INFRASTRUCTURE PLANNING AND DATA CURATION

A COMPARATIVE STUDY OF INTERNATIONAL APPROACHES TO ENABLING THE SHARING OF RESEARCH DATA

VERSION 1.6

30 NOVEMBER 2008

Prepared by:
Raivo Ruusalepp
Estonian Business Archives Consultancy

EXECUTIVE SUMMARY

The current methods of storing research data are as diverse as the disciplines that generate them and are necessarily driven by the myriad ways in which researchers need to subsequently access and exploit the information they contain. Institutional repositories, data centres and all other methods of storing data have to exist within an infrastructure that enables researchers to access and exploit the data, and variant models for this infrastructure can be conceptualised. Discussion of effective infrastructures for curating data is taking place at all levels, wherever research is reliant on the long-term stewardship of digital material. JISC has commissioned this study to survey the different national agendas that are addressing variant infrastructure models, in order to inform developments within the UK and for facilitating an internationally integrated approach to data curation.

The study of data sharing initiatives in the OECD countries confirmed the traditional perception that the policy instruments are clustered more in the upper end of the stakeholder taxonomy – i.e. at the level of national and research funding organisations – whereas the services and practical tools are being developed by organisations at the lower end of the taxonomy. Despite the differences that exist between countries in terms of the models used for research funding, as well as the levels at which decisions are taken, there is agreement on the expected strata of responsibility for applying the instruments of data sharing. This supports the structure of stakeholder taxonomy used in the study.

POLICY SUPPORT FOR DATA SHARING

The lack of a universal model for data sharing policies appears to be a fundamental consequence of research funding models differing between individual countries. This study found no evidence of either a universal model or agreement on what a data sharing policy should include.

On an **international level**, the key players (organisations like OECD, UNESCO, EU and interest groups like CODATA, ESFRI) have concentrated their policy statements around the principle of open access to publicly funded research outputs. While OECD, UNESCO and CODATA have policies explicitly for data sharing, the European Commission is looking at data sharing issues in the broader context of open access to public domain information.

No **national level** policies or strategic documents that explicitly mandate the sharing of research data were found. Nevertheless, the provision of access to research data is seen as a vital element of the general research infrastructure, and all research infrastructure development strategies acknowledge the need to develop the means for accessing data. Applying Open Access principles to data is discussed at the national level in Germany.

The main burden of developing and implementing data sharing policies is currently being carried by **research funding agencies**, with an expectation (but not a mandate) that individual research institutions and departments will follow these up with their own policy statements. Measures to motivate researchers into sharing their data incorporate conditions being attached to funding schemes or the provision of data sharing policies backed up by services offered to recipients of funding. The prospect of a more pro-active stance in mandating the sharing of data is evidenced in the recent initiatives of funding agencies to agree on common principles for data sharing.

Typically, but not in all cases, the funding agency policies draw on the following incentives and enablers:

Policy Enablers	Aspects Covered
International level examples and statements	General policy statements
National strategic planning documents and mandates	Obligation / mandate to share data
Research associations' statements and codes of ethics	Division of responsibilities between stakeholders
Open Access principles	What data sharing channels should be used
Government funding for research infrastructure	How can the costs involved in data sharing be covered
Government audit and watchdog offices' reports and requirements	What sanctions can be applied if the data sharing requirements are not being met
	Data access principles and protection of data subjects' rights
	Conditions of exclusive use of data

The emerging **institutional policies** still remain *ad hoc* and do not appear to be well coordinated. To develop uniform data sharing policies and put them into practice, the institutions will currently require significant help and guidance.

DATA SHARING INFRASTRUCTURE PROVISION

Policies alone will not result in a higher use of research data. Optimum accessibility and usability of data presuppose a trajectory of proper organisation and curation of data, with access services and analysis tools that provide the researchers with added value.

Proposals for **national data services** have opted for a distributed, umbrella-type approach where the national service provides the environment for repositories – common principles and standards that data repositories in the country apply, and develop tools that facilitate interaction between repositories. The main expected outcomes are better data curation and dissemination services that are based on shared tools and principles.

Data archives and centres funded directly by **research funding agencies** are the dominant class of data repositories. But there is a variance in how data curation and sharing infrastructure is offered and models of how these are used in different research domains. In domains such as the social sciences and medicine a strong tradition exists for depositing data in national data centres, which are usually directly funded by the funding agencies; in astronomy, biomedicine, earth sciences and physics, data centres with a profile of international dissemination are favoured by researchers. The first examples of funding agencies relying on a network of institutional data repositories are emerging (e.g. AHRC in the UK that stopped funding the AHDS and is relying on institutional service provision, or the Helmholtz Society in Germany), whilst some data centres are offering services to more than one funding agency (e.g., ICPSR in the US). Nonetheless, differences still remain in the degree to which funding agencies take responsibility for data sharing, as well as the extent to which they communicate data sharing principles to their research community.

Institutional repositories have until recently put emphasis on the deposit of textual research output. The scope of these repositories is gradually being extended to cover research data as well, but the overall number of stored datasets is very low. Institutional data repositories hold promise for the future with the advantage of being close to researchers, but are at present entangled in a maze of shortages in expert know-how and resources, unclear responsibilities for maintaining the repository (e.g. university library vs IT services), and insufficient institutional policy support. The business case for supporting a data repository is not yet clear for many research institutions.

DATA FEDERATION AND ACCESS SERVICES

In the increasingly international and interdisciplinary context of research, locating data in disparate repositories in different countries, gaining access to them through a web of licence agreements in different languages, and re-using them in a multitude of file formats can be a daunting task. These barriers are not easy to overcome – the sheer diversity of data makes it difficult to design tools with the range and ability to accommodate and translate between the distinctly different data needs of the various domain communities.

To bridge these gaps, a significant portion of data sharing infrastructure funding is being allocated to developing technical solutions for data federation from different repositories in one research domain and across domains. Portal services are emerging that harvest metadata from disparate data repositories and allow the creation of entire cross-sections of research output on national or research domain levels. Digital repository system tools are appearing that allow the integration and management of textual, multimedia and data object collections.

These services are predominantly developed by short-term projects, which inevitably are faced with the transition to a sustainable service environment, with a long-term financial and business structure (e.g. the CARMEN project). Development of data access tools and services has started to receive government funding and backing in several countries (e.g., the US, Australia, Netherlands, France).

DATA SHARING SERVICES IN SUPPORT OF THE RESEARCH PROCESS

To support collaboration between research groups, tools are emerging for the dissemination and sharing of data between disparate groups across diverse disciplines. The data often need to be shared between small

and medium sized laboratories and institutes that may have very different computing environments and levels of IT expertise. To help with automation of the research process and reduce the effort that goes into data conversion, various virtual research environments and researchers' toolbox solutions are being developed. These are predominantly project-based initiatives at this stage, but in the case of Germany and Japan have the backing of a nation-wide platform.

RESEARCHER SKILLS FOR DATA SHARING

Data publishing to a standard that facilitates re-use requires the effective planning and management of data throughout the life-cycle of a project. Studies in the UK and Australia have demonstrated low awareness of policies and requirements, and a lack of adequate data management skills among researchers. Similar conclusions have been drawn from digital library user surveys. Researchers require guidance in translating policy requirements, including open access policy, into operational tasks for which they can plan and take responsibility.

Examples of data management plans that are increasingly required as conditions for receiving funding have been produced in Australia. Good examples of data management and curation manuals have been developed by the UK Data Archive, DCC and ICPSR.

SUMMARY OF THE ANALYSIS

POLICY SUPPORT FOR DATA SHARING

Collaboration in the development of data sharing and curation policies can be seen on all levels: international interest groups are issuing joint policy statements, national co-ordination offices are being set up, research funding agencies are agreeing to shared principles, universities form self-organised groups to support the open access principles. Different contributors in policy collaboration have their own agendas to press (e.g. open access supporters and publishers), hence, a clear vision and co-ordination of effort is required. Policies are an important outcome of this collaboration, but as RIN in the UK and GAO in the US have pointed out, the policies have to be accompanied with effective mechanisms for checking how they are being implemented.

National level strategy documents are being developed to set priorities for government spending on research infrastructure development. The collaborative effort of developing such policies has usually been led by an umbrella organisation (e.g. national research foundation, academy of sciences) or in some cases as a bottom-up process by initiatives from research communities (e.g. RIN in the UK). Because of differences in data collection, use and management practices in different domains of research, national level policies remain too general to be useful in practice. The main benefit of national level strategic documents is the identification of roles for developing further, more specific policies and data sharing services.

Funding agencies are better positioned to follow up on how research projects fulfil their policy requirements, but the practice is variable. Researchers in disciplines that have large centralised data centres benefit from the availability of expertise and resources for data curation, whereas other funding agencies often do not have efficient mechanisms in place for ensuring that their policies are being followed. Several calls for more uniform data sharing policies have been made (e.g., RIN in the UK, GAO in the US), especially to facilitate shared principles across interdisciplinary research boundaries. A significant agreement of common principles and standards amongst the funding agencies for widening access to research data is being stimulated by statements from international groups including the Open Access movement.

A natural focal point where higher-level policy requirements and incentives for researchers to share their data meet is at the institutional and departmental level. Therefore, creating data management and sharing policies on an institutional and/or departmental level would be the rational choice. These policies could still follow the broad requirements of national agencies and the research domain, but they would be designed for operation in the context of individual research projects. Institutions themselves have a vested interest in sharing data, and in some jurisdictions may share or own the intellectual property rights to the data, but institutional data sharing policies are not yet very common. Whilst growing awareness of the open access principles is increasing interest in methods for data sharing, most of the existing institutional level policies for openly sharing research outputs do not yet incorporate research data.

COSTS OF DATA SHARING INFRASTRUCTURE

Data management and sharing needs long-term vision and long-term support, that individual institutions and projects alone cannot provide. The significant costs of data sharing infrastructure provision have mostly been borne by national governments who continue to support directly the (centralised) services and participate in funding research domain level and institutional services. With new models for sharing research data appearing, the question is arising about whose funds could or should be used for developing and maintaining the services on institutional, project and individual researcher levels. The cost figures of data sharing are also vital for budgetary planning purposes on all levels. Yet real cost figures are hard to obtain as data sharing is 'bundled' with other services, most often with archiving and the preservation of data.

To estimate the cost of curating and making data available for re-use institutions first need to take stock of their data resources. Tools like the DCC's Data Audit Framework (DAF) help with the identification of data assets and ULCC's DAAT and DCC/DPE DRAMBORA are of value in assessing what risks are being faced in curating them. These tools do not cover the issues of data quality that are essential in appraising the value of data assets and establishing the scope of data curation activities. Policies and requirements that are being put in place apply to new data being generated from research. There is a host of existing data that potentially have a considerably larger need for collection, curation and dissemination. This cannot be achieved without significant cost and effort (examples of NDAD and ICPSR projects show that this retro-curation or even digital archaeology is extremely costly).

