification of openness in practice and the need for more nuanced understanding of openness at the level of policy making.

> *enormous. And the detectors at the edges have a resolution of microns. I would have no idea how to go about building it. So many thanks to Salvatore and colleagues. And it gives me a feeling of ownership. I shall be looking for my own sponsored hadron (I've never seen one). So this is "Big Science" - big in mass, big in spending, big in organisation."* (Peter Murray-Rust, from his blog January 2008. http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?m=200801)

One way in which the 'data deluge' is being addressed is through the development of long-term information infrastructures, such as those being developed under the auspices of e-Science programmes in the UK and Europe (more commonly referred to as e-Research in the UK), or the Cyberinfrastructure programme in the US. Investment in these infrastructures is significant and often supported by both government and industrial funding. In the UK, e-Science was initially spearheaded by the Core Programme[10], which was a common effort between the Department for Trade and Industry (DTI), the Engineering and Physical Sciences Research Council (EPSRC) and other UK Research Councils to provide a major funding push to kickstart a UK e-Science initiative (Fry, Schroeder and Den Besten, forthcoming).

The data deluge has been driven by the rapid increases in computer power and memory and communication bandwidths. However, to make use of the data, Hey and Trefethen (2003) argue that access, integration and curation of data are as important as storage and computing facilities.

The traditional research cycle can thus be expanded (Lyon 2007), to include data management functions and beyond this value added functions such as data linking, annotation and visualisation. A step further leads to new knowledge creation and it is from these latter stages that a number of benefits arise as data are put to new uses beyond their original context of creation. Realising such potential benefits requires input both from researchers and policy makers. An initial outlay is, however, essential to maximise effectiveness and the scale varies between disciplines.

Data sharing in the biosciences is well developed, due in large part to the data sharing policies of funding bodies and in particular as a legacy of the Human Genome Project. In fact, some areas of research would not exist without the evolution of data sharing practices. Ball *et al* (2004) illustrate this with the example of GenBank, which grew from holding 606 gene sequences in 1982 to over 30 million in 2003. Holding such data in one place and in the same format has made possible comprehensive genomic sequence analyses. The availability of such large-scale data sets has also facilitated the development of tools, such as base-calling software, and tools for determining error rates. This has improved the quality of data produced by researchers. Working to agreed standards and using established tools facilitates collaboration between widely dispersed groups. However, Ball *et al* (2004) note that to be fully effective, sharing of high-throughput data, such as gene sequences, requires a 'robust, useful and dynamic informatics infrastructure', and warn against apparently cost-effective 'quick and dirty' approaches.

In the biosciences, additional drivers for an established data sharing infrastructure include the high volume of data and the large number of users, often in widely dispersed collaborative

[10] Details of which are available at: http://www.epsrc.ac.uk/ResearchFunding/Programmes/e-Science/default.htm

groups. There is also a culture which sees data sharing as 'fundamental' (Ball et al 2004, Calvert and Williams 2008). However, in the qualitative social sciences views of data sharing are somewhat polarised (Parry and Mauthner, 2005; Moore, 2007) and a sizeable body of literature has built up about reuse of data, which is rooted in sociological methodology and theory, and for data sharing highlights the need for advocacy which is targeted to disciplinary needs.

In sharp contrast to Ball et al's view of data sharing as 'a fundamental tenet of good science', Parry and Mauthner (2005) question whether there is any demand for qualitative data. They suggest that the number of data sets held by Qualidata is merely the result of ESRC requirements to offer datasets and not a function of demand. Indeed, the 'injunction' to offer data is itself seen as a reason for resistance to data sharing (Moore, 2007).

The 'desirability and feasibility' of secondary analysis of qualitative data has been questioned, one of the main challenges being the risk of decontextualisation as well as the research value of archived data (van den Berg, 2005). Gillies and Edwards investigated the barriers to re-use, noted as being 'epistemological, methodological, practical and ethical' (Gillies and Edwards, 2005). The root of these problems lies in the fundamental difference between quantitative and qualitative data: quantitative data exists independently from the researcher, but qualitative data depends on interactions between the researcher and interviewees. This relationship is evident throughout the data collection process and perhaps explains some of the reluctance to share. Much researcher effort is expended in the pre-collection and collection stages, and arranging participants and designing data collection instruments are viewed as part of their intellectual input, strongly steered by the research question *(see Qualidata case study, Appendix 2, Section 8.2)*.

However, advocates of qualitative data re-use consider the debate to have moved on from whether it is possible to reuse to 'how to'. For example, Qualidata has developed requirements for additional contextual information to accompany datasets. Although Parry and Mauthner (2005) dispute that provision of such information overcomes barriers, they do admit a role for reuse by the researchers who originally collected the data. Here there is a similarity in purpose, and benefits, to the sciences where ' rarely do researchers exploit the full potential of high through-put data sets upon initial publication' (Ball *et al* 2004). This highlights the need for data curation, irrespective of who may have access, and activities are required by the researcher as well as data centres. For example, a key concern in data sharing is confidentiality and ethics, and issues of consent need to be addressed at an early stage in the project – 75% of Qualidata rejections are because of consent or confidentiality issues *(see Appendix 2, Section 8.2)*. In the qualitative social sciences anonymising transcripts for deposit is more stringent than that done for use by individual researchers or teams. It is also argued that removing identifiers also removes context reducing the usefulness of the data, as well as being a very time consuming process (Thomson *et al*, 2005). Qualitative data is usually recorded and then a transcript produced. Patzold (2005) argues that this is no longer necessary, as software is available which allows non-sequential access. Original audio sources may be preferred by some researchers, for example linguists, discourse analysts and oral historians. However, anonymising audio requires more than removal of any content, as voice may also be identified.

For the biosciences and some fields of research in the qualitative social sciences, data is 'reused' by carrying out further analysis or investigating new questions. However, in some areas it is beneficial to integrate data sets to improve the reliability and validity of results. This is often

seen in medical research where individual studies into a certain condition cannot recruit enough subjects (Belmonte, 2007). Integration of data sets from several studies has advantages but also barriers. One of the studies in the DCC SCARP project is of the Neuroimaging Group at the Department of Psychiatry, University of Edinburgh. The group aims to address data storage and curation needs by assessing risks and developing tools and standards to address these risks.

Curation challenges in neuroimaging

Neuroimaging focuses on finding neurobiological explanations of psychiatric disorders by means of MRIs and fMRIs. However, the amount of data collected by individual institutions is relatively small and integration of data sets can increase the reliability of research findings. However, there are barriers: different equipment is used at different centres, subjects are recruited from different populations, and different scales are used for symptoms. In addition, although data can be readily reused, it may be misinterpreted without sufficient context. So there is a requirement for standards. To address some of these issues, the Neuroimaging Group at Edinburgh have:

- developed scripts to remove identifiable parts of images to increase shareability: previously some sharing was limited to collaborative groups as images could potentially be identified using face recognition software.
- contributed to scanner in homogeneity correction so data from different MRIs can be pooled
- started to develop an ontology of psychosis symptoms to bridge scales used in different centres.

As a further step in integration, the data available includes not only scans but other information such as:

- social and economic classification
- family history/life events
- alcohol and drug use
- clinical and behavioural data in diagnoses and case history, psychiatric assessment, IQ and other cognitive tests
- genetic data

The benefits of these developments include new and more reliable analyses, however they also increase metadata requirements, termed ‘provenance metadata’, and so highlight that for the full benefits of data sharing to be realised certain costs must be met.

Whyte, A., Job, D., Giles, S. and Lawrie S. (2008), Meeting curation challenges in a Neuroimaging Group, *The International Journal of Digital Curation* 1(3), pp.171-181

7.3 Data sharing policies

The amount of literature on data sharing reflects not only the technological advances, but also the social, economic and political changes. Policies and principles surrounding open access to research data were developed nationally and internationally, including the OECD Principles and Guidelines, which noted that :

> *Sharing and open access to publicly funded research data not only helps maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research.* (OECD, 2007)

In the UK, the guidelines issued by the RIN echoed the OECD :

> *Ideas and knowledge derived from publicly funded research should be made available and accessible for public use, interrogation and scrutiny, as widely, rapidly and effectively as practicable* (RIN, 2008a)

Policies on open access reflect the increasing call for accountability in the use of public funds. The Research Councils are striving to demonstrate the economic impact of the research they fund (RCUK, 2007 and 2007a) and commissioned a study to investigate impacts across disciplines (PA Consulting, 2007). From 18 case studies, the findings suggest impacts can be grouped into four headings:

- development of human capital
- business and commercial
- policy
- quality of life

Data sharing may contribute to these impacts in direct and indirect ways:

- by increasing skills – both for researchers and for data scientists, which has benefits for academia and industry;
- by making outputs available for exploitation in new ways;
- by increasing speed and efficiency of research by cutting down on repeated data collection – this has particular potential in the biomedical sciences.

Although data sharing is not specifically mentioned in the broad impact studies, individual research councils may include data management initiatives in their own evaluations. For example, a report for the ESRC recommending ways of evaluating its impact suggested using intermediate outputs, including data sets, in evaluation as ultimate impacts were time delayed and difficult to measure, particularly in the areas of research undertaken by ESRC (Frontier Economics, 2007). The evaluation model presented demonstrated how funded activities such as research and training fed into outputs which were used by the public and private sectors and by academia. These ultimately led to outcomes such as higher GDP, improved quality of life and higher earnings. The report also indicated that it was possible to collect evidence of the usage and value of data sets. The impact of data sets is a relevant metric as ESRC has required award holders to offer data to UKDA for more than a decade, greatly predating the RCUK position statement of access to research outputs (RCUK, 2006). In 2005, RIN undertook a study of researchers funders policies concerning research outputs (RIN 2007). At that time, only AHRC, ESRC and NERC mandated deposit, but MRC and BBSRC had policy reviews in place. The BBSRC policy was published in 2007, following consultation with the community. It provides useful information for researchers, in addition to outlining policy, for example background, an implementation guide and examples of resources (data centres) for researchers to deposit and access data (BBSRC, 2007). The policy requires applicants to include a statement on data sharing in proposals. The data sharing policy is targeted at data from high

volume experimentation and low throughput data from long time series or cumulative experiments. It encourages deposit of other data where there is 'strong scientific need and where it is cost effective'.

A recent cross-council initiative summarises the benefits of data sharing and emphasises the need for data management throughout the project life cycle. The Rural Economic and Land Use Project (RELU) notes that:

> *"Data are a valuable resource. When managed and preserved properly, they can be used and re-used for future scientific and educational purposes. Sharing data facilitates research often beyond the scope of the original research, encourages scientific inquiry, avoids duplicate data collection and provides resources for training."*
> *(http://www.data-archive.ac.uk/relu/reluadvice.asp)*

To promote good data management, RELU has a Data Management Policy and has also established a Data Support Service based at UKDA *(see also Appendix 2, Section 8.2.6)*.

Not all disciplines are served by a national data centre such as ESDS/UKDA, so there remains a role for individual institutions to manage data produced by their research programmes. The JISC-funded Data Audit Framework (http://www.data-audit.eu) aims to raise awareness in higher education institutions of the need for data curation strategies and provides an online tool to assist with data management activities.

A scoping study at Oxford University revealed the diversity of practices within the institution (Martinez-Uribe, 2008). Researchers, heads of departments and technical/administrative staff were interviewed about their data management practices and needs. The results revealed variable activity across the University, with large projects generally having a high standard of data management, but small scale projects dependent on individual researchers' knowledge and goodwill. The respondents felt they had no or very little support and were often dependent on 'known contacts' within the University. The top requirements for support were for producing a data management plan, storage for data, infrastructure for publication and preservation and funding to support better practices within departments. Although respondents agreed that publicly funded projects should make data available, the majority had not done so. The barriers were those noted elsewhere, including problems with ethical clearance, the extra work required to deposit, and feeling of ownership of the data.

The Oxford study formed part of the feasibility study for a UK Research Data Service, a HEFCE funded project to investigate current provision in data management and the costs of a national data service for UK higher education (SERCO, 2008). The study included a survey of researchers and follow up interviews, which revealed requirements for a data service:

- advocacy: promotion of data sharing benefits and information for researchers
- co-ordination of existing repositories
- local provision issues, such as preservation and storage
- skills and training requirements for data management
- upgrading data collections, including selection criteria.

Although investigating a service for UK higher education, the interim report also draws on international initiatives, and notes that although the UK is currently in a good position with

regard to research data management, this is a fast moving area and it is essential that the UK keeps pace.

7.4 Who benefits?

Funding policies tend to stress the 'public good', as in the example below from the Research Councils UK (RCUK):

> *Ideas and knowledge derived from publicly-funded research must be made available and accessible for public use, interrogation and scrutiny, as widely, rapidly and effectively as practicable.* (RCUK Position Statement 2006)

In practice, however, there are several beneficiaries, and it is arguable these ultimately feed into benefits for society, albeit indirectly.

7.4.1 Researchers

Surveys of researchers concerning open access tend to focus on published outputs. However, Swan (2008) included a discussion of researchers' concerns in a recent study for the JISC. From interviews and focus groups, it emerged that researchers have difficulty finding relevant data resources and using data sets, because, for example, access may be denied, or manipulation requires skills and/or software. This emphasises the need for an infrastructure to address access issues, provide value added services, and provide training.

How far such problems have prevented researchers using or depositing data would require a wider ranging survey. Interim results from the UKRDS study (SERCO, 2008) show that 36% of researchers in the institutions studied (Leeds, Leicester and Bristol Universities) were not aware of grant or legal requirements to retain data. Although the majority did share data, this was mainly an informal arrangement within teams. However, 43% did use data centres (presumably to download data) and 19% shared data (presumably deposited). This information is based on an interim report, so full details of the survey (number of potential respondents, response rate etc) are not available.

A survey of three Australian universities (Henty *et al*, 2008) investigated researchers' data management activities. This survey found 8.6% willing to share openly, 44% via negotiated access, 6.4% after the end of a project and 2.3% only some years after the end of a project. This tends to support the UKRDS survey, which found informal sharing common amongst researchers, who wish to keep some control of their data.

Further evidence for informal practices comes from a study of data management practices at Oxford University. This used interviews and a workshop to obtain the views of researchers, heads of departments, technical and administrative staff across all divisions (social sciences; maths, physics and life sciences; humanities and medical sciences) (Martinez-Uribe, 2008). The study found that most researchers shared data, but usually in an informal way, for example by email or website (usually protected). However, there were problems in sharing large-scale data sets or sensitive data requiring special permission. It was not common practice to deposit in a domain specific archive, and reasons given included the extra work required and feelings of ownership. However, the importance of data and the expense involved in collection were recognised and there was agreement that data produced from publicly funded research should be publicly available. Also of relevance was the variance of data management practices across

the university and respondents expressed a need for support at all stages of the project lifecycle.

The UKRDS study shows an uneven distribution between use and deposit of data. There is some cost in terms of effort and time to prepare data for deposit and this may be viewed as a burden where no re-use is envisaged. Unless deposit is mandated by funders, researchers have little motivation to deposit. This limits the data available for reuse and so tends to set up a vicious circle. A major barrier which is agreed by all academics is that data are not part of the formal academic reward and recognition system which typically relies on journal publications. Although data centres request that use of data sets is acknowledged, there is no method of tracking 'citations'. Such a method would benefit researchers, but also provide important metrics for data centres to demonstrate impact.

7.4.2 Academic communities

There is evidence of informal sharing amongst research teams *(see above),* but the benefits of data sharing become apparent when data is shared beyond teams. This can result in new collaborations and new research questions being applied to existing data. For example, the *Journal of Applied Developmental Psychology* published a special issue in 2007 to highlight how the Study of Early Child Care and Youth Development data sets had been used by a researchers to address a range of research questions not envisaged in the original study plan (Friedman, 2007). In the introduction to the issue, Friedman notes that when the study was initiated, data sharing was not part of the culture in psychology, and even at the time of publication it was still not 'highly valued'. The aim of the special issue, therefore, was to demonstrate the potential of secondary analysis.

Although data sharing has been more common amongst scientists, this has often been on an informal basis. Establishing a data centre to maximise sharing potential may bring benefits unforeseen at the time of inception.

The Protein Data Bank

The establishment of the Protein Data Bank began in 1970, until then macromolecular structure data had been exchanged among research laboratories using punched cards: a typical protein structure required more than 1000 cards. A central repository would, therefore, greatly facilitate the exchange of data. Informal meetings at a symposium in 1971 led to international and multidisciplinary co-operation to form the Protein Data Bank. In 1976 the PDB archive contained 23 structures and 375 data sets had been distributed to 31 laboratories, but the work of the service in the first decade was characterised by promotional activities to engage the community. The 1980's heralded rapid advances both in biology and in computing technologies resulting in growth of the database to its current size of over 54,000 structures, which is growing all the time. But as well as the number, the diversity of structures led to methods which allowed the structures of novel proteins to be determined and established the field of structural genomics.

From the 375 datasets distributed annually in the early years of PDB, daily file downloads now average over 200,000; in addition to the main PDB there are derivative databases that catalogue the data in different ways for ever more specialised applications.

Berman, Helen M. (2007), The Protein Data Bank: a historical perspective, *Acta Crystallographica Section A, Foundations of Crystallography* A64, pp88-95

The availability of large, good quality data sets is a valuable educational resource in both the sciences and social sciences. For example, the percentage of undergraduates using Qualidata has grown from 7% in 2003-04 to 22% in 2006-07 (ESDS, 2004 and 2007). CERN publish specially filtered data sets for educational purposes which allow students to work with original data (Wouters and Reddy, 2003).

Establishing a data sharing culture can also add to researchers' skills. Whyte *et al.* (2008) term this 'human infrastructure' and note how in the Neuroimaging Group at Edinburgh University, the junior researchers learn data management skills:

> *Their learning process is highly participatory, requiring students to contribute skills to others' projects, and to reuse datasets so they can gain sufficient experience to acquire their own data.*

Whyte *et al* (2008) also note that this learning experience has a key role in providing skilled curators, an areas where there is currently a 'dearth of skilled practitioners' (Lyon, 2007).

7.4.3 Institutions

An increasing number of higher education institutions have repositories, but the majority focus on building collections of published output and few contain data (Lyon, 2007). However, there is a role for institutions to provide local support to researchers in data management (Martinez-Uribe, 2008) and provide curation facilities where there is no domain specific data service. Lyon notes the costs for institutions, which include financial costs involved in providing a robust infrastructure, but also cultural and legal barriers. Beagrie (2008) provides examples of costs, but also notes the promotional benefits to institutions. By maintaining a collection of data, the research becomes more visible and may enhance future funding and collaborative opportunities, as well as attracting future staff and students. The promotional potential of data sharing is not restricted to higher education. Casey (2003) discusses the benefits of making biodiversity data available to the public. This included a variety of data about organisms, much of which is held in museums and botanical gardens. Casey cites an example of the increased number of enquiries received by one museum (Museum of Vertebrate Zoology in the USA). In the first year of providing public access to data, the website fulfilled nearly 42,000 specimen queries compared with 95 queries dealt with manually the staff the previous year.

7.4.4 Society

Although stated in policies, 'public good' is not a popular argument for academics, even for supporters of data sharing.[11] The scepticism surrounding 'public access' is well expressed by the leader of the Experimental Physics Division at CERN :

> *"Some attempts have been made to say this is public data, why not let the general public participate in it. There is clearly nothing against this from the general point of view. The question is more: is this useful in the sense that you increase knowledge transfer or would you rather increase confusion."* (Quoted in Wouters and Reddy (2003))

The general public may have little interest in raw data, but they may have an involvement in research and so can be considered stakeholders in data sharing. For example, as participants in research they have a right to be fully informed. A recent survey for the Medical Research Council of public attitudes to medical research and the use personal health information found

[11] Personal communications during data collection stages of this project.

that people were more positive about medical research if sufficiently informed (MRC, 2007). Conversely, the information provided must not be too complex as this gave the impression of research as a 'closed shop'. The concepts of consent and anonymity were not readily understood, at least not in the same terms as researchers understand them. This has implications for data sharing as confidentiality and ethics are key concerns. One of the aims of a data management plan is to ensure appropriate consent is received at the outset, so there is an opportunity to increase public involvement by making consent issues transparent.

The public benefit directly and indirectly from research and the potential impacts have been the focus of studies by the Research Councils in the UK *(see Appendix 1, Section 7*). Martin and Tang (2007) extend the research impact cycle through which research leads ultimately to socio-economic benefits, by suggesting that additional benefits may 'flow' to the community via additional channels throughout this cycle. They use the example of biomedical research to demonstrate possible benefits in addition to the goal of reduced mortality, these include:

- direct costs savings from new or less costly medical treatments,
- the value to the economy of a healthy workforce,
- gains to the economy from product development, employment and sales
- value to society of health gains.

Enhanced access to research outputs has the potential to increase the efficiency of research and development by speeding up the research process and providing opportunities for wider applications of research (Houghton *et al*, 2006). This could reduce both the time taken and the cost to obtain benefits.

7.4.5 Costs

To achieve the benefits of data sharing certain costs must be met. Beagrie *et al.* (2008) reviewed a number of cost models and outlined a cost model and guidance for UK universities which distilled and synthesised from the best available models.

7.4.6 Summary

As shown in the literature review, there is general agreement about the potential benefits to be gained from data sharing. However, to achieve these benefits requires some investment in infrastructure for effective curation and preservation of data to maximise accessibility and re-use. Beagrie *et al.* (2008) reviewed a number of cost models and outlined a cost model and guidance for UK universities which distilled and synthesised from the best available models. However, the socio-cultural costs should not be ignored, in particular the disciplinary differences in requirements for data-sharing. Some of these barriers require input beyond financial investment, and co-operation of key stakeholders in UK Higher Education, but addressing these will reinforce benefits. Table 7.1 summarises the potential costs-benefits identified in the literature and some of which are illustrated by the case study investigations *(see Appendix 2)*. Section 4 and Appendix 3 provide a framework for making a business case for data sharing initiatives.

Table 7.1 Costs-benefits

Costs	Benefits
'Hard' (Financial)	
• Provision of infrastructure for curation and preservation of data: o Staff o Hardware/software for storage and retrieval of data o Development and application of tools and standards for organisation and accessibility of data o Training for data scientists o Outreach and training to engage the research community	• Maximised investment in data collection o Widens access where costs prohibitive for individual researchers/institutions o Potential for new research questions from existing data, especially where data aggregated and integrated o Reduces duplication of data collection costs • Increases research impact o New collaborations o Increases speed of research and time to realise impacts • New knowledge based industries
Socio-cultural	
• Disciplinary differences o Data ownership o Culture o Time and effort to deposit • Ethics and confidentiality • Recognition and reward for researchers and data scientists • Skills required for re-use	• Transparency in research funding • Use of data sets in education enhances data awareness of students • Access to range of data enhances researchers' skills • Tools and standards have potential to increase data quality • Visibility and promotion of institutions and researchers

8 Appendix 2: Case studies

8.1 Introduction

The aim of the case study phase of the project was to provide qualitative and quantitative evidence of data sharing in different disciplines. Interviews with service providers and users of established data centres in bioinformatics and social sciences illustrated attitudes to data sharing, as well as examples of reuse. Quantitative material, for example staff and infrastructure costs, was obtained from documentary evidence including Annual Reports and information provided by the data centres directly.

Although limited in scope because of the timescale of the project, the case studies did provide a rich picture of data sharing needs and practices in these two contrasting disciplines, illustrating that any cost-benefit analysis must consider the cultural as well as the financial elements.

8.2 ESDS Qualidata

8.2.1 Introduction

The purpose of this case study is to illustrate how the dynamics of costs/benefits that underlie data sharing play out in the social sciences. The case study used documentary evidence and interviews with:

- The Associate Director and Head of ESDS Qualidata to supplement printed documentation and obtain a personal perspective on the current activities and future plans of Qualidata
- Two researchers who had deposited data to Qualidata, as a requirement of their ESRC grant. Both had deposited one data set, and could be described as novice 'users' of the Qualidata service. Their respective areas of research were conversational analysis (using spontaneous conversations requiring highly specialised transcription), and women in science (mixed method – survey, and interview/focus group).
- One user of Qualidata datasets; this respondent had also deposited data and was an experienced user. His area of research was primarily social and cultural geography (use of oral and life history tapes and transcripts, interviews and other primary data collection).

Although the timescale of the project limited the number and scope of interviews, several key issues emerged surrounding the sharing of data in social sciences, not least the wide range of disciplines ESDS Qualidata and its parent centre must serve. The examples obtained from respondents serve to indicate the potential benefits, but also to highlight the barriers to be overcome for a data sharing culture to evolve in the Social Sciences.

8.2.2 ESDS Qualidata – outline of the service

> *To support high quality research, teaching, and learning in the social sciences by acquiring, developing and managing social and economic data and related digital resources; and by promoting, disseminating and supporting the use of these resources as effectively as possible.* (ESDS Mission Statement)

ESDS was established in January 2003 with the aim of developing resources for social science research and learning in UK Higher and Further Education. The service is funded by the

Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC), initially from 2003 – 2007, with extension of funding to 2012 following a successful mid-term review in 2005.

ESDS is a distributed service based on collaboration between four centres: UK Data Archive (UKDA), Institute for Social and Economic Research (ISER), Manchester Information and Associated Services (MIMAS) and the Cathie Marsh Centre for Census and Survey Research (CCSR). ESDS has two central functions – Management and Accessions and Preservation. In addition it has four specialist divisions providing targeted value added services: ESDS Government, ESDS International, ESDS Longitudinal and ESDS Qualidata. Although this case study focuses on ESDS Qualidata, the difficulty in isolating financial costs for each unit was stressed by the Head of Qualidata. The core activities of data acquisition, processing and preservation, as well as user registration are dealt with centrally. Qualidata provides specialist support for users and depositors of qualitative data or mixed datasets containing qualitative data and these are discussed later in the case study. Qualidata is hosted by the UKDA at Essex University. Therefore, depending on the activities under discussion, reference is made to ESDS, UKDA and Qualidata throughout this case study.

8.2.3 Data holdings

Holders of ESRC grants must offer resulting data to the UKDA, and qualitative data is offered to Qualidata. Not all data can be accepted, one of the major barriers is consent, although Qualidata works with Ethics Committees to overcome this. In 2007/2008, 42% of submissions could not be accepted because of consent issues: 30% were rejected and a 12% obtained a waiver. Although both ESRC and ESDS guidelines stress the need for researchers to consider data deposit implications at the outset of projects, for example so that appropriate consent can be written in, this is clearly still a major barrier in sharing qualitative data. Table 8.1 provides full breakdown of submissions, with reasons for rejection or waiver.

Table 8.1 Submissions to Qualidata in 2007/2008 and reasons for rejection or waiver

	Submissions to Qualidata 2007/2008			
Reason	Accepted/No	Rejected/No	Waiver/No	Total/No
	48	*40*	*13*	*101*
Consent/confidentiality		30	12	42
Copyright/IPR		2	0	2
Format		2	0	2
Poor quality		1	0	1
Publication sufficient		5	1	6

Source: Qualidata

Although rejections on the grounds of quality or format are low, deposited data still require input from Qualidata staff to enhance accessibility. Much of the work is currently manual, so there can be no economies of scale and Qualidata is constrained by its budget. However, a self archive facility is currently being trialled, and Qualidata also refer potential depositors to other services. When considering benefits/costs, especially across several disciplines, the data holdings in Qualidata may appear low compared to data centres concentrating on quantitative data. This is

a feature of the labour intensive nature of the data – both to collect (i.e. a cost to the research) and to make available for re-use (a cost to researchers and Qualidata). However, the statistics presented in the ESDS Annual Reports[12] show steady growth, as shown in Table 8.2.

Table 8.2 Qualidata acquisitions by year

Reporting year	New datasets	New editions (e.g. enhancements)	Total
2003/2004	11	1	12
2004/2005	24	7	31
2005/2006	24	27	51
2006/2007	63	0	63

Source: ESDS Annual Reports 2003/4 – 2006/7

8.2.4 Qualidata users

A particular feature of Qualidata is the diverse nature of the user base. Statistics are provided for the status of users *(Table 8.3)* and also the discipline *(Table 8.4).*

Table 8.3 Users of Qualidata by status

Status of user	2006-07 %	2003-04 %
Academic	30	61.1
Postgraduate	40	31.9
Undergraduate	22	6.9
Public sector	2.1	-
Commercial	0.7	-
Other	4.6	-

Source: ESDS Annual Reports 2006-07 and 2003-04

Table 8.4 Users of Qualidata by discipline

Discipline of user	2006-07 %
Business/accounting/finance	16.0
Economics	7.7
Geography	3.5
Health	5.3
History/humanities	7.0
Politics/International Studies	7.7
Psychology	7.4
Social Policy	2.8
Sociology	21.8
Statistics/methodology	2.8
Other	17.9

Source: ESDS Annual Reports 2006-07.

[12] ESDS, Economic and Social Data Service Annual Reports, available from : http://www.esds.ac.uk/news/publications.asp, accessed August 2008.

There has been an increase in the range of users from 2003-04, when only 4 disciplines were represented (Sociology, Social Policy, Politics and Psychology, plus Support Services) although categories differ slightly. For example, at that time, 33% of users were 'support services', which is not a category for 2006-07 statistics.

Debates in the literature have questioned the feasibility of data reuse in some areas of the social sciences, and some of the reasons for this emerged during interviews for this case study.

Dr A
Area of research: Conversational analysis
Deposited data from a 3 year ESRC project under ESRC mandate
'It's part of being an academic, that you find a topic, you negotiate access, you think about the data sources and then you go and find them – it's a long process, so it's a bit naïve to think an academic should then give their data away'.

This area of research requires spontaneous conversations rather than interviews, and subsequent transcription is highly skilled. Qualidata does not currently hold data of this sort and so is not a valued resource for this Dr A. There are some data sets available for a fee and also some informal sharing within the community, which is relatively small. The specialised nature of the data and competitive environment make data sharing unattractive.

'There is absolutely nothing of use to me on Qualidata, which makes me question why should I give away the data that has taken years to collect and transcribe'.

A further key feature of this kind of research is that topic areas are often highly sensitive, and this creates further problems for sharing, so for Dr A the barriers to data sharing are:

- Researcher spends considerable amount of time gaining access, even before data collection, so the feeling of 'ownership' of data is very strong ' *negotiating access – took me well, years really.. I would say in total about two years to negotiate the access I needed'.*
- Competitive field with established areas of expertise '*I don't really want some other academic coming along and using my data to build a career on'.*
- Large data volumes requiring specialised transcription which is very much part of the intellectual input and not comparable with typists' transcriptions.
- Size of data sets and sensitive nature mean the level of anonymisation required for deposit is time consuming
- Some data could not be made available because of consent issues; the depositor estimated about 40% was deposited, this being typists' transcription.

For this area of research it was felt that Qualidata does not provide relevant or useful data, so deposit was felt to be a burden throughout the process This is at the extreme of the cost – benefit continuum

'I can't imagine any benefits – what benefits could there be for me?'

However, it should be noted that there is some exchange of data within this research area, for example data sets are available for a fee and there is some informal sharing within the community, which is relatively small. Another academic in the same field was also interviewed, although he had not used or deposited data. He noted the high cost of data collection, and saw a role for a data service to undertake data collection

'Individual departments put a lot of their time and resources into collecting something which could be collected centrally in a much more cost efficient way and made widely available to academics…I'd like to see data which is not collected under a specific research project.. these kinds of resources would offer a huge opportunity [for social scientists]'.

8.2.5 Value added services

Depositors are requested to provide supporting documentation to give as much background to the datasets as possible. Examples of documentation provided include:

- Grant documentation such as original application, end of award report
- Methodology
- Data collection instruments (e.g. interview schedule/guide, questionnaire)
- Confidentiality/consent documents (e.g. communication with participants, consent forms)
- Matrices
- Tree diagrams
- Equipment used
- References to publications and reports based on the study.

Interview data is usually submitted as typed transcripts which are appropriately marked; Qualidata provide a template for this. Additional processing is then undertaken by the central Access and Preservation team at ESDS. The aim is to make the data available as quickly as possible, together with a comprehensive catalogue record.

Enhanced datasets

Qualidata select certain datasets for additional processing. This may done to provide a thematic resource, for example Health, or to further enhance a high profile data set. This work is labour intensive, but has several benefits:

- It provides projects of interest for data processing staff, to develop their skills and enrich their work
- It promotes data sets, and so increases use
- Resources are used for workshops, further promoting the services and increasing data awareness skills

A key resource is *The Edwardians,* originally reel to reel tapes and typed manuscripts, this collection is now available as an online resource, and one of Qualidata's most used data sets.

Edwardians online

"Edwardians was originally reel to reel tapes and transcripts done on a typewriter, so very unsearchable, very hard to use except by going to Essex, and they spend a lot of time doing 'Edwardians online'" Professor B

ESDS Qualidata Online moves beyond catalogue searching and data download to allow web-based free-text and filtered searching, browsing and retrieval of research data in real time.

Increasingly, data in the system include not only traditional interview transcripts, but also audio and image files. One of Qualidata's most used resources *The Edwardians: Family Life and Work Experience Before 1918* is available on the online service.

The major part of the collection comprises life story interviews originally collected as part of the study. The interviews were undertaken in the early 1970s and formed the basis of the first national oral history project in the United Kingdom, as well as the basis for Professor Paul Thompson's, 'The Edwardians, The Remaking of British Society', (1975, 1992). A total of 444 interviews were recorded on reel-to-reel audio tape and later transcribed as typed, paper documents. The interviews were open-ended (guided by a schedule) and of between one and six hours duration.