DATA SHARING INFRASTRUCTURE PROVISION

There is no obvious need to dismantle the existing data curation and sharing infrastructure, which in the UK is mostly based on data centres supported by research funding organisations. In the main, universities and research institutes are either not ready, or it is not appropriate for them, to take this task away from centralised data services. Clear policies, more resources and more skills are needed to allow universities to enter the data management and curation realm, but with the choice of data sharing channels expanding, the physical location of data becomes less and less relevant: access and dissemination services can harvest data from a variety of repository environments. Given that one method of data sharing does not preclude the use of other ones, and ultimately it is the researcher who decides which (additional) channels to use for dissemination, the university and institutional data repositories and social networking web services may become more popular in the future. It would still be in the interest of research funding agencies to ensure that data created with their funding are released to the public domain, are adequately described, curated over time, and the necessary data security rules are effectively applied. The centralised data repositories will continue to provide such a data control regime, but they should consider implementing mechanisms that allow institutional repositories to harvest metadata and link to actual data in their repositories, giving the institutions an opportunity to disseminate the data as linked resources.

In the longer term, however, there is a need to produce and adopt universal rules for data description, to define minimum data curation services, and to identify rules for data security that are designed for use across different disciplines. The implication here is for more collaboration and the provision of more practical tools for use at the institutional level, with lessons learned from the experiences of established data curation institutions and centres.

RESEARCHER SKILLS FOR DATA SHARING

Whether and how a research project's data will be shared in practice depends upon the prevailing attitudes and cultures in the research domains. The policies and conditions of grants awarded by the funding agencies are but one among the many reasons for researchers to decide on sharing their research data. Researchers' awareness of these data sharing policies has been reported to be low (UKRDS survey returned only 66% positive responses) and should be improved, yet researchers have other incentives and requirements to share their data: principally, they want to publicise the results of their research, which in some cases includes data; some publishers require data underlying an article to be made available on request to other researchers; codes of ethics and agreements may require data sharing with other research projects in the same area; adherence to open access principles, which increasingly are applied not only to printed materials, but all other types of research outputs, and other informal or internal agreements can motivate data sharing.

The research world is strongly focussed on publication as an outcome – publications, citation rates and impact factors are the traditional research assessment indicators. First attempts are being made (e.g. in Germany) to change the research assessment rules to also include data publications. If a general agreement was to be reached on this then researchers would have an increased incentive to share data and adhere to various policy requirements, although mechanisms for ensuring the quality of data publications would first have to be put in place.

Decisions affecting the practice of good data management at the level of an individual research project are influenced by many factors in addition to data sharing policies. The act of creating data management plans (required increasingly to accompany new project funding proposals) has the potential for incorporating structured guidance on how research data should be managed throughout their lifecycle. Examples of data management plans have been published in Australia and the US. A further step could be to link the data management requirements with the data curation models that are being produced in several countries.

Recommendations for JISC and other stakeholders for further work in the area of research data sharing are presented in Chapter 5.

CONTENTS

1. INTRODUCTION

The value of research data lies in their use. Data are recognised on all levels as a valuable resource that should be made publicly available and maintained over time to ensure that their potential value is realised. Access to research data will not only ease the validation of published findings but will also permit reconstructing the research event, reanalysis, reducing potential costly duplication, and opening new research opportunities and creation of new knowledge.

Policies and strategies to achieve these gains through improved access to research data are being put in force on all levels – from international organisations and interest groups, national science organisations, research funding bodies and professional associations, to individual universities and research projects. In January 2004 national governments including that of the UK signed the *OECD Declaration on Access to Research Data from Public Funding*[1] highlighting opportunities and agreeing to work towards commonly agreed principles and guidelines. National level strategy documents are being developed to set priorities for government spending on research infrastructure development. Research funding agencies are collectively agreeing on open access principles to research outputs and publishing data sharing policies that guide the requirements funded research projects have to meet.

However, policies alone will not result in a higher availability and use of research data. Optimum accessibility and usability of data presuppose an infrastructure for proper organisation and curation of data, with access services and analysis tools that provide the researchers with added value to raw data. Infrastructure to support research is receiving much attention and major funding is being made available to develop it. The European Commission calls it e-infrastructure and defines it as a new research environment in which all researchers have shared access to unique or distributed scientific facilities (including data, instruments, computing and communications), regardless of their type and location in the world.[2] Data repositories that provide data curation and sharing are a vital component of this support infrastructure that in the UK is referred to as 'e-science' or 'research infrastructure', in the US as 'cyberinfrastructure', and as 'e-research' in Australia. The new research infrastructure services have to take into account the increasing trend towards establishing institutional repositories for collection and dissemination of research output, including data.

The JISC report *Dealing with Data*[3] conceptualised two high-level data flow and sharing models: the 'Domain Data Deposit Model' and the 'Federation Data Deposit Model'. An additional 'Open Science Model' has been proposed to take account of widely practiced depositing of data straight onto web-based resources such as blogs and wikis. The so called 'big science' data sharing model that relies on grid technology should also be added to this list. New tools that form building blocks of the Virtual Research Environment are being developed and are likely to emerge as another data sharing model where data stores are directly integrated with automated research workflows.

The technology that supports all these new opportunities continues to evolve rapidly, though researchers' attitudes to data creation and dissemination and skills to implement all the requirements are not keeping pace in all disciplines. Ultimately it is the researcher who needs access to data in repositories and also who decides which channels to use to disseminate the data resulting from a research project. Preliminary results from a worldwide survey of researchers conducted by the PARSE.Insight project show that researchers:[4]

Researcher activity	Proportion
Collaborate in their discipline and across disciplines	65%
Make use of data from other disciplines	39%
Have needed digital research data gathered by other researchers that was not available	80%

Table 1. Researchers' data sharing needs.

[1] OECD, *Declaration on Access to Research Data from Public Funding* (2004)
[2] http://cordis.europa.eu/fp7/ict/e-infrastructure/home_en.html
[3] Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007)
[4] PARSE.Insight, *Digital preservation and scientific data infrastructure: a preliminary summary of evidence on scientific community needs* (November 2008), p. 3

Researchers would like to (re-)use data from both their own and other disciplines because this is likely to produce more and better science. The 80% of researchers who have wished to access digital research data gathered by other researchers which turned out to be unavailable, as reported in the PARSE.Insight survey, is a sign that the practice of data sharing has a lot of room for improvement.

Researchers are also finding it difficult to discern the finesses of some of the new data related terminology. The concept of 'publishing datasets' means different things to different researchers. Some see a natural analogue with the traditional means by which scholarly papers are published; that is, the information is 'fixed' in a particular form at a particular point in time. There is, for some, also an implication that the information has been through a quality control process. The point of publication may be perceived, therefore, as a line in the sand.[5] The meaning of 'data sharing' also has several interpretations – does it mean depositing data to a public data repository, or giving access to a subset of data to, for example, colleagues from other research institutions? Researchers are making their preliminary or 'work in progress' data available before the project is finished. When can data be said to have been shared? From the researcher's point of view, not only when data are published in the traditional sense.

The scope of the 'data sharing' concept differs also in the rhetoric of policy makers. The concept of data sharing that has its roots in data archives, as most of the discussion in UK does,[6] always includes the aspect of curation of data over time and taking the long-term view when data sharing infrastructure is argued for. Some of the open access discussion to research outputs takes a more narrow view of data sharing, calling it often 'data publishing', most likely inspired by the contrast of access to free e-prints and fee-based journals, and is concerned primarily with removing barriers to making data freely accessible.

For research data to be and remain shareable over time, both digital curation and access services will be necessary. This presumes that good data management practices are employed from the start of every research project to ensure adequate curation of data throughout their lifecycle. The current state of data sharing discussion does not, however, always convey this message to the researchers.

This study compares the different national agendas in the OECD countries that are addressing variant data sharing infrastructure models, in order to inform developments within the UK. The study report was prepared by Raivo Ruusalepp (EBAC), with contributions from Graham Pryor (DCC), between May and November 2008.

1.1 RATIONALE

JISC has commissioned this study as part of a suite of work that together takes forward key recommendations from the *Dealing with Data* report, and which together form an integrated whole. This suite of work includes:

- Data Audit Framework,[7] which will specify good practice in undertaking an audit of institutional and departmental data collections, awareness, policies and practice and infrastructure for data curation and preservation, and will enable institutions to undertake such audits.
- Pilot implementations of the Data Audit Framework in a range of institutions,[8] resulting in descriptions of the data collections, levels of awareness, policies, practices and infrastructure in or on behalf of parts of those institutions.
- A project to examine and make recommendations on the role and career development of data scientists and curators.[9]
- Summer Schools run by the Digital Curation Centre in 2008 to help curators and data scientists share good practice and learn new skills.

1.2 APPROACH

The purpose of this study is to compare international approaches to conceptualising, implementing, maintaining and developing infrastructure arrangements relevant to the sharing of research data. Through an

[5] RIN, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* (2008), p. 24
[6] For an early example see: Royal Statistical Society, UK Data Archive, *Preserving and Sharing Statistical Material* (2002)
[7] http://www.data-audit.eu/
[8] http://www.data-audit.eu/users.html
[9] Alma Swan, Sheridan Brown, T*he Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs* (2008)

investigative analysis of a cohort of OECD countries, this study aims to identify the prevailing and predicted landscape for data sharing infrastructures. The concrete objectives for this study were set by JISC as:

- Examine the data infrastructure strategies (as they pertain to staff working in HEI's or equivalent) of a number of OECD countries and establish the variance (or otherwise) in their approaches.
- Establish the rationale for the existing infrastructure arrangements as articulated by the bodies responsible for funding them.
- Make some assessment of the effectiveness of infrastructure provision.

The scope of the geographical coverage of the report was set by JISC as 'Organisation for Economic Co-operation and Development (OECD) member countries'. The thirty members of the OECD are:

Australia	Austria	Belgium
Canada	Czech Republic	Denmark
Finland	France	Germany
Greece	Hungary	Iceland
Ireland	Italy	Japan
Korea	Luxembourg	Mexico
Netherlands	New Zealand	Norway
Poland	Portugal	Slovak Republic
Spain	Sweden	Switzerland
Turkey	United Kingdom	United States

Table 2. The OECD countries.

The main focus of the report is on Australia, the UK and the US. The latter two are from amongst the top five largest world economies (with Japan, China and Germany), whilst Australia ranks fourteenth.[10] Australia, the UK and the US are known to be key proponents of research data strategies. This focus was taken to introduce a geographically global coverage and to establish a significant economic/political presence.

The information collection was undertaken through desk research, targeted interviews and e-mail communication. The analysis of collected information looked at the top-down drivers for establishing data sharing infrastructures – policies, strategies and development plans; and compared examples of typical data sharing infrastructure provision in OECD countries. The project timeline did not allow the creation of a full inventory of all data sharing initiatives across all the OECD countries. Therefore, the focus was set on identifying the best practice examples that cover the range of approaches taken across most research disciplines.

For analysis purposes the various data sharing initiatives were structured into a five level taxonomy:

- International and trans-national initiatives.
- National and government-initiated initiatives.
- Research domain and funding agency initiatives.
- Institutional initiatives.
- Individual project initiatives.

This structure is used across all the data sharing topics covered and represented in the presentation of the analysis results in this report.

[10] According to the International Monetary Fund list of countries by GDP.