The online project created a digital multimedia resource that integrated a wealth of existing primary and secondary materials, with the following materials made available online for initial evaluation:

- a catalogue of 444 interview summaries
- 5 full electronic interview transcripts
- thematic browsing of interview transcripts
- a collection of digital sound clips
- a set of contextual images of Edwardian life
- background information and press reviews on the original Paul Thompson study
- details of publications based upon secondary studies of the collection
- an account of the digitisation methodology

Source: http://www.esds.ac.uk/qualidata/online/data/edwardians/introduction.asp

"the project would not have been possible without this resource – we wouldn't have done a comparison of' then and now' kind of thing, and that's been very useful to do" Professor B

ESDS Qualidata also undertakes research and development work on methods of dissemination. In addition to core funding, Qualidata have been successful in securing additional funding for Research and Development. For example, UKDA won two ESRC awards in 2005-6 for qualitative data enhancement. QUADS (Qualitative Archiving and Data Sharing Scheme) has advanced practice in the area of sharing audio-visual data, and SQUAD (Smart Qualitative Data: Methods and Community Tools for Data mark up) modified natural language processing tools for preparing and anonymising qualitative data.

These developments will increase the type and number of data sets available and may encourage new users, for example the interview with Dr A *(above)* highlighted wariness of data sharing when the process was one way. Conversely, availability of useful data establishes a 'virtuous circle' of reuse and deposit.

Professor B
Area of research: Social and cultural geography
Uses data sets from Qualidata and other sources
Deposits data sets to Qualidata and other dissemination activities for sharing of data
Professor B demonstrates that the most benefit is gained from data sharing when it is a two way process: the availability of relevant data encourages deposit and further use establishing a 'virtuous circle'. In this case, availability of data opened up a new research approach.

I was doing contemporary work on Bradford in an Industrial area, which is an Asian area and a lot of fuss about whether Asians were fitting in or not, and I came across some life history tapes, if you look at life histories of people who worked in the area it's got Italian people, Polish people, Hungarians, Ukrainians and for them the issue was were they fitting in, were they.. a whole rerunning of this thing, so showing some sort of continuity is a really valuable thing to do I think – it has a real purchase on the present day to show that the current population who consider themselves the 'white Bradford population' under threat from Asians were themselves under threat only a generation or so ago. So that's what got me into it – looking at an area historically made a huge difference to how you understand it today.

Subsequent work has focussed on giving a historical and contemporary view of topics, which has been successful in attracting research funding and career development.

As well as reusing data, Professor B goes beyond mandatory deposit to seek innovative ways of disseminating research outputs, this has also been successful in winning grants. For example, one proposal to ESRC included around £15,000 to produce an educational, interactive website based on the results of research into food commodity chain.

We thought it was risky, we put in, I think £15,000 into building this educational website and we crossed our fingers and hoped it wouldn't go against us and most of the referees and board members said 'wonderful think more people should do this', so it definitely worked in our favour

Although data sharing has led to successes, he admits that there is little value attached to data sharing among the wider research community, and does recognise the barriers, in particular the reluctance of some researchers to reuse data they have not collected. This, he notes, is partly protecting academic status, but there are also practical barriers. Appropriate sharing of data requires skills:

Need to familiarise yourself with context… don't want to encourage people to just cut and paste from archives without knowing how to use them properly.

The role of Qualidata in promoting good practice is recognised, for example specifying the documentation required to accompany data sets and the enhancement of key data sets.

Main [value added service] is making it accessible, plus material to contextualise – this is absolutely crucial, without this would be in the dark, so that is the most important stuff they do. When depositing it is a nuisance at the time, but it is invaluable to users.

Training

Training activities promote the service and ensure material offered is in an appropriate form. This saves time and effort both for the researchers and Qualidata, potentially increasing the number of quality data sets available. Examples of training activities range from online materials to assist potential depositors, such as a model transcript and guidance on consent and confidentiality issues to workshops and courses. The 'Milestones 2007-2012' for Qualidata include five workshops per year, which may be in collaboration with other organisations.

Education

The increase in student users of Qualidata can be seen from Table 8.3. In 2003-04 only 6.9% of users were undergraduates, but this rose to 22% in 2006-07. This marks a shift in policy, as noted by the Head of Qualidata

> *"our remit changed a lot over the years, [we] didn't used to work with teachers, only high profile researchers, but now increasingly post grads, young researchers".*

Classroom usage is also recorded and in 2005-2006, 6,297 students used Qualidata resources in classroom activities (ESDS Annual Report 2005-06).

As well as provision of data sets, Qualidata provides additional educational resources, such as teaching packs, for example *Teaching qualitative interviewing.* The pack summarises interview types and illustrates each type with extracts of datasets. Such educational activities increase the skill base within the social science community, and will have a future impact on data sharing. More immediate benefits are time and effort saved for teachers and lecturers.

Outreach

Training and education serve as outreach activities, and in addition the 'Milestones 2007-2012' include seven annual or biannual (set) events, such as national and international conferences to increase the profile of Qualidata. Qualidata also attracts international visitors. It was noted by the Head of Qualidata that the value of such activities are difficult to quantify:

> *"We are still arguing with ESRC over the right kind of KPIs – another thing is reputation – if you have a lot of visitors it says a lot about what you do, there's an awful lot of visitors to see how we do things – really from all over the world".*

It is also necessary to provide evidence of the impact of data sharing to funders, researchers and other stakeholders. However, this is difficult to do, as there is currently no mechanism for tracking citations. Users of data are asked to send lists of publications to Qualidata for inclusion on the catalogue, but it not everyone does. The head of Qualidata stressed the potential benefit of introducing mechanisms to identify and track data

> *"I do think there should be more value in what's a dataset and how its published … a nice DN number – something you can cite and it needs to be recognised by whatever the new RAE is, to say 'yes this is something you publish, it's an output, and once this gets agreed I think there'll be more kudos attached".*

Although citations to data cannot be tracked, ESDS do record usage, and notes that this can be made available to depositors on request. However, the depositors interviewed for this case study were unaware of this, and both noted it was something they would like to know.

> *"Given that the site must know if there have been any users, they could let the depositors know – that would be quite nice to know".* (Dr A, depositor)

Another researcher who had deposited data as mandated by ESRC was unaware of this provision, but in general felt some benefit from depositing, not least in awareness of the process for further ESRC bids.

Researcher C
Area of research: Women in science and engineering
Deposited data to Qualidata, as mandate of ESRC award

Researcher C deposited data resulting from a ESRC funded project eighteen months ago. This raised awareness of Qualidata which helped in preparing a recent proposal to ESRC. For example, ESRC requires applicants to check Qualidata for existing, relevant datasets at the proposal stage and also includes questions about deposit of data at the end of the project.

I think that's definitely one of the feelings you get from ESRC when writing proposals that they will look favourably if you say you are going to deposit data.

Preparation of data for deposit has also changed her research practice to some degree, in particular around the data collection process

From my experience of doing it before, especially for qualitative data – the interviews, I didn't think about the fact that we had to deposit data until the project finished – in terms of anonymising data, so I would definitely put more effort into doing that – anonymising as I went along rather than leaving it to the end.

Involvement in the process has increased awareness of the benefits and increased skills, for example in proposal preparation. However, these were not obvious to her, and could be better promoted by ESRC and ESDS.

8.2.6 Funders

Research funders also have costs-benefits associated with investing in data sharing. The funders of ESDS are ESRC and JISC. Both provide funding to researchers, but also to initiatives such as ESDS; in turn, they must bid for funds from Central Government and so must demonstrate the impact of the research they fund. In recent years, there has been emphasis on access to research outputs, not least because it is funded by public money and there is an increased awareness of accountability. However, ESRC could be seen as ahead of its time, as it required grant holders to offer data to UKDA (qualitative data to Qualidata) before the open access policies prompted by the RCUK Position Statement (RCUK, 2006).

> *The ESRC requires all grant-holders to offer for deposit copies of both machine-readable and non machine-readable qualitative data to the ESDS Qualidata unit at the UKDA within three months of the end of the grant.* (ESRC Research Funding Guide, Annex C)

Unfortunately, the policy does not have a high profile on the ESRC website and the benefits of data sharing are not clearly promoted to researchers, for example as part of an overall data management plan. This may lead to deposit being seen as a burden, although ESRC does stress the need to consider data deposit at an early stage in the project, and will provide funding for preparation of data. A recent cross-council initiative, RELU, does focus on data management, including the implications of data sharing, and ESRC is a partner in this. The data support service associated with the programme is based at UKDA.

Rural Economic and Land Use Project (RELU)
Data are a valuable resource. When managed and preserved properly, they can be used and re-used for future scientific and educational purposes. Sharing data facilitates research often beyond the scope of the original research, encourages scientific inquiry, avoids duplicate data collection and provides resources for training.
(http://www.data-archive.ac.uk/relu/reluadvice.asp)

This is a cross council initiative involving ESRC, BBSRC and NERC to fund interdisciplinary research into the social, environmental and technological challenges faced by rural areas.

A key feature of the programme is the emphasis on sharing the outcomes between disciplines and highlighting the potential benefits of data sharing. The programme has a Data Management Policy and has also established a Data Support Service based at UKDA. This aims to ensure that the data management activities of individual projects are carried out effectively by:

- Providing information and guidance to researchers on data sharing issues

- Providing information and recommendations to RELU management on data management issues
- Providing a data management plan for completion by RELU award holders. This includes details of data requirements (i.e. from existing sources), data to be collected, quality assurance and back up procedures, plans for long term management and archiving of data, copyright and intellectual property rights issues.

RELU fund a one full time and one part time member of staff, based at UKDA to monitor data management plans, hold workshops and visit researchers.

8.2.7 Summary of costs-benefits associated with data sharing in the qualitative social sciences

The ESRC is a major funder of research in the Social Sciences and requires its award holders to offer data to UKDA (qualitative data to ESDS Qualidata). However, a culture of data reuse is not fully established in the research community. One of the key roles for Qualidata is to develop and encourage a data sharing culture by improving methods of organisation and dissemination of data and by education and training.

Acquisition of data in this discipline is labour intensive, both during the original data collection stages and during data deposit. As a result, researchers feel a strong sense of ownership and a reluctance to share; this is compounded by what they view as extra work to prepare data for deposit. Furthermore, the manual processing required by Qualidata limits the number of datasets that can be disseminated. Other constraints are the confidentiality and consent issues which are a feature of much social science research: in 2007/2008, 30% of data sets offered were rejected because of confidentiality/consent issues. These factors limit the number and range of datasets and reinforce the reluctance to share and reuse data, as it is felt to be 'one way'. However, Qualidata undertake research activities, with additional grants, to develop methods of automating aspects of data processing, including methods of anonymising data. This will result in a greater number of data sets being accepted, increase processing rates, and also provide data in different media. Together with outreach activities, the increasing range and number of data sets available may encourage new users within the research community.

Therefore, this case study demonstrates that investment of both financial and human resources are required to reap any benefits from data sharing, and that such benefits may be delayed. However, the study has also provided some positive examples of how data can be reused in the social sciences to give a fresh perspective on research questions.

8.3 EBI

8.3.1 Introduction

The purpose of this case study is to illustrate data sharing practices in the biosciences and provide examples of the benefits of data re-use. The case study is based on documentary evidence drawn mainly from annual reports and interviews with the following:

- The Associate Director of the European Bioinformatics Institute (EBI) to provide an overall picture of the role of the EBI in the biosciences and give a historic perspective of the creation of networked databases in the biosciences based on his experience establishing the EMBL data library.

- The team leader for Vertebrate Genomics at the EBI to give the researcher's perspective on working with raw sequence data and also insights into developing value-added services through the Ensembl Project, which is a software system that produces and maintains automatic annotation on selected eukaryotic genomes.
- The group leader for ArrayExpress at the EBI again to provide both the researcher's perspective of using large-scale networked databases and the perspective of providing value-added services.
- The Head of Outreach and Training at the EBI to get an overview of personnel at the EBI and insights into skills training.
- The group leader for the Human Genome Analysis Group and team leader of The Ensembl project at the Sanger Institute to obtain a data processing perspective and the costs associated with it.
- The Head of IT at the Sanger Institute to obtain a picture of the infrastructure costs and establish what percentage of data production and processing costs are associated with infrastructure costs.

The activities of the European Bioinformatics Institute are split three quarters on data processing and developing value added services and one quarter on research. This is reflected in the spread of employees across services, team roles and research group roles. All of the interviewees, with the exception of the Head of IT at the Sanger Institute, provided data on the benefits/costs of data sharing from both the perspective of data providers and data users. The Sanger Institute, which is located on the same campus as the EBI, is an important data provider to the EBI, especially for DNA sequence. In order to obtain an estimate of the costs of data production, which the illustrative benefits could then be offset against, it was necessary to include interviews with personnel from the Sanger Institute. These interviewees were selected on the basis of recommendation of the Associate Director of the EBI. The EBI case study, therefore, was more distributed in nature than originally anticipated, which reflects in turn the rather distributed and collective nature of the institute's organisation. For example, Ensembl, one of the many publicly accessible value-added services delivered through the EBI, is jointly developed and maintained with the Sanger Institute.

The examples obtained from participants serve to indicate the potential benefits of data sharing, but also highlight some of the unforeseen costs, including non-financial costs, and potential positive impact on researchers' careers of making data available for reuse.

8.3.2 EBI outline of service

> *"To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress. To contribute to the advancement of biology through basic investigator-driven research in bioinformatics. To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators. To help disseminate cutting-edge technologies to industry."* (European Bioinformatics Institute Mission Statement)

The EBI hosts the major core biomolecular data resources in Europe. These include EMBL-Bank (nucleotide sequence), Ensembl (genomes), ArrayExpress (microarray-based data), UniProtKB (protein sequence and functional information), MSD (macromolecular

structures) and InterPro (protein families, motifs and domains). The amount of data deposited and processed has grown exponentially across a number of these services and high-throughput methods are being developed to deal with the 'data deluge' that the EBI is currently experiencing. The emphasis at the EBI is to focus on delivering appropriate services and to support evolving user demand. The reports by Lyon (2007) and Beagrie *et al* (2008) also provide an account of the services provided by the EBI from the perspective of data roles and responsibilities and costs of data repositories respectively.

The EBI is part of an International biomolecular information landscape and collaboration plays a significant role in its activities, including contributing to the public record of science and the development of data standards. In particular, it collaborates with the neighbouring Sanger Institute in the delivery of specific value-added services. Furthermore, the EBI obtains a substantial percentage of its data from the Sanger Institute, which is why it plays a cameo role in the case study. Other data centres that are part of the EBI's broader landscape are the National Center for Biotechnology Information (NCBI) hosted by the National Institutes of Health and the National Library of Medicine in the USA and the DNA Data Bank of Japan.

The case-study below highlights the interaction between costs-benefits relating to data sharing at the institutional level, rather than at the level of individual services. Three specific value-added databases are highlighted in order to illustrate the potential benefits to the institution, the research community and individual researchers: Ensembl, ArrayExpress and the 1000 Genomes Project.

8.3.2.1 Funding

The EBI's main funders are the European Molecular Biology Laboratory (EMBL; an intergovernmental organisation to which 20 member states and one associate member state contribute funds) and the European Commission. Within the UK, the EBI is mainly funded by the Wellcome Trust, the MRC and the BBSRC. It also receives funding from the US National Institutes of Health. All of the relevant UK funding agencies have open data policies, though the extent to which data sharing is actively supported varies. For example, since 2006, the MRC has stipulated that grant applications must include costed plans for preparing and documenting research data for preservation and sharing. Additionally, as part of the end of grant reporting process, the MRC expects projects they fund to report on data management and sharing activities. The Wellcome Trust's policy on data sharing is underpinned by a philosophy of 'open science', but whilst the policy notes that researchers should consider management and sharing of data at the proposal stage, deposit does not seem mandatory in all cases. However, it does state that:

> *"In specific cases where applications for Trust funding involve the creation or development of a resource for the research community as the primary goal, or involve the generation of a significant quantity of data that could potentially be shared for added benefit, the Trust will:*
>
> - *require that the applicants provide a data management and sharing plan as part of their application; and*
> - *review these data management and sharing plans, including any costs involved in delivering them, as an integral part of the funding decision."*

This latter, more specific, stipulation directly relates to the EBI, which is funded to provide a resource to the scientific community. Lyon (2007) points out that it is difficult to assess the

impact of Wellcome Trust funding on research, because it is acknowledged in a variety of ways in outputs. The creation of an database or a patent, for example, is viewed as a legitimate output of funded research.

The BBSRC also stipulates that all grant applications should include a data management plan or provide explicit reasons why data sharing is not possible or appropriate. Similarly, the Sanger Institute is underpinned by an 'open science' philosophy. It is funded by the Wellcome Trust and National Institutes of Health in the USA and both institutions support data sharing. The Sanger Institute originates from the Human Genome Project and the generic principle of making publicly funded research and development freely available drive service provision principles, both at the Sanger Institute and the EBI:

> *"…so you don't have a monopoly over that data and exclude others. I think it's completely clear from post-analysis of usage of genome data that the principle [of open data] has worked. You can see people doing research that they would not have been able to do because they would not have had the resources to collect the data themselves".* (Head of Human Genome Analysis Group, Sanger Institute)

Funding policies, therefore, have been highly influential in establishing responsibilities with regard to publicly funded data. Despite proactive data sharing policies within the biosciences, there are still uncertainties about building long-term infrastructure. It is recognised by the bioscience community that the infrastructure to support data sharing is difficult to fund and there are many questions around long-term sustainability of such infrastructure.

The EBI core budget (from EMBL funds) for 2007 was 16.2 million Euros, which accounts for just over 50% of the overall budget. External funding came to 15 million Euros (48% of total budget).

8.3.2.2 Data holdings

Overall, growth in data holdings at the EBI is exponential. The databases vary in scale and scope, and are having to cope with the submission of ever-larger datasets. For example, the 1000 Genomes Project[13], an international research consortium that includes scientists from both the Sanger Institute and the EBI, amongst other institutions, is producing

> *"somewhere between enormous and terrifying amounts of data".*
> (Team Leader Vertebrate Genomics, EBI, 2008)

The data management task alone associated with the 1000 Genomes Project is, according to the EBI, beyond the capacity of any individual researcher or research group.

The popularity of a particular database is often linked to its maturity. At the time of the case study, the fastest growing data type, in terms of accessions, was DNA sequence. DNA sequence is collected, curated and made publicly available by an international collaboration between the EBI, the NCBI and the DNA Data Bank of Japan. This joint collection acquires about three new sequences per second, twenty four hours a day, three hundred and sixty five days a year.

[13] Details can be found at http://www.1000genomes.org/page.php. Accessed 9th October 2008.

Table 8.5 Acquisitions 2006 and 2007[14]

Type of data	New acquisitions
'DNA Sequence World' (Fastest moving database)	3 new sequences per second, 24 hours a day, 365 days a year
EMBL-Bank entries end 2007	38,000,000 entries processed
EMBL-Bank entries end 2006	18,000,000 entries processed
New genomes end 2007	450 new genomes
New genomes end 2006	200 new genomes
Complete genomes end 2007	650 complete genomes
Complete genomes end 2006	467 complete genomes

Historically, the cost of the long-term storage of raw data has been only about 2% of the costs of reagents used to produce it and this does not take into account the depreciation of equipment and the costs of collecting samples. Some samples are difficult to collect, those used in population studies for example, and therefore may be irreplaceable. So, the saved costs of data production are substantial.

Illustrative data production and storage costs

The following table is illustrative of the storage costs associated with trace data, analysable filtered data, at the Sanger Institute.

Table 8.6 Storage capacity and costs of trace data produced by the Sanger Institute[15]

Trace repository 2005	20 terabytes capacity
Trace repository 2008	80 terabytes capacity
Storage cost 2008	£1,000 GBP for 1 Terabyte of data
Storage capacity 2008	2 and a half petabytes

To put this in perspective the following table illustrates growth in raw data holdings and its associated storage capacity at the Sanger Institute.

Table 8.7 Growth in raw data holdings and storage capacity at the Sanger Institute

Holdings/volume 2008	90 genomes worth of data produced each week - 4,680 over 12 month period
Growth over 15 years to 2005	300 terabytes capacity
Growth between 2005 and 2007	almost 1 petabyte capacity
Sept 2007 to July 2008	1 and a half additional petabytes capacity
Growth per week 2008	3 to 4 additional terabytes capacity

[14] Source: Annual Scientific Report 2007 European Bioinformatics Institute. Available at http://www.ebi.ac.uk/Information/Brochures/index.html. Accessed 9th October 2008.

[15] Source: Interviews with participants from the Sanger Institute.

Illustrative data production and storage costs

The following table is illustrative of the storage costs associated with trace data, analysable filtered data, at the Sanger Institute

Table 8.6 and Table 8.7 translate into usage at the Sanger Institute in the following way.

Table 8.8 Usage at the Sanger Institute

Data downloads 2001 per annum	4.6 terabytes
Data downloads 2001 per month	89 gigabytes each week
Web hits per week 2001	300,000 hits
Web hits per week 2004	8,000,000 hits
Web hits per week 2008	16,000,000 hits

8.3.2.3 Usage

> *"Nowadays, the data collections that the EBI produces are so much at the heart of the work that anyone does in the life sciences. It's not that the EBI holds data and there's suddenly some macro discovery based on what the EBI has made available. It's that everyone who works in the life sciences works with the data we hold all of the time".* (Associate Director, EBI, 2008)

In terms of the types of research for which the data are used, the prototypical example given by researchers at the EBI is that researchers working within a certain living system might find a genetic difference between two organisms that they find interesting for some reason. For example, they might find a gene that has a particular variation in it and one interesting variant might have a disease associated with it or some kind of pathology or abnormality.

> *"If you find a genetic variant that is associated with a given disease you may not know what the gene does. You know it produces a protein, but you do not actually quite know how it does this, you just know that there is this difference. What people would then do is go and search the databases for things similar to that gene to see if there is something similar to it that somebody has actually done some experimentation on and worked out what molecule it produces and what processes it was involved in".* (Associate Director, EBI, 2008)

The above is an example of a clinical application of the kinds of services provided by the EBI, so by working out what processes a gene is involved in, the pathways of reactions in a biological system, a researcher can work out ways of intervening.

> *"That process of determining what a gene does by laboratory methods could easily be two years work".* (Associate Director, EBI, 2008)

Users are not necessarily interested in the raw, unprocessed, data and filtering data is a key aspect of data curation. It is the intermediary data or summary datasets that enable researchers to re-analyse the data. According to the Associate Director of the EBI, if the data were held in an archive that mainly functioned as a standalone scientific record, the data would be of very little value. It is the data integration task, coordinating data of different types and from different sources, that is important to users. This is because the databases are a composite of thousands

of investigations and therefore afford the types of use described above that otherwise would not be possible if the data were scattered across thousands of individual investigators.

> *"If you take your gene and search the available databases to find something similar to it, which you normally would nowadays because the databases are so comprehensive, you could probably within thirty minutes get a good understanding of what your gene is about. Basically, you are talking about, by consulting the databases, cutting down from a couple of years work to half an hours work, so that's a fairly good economic gain".*
> (Associate Director, EBI, 2008)

The EBI measure their usage in a number of ways as illustrated in Table 8.9.

Table 8.9 Usage EBI 2005-2008[16]

Type of use metric	Level of use
EBI Website average hits/day 2007 (average based on first 9 months)	2,260,965
EBI Website average hits/day 2006 (average based on first 9 months)	1,481,375
EBI Website average hits/day 2005	1,320,756
Global unique users 2008	300,000 to 1,000,000 unique IP addresses
Unique hosts served per month 2007	339,629 unique hosts
Unique hosts served per month 2006	293,000 unique hosts
Requests Sept 06-Sept-07	2,328,857
Requests Sept 05-Sept-06	1,481,000

8.3.2.4 Infrastructure costs

Given that the Sanger Institute is an important supplier of data to the EBI, the estimate of infrastructure costs is based mainly on interviews with participants from the Sanger Institute.

The costs of data production and the underlying infrastructure to produce, process and store that data need to be viewed in light of the industrial scale of the data being produced.

> *"We had our first terabase party the other day – a new sequencer that turned out a huge amount of data in a short time and we didn't expect that to be done until Christmas. So, the data acquisition rates are increasing dramatically at the moment".*
> (Head of IT, Sanger Institute, 2008)

The scale of the science is also expensive. Sequencing technologies are designed for small labs and is it possible to produce a lot of data with one or two small machines. The massive rate with which the Sanger Institute is currently producing data is attributable to the fact that thirty new sequencing machines were recently purchased.

IT budgets, however, are flattening and economies of scale are being gained. Reductions in the costs of hardware and the increased storage that new technology affords has meant that IT costs have been falling. For example, a typical four-processor server cost 50,000 GBP back in

[16] Source: Case study interviews and Annual Scientific Report 2007 European Bioinformatics Institute. Available at http://www.ebi.ac.uk/Information/Brochures/index.html – accessed 9th October 2008.

2001, but in 2008 more powerful servers cost 2,500 GBP. Similarly, in the early years of the Sanger Institute, twenty terabytes of storage used to cost in the region of 50,000 to 100,000 GBP, whereas now the cost of one terabyte is one thousand pounds.

> *"So we can buy and awful lot more for less money. The budget has recently been increased to account for new generation sequencing technologies, but we hope to run on a flat budget with increased IT real estate".* (Head of IT, Sanger Institute, 2008)

The falling costs of storage, however, are not keeping pace with the increasing rate of data generation; this means that significant funds will have to be invested in bioinformatics infrastructure over coming years to secure the future of Europe's biological data. The European Commission's Framework 7 Capacities Programme for Research Infrastructures is currently funding ELIXIR, a preparatory phase project whose goal is to secure funding commitments from government agencies, charities, industry and intergovernmental organisations throughout Europe, to strengthen and sustain a world-class infrastructure for the management and integration of information in the life sciences.

> *"Right now the value of that investment has been reaped many times over, because we now produce in the region of ninety genomes worth of data each week".*
> (Head of IT, Sanger Institute, 2008)

The Sanger Institute's core IT Team consists of thirty members of staff and include database administrators, web administrators, desk top support and admin. Informaticians are not included in the IT budget, but as a point of comparison with the EBI the Sanger Institute employs 850 staff and of those 150 are informaticians.

The Sanger Institute has estimated that in the first six months of 2008, it produced between five to ten times the amount of data previously held in the publicly accessible DNA databases (EMBL-Bank, GenBank and DDBJ). Based on benchmarks set by services such as GenBank, the Sanger Institute has calculated that the volume of data produced in the first half of 2008 is equivalent to sixty to seventy years' worth of sequencing data.

> *"It's not just a question of producing data of course, when you have a single genome you need to produce more genomes to see what the differences are. When you have large-scale population studies, then you start to reap the benefits. So really we are at a step-change in the life sciences, particularly genomics, in terms of how the data are produced. The value of the results is really going to change over the next years - even months".* (Head of IT, Sanger Institute, 2008)

The 'step-change' in rates of data production are also reflected in usage rates. In 2001, researchers downloaded 4.6 terabytes of data per annum from the Sanger Institute and in 2005, this figure increased to 18 terabytes of data per annum. Web-based services are core to service delivery of data centres such as the Sanger Institute and in the first six months of 2008 the number of web hits per week reached sixteen million.

8.3.2.5 Staffing

Staffing was a major factor in the overall costs of data management at the EBI, with curation or bioinformatics staff being spread across the various value-added databases. According to the EBI's 2007 Annual Scientific Report the institute had 285 employees at the end of 2007. Distribution across roles is shown in Table 8.10.

Table 8.10 Number of staff by role

Bioinformaticians	5
Coordinators	27
Curators	46
Database administrators	7
System administrators	12
Software developers	82
Scientists (includes postdocs)	31
PhD students	30
Other (includes 4 outreach and training)	45
Total	285

It was difficult to determine the proportion of curation staff per value-added database as personnel are typically split across activities.

8.3.3 The community view

8.3.3.1 Incentives

> *"In my talks I almost always include an acknowledgement of the data release policies and the open data policies that have allowed the field of bioinformatics to grow".* (Team leader vertebrate genomics, EBI, 2008)

Data access policies have evolved as well, so that databases at the EBI have the ability to accept data and mark it 'hold until publication', which means that reviewers can login and see the data. This speeds up the review process, since reviewers do not then have to go to an individual's web site and potentially compromise the anonymity or anything else. Once the paper is published, the data curators literally flip a switch and the data becomes publicly available. The evolution of data sharing policies has meant that any researcher can access the type of large-scale networked databases that the EBI and Sanger Institute provide and use them to develop new methods or almost anything else. Access on this scale has led to rapid and revolutionary growth of fields in the biosciences.

> *"To me that's the baseline fundamental aspect – if the data sharing policies and practices were not in place there would not be bioinformatics concentrations in computer science departments across the world. All of those would, instead, have to be associated with collaborations with independent researchers".* (Team leader vertebrate genomics, EBI, 2008)

From an individual researcher's perspective, the most effective incentives to share and re-use data are not necessarily those derived from funding policies. For example, popular databases such as ArrayExpress are doubling every 14 months. With over 50% of journals in the 'omics' and bioinformatics fields mandating that the underlying data need to be submitted to a specified data centre, it is highly likely that a researcher trying to publish an article in a journal will submit their data.

> *"I think the main reason for uploading the data is because of journal requirements. The funders have these policies that all data have to be public, very often, but they don't have any sticks to really influence the researchers. All they can do [funders] is not give*

the next grant, they cannot take money back really, and they would not do that because it is very difficult to quantify. So, journals are the major driving forces here". (Group Leader ArrayExpress, EBI, 2008)

8.3.3.2 Recognition and reward

Competition is, of course, a fundamental part of research (regardless of whether is it is the biosciences, physical sciences or the social sciences) and, therefore support for data sharing is not absolutely universal even in the community that the EBI serves. Many researchers, understandably, would like to hold the data back just a little while, so that they can get their papers submitted and receive due recognition.

The team leader in vertebrate genomics highlighted that the so called 'Fort-Lauderdale Agreement' (Wellcome Trust, 2003) contains a number of contradictions in it, yet its underlying ethos seems to have permeated through the bioscience community.

"It basically says 'everyone please be nice and everyone please recognise that releasing data in a pre-publication way can sometimes have adverse consequences, but those are worth it compared to not releasing the data'". (Team leader vertebrate genomics, EBI, 2008)

Most of the funders within the EBI community state explicitly the responsibility of data users to cite acknowledge the sources of their data and abide by the terms and conditions under which they accessed the original data. Yet, citation practices varied greatly and only partial attempts have been made to keep a record of citation on a service by service basis. Even where a data accession number is cited in a journal article it is difficult to monitor impact in this way due to variation in data citation practices across users. For example, even though each database may specify how data should be cited more often than not it is the homepage of the EBI itself that is referenced in scholarly outputs.

8.3.3.3 Enabling discovery

When participants at the EBI and the Sanger Institute were asked how the public availability of large-scale networked databases had benefitted the biosciences, the universal response was simply that without them the field of bioinformatics would probably not exist.

"It is always important to remember that almost the entire field of bioinformatics has grown up because of data sharing". (Team Leader for Vertebrate Genomics, EBI, 2008)

In this sense, many of the benefits of data sharing have been revolutionary in nature, and therefore challenging to develop quantifiable metrics for. Other benefits have been more incremental and perhaps lend themselves more easily to being measured since they fall within the ambit of traditional indicators or research practice.

The main revolution appears to have happened in Genome Sequencing.

"Before sequencing we had completely the wrong estimate as to how many genes we had. Functional genomics is the next wave, so questions like 'how many genes are expressed always and everywhere, how many housekeeping genes and how many genes are specific for 1,2,3,4 particular conditions'? You cannot ask these questions without having this overview and combining data from lots of experiments". (Group Leader ArrayExpress, EBI, 2008)

ArrayExpress and the bench biologist – a revolution

> *"We have to distinguish between a database that is just built for archiving data and associated publications, like an archive of scientific records, and a value-added database that tries to extract in a systematic way some new knowledge from all this archived data and combined data".* (Group Leader ArrayExpress, EBI, 2008)

ArrayExpress has two databases, one that is the archive part – the ArrayExpress archive or repository, and a value-added database that overlays the archive or repository – the *Atlas of Gene Expression*. Resources such as the *Atlas of Gene Expression* allow a bench biologist to search across the results of thousands of individual studies with a simple search query

> *"and you can't have that unless you have this type of data collection".*
> (Group Leader Micro Array Express, EBI, 2008)

Without the culture of data sharing in the biosciences and the availability of value-added databases it would typically take a bench biologist two years of laboratory-based research to discover what gene is expressed in a particular disease.

> *"… so you have to do your own experiment effectively, or maybe telephone colleagues and obtain information from them, but now you can come to a database and just ask 'give me all the genes that are expressed in a particular leukaemia', and you will get the answer in one click".* (Group Leader ArrayExpress, EBI, 2008)

Not only have the availability of data, investment in infrastructure, and development of new methodologies and their accompanying tools saved bench biologists reagent costs, equipment costs and costs in terms of time saved, but it is likely that these factors would probably be prohibitive to bench biologists doing the experiments themselves.

> *"The ultimate goal is to enable new research that would not be possible without having all this data in one place and I think we are getting there".*
> (Group Leader ArrayExpress, EBI, 2008)

Incremental benefits are likely to be overshadowed by the revolutionary benefits of data sharing. One such incremental benefit is the improved opportunity for researchers to verify their own research by looking at the results of other studies. By combining thousands of studies in a value-added database such as the ArrayExpress *Atlas of Gene Expression,* a global picture of experiments can be built up as a shared community resource. Understanding the types of principally different responses of biological organisms to different conditions is something that would probably not be achievable based on independent lab-based experiments alone.

> *"It is based in essence on combining thousands of different studies together in one go".*

Another example of an incremental benefit is the standardisation of data. The Team Leader for Vertebrate Genomics at the EBI described the Ensembl gene sets as the absolute standard set for many species, which means that any researcher working on those particular gene sets would use the Ensembl database.