1.3 LIMITATIONS OF THE STUDY

The discussion of data sharing issues is happening on all levels and in many different contexts – from data curation and access to publicly funded information to legal and ethical constraints. This analysis focused primarily on analysing the enablers of research data sharing, and less on the various barriers. Technical and auxiliary topics that are indirectly linked with data sharing infrastructure provision, like digital preservation, IPR, data quality, data formats, metadata and data documentation, etc. have not been considered in this analysis. However, the instances of best practice referenced in this report also demonstrate the approaches used to overcome also these barriers to effective data sharing. The scope of the project and limited ease of access to adequate information did not permit including commercial data in the analysis.

With an explicit scope of OECD countries, this analysis is representative of only a sample of all the projects ongoing around the world. Given that over 300 data sharing initiatives were identified within this cohort of countries, further sampling was applied to identify initiatives that are representative of the specific needs in different research domains, and demonstrate new and innovative approaches. The full list of significantly overlapping approaches taken and similar solutions developed and applied in many countries and disciplines would defeat the purpose of this study.

Given the prolific reporting on various data sharing issues within the UK – eight publications from the last three years alone are listed among the referenced materials of this report (see Appendix. References) – the examples included in the study report have been chosen from outside the UK, where a choice was available.

Where possible, preference was given to existing services, rather than projects that are still only developing new tools or plans for new services.

1.4 TARGET AUDIENCE

The intended audience of the study report includes:

- Funding bodies (including the JISC and its strategic partners).
- Research Councils.
- Data services and data centres.

1.5 STRUCTURE OF THE REPORT

The study begins with analysing and comparing the policy drivers for data sharing on international, national, research domain and institutional level (Chapter 2). Summaries of main findings in each category of policy documents are presented at the end of respective sub-chapters. The analysis of current data sharing infrastructure provision (Chapter 3) begins with a brief overview of funding and resources made available for research infrastructure. Summaries of various national, domain and institutional research data infrastructure arrangements (Chapter 3.2) are followed by an analysis of data access services under development in most countries (Chapter 3.3). Examples of emerging virtual research environment services in connection with data sharing (Chapter 3.4) and a brief discussion of researchers' skill development needs (Chapter 3.5) follow. The main findings on all these topics are spelled out and discussed in the final analysis (Chapter 4), followed by a set of recommendations for further work that have emerged from this study (Chapter 5). A full list of references and materials used in the study is presented in the report appendix.

1.6 ACKNOWLEDGMENTS

Authors of the study would like acknowledge the contribution made to this report by all respondents and interviewees internationally.

2. POLICY AND STRATEGY FOR RESEARCH DATA SHARING

The rapidly mounting body of research data represents both a massive investment of public funds and a potential source of knowledge that can be used not only in further research, but also to produce innovative value-added services and goods. Today, a wider population is seeking better access to educational and cultural knowledge; commercial opportunities and the wider spread of information can have positive economic and social benefits. Data access provision will not happen without policy support and funding of appropriate infrastructure.

2.1 TRANS-NATIONAL INITIATIVES AND THEIR DATA POLICIES

Co-ordinated efforts at national and international levels are needed to broaden access to data from publicly funded research and contribute to the advancement of scientific research and innovation. To promote improved scientific and social return on the public investments in research data, countries have established a variety of laws, policies and practices concerning access to research data at the national level. In this context, the **Organisation for Economic Co-operation and Development (OECD)** recognised that international guidelines would be an important contribution to fostering the global exchange and use of research data.

The OECD first discussed data sharing issues in its workshops in 2000,[11] and then set up a working group to draw up commonly agreed principles to guide access to publicly financed research. Collaboration with a similar working group at CODATA was established. The set of principles was discussed in 2004 when the OECD Committee for Scientific and Technological Policy (CSTP) met at a Ministerial level to consider *Science, Technology and Innovation for the 21st Century*. The Ministers recognised that fostering open access to and wide use of research data will enhance the quality and productivity of science systems worldwide, and approved the *Declaration on Access to Research Data from Public Funding*.[12] The communiqué also invited the OECD 'to develop a set of OECD guidelines based on commonly agreed principles to facilitate optimal cost effective access to digital research data from public funding to be endorsed by the OECD Council at a later stage'. This request was taken up by OECD's CSTP, which launched a project by asking a group of experts to develop a set of principles and guidelines. The experts drafted a first set of principles and guidelines and engaged in several rounds of consultation with research institutions and policy making bodies in the OECD member countries to achieve a consensus. A workshop involving key stakeholders was held in Paris in 2006, which also contributed to this process. The work leading up to the final draft revealed that international frameworks to facilitate access were still lacking in the member countries, but also that improved access was generally seen as benefiting the advancement of research, boosting its quality and facilitating cross-disciplinary research co-operation. The principles and guidelines that resulted from this extensive consultation process were approved by the OECD's Committee for Scientific and Technological Policy in October 2006. The Principles and Guidelines were attached to an OECD Recommendation and endorsed by the OECD Council on 14 December 2006.[13] The OECD expects that member countries take the Principles into consideration when developing policies and good practices related to the accessibility, use and management of research data.

The OECD Principles recognise that:[14]

> Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.

The following principles are considered important in developing access arrangements for research data from public funding:

- Openness: access to research data for the international research community at the lowest possible cost;
- Flexibility: take into account characteristics of different research fields, legal systems, cultures and regulatory regimes;

[11] See: OECD, *The Global Research Village Conference* (2000)
[12] *Science, Technology and Innovation for the 21st Century*. Final Communiqué of the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level (2004)
[13] Cf. Dirk Pilat, Yukiko Fukasaku, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007)
[14] OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007)

- Transparency: information on data be made available through the Internet;
- Legal conformity: conform to the national legal requirements on national security, privacy, intellectual property rights;
- Formal responsibility: promoting formal institutional practices pertaining to authorship, usage restrictions, financial arrangements, ethical rules, licensing terms, liability and sustainable archiving;
- Professionalism: observe relevant professional standards embodied in the codes of conduct of the scientific communities involved;
- Interoperability: pay due attention to relevant international data documentation standards;
- Quality: adopt good practices for methods, techniques and instruments employed in the collection and archiving of data;
- Security: pay attention to the use of techniques and instruments to guarantee the integrity and security of research data;
- Efficiency: improving overall efficiency of scientific research by avoiding duplication of data collection efforts;
- Accountability: evaluation of access arrangements by user groups, responsible institutions and funding agencies;
- Sustainability: taking measures to guarantee long term access to data.

In 2008, at the Ministerial Meeting on the Future of the Internet Economy, the OECD released the report, *Shaping Policies of the Future of the Internet Economy,* which includes the *OECD Principles and Guidelines for Access to Research Data from Public Funding* as an annex and stresses the importance of open access to research data:[15]

> The Internet and ICTs are profoundly changing how research and creative activity are undertaken (e.g. distributed research, grid and cloud computing, virtual simulation, virtual worlds), with potentially major impacts on innovation and growth. They are fostering new types of market-based entrepreneurship and encouraging people outside traditional institutions and hierarchies to collaborate to produce content, services and goods. The Internet enables the rapid diffusion of codified knowledge and ideas, thereby linking science more closely to business, and facilitates the development of informal creative networks. Central to this is open access to the vast amounts of information and data available over the Internet.

OECD Recommendations are considered to be a 'legal instrument', but not legally binding (a 'soft law'),[16] that set out collective standards or objectives to be implemented by OECD member governments. Although the Recommendations do not give precise guidance on implementation of the principles, they do represent an expression of strong political commitment on the part of member governments and are expected to be carried through. The OECD Principles have, nevertheless, set a high-level example and are frequently referenced in national and research funding organisations' data sharing policies (see below).

The **United Nations Educational, Scientific and Cultural Organisation (UNESCO)** General Conference invited its Director-General in 1997 to undertake action 'to facilitate access to information in the public domain with relevance to UNESCO's fields of competence'. Draft guidelines to advise member states on policies for the development and promotion of public domain information, taking account of both national needs and international practices were developed in 2003.[17] These underwent an international review process and were discussed at an International Symposium on Open Access and the Public Domain in Digital Data and Information for Science. The final guidelines were published in 2004 as the *Policy Guidelines for the Development and Promotion of Governmental Public Domain Information*.[18] The Guidelines support open dissemination of research data, among other public information:

> The open availability of publicly funded scientific data and the public domain status of unprotected factual information are one of the cornerstones of basic research.
> Open and efficient access to public scientific and technical information funded by the public sector, subject to applicable national security controls and the rights of others deriving from obligations of confidentiality, intellectual property and privacy protection, fosters excellence in research and effective use of public research and development funds.

The UNESCO Guidelines are expected be broadly useful to decision- and policy-makers at the national and international levels. However, they are meant to be strictly advisory and are not intended as a prescriptive or normative instrument.

---

[15] OECD, *Shaping Policies for the Future of the Internet Economy* (2008), p. 18
[16] Cf. Yukiko Fukasaku, *International Initiatives in Data Sharing: OECD, CODATA and GICSI* (2007)
[17] Paul Uhlir, *Draft Policy Guidelines for the Development and Promotion of Public Domain Information* (2003)
[18] Paul Uhlir, *Policy Guidelines for the Development and Promotion of Governmental Public Domain Information* (2004)

In 2003 UNESCO published its *Draft Charter on the Preservation of the Digital Heritage* that also makes a provision for research data:[19]

> Measures should be taken to [...] encourage universities and other research organizations, both public and private, to ensure preservation of research data.

The measures in the UNESCO Charter are expected to be implemented by the member states.

The **European Union** has declared in its *Lisbon Agenda* that the EU aims 'to become the most competitive and dynamic knowledge based economy in the world' by 2010. To facilitate this, a European 'internal market' for research should be created, where researchers, technology and knowledge freely circulate.[20] The European Commission has positioned itself as both a policy-making body that launches policy debates at the European level, and a research funding body (see Chapter 3.2.1 below).

The European Commission opened up a significant debate over open access to research results through its *Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe*.[21] The report focussed on research publications models, but also recommended that member states should 'guarantee public access to publicly-funded research results shortly after publication'. The Study was opened for public consultation and received numerous reactions that were synthesised into the EC *Communication on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (COM (2007) 56).[22] This document discusses the reasons for dissemination of scientific information, pros and cons of current practices and outlines an action plan for the European Commission. The Commission will also 'encourage universities, research organisations, research funding bodies and scientific publishers to exchange information on good practices in relation to new access and dissemination models for scientific information'. The Communication met a lively response by various stakeholders and was the main topic of discussions at the stakeholder conference in Brussels.[23]

In a parallel process, the European Research Area programme published its Green Paper,[24] which included a recommendation to study the 'need for EU-level policies and practices to improve and ensure open access to and dissemination of raw data and peer-reviewed publications from publicly funded research results'. The European Commission carried out an independent analysis of the responses received from the public consultation of the Green Paper. A large proportion of the respondents were of the view that data or publication repositories should be available at both national and EU levels. At the same time, generally free availability of research data was supported by just over two thirds of the respondents, and issues of ownership, privacy and quality of data were highlighted, scepticism about the usefulness of raw data to anybody other than experts was expressed, and warnings of dangers of misinterpretations by laymen were given.[25]

The Council of the European Union then summarised the above documents and discussions of their principles in its *Conclusions on Scientific Information in the Digital Age: Access, Dissemination and Preservation*.[26] This document outlines specific tasks for both EU member states and for the European Commission:

> The Council invites the member states to:
> - reinforce national strategies and structures for access to and preservation and dissemination of scientific information, tackling organisational, legal, technical and financial issues;
> - enhance the co-ordination between member states, large research institutions and funding bodies on access, preservation and dissemination policies and practices;

> The Council invites the European Commission to:
> - support and contribute to improving policy co-ordination and to fostering a constructive debate and exchange of information between stakeholders.