> *"… for example the Platypus gene set - if a researcher wanted to do research on platypus genes or evolution of genes from the [monotreme or the mammalian lineage] – they would use the Ensembl genes. There is no choice whatsoever about that. In that we provide a measure of standardisation and uniformity to all the other researchers that would do that research".* (Team Leader for Vertebrate Genomics, EBI, 2008)

Without services such as Ensembl, there would be a lot of duplication in the biosciences, though the increased opportunity for verification shows that some duplication is beneficial. For example, verification is beneficial for maintaining consistency in data quality. The fact that Ensembl has produced protein-coded gene sets for upwards of forty species means that there is significant uniformity across all of the completely sequenced genomes. This is important for comparative genomics, or any type of genome analysis, because poor data or data with variable quality will significantly 'spike' or ruin the analysis.

> *"Almost nobody is funded to do the type of things that we do and so I think that is one of our significant contributions – enabling certain types of genomic analysis to be done in a uniform and reproducible way".* (Team Leader for Vertebrate Genomics, EBI, 2008)

New discoveries from previously disregarded data

"I think a more direct benefit is the way we are able to collect datasets that people have deposited or produced and left in the public domain in an open access sort of way and use those for interesting discoveries. To give one example, recently from the HapMap Project, which was a project to create a Haplotype map of the Human Genome. The data for that project were all released – there were a number of genotypes on the Array that had failed and they had just been screened out of the process to call the genotypes. People went back and started [looking at the arrays] and what they discovered that in many individuals the exact same regions had failed, the exact same SNPs (single nucleotide polymorphisms), which was statistically unlikely to happen if the failures were distributed randomly and upon further analysis they discovered that these were associated with copy-number variations through the Genome. So, here was a case of data that the initial investigators thought were junk – and only later on did people realise that the data that other researchers were filtering out was actually the signature of significant human copy-number variation, which is now a fairly large effort to understand human genetic variation". (Team Leader for Vertebrate Genomics, EBI, 2008)

8.3.3.4 Visibility of research institutes

In terms of the benefits of data sharing for institutions, the experience of the EBI indicates that those scientists or institutes that are linked with major data centres are the most highly visible amongst their scientific communities.

> *"If you look at the most impactful scientists and the most impactful research institutes, at least in biology, those associated with the genome centres and the major genome institutes have been among the most impactful scientists for the last half-dozen years".* (Team leader vertebrate genomics, EBI, 2008)

The advantage for researchers working at centres like the EBI and the Sanger Institute is that not only do they have access to first-class resources, they can access the data and do their own analysis, and build a research profile within the community based on that, but also they facilitate very large numbers of other scientists to also do research on that data. This saves them both the time and cost of having to collect the data themselves. In addition, the value-added layer of data integration and development of search and analytical tools that are also being developed by the EBI and the Sanger Institute are important.

It is possible that the availability of shared data collections also leads to collaboration, such as collaboration between researchers who produce a data set and some other scientists who see a potential new application of the data.

8.4 Findings from the case studies

8.4.1 Convergence and divergence across disciplines

The two case studies were very different in characteristics, in terms of the types of disciplines they represent, the users they support, the types of data they produce and curate, and the value-added services that are important to the communities they serve.

In certain respects, therefore, these two case studies represent diverse examples. Rather than comparing like-with-like we have chosen two extreme examples to illustrate the points of divergence and convergence in thinking about ways in which benefits might be offset against costs and the disciplinary characteristics that underlie the interaction between costs-benefits.

Table 8.11 shows points of divergence and convergence which emerged from the case studies. These provide examples of the issues to be addressed in developing data sharing, and where barriers need to be overcome and benefits may emerge.

Table 8.11 Points of divergence and convergence for EBI and Qualidata

	EBI	Qualidata
	Divergence	
Data production	Born digital	Manual collection and processing
Data processing and cleaning	Automated and manual	Largely manual
Key value added services	Development of tools and standards	Outreach, training and re-presentation of datasets
Culture	Data sharing embedded in culture	Data sharing culture evolving; debates over value of reuse in some areas.
	Convergence	
Outreach and training	Provides skills for researchers which further establishes data sharing culture Provides skills for data scientists and curators with potential new industries in data management	
Reward and recognition	Requirement for method of 'tracking' data set use Production and deposit of data sets to contribute to academic reward and recognition structure Opportunities for increased visibility for researchers and institutions, leading to new opportunities	

8.5 Interview schedules

8.5.1 Service providers

8.5.1.1 Acquisitions

1. Data holdings
2. Growth in contributions
3. Storage costs
4. Usage statistics:
5. How are data deposited?

6. What are the criteria for accepting data e.g. quality control mechanisms?
7. What incentives are given/should be given to encourage submission of data?
8. Should funding agencies require data submission as a condition of grant awards?
9. What is your constituent scientific/scholarly community for depositing and re-using data?

8.5.1.2 Staff

10. How many staff are employed?
11. Break down of staff numbers by role?

8.5.1.3 Value added services

12. What types of value added services are provided?
13. What percentage of personnel time/costs are committed to these? (Broken down by data cleaning, anonymisation, metadata creation, and integration of datasets)
14. What other aspects of data preparation are necessary in order for the data to be re-used and what are the costs/time associated with these?
15. What training, if any, is necessary for the continuing professional development of staff and what costs are associated with these?
16. What standards, automatic metadata generation algorithms, ontologies, retrieval tools have you developed in order to provide services? Can you estimate the costs associated with these?
17. Have these been appropriated by other data services/centres or members of the wider scientific/scholarly community?
18. What analytical tools have you developed in order that the data can be processed? Can you estimate the costs associated with these?
19. Similarly, have these been appropriated by those who deposit or withdraw from the data centre/service?

8.5.1.4 Visibility/promotion

20. What recognition/incentives, if any, do you receive for these value added activities e.g. funding?
21. Who do you perceive to be the main audience(s) for the service/centre?
22. How do you promote the centre/service? Are there costs associated with these activities?
23. What training do you provide for users who deposit data? Can you estimate the costs associated with these?
24. What training do you provide for users who withdraw data? Can you estimate the costs associated with these?
25. What outreach activities have you conducted in order to engage users that deposit or withdraw data? Can you estimate the costs associated with these?

8.5.1.5 Contribution to science/scholarship

Could you provide us with some examples of ways in which the data service/centre has had a positive impact on the scientific/scholarly communities it aims to serve (including research councils and other funding agencies)?

8.5.2 Service users

8.5.2.1 General background

Costs-benefits of open sharing of research data

Introduce self. I am currently involved in this JISC funded project to estimate the costs/benefits of open sharing of data. Part of the project includes case studies in different disciplines to obtain the perspectives of service providers, depositors to the service and users of the service. X is one of the data centres included in the study. To build a full picture of the costs-benefits we would like as much hard (quantitative) data as possible, but are also very interested in any illustrative examples of the benefits you, or any colleagues, have derived from sharing data. Any examples used in the final report will be treated with confidentiality and any direct quotes will be distributed to participants for consent.

Establish whether they deposit data, use (down load data) or both – which first (did one encourage the other)

8.5.2.2 General background questions

How would you describe your field of research (name of sub-field, object of research, character of field e.g. mono/inter/trans-disciplinary, fast moving, competitive)?

What type of data does your research use/generate?

To what extent is your field of research 'data centric' and how important are databases, data centres or data services to your research? (i.e. existing data)

What are the main outputs of your research?

8.5.2.3 Depositors to data centres/services

Thinking about data that you deposit in data centres (share)

Creation

1. How are the data generated e.g. use of experimental apparatus, interviews?

2. What effort goes into generating this data?
   a. number of hours
   b. percentage of total time
   c. percentage of time doing research

3. Are there certain data that you do not share?
   a. why
   b. how much, e.g. what % of all data generated

Dissemination

4. Which data centres/services do you submit data to, or other methods of 'sharing'

5. How long have you been depositing data (e.g. part of established research culture, funding mandates, institutional policy)

6. What percentage of data generated do you currently deposit/share

7. How much might you share (i.e. some sort of lower and upper bound on what they might do, and why).

8. How frequently do you submit data?

9. How 'raw' is the data that you submit?

10. What additional work is required to prepare the data for submission to third parties for sharing
    a. over and above work on data for use in your own project
    b. over and above 'good research practice'

11. Can you quantify this?
    a. time/hours
    b. time as % of total project, research time
    c. how much effort is involved (scale 1 – no effort, 5 a lot of effort)

12. Is this the requirement of the data service, funding body, disciplinary norms or own preference?

13. Who performs this additional work e.g. self, research team, department, data service/centre?

14. Do you get incentives (e.g. money) to do it
    a. if so how much?
    b. If not, do you know or have any sense of how much its costing you (in money, hours, % of total time, etc.)?

15. Have you or members of your team/academic department acquired any additional skills as a result of making this data available? (Include by-products, for self, for others)

16. Is it possible for you estimate what proportion of your total research output is submitted/shared? (i.e. data, papers, reports etc)

Motivation/benefits

17. What are your motivations for submitting/sharing data (mandatory submission, would you submit anyway)?

18. Have you experienced any (non-technical) issues relating to submitting/sharing data via a data service/centre?

19. Are you aware of the extent to which the data you submit has been re-used?

20. Are you aware of the benefits that others have received from using the data you have submitted? If so, what are they…
    a. Quantifiable (time, money)
    b. Qualitative

21. Have there been any direct benefits to you (i.e. as a depositor) based on the sharing/re-use of your data e.g. funding, citations, collaborations?

22. [May be covered above, if not probe] What recognition and reward, if any, have you received as result of depositing data
    a. from your funders,
    b. employing institution,
    c. peers and disciplinary community for these activities
    d. download impacts (journal articles, reports, data)

23. To what extent does sharing open-data translate to esteem indicators/outputs that can be put on your CV? (citation, impact, skills)

24. When applying for funding is provision (training, time adjustments) made for the acquisition of new skills and the additional work necessary to make the data reusable?
    a. if so, what
    b. how much

25. Do you receive any support from your employing institution for this work?
    a. if so, what form, (e.g. financial, time, recognition)
    b. how much, etc

8.5.2.4 Use of data service to download/reuse data

General practice

1. What data centre(s)/service(s) do you use to search for and/or source data for your research?

2. Experiences of sourcing appropriate data (finding 'centre')

3. What type of data do you use?

4. To what extent is it processed or do you re-use raw data? Match of availability and requirements

5. How easy is it to access and use the data
    a. time
    b. effort
    c. skills

6. How often do you download/access data?

7. Which value added services/byproducts are you aware of and which do you use ?
    a. how much do they add
    b. how easy are they to use, etc.?
    c. What services would you like to see and why?
    d. What services could you do without and why?

8. What new skills have you, or members of your team, had to learn in order to use these datasets, for example in terms of
   a. find (search and discover),
   b. access,
   c. process and analyse

9. What have been the difficulties, either technical or non-technical, in acquiring these new skills?

10. Is there a 'cost' to learning these new skills
    a. time
    b. resources
    c. cost of software or other facilities

11. Who covers these costs (e.g. can claim from grant, institution, none)

12. Overall, how much does it cost, to re-use data.
    a. direct (£)
    b. time (% of project/research time)
    c. effort (1 – 5)

Impact on practice (knowledge creation)

13. In what ways has your research practice changed since the datasets you use have been available?

14. In general, has the greater availability of data sped-up the research process?

15. If so, do they have any sense as to how much (e.g. some sort of % of time)?

16. Has it enabled you to answer existing questions faster, and if so how much faster?

17. Has it generated new research questions? Can you give any value of this?

18. Has the availability of these data sets opened up new fields of research to you? (quantitative, stories)
    a. what areas,
    b. how,
    c. importance (career, institution, research community)

19. Has working with data in this way led to an increase in collaborative work? (quantitative, stories)
    a. how, e.g. network of people
    b. importance to self, to institution, to research community

Impact on inputs

20. What kind of impact, if any, has working with these datasets had on the availability of external funding to continue your research?

21. Do you get any funding to help in the use of shared data? If so, how much?

Impact on outputs

22. Can you describe ways in which you feel access and use of these datasets have had an impact of your research productivity?
    a. % increase in productivity
    b. % research time saved

23 Are you producing more of the same types of outputs or new ones?

24. Has the way that you disseminate your research changed?

25. Do you have more confidence in the reliability of your results?
    a. if so how much (%)?

26. Has it increased the validity of your results in terms of your scientific/scholarly audience and being incorporated in the existing body of knowledge?
    a. if so, how much (e.g. 5 point scale)

27. Have you experienced any difficulties in getting articles published based in the re-use of these datasets? Or Is it easier to get published? How much?

28. Are you aware of any citation impacts

29. Are there known download impacts on either journal articles, research reports, and/or the data?

30. Do you find yourself publishing in the same set of journals or new ones?

Impact on recognition and reward

32. To what extent does using/re-using open-data translate to esteem indicators/outputs that can be put on your CV?

33. Do you receive any support (financial, time, recognition) from your employing institution for this work? If so, what form, how much, etc.?

9 Appendix 3: A framework for making a business case

Whatever the motivation for preserving and sharing research data it is important to have a solid grasp of the costs involved and commitment required. The guidelines suggested by Beagrie *et al.* (2008) provide an excellent foundation for costs, but as Beagrie *et al.* themselves note:

> *An appeal to the Newtonian vision of "standing on the shoulders of giants" may fall short of what is needed to make a persuasive case for adding these costs to already-strained budgets* (Beagrie *et al.* 2008, p16)

To make such a case it is necessary to examine benefits/costs. To that end we outline an approach to making a 'business case' for data curation and sharing, which includes guidance for costs-benefits analyses.

9.1 Costs-benefits

In this section we explore the range of costs, savings and benefits that might be considered, before turning to the issue of information sources and sourcing.

9.1.1 Costs

Beagrie *et al.* (2008) explored the costs of curation of research data. Their study provided a detailed description of cost elements and an activity costing framework focused on costs relating to staff, equipment, travel, consumables, estate and indirect costs of establishing and operating a research data repository (i.e. full economic costing following the guidelines for transparent costing of full economic costs (TRAC fEC)).[17]

The 'Beagrie model's' major activity phases or elements are:

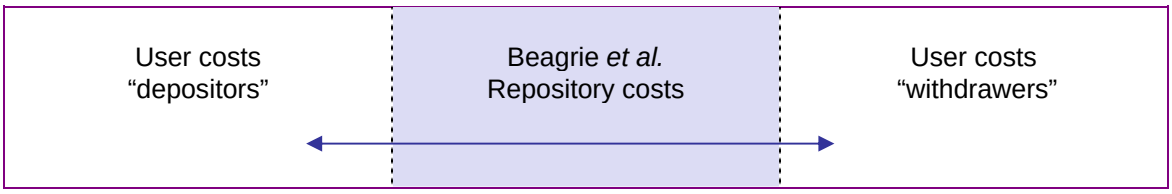
- **Pre-archive** – a phase primarily relating to research projects in universities creating research data for later transfer to a data archive, in which implications for repository costs are considered and data collection/creation designed and implemented with curation and sharing in mind;
- **Archive** – a phase primarily relating to the acquisition/disposal, ingest, storage and management of data, but also expanding into the provision of access and user support; and
- **Support services** – covering administration, common services and estates.

Costs include acquisition, ingest, metadata, access and storage costs, with costs relating to systems, staff, ancillary staff expenses, services and fees, supplies and materials, cost of software tools to process and manipulate data (even though the majority of these tools might be freely available or open source there is an issue of updating them), cost of maintaining and updating the data, costs of quality control procedures, and costs of training data competent scientists. For data centres operated by UK research councils, the authors noted that from the repository perspective the distribution of costs by functions is similar across the councils and research fields, with acquisition and ingest accounting for 42% of overall costs, archival storage and preservation for 23% and access 35%.

[17] See http://www.jcpsg.ac.uk/guidance/index.htm

While the focus of the 'Beagrie Model' is on data repository operation, there are clear interfaces and some overlaps with the activities of repository users, be they researchers who deposit data or those who withdraw it. Nevertheless, for the purposes of exploring benefits it is necessary to extend the 'Beagrie Model' to more fully include these user costs *(Figure 9.1)*.

Figure 9.1 Scope and coverage of costs



Source: Authors' analysis.

9.1.2 Benefits

It is always more difficult to identify and quantify benefits than costs. Benefits may accrue in a variety of ways, including cost savings, efficiency gains, and new opportunities to create value through doing things in new ways and doing new things. These are, successively, more difficult to quantify: not least because they often emerge over time and can only be realised in the future. An obvious starting point is to begin with the most direct and directly measurable, namely cost savings.

9.1.2.1 Cost savings

Possible cost savings can be reaped by the users of data repositories, be they on the input side ("depositors") or the output side ("withdrawers"). Both will also face costs that must be factored into an analysis of costs-benefits.

- **Depositor costs:** The production-side costs of "depositors" include the total costs of production of the datasets held in a repository, any additional research costs relating to researchers working-up the data for use by others, the application of standards and metadata above and beyond that necessary for the project itself, additional costs of anonymisation, and so forth.
- **Withdrawer costs:** The user-side costs of "withdrawers" are also important, with users facing costs in searching for, accessing and assessing data held in repositories, as well as reaping potential savings through reduced data collection demands and realising a range of wider potential benefits.

Table 9.1 outlines these user costs.

Table 9.1 Activity cost elements for repository users

Activity	Description and notes
Depositor costs	
Collection	All the costs associated with the generation/collection of the data, including: • Design and implementation of collection instruments; • Operation of the collection phase; • Recording and initial preservation of data for project purposes; • etc.
Preparation	All the costs associated with the preparation of the data for deposit into a repository that would not otherwise be encountered as part of the research project, including additional efforts relating to: • Standardisation of ontologies, formats, etc.; • Anonymisation and cleaning; • Description, tagging and identification; • Handling issues relating to special associated software, where necessary; • etc.
Deposit	All the costs associated with the deposit of the data with a repository, including additional costs relating to: • Liaising with repository staff, timing and planning for the transfer; • Dealing with issues relating to intellectual property, licensing conditions, research ethics compliance; • Generation of associated user documentation; • Creation of a submission package in accordance with repository guidelines; • etc.
Withdrawer costs	
Search and discovery	All the costs associated with searching for and discovering data held in repositories and available for sharing that might be suitable for the proposed research project.
Access and assessment	All the costs associated with access, investigating and assessing data held in repositories and available for sharing that might be suitable for the proposed research project.

The costs of production of the data, its preservation and access, can also be used in estimating first order benefits, with one benefit of data repositories being that the data would not have to be re-created – making the data's production costs multiplied by the number of times they might have been re-created if not made available through curation and sharing one path towards estimating potential benefits.[18]

Other participants in the research system may also realise cost savings through data curation and sharing:

- **Funders** – the funders of research may be able to fund more research if less is spent on collecting data for specific projects as researchers make more use of existing datasets, effectively increasing total research funding and realising higher economic and social

[18] The 'contingent value' of a given data holding might also be estimated from production and use costs, with minimum value being the cost of production, and use value being the sum of production and curation costs and the amount that people spend using what was produced (in money and time, including overheads).

returns to a given research spend. Funders supporting data curation and reuse may also raise their profile with supporters making it easier/cheaper to raise funding support.

- **Research institutions** – research institutions supporting data curation and reuse may also realise savings through reduced data collection and greater re-use, again effectively increasing research spending, as there will be less time and equipment involved in data generation and collection for a given outcome.
- **Subjects** – the subjects of social and medical research may also realise cost savings, through being subjected to fewer surveys, fewer tests and experiments. This may also reduce research ethics compliance costs for researchers and institutions, as well as raising the quality and reliability of findings where larger sample and trial sizes are made possible through data sharing.
- **Publishers, funders and other participants** – may also realise savings in such areas as preparation of papers, research reports and grant applications, and their peer review, through the ready availability of data and ability to check and substantiate findings.

Fields of research vary greatly, so there may be other costs and cost savings relevant to a particular proposal. These should emerge in discussion and can be added as required.

9.1.2.2 Other benefits

Benefits beyond cost savings are more difficult to quantify, with a wide range of possible impacts imaginable but greater difficulty in attributing more distant impacts to specific causes. Here we note some of the possible benefits, focusing on those that might be measurable.

As noted, key dimensions of benefit might include:

- Opportunities afforded for wider access than has been typical with informal channels for data sharing, including:
  - Access for researchers outside the core HE and public sector research networks, for researchers in industry, government and non-government organisations, thereby enabling greater cross-sectoral collaboration;
  - Ditto for international research and collaboration;[19] and
  - Greater use in teaching and research training in such areas as graduate and post-graduate education projects, etc.

[19] It is widely held that there are advantages to collaborative research and greater use is made of the findings of collaborative work (Katz and Hicks 1997; Katz and Martin 1997; Walsh and Maloney 2001).

Table 9.2 Dimensions of benefit

Area of benefit	Description and notes
In research and education	
Collaboration	Access for researchers outside the core HE and public sector research networks, for researchers in industry, government and non-government organisations, thereby enabling greater cross-sectoral collaboration. Ditto for international research and collaboration.
Education and research training	Greater use in teaching and research training in such areas as graduate and post-graduate education projects.
Indirect cost sharing and reduction	Avoiding survey fatigue and thereby improving response rates, extending sample sizes and trials and thereby improving the validity and quality of research, etc.
New opportunities and uses	Enabling new uses unforeseen at the time of collection, data mining opportunities, etc.
Complete and transparent record	Create a more complete and transparent record of science, with implications for improved detection of fraud and plagiarism, and easier assessment and peer review at the grant application and assessment, publication and research evaluation stages, etc.
Research evaluation and the direction of inquiry	Better align research evaluation with what researchers produce, by providing recognition for a wider range of 'outputs' and contributions than is typical in current publication-centric evaluation programs, and thereby avoiding the possibility of evaluation shaping the modes of scholarly communications and, perhaps, the direction of inquiry.
Visibility and reward	Raise the visibility of researchers, research institutions, repository host institutions and funders, by linking them to valued resources that are an important part of research in a particular field.
Wider economic and social impacts	
New industries and activities	Emergence of re-use 'industries' in particular areas of research and observation, as has happened with geospatial, meteorological and oceanographic data, etc.
New industries and support services	Emergence of support and service 'industries', focusing on providing value adding products and services that enable easier storage, discovery and access to datasets, in turn increasing the value of the data held.

- Opportunities for use and re-use of data, including the reduced cost of collection and duplication noted above, but also sharing/reducing the indirect costs of collection, including:
  - Avoiding survey fatigue and thereby improving response rates;
  - Extending sample sizes and trials and thereby improving the validity and quality of research; and
  - Enabling new uses unforeseen at the time of collection, data mining opportunities, etc.
- Opportunities to create a more complete and transparent record of science, with implications for:
  - Improved detection of fraud and plagiarism; and
  - Easier assessment and peer review at the grant application and assessment, publication and research evaluation stages, etc.

- Opportunities to better align research evaluation with what researchers produce, by providing recognition for a wider range of 'outputs' and contributions than is typical in current publication-centric evaluation programs, and thereby avoiding the possibility of evaluation shaping the direction of inquiry.
- Opportunities to raise the visibility of researchers, research institutions, repository host institutions and funders, by linking them to valued resources that are an important part of research in a particular field. This may result in direct rewards, such as greater funding, as well as qualitative rewards around recognition and peer acknowledgement.
- Opportunities beyond research itself, but nevertheless feeding back into it, such as:
    - The emergence of re-use 'industries' in particular areas of research and observation, as has happened with geospatial, meteorological and oceanographic data, etc.; and
    - The emergence of support and service 'industries', focusing on providing value adding products and services that enable easier storage, discovery and access to datasets, in turn increasing the value of the data held.

Again, it should be noted that fields of research vary greatly and there may be other benefits relevant to the particular proposal.

9.2 Information sources

As time goes by there will no doubt be an increasing number of 'business cases' made for data repositories and reports of their operations and impacts that will be useful inputs to building such a case. Meanwhile, it is likely that those most immediately involved will remain the main sources of information on which to base cost-benefit estimates. These will include existing repository staff and users as well as potential users of the proposed repository. Issues to explore and possible question to ask include the following.

9.2.1 Data centre/repository management and staff

Understanding the nature and activities involved in the proposed or existing repository is essential. By activity phase, the following issues and questions are important.

9.2.1.1 Acquisition

In the acquisition stage, important drivers of costs-benefits will include the nature of the research community being served by the repository, such as the practices surrounding data collection and generation, data sharing and the nature of the data used in that specific field of research and/or discipline. Such information can also be used as a basis for estimates of the potential scale of use and re-use.

Questions might include:

- What is your constituent scientific/scholarly community for deposit and re-use?
- What are the criteria for accepting data (e.g. quality control mechanisms)?
- How are data deposited?

9.2.1.2 Value added services

They key issue relating to the value added services offered by the repository is to get a clear idea of the dividing line between the work and efforts of the researchers in deposit and withdrawal/use on the one hand, and repository staff on the other (and, perhaps, where it

should be), as this is crucial to understanding the division of costs, potential cost savings and cost-benefits.

Questions might include:

- What percentage of personnel time/costs are committed to data cleaning, anonymisation, metadata creation, and integration of heterogeneous data, etc?
- What other aspects of data preparation are necessary in order for the data to be re-used?
- What training, if any, is necessary for the continuing professional development of staff?
- What standards, automatic metadata generation algorithms, ontologies, retrieval tools have you developed in order to provide services?
- Have these been appropriated by other data services/centres or members of the wider scientific/scholarly community?
- Similarly, what analytical tools have you developed in order that the data can be processed and have these been appropriated by members the scientific/scholarly community?

9.2.1.3 Visibility

One avenue to exploring benefits is to explore how data curation, preservation and sharing affects the visibility of the parties within and beyond their respective communities in 'soft' terms and in terms of funding and other rewards.

Questions might include:

- What recognition, if any, do you receive for these activities?
- Who do you perceive to be the main audience for the centre/service?
- How do you promote the centre/service?
- What training do you provide for actual users?
- What outreach activities have you conducted in order to engage potential users?

9.2.1.4 Contribution to science/scholarship

Another avenue into the benefits is to explore ways in which data curation and sharing have had an impact in general or anecdotal terms, and to ask about the thinking behind establishment (even if a formal business case was not made).

Questions might include:

- Could you provide us with some examples of ways in which the data service/centre has had a positive impact on the scientific/scholarly communities it aims to serve (including research councils and other funding agencies)?
- Could you list the major expected costs-benefits, as expected and/or reported from users who deposit or use data (known either anecdotally or statistically)?
- Have you done a cost-benefit analysis or thought about it in those terms (e.g. in making a business case)?
- What is the 'value proposition' and what have you done to prove it?

9.2.2 Users (who deposit)

For the benefits side of the equation, however, it is the users that are most important – be they on the input (deposit) or output (withdraw) side. It is important to understand the nature of the data generated, how it is used and the role data play in their fields of research, as well as how and how much they and/or researchers in their field use data repositories or might do so in the future.

9.2.2.1 General background questions

Focusing on the nature of the data and their roll in the field. Questions might include:

- How would you describe your field of research (name of sub-field, object of research, character of field (e.g. mono/inter/trans-disciplinary, fast moving, competitive)?
- What are the main outputs of your research?
- What type of data does your research use/generate?
- To what extent is your field of research 'data centric' and how important are databases, data centres or data services to your research?

9.2.2.2 Contributors to data centres/services

For the contributors to data repositories (depositors) questions should focus on the qualitative and, where possible, quantitative aspects of their activities, especially why they contribute data to data repositories and what they get out of doing it.

Questions might include:

- **Creation**:
  - How are the data generated (e.g. use of experimental apparatus, interviews, etc.)?
  - What effort goes into generating these data (e.g. time, personnel, and learning new skills)? Can you give specific numbers (e.g. number of hours or percentage of total time or of time doing research)?
  - Do you also make use of data generated by other researchers? If so, how much, in what ways, and do you have thoughts on the value, time saved, etc.?
  - Are there certain data that you do not share? If so, why, how much, what percentage of all data generated, etc.?
- **Dissemination**:
  - Which data services/centres (or other third parties (e.g. journals)) do you submit data to?
  - What types and quantity of data do you submit?
  - How 'raw' is the data that you submit?
  - How frequently do you submit data?
  - What percentage of data generated do you deposit/share? How much might you share if encouraged to in the future?
  - For what period of time have you been sharing your data?

- What additional work is required to prepare the data for submission to third parties for sharing (i.e. over and above necessary work on the data for use within your own immediate projects and/or over and above good research practice, etc.)?
- How much effort is involved, how much time does it take in hours or as a percentage of total project time, total research time, etc.?
- Who performs this additional work (e.g. self, research team, department, data service/centre)?
- Is this the requirement of the data service, funding body, disciplinary norms or own preference?
- Do you get money to do it, and if so how much? If not, do you know or have any sense of how much its costing (in money, hours or as a percentage of total time, etc.)?
- Have you or members of your team/academic department acquired any additional skills as a result of making research data available?
- Is it possible for you estimate what proportion of your total research output is submitted/shared?

- **Motivation/benefits:**
  - What are your motivations for submitting/sharing data (If not mandatory submission, would you submit anyway)?
  - Have you experienced any (non-technical) issues relating to submitting/sharing data via a data service/centre?
  - Are you aware of the extent to which the data you submit has been re-used? If so, how much?
  - Are you aware of the benefits that others have received from using the data you have submitted? If so, what are they and how might they be quantified (e.g. in terms of money and/or time saved, etc.)?
  - Are there any anecdotal stories to tell that could be used as concrete examples?
  - Have there been any direct benefits to you the depositor based on the sharing/re-use of your data (e.g. funding, citations, collaborations, etc.)?
  - What recognition and reward, if any, do you receive from your funders, employing institution, peers and disciplinary community for these activities (teasing out what and from who)?

9.2.3 Third party users (who withdraw)

Similarly with the third party users of data repositories who search for and explore, and/or download and use data made available for sharing, questions should focus on the qualitative and, where possible, quantitative aspects of their activities, especially why they use data from data repositories and what they get out of doing so.

General questions might include:

- What data centre(s)/service(s) do you use to search for and/or source data for your research?

- Are the data all of the same type or heterogeneous in nature?
- Do you download the data and do the analysis on your local machines, or do you perform the analysis remotely and then download the results?
- How easy is it to access and use the data?
- How much time does it take to find and access?
- How much does it cost, either directly or in terms of the time and effort involved in access and retrieval?
- Are there value-adding services available? If so, how much value do they add, how easy are they to use, etc.?
- What services would they like to see and why?
- What services could they do without and why?

9.2.3.1 Impact on practice (knowledge creation)

The impacts of using/re-using data on users' research practices will be central to exploring the benefits they reap from doing so.

Questions might include:

- In what ways has your research practice changed since the datasets you use have been accessible?
- In general, has the greater availability of data sped-up the research process? If so, do you have any sense as to how much (e.g. as a percentage of project time or overall research time)?
- Has it enabled you to answer existing questions faster, and if so how much faster?
- Has it generated new research questions? If so, what is the value to your research?
- What new skills have you, or members of your team, had to learn in order to find (search and discover), access, process and analyse these datasets?
- What have been the difficulties, either technical or non-technical, in acquiring these new skills?
- Is there a 'cost' to learning these new skills (e.g. in time, resources, cost of software, etc.)? If so, can you elaborate?
- Has the availability of these datasets opened up new fields of research to you? If so, how important is it, and how would you say it effects your efficiency/productivity?
- Has working with data in this way led to an increase in collaborative work? If so, how important is it, and how would you say it effects your efficiency/productivity?
- Has the network of people that you collaborate with changed in anyway? If so, how important is it, and how would you say it effects your efficiency/productivity?

9.2.3.2 Impact on inputs

The impacts on research inputs, especially funding, are also important in assessing the qualitative and quantitative benefits experienced.

Questions might include:

- What kind of impact, if any, has working with these datasets had on the availability of external funding to continue your research?
- Do you get any funding to help in the use of shared data? If so, how much?
- When applying for funding is provision (e.g. training, time adjustments, etc.) made for the acquisition of new skills and the additional work necessary to make the data reusable? If so, how much?

9.2.3.3 Impact on outputs

The impacts on research inputs are also important in assessing the qualitative and quantitative benefits experienced.

Questions might include:

- Can you describe ways in which you feel access and use of these datasets have had an impact of your research productivity (e.g. estimated percentage increase in productivity, percentage of research time saved, etc.)?
- Are you producing more of the same types of outputs or new ones?
- Has the way that you disseminate your research changed?
- Are there by-products that are produced as a result of this work?
- Do you have more confidence in the reliability of your results? If so how much?
- Has it increased the validity of your results in terms of your scientific/scholarly audience and being incorporated in the existing body of knowledge? If so, how much?
- Have you experienced any difficulties in getting articles published based on the re-use of these datasets, or is it easier to get published? How much?
- Are there citation impacts (e.g. more cites because there are more people using the same data, etc.)?
- Are there known download impacts on either journal articles, research reports, and/or the data?
- Do you find yourself publishing in the same set of journals or new ones?
- Do you upload any new data, or by-products from working with the open data, to a data service/centre (or other centralized resource)? If so, is there some sort of virtuous cycle, and might you be able to quantify something about levels of use and deposit?

9.2.3.4 Impact on recognition and reward

The possible impacts of data use/re-use on research rewards experienced by users are key to understanding the drivers and benefits.

Questions might include:

- To what extent does using/re-using open data translate to esteem indicators/outputs that can be put on your CV?
- Do you receive any support (e.g. financial, time, recognition) from your employing institution for this work? If so, what form, how much, etc.?

- Do you know of any examples of duplication and the avoidance of duplication that have occurred because of data curation and sharing (e.g. clinical trials, surveys, etc.)? If so, what was the nature of the duplication and what second order impacts might there have been (e.g. survey fatigue and its impacts on response rates, medical treatments and their effects, etc.)?
- How likely is duplicative data collection in your field of research? If likely, how difficult, time consuming, expensive, etc. would duplicate collection be?

Fields of research vary greatly, so there may be other issues and questions relevant to the particular proposal.

9.3 Calculations and examples

In this section we present some hypothetical examples based on our case studies and the type of information that might be generated from answering the questions outlined above. The focus is on the way that the various types of information might be used to explore costs-benefits and build a 'business case' in the most simple and practicable way.