[19] UNESCO, *Draft Charter on the Preservation of the Digital Heritage* (2003), Annex I, p. 3
[20] European Commission, *The European Research Area: New Perspectives. Green Paper* (2007)
[21] European Commission, *Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe* (2006)
[22] European Commission, *Communication on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (2007)
[23] European Commission, Scientific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age. Stakeholder conference 15-16 February 2007
[24] European Commission, *The European Research Area: New Perspectives. Green Paper* (2007)
[25] Johannes Velterop, *Analysis of the Responses to the Knowledge Sharing Questions in the Online Public Consultation on the Future of the European Research Area* (2007)
[26] Council of the European Union, *Council Conclusions on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (2007)

The action plan added in an appendix to the Conclusions provides concrete tasks and deadlines for realising the recommendations.

Another European Commissions initiative – the i2010 programme for the development of the information society – has a Digital Libraries Initiative that has set up a High Level Expert Group on Digital Libraries. The Expert Group is tasked with discussing and finding ways forward on potentially difficult issues such as copyright, public-private partnerships for digitisation and scientific information. The Expert Group has recently published its *Position Paper on Digital Research Data Access and Preservation*,[27] that recommends:

> A general (policy) framework, including sustainable custody and funding/business models, needs to be established by the key stakeholders in science and science information and national and EU policymakers to establish the roles and responsibilities of these in building a European Digital Information Infrastructure that allows the access and re-use of research data and ensures their long term preservation.

The European Commission's approach to open data sharing is step-by-step, sector based, and inclusive of all stakeholder views.

ESFRI, the European Strategy Forum on Research Infrastructures[28] has published its position on digital repositories.[29] The position paper makes recommendations for research data infrastructure availability, permanency, quality, right of use and interoperability.

The interdisciplinary Scientific Committee of the International Council for Science established a **Committee on Data for Science and Technology (CODATA)**[30] in 1966 to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use. CODATA has 24 national members and 16 scientific union members. Its objectives include:

- Improve quality and accessibility of data, especially for developing countries.
- Facilitate international co-operation of data experts and researchers.
- Promote increased awareness of the importance of data sharing.
- Consider data access and intellectual property issues.

CODATA has taken a special focus on preservation of and access to scientific and technical data in developing countries.[31]

> The key players on the international level (organisations like OECD, UNESCO, EU and interest groups like CODATA, ESFRI) have concentrated their policy statements around the principle of open access to publicly funded research outputs. While OECD, UNESCO and CODATA have policies explicitly for data sharing, the European Commission is looking at data sharing issues in the broader context of open access to public domain information.

2.2 GOVERNMENT GENERATED POLICIES AND STRATEGIES

In the **United Kingdom**, the Treasury, Department for Trade and Industry, and Department for Education and Skills published the *Science and Innovation Investment Framework 2004-2014*[32] in 2004, that sets out the Government's goals for UK science and innovation development for this period. The framework document recognised that a significant proportion of the information resources needed for research is now, and increasingly, in digital form. This is helping the sharing and rapid access to research output, but also presents a number of potential risks and challenges. The Office of Science and Innovation (OSI) was tasked to detail a delivery system for a national e-infrastructure for research under the Government's Framework. The OSI formed an e-Infrastructure Working Group to explore the existing provision of the UK's e-infrastructure and to help define its future development. The working group report *Developing the UK's e-infrastructure for Science*

[27] High Level Group on Digital Libraries, *Position Paper on Digital Research Data Access and Preservation* (2008)
[28] http://cordis.europa.eu/esfri/
[29] ESFRI Working Group About Digital Repositories, *ESFRI Position Paper* (2007)
[30] http://www.codata.org/
[31] http://www.codata.org/taskgroups/TGpreservation/index.html
[32] Treasury, Department of Trade and Industry, Department for Education and Skills, *Science and Innovation Investment Framework 2004-2014* (2004)

and Innovation[33] sets out a vision for a national e-infrastructure for research with more concrete requirements presented in six sub-group reports that were established to define roadmaps for specific areas. According to the presented vision, the UK e-infrastructure should allow 'researchers to share their research outputs with others and re-use them in the future' and provide 'researchers with assurance that their outputs will be accessible now and in the future'. The report produced made by the Preservation and Curation Working Group includes recommendation to 'reinforce the government legislation, regulations, codes of practice and policies to require, or emphasise existing requirements for, adequate long term protection and appropriate accessibility of valuable information of all kinds from the science record to the public and private record'.[34] The report from the Data and Information Creation Working Group working group noted that 'the e-infrastructure will need to support a variety of business models if the content is to be made as accessible as possible with limits to access only for good reasons'.[35]

The Joint Information Systems Committee (JISC)[36] is a national agency with a significant role in designing the UK higher education and research data environment. JISC is providing a network service, funds development and services that support learning, teaching and research and works in partnership with the research councils on areas of mutual interest. The aim of JISC's development work in the area of digital repositories and preservation is to bring together people and practices from across various domains (research, learning, information services, institutional policy, management and administration, records management, and so on) to ensure the maximum degree of coordination is achieved. The work funded by JISC relating to repositories and preservation aims to create a managed and standards-aware interoperable network of repositories for the UK higher education community. Ensuring that users have effective ways of discovering and accessing digital materials, for as long as it is useful to do so, is an integral part of this mission.

The first aim in the JISC Strategy 2007-2009 is to 'deliver innovative and sustainable ICT infrastructure, services and practice that support institutions in meeting their missions'.[37] As part of this, JISC considers 'online content an important resource where significant economies of scale can be found by national procurement and delivery. The key ambitions underlying this major development include establishing longer term sustainability, improving accessibility, building a critical mass of electronic content, greater take-up and usage leading to further efficiencies, saving time and improving management.' JISC also 'envisages continuing to develop the two JISC National Data Centres (Edina and MIMAS) as primary service-partners for a range of content and production services; and will seek to improve access to content and resources through a range of resource discovery services and through support for repositories and preservation.'

JISC's vision is of a content layer of scholarly and academic resources that are freely available, well structured and searchable. JISC is working to create an open, seamless environment that will provide appropriate content to the end user at the right time and in the right place. The commitment to openness also extends to JISC's interest in supporting the open access agenda and the exploration of other new business models. Various sectors within UK education and research need to work together in this key area.

To complement the top-down approach, the UK Government initiated the development of nation-wide data management policies from the research communities themselves. The Research Information Network (RIN)[38] focuses on understanding and promoting the information needs of researchers. The work of RIN focuses on five key themes: search and discovery, access and use of information services, scholarly communications, digital content and e-research, collaborative collection management and storage. RIN's *Strategic Plan*, among others, aims to co-ordinate action to ensure that the outputs that researchers produce and need are retained and made available for use in the most effective way. The *Strategic Plan* highlights 'a lack of co-ordination in managing and handling research data and published material, and in making them available for use by researchers. Resources are scattered over innumerable sites with different platforms, and interoperability in using more than a limited range of them is difficult if not impossible to achieve'.[39] RIN is undertaking

[33] Office of Science and Innovation e-Infrastructure Working Group, *Developing the UK's e-infrastructure for Science and Innovation* (2007)
[34] N. Beagrie, *E-infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group* (2007)
[35] OSI Data and Information Creation Working Group, *20/20 Vision: An e-Infrastructure for the Next Decade. Report of the Data and Information Creation Working Group to the e-Infrastructure Steering Group* (2006)
[36] http://www.jisc.ac.uk/
[37] http://www.jisc.ac.uk/aboutus/strategy/strategy0709/strategy_aim_one.aspx
[38] http://www.rin.ac.uk/
[39] http://www.rin.ac.uk/business-plan

evidence-based research and develops policy, guidance and advocacy on that basis. Its series of reports are tackling the data sharing issues from the data creators', data users' and research funders' perspectives which are aptly summarised in the framework document of principles and guidelines *Stewardship of Digital Research Data: A Framework of Principles and Guidelines: Responsibilities of Research Institutions and Funders, Data Managers, Learned Societies and Publishers*.[40] One of the policy objectives of this document states that 'Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used' (p. 3). There is also a call for establishing clear 'roles and responsibilities of different agents, there is the risk of misunderstandings, of wasted efforts, and growing difficulties in managing and providing access to digital research data. Increases in the volume and complexity of data mean that traditional informal arrangements alone are no longer adequate to ensure the effective stewardship of data, and may need to be complemented by more formal codes' (p. 8). The report then specifies the roles of research funders, researchers, and their support services in libraries, archives and data centres. RIN has also analysed research councils' policies for the management of research outputs,[41] whether or not researchers do make their research data available to others, and the issues they encounter when doing so,[42] and training provision for researchers on information seeking and management.[43] Together the RIN studies and principles form a solid body of best practice guidelines that the involved stakeholders could and should follow, to improve service provision around management and sharing of research data. These principles should also be considered when action plans are formulated and funding decisions made by policy makers and research funders.

In the **United States**, the sharing of research data is indirectly supported by law – the US intellectual property law does not allow for intellectual property protection of 'raw facts'. A scientific article could be copyrighted, but the data on which it rested could not. The US law also mandates that even those federal government works that could be copyrighted, fall immediately into the public domain — a provision of great importance given the significant governmental involvement in scientific research.[44]

The National Science Board's (NSB) report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*[45] highlighted the fact that 'there is no single site at which a member of the community can readily locate all applicable or relevant policy statements; that many programs lack an explicit statement of data access and release policy; and that there is little coherence and consistency among the set of existing statements' (p. 37). The report recommended that 'development of a comprehensive set of policy statements for data access and release that provides for consistency and coherence across disciplines while meeting the distinct needs of individual disciplines and communities, that are transparent and readily accessible to the community, and that prevent unnecessary proliferation and duplication of standards could greatly facilitate progress in research, education, and collections management'.