A reusable spreadsheet template for these calculations is available from JISC
(Temporary Test Version http://johnhoughtons.homeip.net/data-repository-template.xls)

9.3.1 Example I: Cost savings

The most direct and readily quantifiable benefits are those relating to cost savings. Here we suggest one approach to exploring the direct and indirect impacts of cost savings using a hypothetical example applicable to individual datasets or repositories.

Data requirements and sources

Data can be sourced and/or estimates generated from existing sources and the experience of existing repositories or through consultation with experts in the field from both the repository management and user communities.

Data requirements are modest and relate to the major cost elements involved, including:

- **Depositor costs:** The costs of data collection/creation and preliminary preparation faced by the depositor (sourced through consultation with users who have deposited or are depositing data);
- **Repository costs:** The costs of preparation and storage faced by the repository management, including an allowance for a share of the overall repository costs (sourced from existing repositories and/or consultation with repository managers and based on the 'Beagrie Model' for full economic costing); and
- **Withdrawer costs:** The costs of search, discovery and access faced by the withdrawers and users of the data (sourced from repository users).

These costs can be set against the savings realised from the reduction or elimination of duplication of collection/creation costs achieved through data curation and sharing.

Step 1: Direct cost savings from an individual dataset or repository

Direct cost savings can arise from the use and re-use of data made available for sharing, as the costs of the data collection/creation are shared across multiple users and the value of the data

realised by each user. A simple calculation is to sum the costs and set them against potential savings, which can be done for individual datasets or a repository as a whole.

Data repository costs vary widely, from modest levels in some fields to many millions of pounds in others (Ball *et al.* 2004), and the examples presented here are purely hypothetical. Take an example in which:

- The cost of data collection/creation (faced by the depositor) is £200,000;
- The cost of additional preparation for sharing (faced by the depositor and/or repository) is £10,000;
- The cost of searching for and accessing the data (faced by the user/withdrawer) is an average of £2,500 per use;
- The annualised cost of storage and access provision for the data concerned (faced by the repository) is £10,000; and
- The data are used/re-used 6 times over the 10 year life-cycle for which they are stored.

Where:

C1 = Cost of data collection/creation (faced by the depositor).

C2 = Cost of any additional preparation for sharing (faced by the depositor and/or repository).

C3 = Cost of searching for and accessing the data (faced by the user/withdrawer).

C4 = Annualised cost of data storage, including a share of overall repository costs (faced by the repository).

L = The expected useful lifespan of the data in years.

N = Number of times the data are likely to be used/re-used over that lifespan.

The costs would be:

*C1 + C2 + (C3 * N) + (C4 * L)*

or

*£200,000 + £10,000 + (£2,500 * 6) + (£10,000 * 10) = £325,000*

The benefits would be:

*C1 * N*

or

*£200,000 * 6 = £1,200,000*

And the benefit/cost ratio would be:

$$\frac{C1 * N}{C1 + C2 + (C3 * N) + (C4 * L)}$$

or

$$\frac{£1,200,000}{£325,000} = 3.7$$

(i.e. the benefits are almost four times the costs)

While hypothetical, this example shows that given the relative costs of research and curation, relatively low levels of use/re-use can justify the activity in terms of direct cost savings alone.

Table 9.3 Example I: Use/re-use leading to research cost savings

Direct cost savings from data use/re-use		Value
Data requirements:		
Cost of data collection/creation (faced by the depositor)	C1	£200,000
Cost of any additional preparation for sharing (faced by the depositor/repository)	C2	£10,000
Cost of searching for and accessing the data (faced by the user/withdrawer)	C3	£2,500
Annualised cost of storage of the data concerned (faced by the repository)	C4	£10,000
The life of the data in years	L	10
Number of times the data are used/re-used over the life-cycle	N	6
Direct cost savings (Step 1):		
Direct Costs = C1 + C2 + (C3 * N) + (C4 * L)	DC	£325,000
Direct Benefits = C1 * N	DB	£1,200,000
Direct benefit/cost ratio = (C1 * N) / (C1 + C2 + (C3 * N) + (C4 * L))	DBCR	3.7
Indirect cost savings (Step 2):		
Effective additional R&D spending = DB - DC	ARD	£875,000
Additional returns to R&D from that spending @ 20% = ARD * 0.20	AR	£175,000
Indicative total benefits = DB + AR	TB	£1,375,000
Indicative total benefit/cost ratio = TB / DC	TBCR	4.2

Source: Authors' analysis.

Step 2: Indirect cost savings from an individual dataset or repository

These research cost savings provide the basis for increased research activity, to the benefit of funders, institutional supporters, researchers and the users of research in government, industry and society (Houghton *et al.* 2006, p34). For the purposes of estimation it is reasonable to assume that the research costs saved would be spent on additional research (i.e. that there would be no substitution at the margin), thereby effectively increasing available R&D funding by the amount of the saving.

So in the example outlined above, R&D spending would, effectively, increase by:

*(C1 * N) - (C1 + C2 + (C3 * N) + (C4 * L))*

or

£1,200,000 - £325,000 = £875,000

Returns to R&D vary considerably between fields of research.[20] If we assume that the R&D expenditure savings went back into the overall pot of public research funding, then at the

[20] In one of the most thorough summaries of the literature, Martin and Tang (2007, pp6-7) noted that there have been numerous attempts to measure the economic impact of publicly funded R&D, all of which show a large positive contribution to economic growth, with studies spanning more than 30 years finding a rate of return to public R&D of between 20% and 50%. Similarly, Arundel and Geuna (2004, p3) noted that estimates of the rate of return to publicly funded research ranged between 20% and 60%. Based on a review of the literature, therefore, a very conservative estimate of average social returns to publicly funded research would be 20%.

conservative estimate of 20% social returns would be worth an additional £175,000[21] calculated as:

*£875,000 * 20% = £175,000*

In our hypothetical example, this would increase the attributable benefits to almost £1.4 million over the life-cycle, calculated as:

£1,200,000 + £175,000 = £1,375,000

Thereby, lifting the benefits to more than four times the costs, calculated as:

£1,375,000 / £325,000 = 4.2

Due to the lag between research expenditure and impacts these indirect impacts are no more than indicative *(see footnotes),* and it is important to note that the benefits accrue over time while many of the costs accrue up-front, and most costs accrue to producers and repository operators while many of the benefits accrue to external or third-party users.

9.3.2 Example II: Potential benefits

One might also expect there to be increased returns to expenditure on research that is curated and shared, as the curation and sharing add value to it.

In addition to cost savings, we noted a number of potential benefits from data curation and sharing relating to collaboration and enhanced outcomes, better education and research training, new opportunities and uses, a more complete and transparent record of 'science', potentially more sensitive and less invasive research evaluation, and greater visibility and reward. All of these impact the quality and efficiency of research over time.

However, impacts will vary from case to case and are by their nature diffuse and uncertain.[22] Nevertheless, in order to get some sense of the possible value of these potential benefits one can explore scenarios based on examples from existing data repositories or expectations about the proposed repository. At their simplest, estimates of the value of these wider impacts might be based on percentage point increases in returns to R&D expenditure at the institutional, disciplinary or sectoral levels.

Data requirements and sources

The only data requirements for such scenario building are repository costs and approximate expenditure on the research contributing data to the repository.

---

[21] These are recurring gains, albeit lagged to account for the time between the conduct of research and its impacts. Such returns can be expressed in Net Present Value (NPV), lagged and recurring over the useful life of the knowledge. However, NPV calculations are sensitive to the discount rates applied. For example, lagged 10 years and recurring for 10 years thereafter the £175,000 would be worth around £130,000 (NPV) using a very conservative discount rate of 10% per annum, and around £240,000 (NPV) at 5% per annum. While one might choose a real discount rate of 7%, returning around £187,000, we have taken the view that for the sake of simplicity and transparency we will simply take the original number (i.e. £175,000) as indicative of the value of the returns.

[22] Those exploring costs and benefits in order to develop a business case for a data repository may wish to go no further than the cost savings scenario outlined in Example I (above). However, we explore one possible approach to estimating these more diffuse additional benefits as a guide to their possible scale.

- Estimates of annualised repository costs can be derived as outlined above, and should be based on a full economic costing (i.e. the 'Beagrie Model' using TRAC fEC).
- Where the data repository is institutional it is necessary to estimate the share of institutional research likely to contribute data. Institutional research spending data are available from institutional research offices and HESA in the UK.
- Where the data repository is disciplinary it is necessary to estimate the share of disciplinary research likely to contribute data. Disciplinary research spending data can be derived from research council grants funding in that particular field from the individual research councils or RCUK in the UK, or from breakdowns of research funding by field reported by HESA.[23]

The other parameter required is an estimate of the impacts of data curation and sharing expressed as a percentage point change in returns to R&D expenditure. This can be no more than an estimate based on a sense of the potential contribution of the data to further research and its uses on the one hand, and reducing research costs and thereby increasing the efficiency of subsequent research on the other *(see box below for details)*.

Estimating the impacts of data curation and sharing as a percentage point change in returns to R&D expenditure

Of course, there is no way of knowing what the future returns to research spending will be. Informed (guess)timation is the only possible course, and making it informed depends on an honest assessment based on as much information as possible. This can begin with a simple scenario based on what is known.

- We know that average returns to publicly funded R&D are of the order of 20% to 60% (Arundel and Geuna 2004; Martin and Tang 2007);
- It is generally accepted that for journal articles there is an Open Access citation advantage of the order of 25% to 250% (Hajjem *et al.* 2005; EPS *et al.* 2006); and
- We know that researchers spend around 20% to 30% of their time reading and writing (Tenopir and King 2000 and subsequent tracking studies; Houghton *et al.* forthcoming).

If data curation and sharing led to a similar increase in use, and the generation, collection and analysis of data occupied as much research time as reading and writing, then:

- At the most conservative end of the scale, we might see a 25% increase in use (equivalent to a single re-use of 1 in 4 project datasets) of something that occupied 20% of research time, leading to a 5% increase in the use of the stock of research knowledge;[24] and
- At the upper end, we might see a 250% increase in use (equivalent to each project dataset being re-used two and a half times during their life-cycle) of something that occupied up to 30% of researcher time, leading to a 75% increase in use.

Given average returns of the order of 20% to 60%, then:

- At the most conservative, there would be a 1 percentage point increase in returns (i.e. a 5% increase of 20%);

[23] It should be noted that to the extent that research is sometimes not fully costed, funding will be lower than effective expenditure. For example, in UK higher education the gap between reported funding and expenditure is around 15% to 20%. Hence, estimates based on funding will tend understate returns based on expenditures and should, therefore, be seen as conservative lower bound estimates.

[24] The estimated increase in returns is based on using the share of research time as a proxy for the share of the stock of knowledge generated by research.

- At the upper end, there would be a 45 percentage point increase in returns (i.e. a 75% increase of 60%); and
- Crossing over, there would be a 3 percentage point increase in returns from a 5% increase in returns of 60%, and a 15 percentage point increase in returns from a 75% increase in returns of 20%.

So, taking the conservative end of possibilities suggests plausible increases in returns to R&D of, perhaps, 1-3 percentage points.

9.3.3 Example IIa: An institutional data repository

For an institutional data repository at a university or public research centre, the wider benefits can be explored by looking at estimated impacts on annual returns to R&D expenditure.

Taking a hypothetical example of a university that establishes a data repository and encourages the deposit of all research data that could usefully be shared, estimating that that might mean curating research data from around one-third of its research activities. Institutional research expenditure is £50 million per annum, and curation and sharing is expected to increase returns to R&D by 3 percentage points.

Where:

UniERD = Institutional expenditure on R&D.

RSD = Research share deposited.

PPR = Percentage point increase in returns to R&D resulting from data curation and sharing.

I-ARC = Annual institutional repository costs.

The impacts would be worth around £495,000 per annum in increased returns,[25] calculated as:

*(UniERD * RSD) * PPR*

or

*(£50,000,000 * 33%) * 3% = £495,000*

If the operational costs of the repository averages £250,000 per annum, then the estimated annual benefits would be double the annual costs, calculated as:

*((UniERD * RSD) * PPR) / I-ARC*

or

£495,000 / £250,000 = 2.0

[25] Again, these are recurring gains, albeit lagged to account for the time between the conduct of research and its impacts. Such returns can be expressed in Net Present Value (NPV), lagged and recurring over the useful life of the knowledge. However, NPV calculations are sensitive to the discount rates applied. For example, lagged 10 years and recurring for 10 years thereafter the £495,000 would be worth around £367,000 (NPV) using a very conservative discount rate of 10% per annum, and around £679,000 (NPV) at 5% per annum. While one might choose a real discount rate of 7%, returning £528,000 million, we have taken the view that for the sake of simplicity and transparency we will simply take the original number (i.e. £495,000) as indicative of the value of the returns.

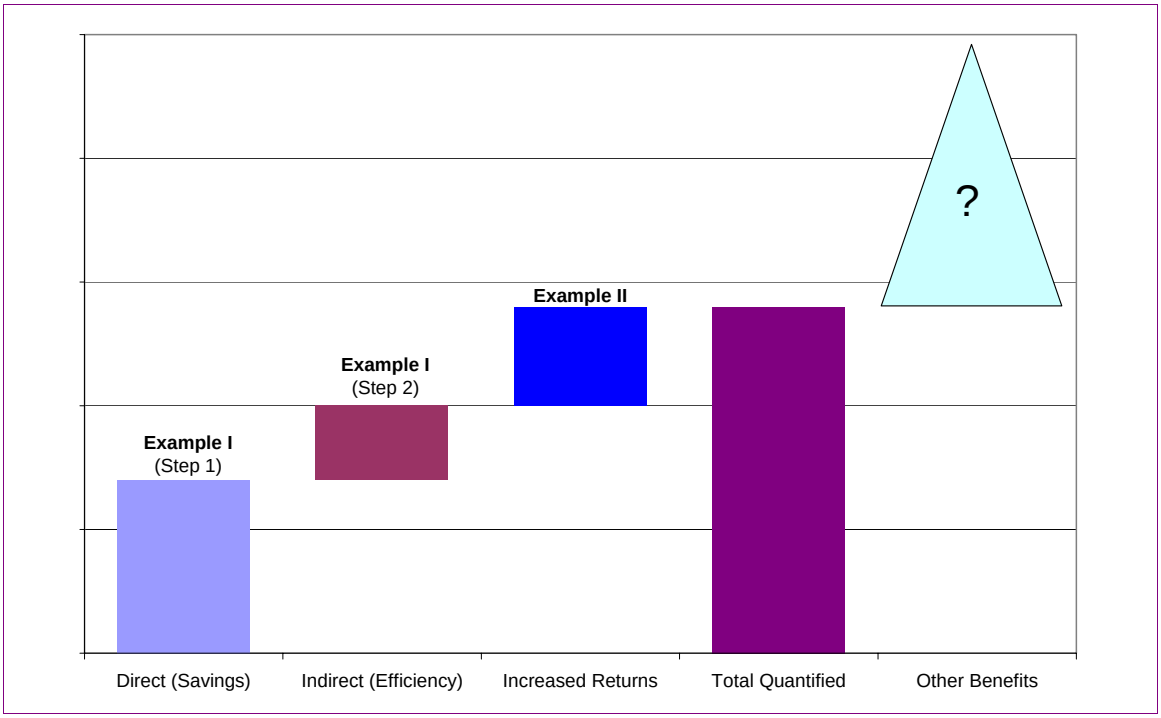
Table 9.4 IIa: An institutional repository

Impacts on returns to R&D		
Data requirements:		
Annual repository costs	I-ARC	£250,000
Annual institutional research spending	UniERD	£50,000,000
Research share deposited	RSD	33%
Percentage point increase in returns	PPR	3%
Impacts:		
Impacts (Additional annual returns) = (UniERD * RSD) * PPR	HB	£495,000
Benefit/Cost ratio = HB / I-ARC	BCR	2.0

Source: Authors' analysis.

This does not account for the additional depositor and withdrawer user costs, but is in addition to the benefits from the cost savings and efficiency gains outlined in Example 1.

Figure 9.2 Types of benefits explored in examples



Note: Not to scale.

Source: Authors' analysis.

9.3.4 Example IIb: A disciplinary data repository

Similarly, an agency establishes a data repository and encourages deposit of all research data that could usefully be shared, estimating that that might mean curating research data from around one-third of its research activities. Disciplinary research expenditure is £250 million per annum, and curation and sharing is expected to increase returns to R&D by 3 percentage points.

Where:

DisERD = Disciplinary expenditure on R&D.

RSD = Research share deposited.

PPR = Percentage point increase in returns to R&D resulting from data curation and sharing.

D-ARC = Annual disciplinary repository costs.

The impacts would be worth around £2.5 million per annum in increased returns,[26] calculated as:

*(DisERD * RSD) * PPR*

or

*(£250,000,000 * 33%) * 3% = £2,475,000*

If the operational costs of the repository average £500,000 per annum, then the benefits would be 5 times the costs, calculated as:

$(DisERD * RSD) * PPR / D\text{-}ARC$

or

$£2,475,000 / £500,000 = 5.0$

Again, it should be noted that this does not account for the additional depositor and withdrawer user costs, but is in addition to the benefits from the cost savings and efficiency gains outlined in Example 1.

Table 9.5 IIb: A disciplinary repository

Impacts on returns to R&D		
Data requirements:		
Annual repository costs	D-ARC	£500,000
Annual institutional research spending	DisERD	£250,000,000
Research share deposited	RSD	33%
Percentage point increase in returns	PPR	3%
Impacts:		
Impacts = (DisERD * RSD) * PPR	HB	£2,475,000
Benefit/Cost ratio = HB / D-ARC	BCR	5.0

Source: Authors' analysis.

9.3.5 Example III: A system of data repository in UK HEIs

This same calculation could be done for the higher education sector, based on estimates of the costs of a system of data repositories.

[26] Again, these are recurring gains, but lagged to account for the time between the conduct of research and its impacts. As noted above, we have taken the view that for the sake of simplicity and transparency we will simply take the original number (i.e. £2.5 million) as indicative of the value of the returns.

For example, if all 168 UK universities were to establish a data repository, estimating that that might involve curating research data from around one-third of higher education research activities, then with HERD at £6.1 billion in 2006 and curation and sharing expected to increase returns to R&D by 3 percentage points, the impacts would be worth £60 million per annum in increased returns, calculated as:

*(£6,062,000,000 * 33%) * 3% = £60,013,800*

If the operational costs of the repositories averaged £200,000 per annum per repository and there was one in each university, then the benefits would be 1.8 times the costs.

Table 9.6 Example III: A system of repositories in UK HEIs

Impacts on returns to R&D		
Data requirements:		
Annual repository costs	HE-ARC	£200,000
Annual institutional research spending in 2006	HERD	£6,062,000,000
Research share deposited	RSD	33%
Percentage point increase in returns	PPR	3%
Number of repositories in the system	NR	168
Impacts:		
Impacts = (HERD * RSD) * PPR	HB	£60,013,800
Benefit/Cost ratio = HB / (HE-ARC * NR)	BCR	1.8

Source: Authors' analysis.

9.4 Limitations and caveats

There a number of issues to consider in the interpretation of these hypothetical examples.

Attribution and impacts: It is very difficult to attribute particular impacts and outcomes to a particular research project or stream of research expenditure. There will, of course, be major breakthroughs traceable to particular projects, but research is cumulative, building on past research. A full attribution would need to trace the work back to its origin and attribute a value to the contributions arising from each step along the way. Realistically, this cannot be done. Consequently, while we may know the average return to R&D spending at an aggregate level, it tells us nothing about returns to a specific project.

Returns to R&D: Returns to R&D vary considerably from field to field. At the aggregate level social returns to publicly funded research are typically of the order of 20% to 60% per annum (Arundel and Geuna 2004, p3), but at the project level returns are too varied to be predictable. Conservative estimates of average returns can be interpreted as indicative of the possible orders of magnitude of returns one might expect from reasonably aggregated expenditures, but should be treated with great caution at the disciplinary level and even greater caution at the institutional level (e.g. returns to medical research may be much higher than those to humanities research, and this may make a substantial difference when considering disciplinary data repositories and/or where institutions have different research mixes).

Lags between research and its impacts: There are lags between research expenditure, the promulgation of findings, and the application and use that are both considerable and varied.

Lags can be long in some fields, perhaps up to 20 or 30 years, and short in others, perhaps 1 to 2 years or less. Mansfield (1991) reported that for US firms the average lag between the publication of academic research and the timing of subsequent commercial innovation relying on it was seven years. Allowing a further three years for the lag between project activity and publication/promulgation (e.g. by making the data available) might add a further three years. This suggests that an average lag of around 10 years may be indicative, but it may vary significantly between disciplines and fields of research (e.g. lags in electronics research may be much shorter than in geology).

Recurring gains: Returns are recurring annual gains from one years' research spending which accumulate over time, subject to rates of accumulation and obsolescence of the knowledge generated and to rates of growth in R&D spending. They can be expressed in Net Present Value (NPV), but such calculations are very sensitive the discount rates and can be no more that indicative.

Handling the time, lags and recurring gains: Estimations of impacts on returns that take account of time, lags and the recurrence of gains are sensitive to the discount rate used. For example, in Example I *(above),* if lagged 10 years and recurring for 10 years thereafter the £175,000 in increased returns would be worth around £130,000 (NPV) using a very conservative discount rate of 10% per annum, and around £240,000 (NPV) at 5% per annum. While one might choose a real discount rate of 7%, returning around £187,000, we have taken the view that for the sake of simplicity and transparency we could simply take the original number (i.e. £175,000) as indicative of the value of the returns.

10 Appendix 4: Bibliography

Arundel, A. and Geuna, A. (2004), Proximity and the use of public science by innovative European firms, E*conomics of Innovation and New Technology* 3(6), pp. 559-580.

Ball, C.A., Sherlock, G. and Brazma, A. (2004), Funding high-throughput data sharing, *Nature Biotechnology* 22(9), pp. 1179-1183.

Batty, M. (2005), Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models and Fractals. London: The MIT Press.

BBSRC (2007), Data sharing policy. Available at http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf , accessed August 2008.

Beagrie, N., Chruszcz, J. and Lovoie, B. (2008), Keeping research data safe: a cost model and guidance for UK Universities. Report for the JISC by Charles Beagrie Ltd. Available at http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf, accessed September 2008.

Belmonte, M K et al (2007), Offering to share: how to put heads together in autism neuroimaging, *Journal of Autism and Developmental Disorders* 38, p. 2-13.

Berman, Helen M (2008), The Protein Data Bank: a historical perspective, *Acta Crystallographica A: Foundations of Crystallagraphy*, A64, pp. 88-95.

Borgman, C., Wallis, J C. and Enyedy, N (2006), Building digital libraries for scientific data: an exploratory study of data practices in habitat ecology IN Gonzalo, J et al (eds), *ECDL* 2006, pp. 170-183.

Borgman, C.L. (2006), What can studies of e-learning teach us about collaboration in e-research? Some findings from digital library studies, *Computer Supported Co-operative Work,* 15, pp. 359-383.

Calvert, J. and Williams, R. (2008), Data sharing in the biosciences: workshop report. Available from http://www.genomicsnetwork.ac.uk/innogen/publications/briefingsreports/title,7950,en.html accessed October 2008

Carlson, S. and Anderson, B. (2007), What are data? The many kinds of data and their implication for data re-use, *Journal of Computer Mediated Communication.* Available from http://jcmc.indiana.edu/vol12/issue2/carlson.html., accessed July 2008.

Casey, K (2003), Issues of electronic data access in biodiversity IN Wouters, P and Schroder P, *The Public Domain of Digital Research Data: Promise and Practice in Data Sharing.* Amsterdam: NIWI-KNAW, pp. 41-64.

David, P A (2006), Towards a cyberinfrastructure for enhanced scientific collaboration: providing its 'soft' foundations may be the hardest part. IN Foray, D. and Kahin, B. (eds), *Advancing knowledge and the Knowledge Economy*. MIT Press 2006.

Delson E et al (2007), Databases, data access and data sharing in paleoanthropology: first steps, *Evolutionary anthropology* 16, pp. 161-163.

Electronic Publishing Services and the Department of Information Science, Loughborough University (2006), *UK scholarly journals: 2006 baseline report – An evidence-based analysis of data concerning scholarly journal publishing*, Research Information Network, Research Councils UK and Department of Trade and Industry. Available at http://www.rin.ac.uk/data-scholarly-journals, accessed February 2008.

ESRC Data Sets Policy, http://www.esrc.ac.uk/ESRCInfoCentre/Images/ESRC_Research_Funding_Guide_June_2008_tcm6-9734.pdf, accessed September 2008.

Friedman, S. L. (2007), Finding treasure: data sharing and secondary analysis in developmental science. Introduction to special edition of the *Journal of Applied Developmental Psychology* 28, pp. 384-389.

Frontier Economics (2007), Evaluating the impact of ESRC funding: a report prepared for the Economic and Social Research Council, October 2007. Available from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Evaluating%20the%20Economic%20Impact%20of%20ESRC%20Research_tcm6-25908.pdf, accessed September 2008.

Fry,J., Schroeder. R, and Den Besten, M. (forthcoming), Open science in e-science: Contingency or policy, Journal of Documentation, 65(1), 2009.

Gillies, V. and Edwards, R. (2005), Secondary analysis in exploring family and social change: addressing the issue of context, *Forum: Qualitative Social Research,* 6(1). Available at http://www.qualitative-research.net/index.php/fqs/article/view/500/1076, accessed October 2008.

Hajjem, C, Harnad, S. and Gingras, Y. (2005), Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact, *IEEE Data Engineering Bulletin* 28(4), pp. 39-46.

Henty, M., Weaver, B., Bradbury, S. and Porter, S. (2008), Investigating data management practices in Australian Universities. Report to the Australian Partnership for Sustainable Repositories, July 2008. Available at http://www.apsr.edu.au/investigating_data_management, accessed September 2008.

Hey, A. J. C. and Trefethen, A. (2003), The data deluge: an e-science perspective IN Berman, F., Fox G C and Hey A J C (eds), *Grid Computing: making the global infrastructure a reality.* Chichester: John Wiley.

Houghton, J.W., Steele, C. and Henty, M. (2003) *Changing Research Practices in the Digital Information and Communication Environment*, Department of Education, Science and Training, Canberra.

Houghton, J.W., Steele, C. and Henty, M. (2004) Research Practices and Scholarly Communication in the Digital Environment, *Learned Publishing,* 17(3) pp. 231-249.

Houghton, J.W. Steele, C. and Sheehan, P.J. (2006), Research Communication Costs in Australia, Emerging Opportunities and Benefits. Report to the Department of Education, Science and Training, Canberra. Available at http://dspace.anu.edu.au/handle/1885/44485, accessed February 2008.

Houghton, J.W. and Sheehan, P.J. (2006), *The Economic Impact of Enhanced Access to Research Findings*, CSES Working Paper No.23, Victoria University, Melbourne (July 2006). Available at http://www.cfses.com/documents/wp23.pdf, accessed November 2008.

Houghton, J.W. et al. (forthcoming), Economic implications of alternative scholarly publishing models: Exploring the costs and benefits, A report by The Centre for Strategic Economic Studies, Victoria University and The Departments of Information Science and Economics, and LISU, Loughborough University for the JISC.

JISC Data Audit Framework, http://www.data-audit.eu, Accessed August 2008

JISC Digital Repositories Programme, http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005.aspx, accessed October 2008.

The "Joint Standards Study" (2005), Large-scale data sharing in the life sciences: data standards, incentives, barriers and funding models. Report prepared by the Digital Archiving Consultancy (DAC), Bioinformatics Research Centre, Glasgow University and the National e-Science Centre for the BBSRC, MRC, NERC, JISC, Department of Trade and Industry and the Wellcome Trust, August 2005. Available at http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552, accessed September 2008.

Katz, J.S. and Hicks, D. (1997), *How much is collaboration worth? A calibrated bibliometric model*, Brighton: University of Sussex (SPRU).

Katz, J.S. and Martin, B.R. (1997), What is Research Collaboration?, *Research Policy 26*, pp. 1-18. Available www.sussex.ac.uk/Users/sylvank/pubs/Res_col9.pdf accessed March 2003.

Lyon, Liz (2007), Dealing with data: roles, rights, responsibilities and relationships, UKOLN. Available at http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf, accessed September 2008.

Mansfield, E. (1991) Academic research and industrial innovation, *Research Policy* 20(1), pp. 1-12.

Martin, B. R. and Tang P. (2007), *The benefits from publicly funded research*. Science and Technology Policy Research Working Papers, No. 161, University of Sussex. Available at http://www.sussex.ac.uk/spru/documents/sewp161.pdf, accessed September 2008.

Martinez-Uribe, Luis (2008), Findings of the scoping study interviews and the research data management workshop: scoping digital repository services for research data management, A project of the Office of the Director of IT, University of Oxford. Available from http://www.ict.ox.ac.uk/odit/projects/digitalrepository/, accessed September 2008.

Medical Research Council (2007), The use of personal health information in medical research: general public consultation. Report by Ipsos MORI for the MRC, June 2007. Available at http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC003810, accessed September 2008.

Moore, Niamh (2007), (Re)Using Qualitative Data?, *Sociological Research Online,* 12(3). Available at http://www.socresonline.org.uk/12/3/1.html, accessed July 2008.

MRC, Data sharing initiative. http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/AimsoftheMRcDataSharingandPreservationInitiative/index.htm. accessed July 2008.

Murray-Rust, Peter http://wwmm.ch.cam.ac.uk/blogs/murrayrust/, accessed August 2008.

Patzold, H. (2005), Secondary analysis of audio data: technical procedures for virtual anonymisation and modification, *Forum Qualitative Social Research,* 6(1). Available at http://www.qualitative-research.net/index.php/fqs/article/view/512/1106, accessed October 2008.

Murray-Rust, Peter (2008) Open Data in Science http://www.dspace.cam.ac.uk/handle/1810/194890 ,accessed September 2008.

OECD (2007), Principles and guidelines for access to research data from public funding, Organisation for Economic Co-operation and Development 2007. Available at http://www.oecd.org/dataoecd/9/61/38500813, accessed July 2008.

PA Consulting Group (2007), Study on the economic impact of the Research Councils. Report to RCUK by PA Consulting Group and SQW Consulting, October 2007. Available from http://www.rcuk.ac.uk/innovation/impact/default.htm, accessed September 2008.

Parry, Odette and Mauthner, Natasha (2005), Back to basics: who re-uses qualitative data and why?, *Sociology* 39, pp. 337-342.

RCUK (2007) Increasing the economic impact of the research councils, January 2007. Available from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/publications/ktactionplan.pdf, accessed September 2008.

RCUK (2007a) Excellence with impact, October 2007. Available from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/economicimpact/excellenceimpact.pdf, accessed September 2008.

RCUK (2006) Position statement on access to research outputs (updated), Available from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf, accessed July 2008.

RELU http://www.relu.ac.uk, accessed June 2008.

RELU Data Support Service http://www.data-archive.ac.uk/relu/, accessed June 2008.

RIN (2008), To share or not to share: publication and quality assurance of research data outputs. Available at http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf, accessed September 2008.

RIN (2008a), Stewardship of digital research data: a framework of principles and guidelines. Available at http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf, accessed September 2008.

RIN (2007), Research funders' policies for the management of information outputs. Available at http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20-%20Final%20Report.pdf, accessed September 2008.

SERCO (2008), The UK research data service feasibility study: interim report, July 2008. Available from http://www.ukrds.ac.uk/. Accessed September 2008.

Swan, A (2008), The Big Picture and researchers' key concerns within the scholarly communications process: report to JISC Scholarly Communications Group, March 2008. Key Perspectives.

Tenopir, C. and King, D.W. (2000), *Towards Electronic Journals: Realities for Scientists, Librarians and Publishers,* Special Libraries Association, Washington D.C.

Thompson, P (2004), Pioneering the life story method, *International Journal of Social Research Methodology* 7(1), pp. 81-84, see ESDS Qualidata Online http://www.esds.ac.uk/qualidata/online/data/edwardians/introduction.asp

Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T. and Estabrooks, C.A. (2005), Central questions of anonymization: a case study of secondary use of qualitative data, *Forum Qualitative Social Research,* 6(1). Available at http://www.qualitative-research.net/index.php/fqs/article/view/511/1102, accessed October 2008.

Van den Berg, H. (2005), Reanalyzing qualitative interviews from different angles: the risk of decontextualization and other problems of sharing qualitative data, *Forum Qualitative Social Research,* 6(1). Available at http://www.qualitative-research.net/index.php/fqs/article/view/499/1074, accessed October 2008.

Walsh, J.P., and Maloney, N.G. (2001), Computer Network Use, Collaboration Structures and Productivity. IN P. Hinds and S. Kiesler, (eds.) *Distributed work,* Cambridge: MIT Press. Available http://tigger.uic.edu/~jwalsh/Collab.html ,accessed October 2008.

Wellcome Trust (2003), Sharing Data for Large-scale Biological Research Projects: A System of Tripartite Responsibility. Report of a meeting organized by the Wellcome Trust, held on 14-15 January 2003 at Fort Lauderdale, USA.

Whyte, A., Job, D., Giles, S. and Lawrie, S. (2008), Meeting curation challenges in a Neuroimaging Group, *The International Journal of Digital Curation* 1(3), pp. 171-181.

Wouters, P. and Reddy, C. (2003), Big science data policies. IN P. Wouters and P. Schroder (eds), *The Public Domain of Digital Research Data: Promise and Practice in Data Sharing.* Amsterdam: NIWI-KNAW

Special issues of journals:

Journal of Applied Developmental Psychology 28 (5-6), 2007, New findings from secondary data analysis: results from the NICHD study of early childcare and youth development

Sociological Research Online 12(3), 2007, Reusing qualitative data

Forum Qualitative Social Research 6(1), 2005, Secondary analysis in qualitative data