The National Science Foundation (NSF) instituted an Office for Cyberinfrastructure (OCI) in 2005 and through its Cyberinfrastructure Council has initiated a comprehensive strategic planning process to guide the agency's investments in cyberinfrastructure. The OCI's *Cyberinfrastructure Vision for 21st Century Discovery*[46] sets out the vision the NSF will pursue in making research data accessible: 'science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialist and non-specialist alike, are openly accessible while suitably protected, and are reliably preserved'. To realise this vision, the NSF's goals for the period of 2006-2010 are twofold: to catalyse the development of a system of science and engineering data collections that is open, extensible, and evolvable; and to support development of a new generation of tools and services for data discovery, integration, visualization, analysis and preservation. The resulting national digital data framework will be an integral component in the national cyberinfrastructure framework. In pursuing its vision, NSF will adhere to the following principles in the area of data sharing:

[40] RIN, *Stewardship of Digital Research Data: A Framework of Principles and Guidelines: Responsibilities of Research Institutions and Funders, Data Managers, Learned Societies and Publishers* (2008)
[41] RIN, *Research Funders' Policies for the Management of Information Outputs* (2007)
[42] RIN, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* (2008)
[43] RIN, *Mind the Skills Gap: Information-handling Training for Researchers* (2008)
[44] Science Commons, *Towards a Science Commons* (2007)
[45] NSB, *Long-lived Digital Data Collections: Enabling Research and Education in the 21st century* (2005)
[46] NSF, *Cyberinfrastructure Vision for 21st Century Discovery* (2007)

> Strategic investments in cyberinfrastructure resources and services coupled with enabling policy and organizational framework are essential to continued U.S. leadership in science and engineering.
> The integration and sharing of cyberinfrastructure assets deployed and supported at national, regional, local, community and campus levels represent the most effective way of constructing a comprehensive cyberinfrastructure ecosystem suited to meeting future needs.
> A collaborative cyberinfrastructure governance and coordination structure that includes representatives who contribute to basic cyberinfrastructure research, development and deployment, as well as those who use cyberinfrastructure, is essential to ensure that cyberinfrastructure is responsive to community needs and empowers research at the frontier.

The goals the NSF has set itself for the current budget period (2006-2010) include support for state-of-the-art innovation in data management and distribution systems, including digital libraries and educational environments that are expected to contribute to many of the scientific breakthroughs of the 21st century. The NSF wants to take advantage of innovation in large data management and distribution activities sponsored by other agencies and through international efforts.

NSF is recognising that policy and management issues in data handling occur at every level, and there is an urgent need for rational agency, national and international strategies for sustainable access, organisation and use. Formal policies must be developed to address data quality and security, ethical and legal requirements, and technical and semantic interoperability issues that will arise throughout the complete process from collection and generation to analysis and dissemination. To govern the development of its services, the NSF is updating its policies on data management and sharing. The NSF's position on data is straightforward: 'all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved'.[47] The NSF will continue to promote open access to well-managed data and will further redesign its policies to 'mitigate existing sociological and cultural barriers to data sharing and access, and to bring them into accord across programs and ensure coherence'. This will lead to the development of a suite of harmonised policy statements supporting data open access and usability (cf. also Chapter 2.3 below).

Several inter-agency groups exist in the U.S. that are in the process of developing policy frameworks for the management and sharing of data. The high-level National Science and Technology Council's Committee on Science set up an Interagency Working Group on Digital Data (IWGdd) in 2006 that has members from most government departments, memory institutions, NASA, the NSF and several councils.[48] The purpose of the IWGdd is to develop and promote the implementation of a strategic plan for the Federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering.[49] The strategy of the working group is to 'create a comprehensive, transparent, evolvable, and extensible policy, management, and organisational framework that makes reliable, effective access to the full spectrum of public digital scientific data a driving force for science and American leadership in science and in a competitive, global information society'. The IWGdd comprises four framework groups: policy, sectors, technology and infrastructure. The final report of the working group should be published in Spring 2009, but the preliminary recommendations have already been presented. The key recommendation is that a national coordinating body should be created for digital scientific data preservation and access that takes a lead in efforts to maximize the accessibility and utility of digital scientific data. Other recommendations include:[50]

- All appropriate departments and agencies have a comprehensive, publicly-available, digital scientific data policy.
- The Agency Digital Data Policy be administered by a designated, cognizant senior science official.
- Each project that will generate preservation data have a data management plan.
- Departments and agencies gather and share information related to costs and best practices for preservation, protection, dissemination, curation, and migration.
- Education and training activities be integral to all of the federal science data investments.
- The national coordinating body promote coordination of education and training among Federal departments and agencies and in partnerships with the education, research, and technology sectors.
- The national coordinating body promote data science and management as a career path with appropriate recognition and rewards structures.

[47] NSF, *Cyberinfrastructure Vision for 21st Century Discovery* (2007), p. 29
[48] NSTC COS, *Terms of Reference of the Interagency Working Group on Digital Data* (2006)
[49] Cita Furlani, Chuck Romine, Chris Greer, *Briefing: Interagency Working Group on Digital Data* (2008)
[50] Cita Furlani, Chuck Romine, Chris Greer, *Interagency Working Group on Digital Data* (2008)

To guide the implementation of its recommendations, the IWGdd have developed a scientific data lifecycle model:

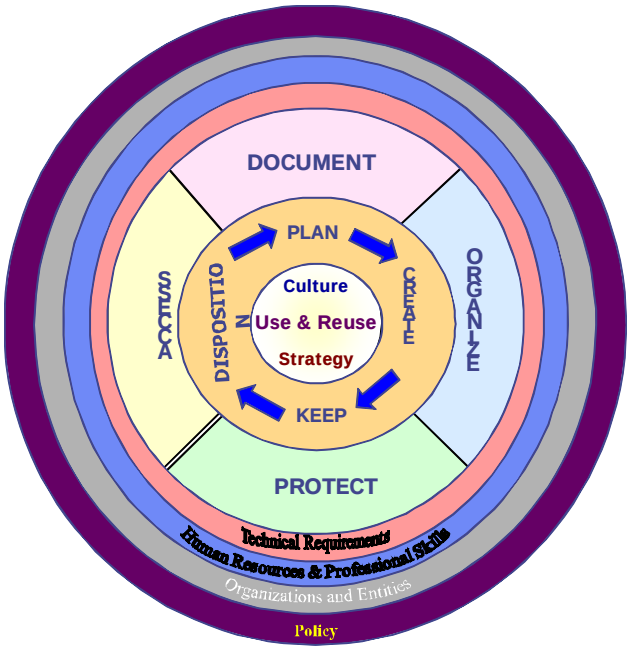

Figure 1. Scientific data lifecycle model (Interagency Working group on Digital Data, 2007).

The U.S. National Academies is in the process of setting up a Board on Research Data and Information under its Policy and Global Affairs Division to address emerging issues in the management, policy, and use of research data and information at the national and international levels.[51] The Board will consist of academic and industry participants and a growing core of data managers will be expected to work with research scientists in development of protocols and practices for data management. The Board's initial standing committee would include the U.S. National Committee for CODATA, a committee already constituted under the National Academy of Sciences.[52] The Board is intending to provide independent and objective advice through studies and reports, reviews of programs, and assessment of priorities concerning research data and information activities and interests of its sponsors.

In **Australia** the government issued a national strategy *Backing Australia's Ability – Building our Future through Science and Innovation* in 2004, and has subsequently published two versions of the *Strategic Roadmap for Australian Research Infrastructure* (2006 and 2008).[53] The latest of these lists components of effective 'transition from research to eResearch' with a focus on data federation, seamless collaboration and effective sharing of resources. Principles it espouses include:

- To ensure increase in shared data and its re-use, the concept of an Australian Research Data Commons will be used and additional resources directed towards: identification, registration and searching services; and policy development to agree and, where possible, simplify the arrangements around data, so that reuse and data integration are socially, legally and technically feasible.
- Creation of Shared Spaces as collaboration spaces that are managed in their own right and span the enterprise spaces provided by individual research organisations.

The Australian Research Council (ARC) updated its Funding Rules for funding commencing in 2008 to include the following condition:[54]

> The ARC [...] encourages researchers to consider the benefits of depositing their data and any publications arising from a research project in an appropriate subject and/or institutional repository wherever such a repository available to the researcher(s). If a researcher is not intending to deposit the data from a project in a repository within a six-month period, he/she should include the reasons in the project's Final Report. Any research outputs that have been or will be deposited in appropriate repositories should be identified in the Final Report.

[51] Paul F. Uhlir, *Board on Research Data and Information. Draft Prospectus* (2007)
[52] http://www7.nationalacademies.org/usnc-codata/
[53] http://www.innovation.gov.au/ScienceAndResearch/Documents/Strategic%20Roadmap%20Aug%202008.pdf
[54] ARC, *Discovery Projects: Funding Rules for Funding Commencing in 2008* (2008), Chapter 1.4.5

This essentially voluntary deposit mechanism has already been commented on by the Australian Government Productivity Commission that released a *Research Report on Public Support for Science and Innovation*. In this report the Commission stated:[55]

> The Commission continues to hold the view that funding agencies should take an active role in promoting open access to the results of the research they fund, including data and research papers. Although the ARC and NHMRC's recent announcement of promoting voluntary access is to be commended, the Commission considers that the progressive introduction of a mandatory requirement would better meet the aim of free and public access to publicly-funded research results.

Similar proposals have been made in **Canada**, where a *National Consultation on Access to Scientific Research Data*[56] report was published in 2005 that recommended:

> Research councils, and all other public-sector research funding agencies and departments require that project and grant applications include a data management plan, as well as specifically identified funding that will ensure quality, integrity, accessibility and accountability. A funding condition should be the inclusion of a well-constructed plan for data acquisition, management, access and preservation. Adherence to such plans should also become a non-competitive performance metric for the project and gateway for subsequent grant applications. Councils should recognize these as added costs to the main thrusts of research projects.

Recently, a Canadian Research Data Strategy Working Group was been set up.[57] This is a collaborative effort to address the challenges and issues surrounding the access and preservation of data arising from Canadian research. It is a multi-disciplinary group of universities, institutes, libraries, granting agencies, and individual researchers. The working group started by conducting a gap analysis between the ideal state of research data stewardship in Canada and a description of the current state and now develops strategies to begin filling in the gaps. Further work of the Group will address challenges relating to infrastructure, research culture, legal/policy frameworks and training. In developing standards and policies suitable for cross-disciplinary and multi-sectoral collection, preservation and access, the working group is recognising the complications arising from different legal jurisdictions (federal/provincial/international), as well as ethical, privacy, and intellectual property rights issues. In developing policies for access to data, the working group is drawing on the objectives stated in the *Canadian Digital Information Strategy* where maximizing access and use is discussed as one of the challenges.[58]

In **Germany**, the Ministry for Education and Research established a commission in 1999 to assess the state of research infrastructure and to come up with recommendations for development. The commission's report *Wege zu einer besseren informationellen Infrastruktur*[59] discusses both infrastructure provision and legal hurdles to sharing scientific microdata and aggregated statistical data at national and international levels, and makes a number of recommendations for improving archiving and access to scientific and statistical data. Recommendations for access provision are usefully presented in several categories: economic drivers, data protection, re-use for statistical purposes, and all argue for free or discounted access to data for academic research purposes.

German research councils, especially the Max-Planck Society, were initiators of the Open Access movement and were instrumental in producing the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*[60] that has subsequently been signed by over 250 research institutions worldwide. Although initially focussed primarily on research articles, the Open Access discussion now fully covers research data as well.[61] The German Research Foundation (Deutsche Forschungsgemeinschaft) has invested in an analysis of the results of applying Open Access principles[62] and recently supported the Digital Information Initiative[63] –

[55] Productivity Commission, *Public Support for Science and Innovation: Research Report* (2007), pp. 240-241
[56] David F. Strong, Peter B. Leach, *National Consultation on Access to Scientific Research Data* (2005)
[57] http://data-donnees.gc.ca/eng/about/backgrounder.html
[58] Library and Archives Canada, *Canadian Digital Information Strategy. Draft* (2007)
[59] BMBF, *Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik* (2001)
[60] *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (2003)
[61] See for example session *Shared Responsibilities in Sharing Research Data: Policies and Partnerships* at the 'Berlin 5 Open Access Conference "From Practice to Impact: Consequences of Knowledge Dissemination"' (2007)
[62] http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/index.html#4
[63] http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/projektfoerderung/initiative_digitale_information/index.html

an alliance of German research organisations – that has set goals for the development of the German research infrastructure for the next four years. Among these are 'to guarantee the broadest possible access to digital publications, digital data and other source materials; and to support collaborative research by means of innovative information technologies'.[64]

The roadmap document of the Digital Information Initiative makes special reference to open access provision to research data and talks about new business and funding models as well as new forms of cooperative financing that are necessary to support this. The roadmap also acknowledges that:

> All scientific institutions see an urgent need for action in order to ensure the systematic backup, archiving and provisioning of scientific data for subsequent (re-)use by third parties. The development of archiving and access strategies for primary research data is, admittedly, in varied stages and is of varying urgency in the different disciplines.

The activities of the Initiative alliance will be directed to three areas:

- Formulate a common data policy in order to promote both the need for action and to demonstrate the usefulness of primary data infrastructures for scientists and scholars.
- Foster cooperation between scientists and information specialists and to offer funding for pilot projects. Such projects should develop subject-specific standards and methods of data curation and archiving; they should also define the division of labour required in the process. These steps have the overall goal of establishing a reliable system of digital archives for primary research data, and to ensure that these remain accessible internationally and their data reusable in various interdisciplinary contexts.
- Establish a system of discipline-specific, internationally networked data repositories for primary research data.

The partners within the Alliance of German Research Organisations have agreed to coordinate their funding programmes in the area of primary research data and, when necessary, to merge or harmonise them. They have also agreed to examine the possibility of establishing common infrastructures for primary data.

The strategies for development of research infrastructure in **Scandinavian** countries make it explicit that archiving and providing access for re-use should be a mandatory condition when receiving funding from national funding organisations. Different levels of practice and available services are acknowledged in different research areas, and funding will be made available to bring all domains up to level in this area.[65]

This study did not identify any national level policy or strategic documents that would mandate sharing of research data. Nevertheless, access provision to research data is seen as a vital element of the general research infrastructure and all countries acknowledge the need to develop the means for accessing data alongside the other components of research infrastructure. The various models of funding research and academic institutions in different countries is the most likely explanation why defining models for making data accessible has been left to funding organisations in different research domains. There has been a significant rise in collaboration among research councils and similar organisations to agree on common principles and standards for access to research data. This has been supported by initiatives from international organisations (e.g., OECD, EC), international interest groups (e.g., CODATA, ESFRI) and the Open Access movement that are also increasingly embracing primary research data.
The examples of national level policy documents that support data sharing are either directed at securing additional funding from governments, or have been produced as a result of the funding becoming available. The collaborative effort of developing such policies has usually been led by an umbrella organisation (e.g. national research council, academy of sciences) or in some cases as a bottom-up process by initiatives form research communities (e.g. RIN in the UK).

2.3 RESEARCH FUNDING ORGANISATIONS' AND DOMAIN LEVEL POLICIES

While the international and national policies and strategies on data management and access to research data provide the general framework for sharing research data, the concrete incentives and guidance for how research data should be archived for re-use should be the responsibility of structures that fund research.

[64] Alliance of German Science Organisations, *Priority Initiative 'Digital Information'* (2008)
[65] For example Norges Forskningsråd, *Verktøy for forskning. Nasjonal strategi for forskningsinfrastruktur (2008 – 2017)* (Research Council of Norway, *Tools for Research. National Research Infrastructure Strategy 2008-2017*) (2008)

Research councils, science associations and foundations, research programmes and boards that fund research projects are expected to motivate researchers to share their data, either by setting conditions to concrete funding schemes or by having general policies and services to offer to recipients of funding. Policies on data sharing of research funding organisations and how they have been put into practice have varied significantly. Only very recently have there been initiatives to agree on common principles for data sharing policies.

In the **United Kingdom**, the partnership of seven research councils – Research Councils UK (RCUK) – issued a *Position Statement on Access to Research Outputs* in 2006, intended to provide a broad framework for detailed guidelines that research councils can develop. These principles state that:[66]

- 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.
- Published research outputs must be subject to rigorous quality assurance, through effective peer review mechanisms.
- The models and mechanisms for publication and access to research results must be both efficient and cost-effective in the use of public funds.
- The outputs from current and future research must be preserved and remain accessible for future generations.

The statement discusses the impact of self-archiving on research publishing, but recognises that 'data underpinning the published results of publicly-funded research should be made available as widely and rapidly as possible'.[67]

A recent RIN report on data sharing brought forward some unresolved issues that need to be tackled:[68]

- Funders, researchers and publishers should seek to clarify the current confusion with regard to publishers' policies with regard to allowing access for text-mining tools to their journal contents.
- Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum.

Not all research councils and charitable funds have not all yet succeeded in implementing all these principles in their policies and strategies. Indeed, a RIN report from 2007 stated that:[69]

> [T]he funders with the most detailed policies on data are those Research Councils which have the most developed infrastructure and resources for curating, preserving and making data accessible. Where the funding body has well-developed provision for curating data it feels better able to have formal policies relating to how data are, or should be, created or collected, managed, preserved, and made accessible.

The UK research councils' data management policies have in recent years been compared and thoroughly analysed in several reports.[70] An update to the comparative tables in these reports is provided below (see Table 3 below), with a focus on data sharing principles. The basic structure of the table has been borrowed from the DISC-UK report *DataShare: State-of-the-Art Review* and amended with additional fields. The table lists funding bodies and source documents of their policy statements; the policy statement is supplemented with the requirements for research projects; what leverage the funding agencies have to enforce their requirements; and the last three columns present information on whether funds are available for preparing and sharing data, deadlines when data is expected to be delivered to data centres or service providers, and whether access restrictions or delaying of data dissemination are permitted for exclusive use by the authors.

[66] Research Councils UK, *Updated Position Statement on Access to Research Outputs* (2006)
[67] Research Councils UK, *Draft Position Statement on Access to Research Outputs* (2005)
[68] RIN, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* (2008), p. 5
[69] RIN, *Research Funders' Policies for the Management of Information Outputs* (2007), pp.55-56
[70] See: Philip Lord, et al. *Large-scale Data Sharing in the Life Sciences: Data Standards, Incentives, Barriers and Funding Models* (2005); RIN, *Research Funders' Policies for the Management of Information Outputs* (2007); Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007); DISC-UK, *DataShare: State-of-the-Art Review* (2007); RIN, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* (2008); Graham Pryor, *Research Council Data Policy Statements*. DCC internal document (2007); UKRDS, *Interim Report*. Version v0.1a.030708 (July, 2008)

Table 3. Comparison of UK research funding organisations' data sharing policy statements.

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
Arts and Humanities Research Council	AHRC *Research Funding Guide*. Appendix 3: AHRC Annexes to the Terms and Conditions of Research Grants. GC 28: Deposit of resources or datasets (2008)[71]	Grant holders must make any significant electronic resources or datasets created as a result of research funded by the Council available in an accessible depository for at least three years after the end of their grant.	For projects in archaeology data must be offered within 3 months of a grant ending.	-	-	Data must remain available for at least 3 years after the grant ending. For projects in archaeology data must be offered within 3 months of a grant ending.	Data must remain available for at least 3 years after the grant ending.
Biotechnology and Biomedical Sciences Research Council	BBSRC *Data Sharing Policy* (2007)[72]	BBSRC expects research data generated as a result of BBSRC support to be made available with as few restrictions as possible in a timely and responsible manner to the scientific community for subsequent research. Applicants should make use of existing standards for data collection and management and make data available through existing community resources or databases where possible.	Since 2007 research proposals are required to include a statement on data sharing.	Conditional award of funding if data sharing statement is inappropriate Adherence to the strategies is monitored through the final report assessment and may be taken into account when assessing future proposals.	Funding to support the management and sharing of research data can be requested.	-	Data should be available with as few restrictions as possible. Data should be retained for a period of 10 years after completion of a project. Ownership of the data generated from the research that BBSRC funds resides with the investigators and institutions.
Economic and Social Research Council	ESRC Data Policy (2000); Datasets Policy. Annex C of ESRC *Research Funding Guide* (2008)[73]	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. This relates not only to datasets arising as a result of primary data collection, but also to derived datasets resulting from ESRC-funded work.	Data must be offered for deposit after grant ends.	Final grant payment can be withheld if requirements are not met.	Funding available for preparation of data for archiving.	Data must be offered within 3 months of a grant ending	Access to data from any ESRC facility will be governed where necessary by user licence agreement.

[71] AHRC, *Research Funding Guide* (2008)
[72] BBSRC, *Data Sharing Policy* (2007)
[73] ESRC, *Data Policy* (2000); ESRC, *Research Funding Guide* (2008)

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
Engineering and Physical Sciences Research Council	EPSRC *Funding Guide* (2008)[74]	EPSRC strongly encourages the exploitation of the results of research. Where results of industrial or commercial value emerge from projects, investigators are required to make suitable arrangements for exploitation and take up by industry.	-	If the final report is not received within the period allowed, the Research Council may recover 20% of expenditure incurred on the grant.	-	-	EPSRC makes no claim to the intellectual property rights arising from research that it supports.
Medical Research Council	MRC *Policy on Data Sharing and Preservation* (2007)[75]	The MRC expects valuable data arising from MRC-funded research to be made available to the scientific community with as few restrictions as possible. Such data must be shared in a timely and responsible manner.	Funding proposals are required to include a strategy for data preservation and sharing.	As part of the end of grant reporting process, MRC funded researchers will be expected to report on data management and sharing activities relating to these plans.	-	Data sharing must be timely.	MRC research data are publicly funded and must be made available for new research purposes in a timely, responsible manner.[76] A limited, defined period of exclusive use of data for primary research is reasonable according to the nature and value of the data and the way they are generated and used.
Natural Environment Research Council	NERC *Data Policy Handbook*, Version 2.2 (2002),[77] under review	Data must be offered to one of NERC's data centres, where it will be maintained and made available for future research: grant-holders are required to offer to lodge with NERC a copy of the data resulting from the supported research when it is completed, together with documentation / metadata describing these data. NERC will then be in a position to make the data available to others (under suitable constraints) for further bona fide research only.	Data must be offered after a 'reasonable period' of exclusive use by researcher/s.	Compliance with mandate a factor when considering further funding applications.	NERC policy is to charge for the provision of data at a rate that is dependent on the use to which the data will be put.	Data must be offered after a 'reasonable period' of exclusive use by researcher/s	NERC policy is to specify formally any consequent restrictions on the use of the data in formal licensing agreements. The Intellectual Property Rights to the data need not be transferred.

[74] EPSRC, *Funding Guide* (2008)
[75] MRC, *Policy on Data Sharing and Preservation* (2007)
[76] MRC, *Principles for Access to, and Use of, MRC Funded Research Data* (2007)
[77] NERC, *Data Policy Handbook* (2002)

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
Science and Technology Facilities Council	STFC *Research Grants Handbook*. Section 8: Exploitation of Results and Dissemination of Information (2008)[78]	It is the responsibility of the Research Organisation and all engaged in the research for which funds and resources have been provided by STFC to make every effort to ensure that any potentially valuable results obtained in the course of the research are exploited and that there is a suitable return to the Research Organisation and its researchers from any such exploitation.	-	-	'Follow on Fund' exists to support commercial-isation of ideas.	-	IPR remains with the research organisation.
Wellcome Trust	Wellcome Trust *Policy on Data Management and Sharing* (2007)[79]	Trust expects the researchers that it funds to maximise the availability of research data with as few restrictions as possible.	Trust requires that the applicants provide a data management and sharing plan as part of their application; and will review these plans, including any costs involved in delivering them, as an integral part of the funding decision.	Where a data management and sharing plan is felt to be inappropriate or insufficient, applicants may be asked to revise their plan before a funding decision is made.	Data management plan must include costs involved in delivering data.	Data to be made available on publication of research, where consistent with any ethics approvals and consents and any intellectual rights.	Data should be available with as few restrictions as possible. Data to be made available on publication of research, where consistent with any ethics approvals and consents and any intellectual rights.

[78] STFC, *Research Grants Handbook* (2008)
[79] Wellcome Trust, *Policy on Data Management and Sharing* (2007)

As the list in Table 3 demonstrates, not all research councils have created an explicit policy on data management and sharing, but most research councils have made it a condition for receiving funding that the data resulting from research are deposited in a designated data service or a recognised public data repository. Concrete deadlines for depositing data are set and recommendations for access restrictions when sharing data are made. Several research councils now also require a data management and sharing plan to be part of the funding application and this is assessed as part of the application. The Wellcome Trust has also developed an exemplary guide on what a data management and sharing plan should include.[80] Data management plans are generally expected to include an estimate of costs of data preparation and sharing.

The Natural Environment Research Council (NERC) has had a data policy since 1996, the year in which the first of its seven data centres was established. Current data policy is contained within a detailed *Data Policy Handbook*, which dates from 2002 and is currently under review. NERC places high importance on the stewardship of data for the benefit of scientific advancement and requires that grant holders offer data to one of its data centres, where it will be maintained in the long term and made available for future research. *The Data Policy Handbook* explains NERC's approach to managing the full data life-cycle (Section 5.1):

> It cannot be emphasised too strongly that any proposal to undertake science which will involve the acquisition of datasets should include at the outset consideration of what is to be done with them once acquired. Valuable scientific or commercial opportunities may be lost if this fundamental principle is neglected.

The Economic and Social Research Council (ESRC) has relied on the UK Data Archive since 1967 as a place of deposit for data from the research it has funded and has included the requirement to deposit data in its funding award terms and conditions since 1976. A clear policy statement to this effect was also included in its 2000 *Data Policy*:

> All award holders are required to offer their computer-readable data for deposit, prepared to a standard which may be used by a third party, within three months of the end of an award.

ESRC's current Datasets Policy is published as an annex to the ESRC *Research Funding Guide*. The original ESRC *Data Policy* (2000) is under review and will be widened in scope to include resources other than data under the new title 'Research Resources Policy'.[81]

Following the decision to cease funding the Arts and Humanities Data Service, Arts and Humanities Research Council (AHRC) grant holders were freed of obligation to deposit their data with the AHDS, except in archaeology where the Archaeology Data Service has been retained. The new AHRC statement, included in the appendices of its *Research Funding Guide* states that grant holders must deposit their data with an accessible depository for at least three years after the end of a grant. Grant holders in the area of archaeology must ensure that any significant electronic resources or datasets created as a result of research funded by the Council, together with documentation, are offered for deposit at the Archaeology Data Service (ADS) within three months of the end of the project.

Both the Medical Research Council (MRC) and Biotechnology and Biomedical Sciences Research Council (BBSRC) have published data policies which 'expect' data to be made available 'to the scientific community with as few restrictions as possible'. The Wellcome Trust has adopted a similar line for data generated under its funding. The BBSRC's policy framework is a substantial document that matches the detail of the ESRC and NERC documents, and recognises the needs of the individual scientific areas within its domain, whereas the MRC make public a brief (two page) statement on key principles, a policy statement and a description of the data sharing strategy statements required from grant applicants. The Wellcome Trust's policy defines its position on data management and sharing, which is closely allied to its published philosophy for open access and builds on its *Guidelines on Good Research Practice*. A single page policy statement is supplemented by a substantial question and answer document that explains the context, requirements and processes covered by the policy. It does not define categories of data but is designed as an instrument to promote the potential availability of all research data produced as a consequence of Trust funding. All three funding organisations require a data management and sharing strategy or plan to be part of the funding application, and offer guidance on data sharing principles, safeguarding the research participants and re-use of personal data.

[80] Wellcome Trust, *Q&A: Wellcome Trust Policy on Data Management and Sharing*
[81] Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007), p. 18

The Engineering and Physical Sciences Research Council (EPSRC) and Science and Technology Facilities Council (STFC) statements encourage data sharing but leave it up to the research organisation to ensure that their research outputs are exploited, where appropriate also commercially.

All research councils and the Wellcome Trust utilise a range of levers for ensuring compliance with their policies. The BBSRC's approach is perhaps the most supportive amongst the councils, including advice on standards, guidance and training, whilst the Wellcome Trust supports its broad demands for data sharing with a generous range of options for funding data management. The ESRC uses both a 'carrot and a stick', providing explicit funding to cover data preparation but withholding final payments of awards until data are satisfactorily deposited. All research councils allow a period of exclusive access to data for the data producers but underline their intention that access will be quickly and openly enabled, unless shown to be inappropriate. The BBSRC provides the most detailed explanation of timescales for different scientific areas. The Wellcome Trust expects, as an absolute minimum, that relevant data is made available upon the publication of research, and fully expect researchers to seek opportunities for sharing data prior to publication. Specific reasons for delaying the release of data are described but there is an expectation that these will be minimised.

All research councils have subscribed to the RCUK *Position Statement on Access to Research Outputs* and have published their own position statements on recommended open access to published research results. None of the research councils have, so far, extended this policy to research data – the concept of open data is not advocated in any of their policies. The research councils' designated data centres and services have an explicit remit to serve the research and teaching communities and data is offered under licence terms, where accessing data requires both registration and an institutional account (e.g., Athens System). Although non-commercial access to data incurs minimal or no charge, for several research areas data licensing for commercial users is a key income generator. The Wellcome Trust requires all its grant-holders to ensure that published original peer-reviewed research is available on an open access platform. The Trust's long-term aim is to support greater online integration between research literature and the data on which it is based.[82]

In 2004-2006 the ESRC led the development of a *National Strategy for Data Resources for the Social Sciences*.[83] The National Data Strategy is a plan to develop and maintain a robust data infrastructure, ensuring that relevant and timely data are available to inform and address future research priorities in the social sciences. Concrete tasks in the strategy include:

- Enhancing research access to sensitive data
- Developing access to business and management data
- Promoting common standards for data access, description and documentation

As part of the UK Data Forum,[84] the strategy is undergoing a review in 2008 and a forward plan of action has been proposed for 2008-2010.[85] Under the auspices of this strategy a high-level expert group between MRC, Wellcome Trust and the ESRC has been tasked to elaborate the principles for sharing of linked bio-medical/socio-economic data and to establish a code of practice espousing these principles. Discussions of data access in a safe setting have been held to analyse different approaches to data licensing and the use of a 'safe-setting' environment for data analysis.[86] A new Secure Data Service will be established, liaising among other things with the Office for National Statistics to interpret the confidentiality provisions of the Statistics and Registration Act 2007.

The on-going UK Research Data Feasibility Study (UKRDS) has conducted a survey of awareness of research councils' policies and grant requirements among researchers and reported in its interim report that:[87]

> Case Study Sites response to this study's Researchers Questionnaire showed that 36% did not know if there were any grant or legal requirements to retain their data. Responses relating to questions about sharing data indicated that most researchers share data (only c. 12% do not make their data available) but that informal

[82] Wellcome Trust, *Position Statement in Support of Open and Unrestricted Access to Published Research* (2008); also: Wellcome Trust, *Open and unrestricted access to the outputs of published research,* http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/index.htm
[83] *National Strategy for Data Resources for the Social Sciences* (2006)
[84] The UK Data Forum, http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/nds/ukdf/default.aspx
[85] The National Strategy for the Development of Data Resources for the Social Sciences. Forward Plan, 2008-2010 (2008)
[86] The UK Data Forum, *Progress on Plans to Implement the National Data Strategy* (2006)
[87] UKRDS, *Interim Report* (2008), p. 10

> peer exchange/networks within research teams and with collaborators are pre-dominant. Only c. 19% of data producers share data via a data centre and in contrast c. 43% of data users make use of data centres (e.g. the European Bioinformatics Institute or the NERC data centres) to access other researchers' data. Researchers noted the difficulty in retaining their data beyond the life of a project, largely because of lack of funding to do so.

The conclusion that can be drawn from these figures is that although the formal responsibility for data has been laid on the grant holder it is difficult to assess how well the research council's policy and guidance is being met. Even where funding councils stipulate potential sanctions if data is not offered for deposit there is still evidence that more could be done to improve researcher awareness and institutional support for researchers.

The benefits of sharing data were described by the **United States** National Academy of Sciences as early as 1985.[88] Subsequently, national scientific organisations and research funding organisations have made a commitment to share and archive research data through their ethical codes (e.g., the American Sociological Association,[89] the American Psychological Association[90]), data sharing policies (e.g., NIH) or grant requirements (e.g., the National Institutes of Justice[91]). Individual programs that fund research in specific areas require data underlying an article arising from their grant to be placed in a public archive (e.g., NSF Division of Social, Behavioural and Economic Sciences[92]).

The National Science Foundation (NSF) has embodied its overall philosophy regarding access to the results of research in the NSF General Grant Conditions (cf. Table 4 below):

> The NSF expects investigators to share with other researchers, at no more than incremental costs and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.

A number of NSF divisions and programs have developed specific data access policy statements that are in keeping with this general philosophy but which also recognise discipline, community, or program-specific needs, limitations, and standards. A sample of these statements can be found in Table 4 below. Where the NSF statement uses 'expect', several programs are 'requiring' data to be deposited or made available for re-use, and also require a data management and sharing plan to be included in the grant application as a condition for receiving funding. Despite referring to the general NSF statement the NSF programs' statements do not appear to form a coherent and consistent set that would cover all NSF funded research data with uniform requirements for archiving and sharing, often leaving the choice up to the research project.

The US National Institutes of Health (NIH) has required data sharing in several areas, such as DNA sequences, mapping information, and crystallographic coordinates since 1996.[93] An NIH data sharing policy has been developed and, akin to the Wellcome Trust and MRC in the UK, this has been supplemented with thorough guidance, examples of data sharing plans and consultation on appropriateness of data sharing and suitable mechanisms for disseminating data.[94] The NIH data sharing policy states:[95]

> Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.

The policy statement applies NIH-wide to basic research, clinical studies, surveys, and other types of research supported by NIH that involves human subjects and laboratory research that does not involve human subjects. Data management and sharing plans are currently required only from applications for large grants (over $500,000 a year). Given the breadth and variety of science that NIH supports, neither the precise content for the data documentation, nor the formats, presentation, or transport mode for data is stipulated. But most NIH divisions and programs have developed their own data sharing statements that include examples of domain- and data-specific needs and requirements. Examples of such statements are included in Table 4 below.

---

[88] Stephen E. Fienberg, Margaret E. Martin, Miron L. Straf, (Eds.), *Sharing Research Data* (1985)
[89] American Sociological Association, *Code of Ethics* (1997)
[90] American Psychological Association, *Ethical Principles of Psychologists and Code of Conduct* (2002)
[91] Cf. National Archive of Criminal Justice Data, http://www.icpsr.umich.edu/NACJD/archiving/
[92] NSF SBE, *Data Archiving Policy*
[93] NIH, Frequently Asked Questions on Data Sharing, http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm
[94] NIH Data Sharing Policy documents, http://grants.nih.gov/grants/policy/data_sharing/
[95] NIH, *Statement on Sharing Research Data* (2003)

The US Association of Research Libraries (ARL) has been discussing the role of research libraries in the e-science framework and from a workshop held in 2006 made the recommendation to:[96]

> Develop policy infrastructure to create a culture of sharing. This recommendation calls for creating data management policies to ensure that the contribution of research data is considered a shared asset, enabling reuse in new research contexts as shared public goods.

ARL has also published its model principles for data that start with:[97]

> 1. Open Access: Research libraries will support open access policies and practices regarding scientific knowledge and e-science. Barriers will be removed that impede or prevent open access to research outputs, and consequently that restrict the potential linkage of outputs to the data upon which research findings are based.
> 2. Open Data: Access to open data is a movement supported by research libraries, taking into consideration the ethical treatment of human-subject data.

Statements on data sharing principles have been formulated by a number of other funding and research organisations – NASA, the National Institute of Justice, the Robert Wood Johnson Foundation, and many other organisations – that do not differ significantly from the examples discussed above and presented in Table 4.

The US Government Accountability Office analysed data sharing policies and practices in the climate change research projects funded by the Department of Energy (DOE), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the National Science Foundation (NSF). GAO surveyed 64 program managers at the four major climate change research agencies and identified 23 different policies that account for about 80 percent of the agencies' climate change research programs and that encourage researchers to make data available. Some of the data generated by this research are stored in online archives, but much remains in a less accessible format with individual researchers. Similar to the UKRDS study in the UK, the GAO report found that: [98]

> While the four agencies have taken steps to foster data sharing, they neither routinely monitor whether researchers make data available nor have fully overcome key obstacles and disincentives to data sharing. Because agencies do not monitor data sharing, they lack evidence on the extent to which researchers are making data available to others. Key obstacles and disincentives could also limit the availability of data. For example, one obstacle is the lack of archives for storing certain kinds of climate change data, such as some ecological data, which places a greater burden on the individual researcher to preserve it. Preparing data for future use is also a laborious and time-consuming task that can serve as a disincentive to data sharing. In addition, data preparation does not further a research career as does publishing results in journals. The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure. Consequently, researchers are less likely to focus on preserving data for future use, thereby putting the data at risk of being unavailable to other researchers.

GAO recommends the agencies explore opportunities in the grants process to better ensure the availability of data to other researchers and determine if additional archiving strategies are warranted. The four agencies generally agreed with GAO report findings and recommendations.

Similar to GAO's and RIN's findings about UK research councils' data sharing policies, the National Science Board's analysis of US science organisations policies concludes:[99]

> Data policies are well established and stable for observational earth science data. This may arise in part because of the existence of a well-established system of world data centers that provide archives for data.

Access to research data and various limitations to it have been the subjects of a number of reports, guidance and policies in the United States. The Council on Governmental Relations has analysed the various limitations on accessing research data in a guide that examines the research institution's obligations irrespective of the funding source and regardless of the type of funding mechanism selected.[100] The National Academy of Engineering is supporting work of the Online Ethics Center that has created guidance on *Responsible*

[96] *To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering* (2006)
[97] Joint Task Force on Library Support for E-Science, *Agenda for Developing E-Science in Research Libraries* (2007)
[98] GAO, *Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research* (2007)
[99] National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (2005), p. 71
[100] Council on Governmental Relations, *Access to and Retention of Research Data: Rights and Responsibilities* (2006)

Collection, Retention, Sharing, and Interpretation of Data.[101] The report is part of the Responsible Research series and lists main principles for of responsible data management with scenarios and case studies for data dissemination attached. The National Human Subjects Protection Advisory Committee has approved a set of recommendations on *Public Use of Data Files*[102] that stress the importance of institutional policies for correct statement on the use of public and classified data in research. Access to clinical trial data is treated in numerous policies; a typical example is the National Heart, Lung and Blood Institute *Data Set Policy*[103] that states:

> Under no circumstances will data relating to an individual be distributed in any way that is inconsistent with his or her informed consent. To ensure that the confidentiality and privacy of study participants are protected, all investigators seeking access to data from NHLBI-supported studies that are in the possession of the Institute must execute and submit with their requests the appropriate standard Distribution Agreement for each study.

Protection of the privacy of personal health information in the U.S. is protected by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Guidance for applying the HIPAA rules to research data has been published by the Department of Health and Human Services.[104]

[101] Caroline Whitbeck, *The Responsible Collection, Retention, Sharing, and Interpretation of Data*. Online Ethics Centre (2006)
[102] National Human Subjects Protection Advisory Committee, *Public Use of Data Files* (2002)
[103] National Heart, Lung and Blood Institute, *Data Set Policy* (2005)
[104] Department of Health and Human Services, *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule* (2004)

Table 4. US research funding agencies' data sharing policy statements.

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
National Science Foundation	NSF, General Grant Conditions (2007)[105]	NSF expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.	-	NSF program management will implement these policies through the proposal review process; through award negotiations and conditions.	Appropriate support and incentives for data cleanup, documentation, dissemination, storage and the like.	Data should be made openly available as soon as possible, but no later than two years after the data were collected.	Adjustments and, where essential, exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate legitimate interests of investigators.
NSF Division of Earth Sciences	EAR Data Policy (2002)[106]	Data may be made available for secondary use through submission to a national data center, publication in a widely available scientific journal, book or website, through the institutional archives that are standard for a particular discipline (e.g. IRIS for seismological data, UNAVCO for GPS data), or through other EAR-specified repositories.	-	-	-	For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as possible, but no later than two years after the data were collected.	In the interest of full and open access, data should be provided at the lowest possible cost to researchers and educators. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. In some programs selected principle investigators can have initial periods of exclusive data use.
NSF Division of Environmental Biology	DEB About Environmental Biology (2002)[107]	Proposals submitted to all programs in DEB must adhere to the general NSF policy on data sharing as described in the Grant Proposal Guide.	Proposals should describe plans for specimen and information management and sharing, including where data and metadata, will be stored and maintained, and the likely schedule for release.	These plans will be considered as part of the review process.	-	-	-

[105] NSF, General Grant Conditions (2007)
[106] NSF, *Division of Earth Sciences Data Policy* (2002)
[107] http://www.nsf.gov/bio/deb/about.jsp

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
NSF Division of Ocean Sciences	*Division of Ocean Sciences Data and Sample Policy*, NSF 04-004 (2003)[108]	Ocean data collected under Federal sponsorship and identified as appropriate for submission to a national data center are to be made available within a reasonable time.	Principal investigators are required to submit all environmental data and inventories of all marine environmental data collected to the designated national data centers. Data are to be submitted according to formats and via the media designated by the pertinent national data center.	-	-	Data should be submitted as soon as possible, but no later than two years after the data are collected and inventories of data collected within sixty days after the observational period/cruise.	Principal investigators and ship-operating institutions are also responsible for meeting all legal requirements for submission of data and research results, which are imposed by foreign governments as a condition of that government's granting research clearances.
NSF Division of Social, Behavioral and Economic Sciences	*SES Data Archiving Policy* (2008)[109]	Grantees from all fields will develop and submit specific plans to share materials collected with NSF support, except where this is inappropriate or impossible.	These plans should cover how and where these materials will be stored at reasonable cost, and how access will be provided to other researchers, generally at their cost.		Included in the data plan.	For appropriate data sets, researchers should be prepared to place their data in fully cleaned and documented form in a data archive or library within one year after the expiration of an award.	

[108] NSF, *Division of Ocean Sciences Data and Sample Policy* (2003)
[109] NSF, *SES Data Archiving Policy* (2008)

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
National Institutes of Health	*NIH Statement on Sharing Research Data* (2003)[110]	The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. NIH expects the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset.	Investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.	Reviewers will not use the data sharing plan in determining scientific merit or priority score for applications. Program staff are responsible for over-seeing the data sharing policy and for assessing the appropriateness and adequacy of the data sharing plan. Program concerns must be resolved prior to making an award.	Applicants may request funds for data sharing and archiving. The financial issues should be addressed in the budget section of the application.	-	-
NIH Department for Cancer Epidemiology and Genetics	DCEG *Data Sharing Policy* (2008)[111]	Data from completed studies with published research findings should be made freely available.	The data to be shared and the manner of sharing should be established by mutual consent of all the collaborators. Data sharing plans must be included in all protocols submitted for Committee for the Technical Evaluation of Protocols Committee review.	-	-	-	Data must be stripped of all personal identifiers and any variables that permit deductive identification of individual subjects to ensure confidentiality. Studies where data collection is ongoing are exempt from data sharing.
NIH National Institute for Allergy and Infectious Diseases	NIAID *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (2007)[112]	All investigators who receive NIH support to conduct genome-wide analysis of genetic variation in a study population are expected to submit to the NIH GWAS data repository descriptive information about their studies for inclusion in an open access portion of the NIH GWAS data repository.	The NIH strongly encourages the submission of curated and coded phenotype, exposure, genotype, and pedigree data to the NIH GWAS data repository as soon as quality control procedures have been completed at the local institution.	-	-	As soon as quality control procedures have been completed at the local institution.	The data are released to qualified researches who wish to collaborate with the investigators.

[110] NIH, *Statement on Sharing Research Data* (2003)
[111] Department Cancer Epidemiology and Genetics, *Data Sharing Policy* (2008)
[112] National Institute for Allergy and Infectious Diseases, *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (2007)

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
American Sociological Association	ASA *Code of Ethics*. Chapter 13.5 Data Sharing (1997)[113]	Sociologists share data and pertinent documentation as a regular practice. Sociologists make their data available after completion of the project or its major publications, except where proprietary agreements with employers, contractors, or clients preclude such accessibility or when it is impossible to share data and protect the confidentiality of the data or the anonymity of research participants.	Sociologists anticipate data sharing as an integral part of a research plan whenever data sharing is feasible. Sociologists who do not otherwise place data in public archives keep data available and retain documentation relating to the research for a reasonable period of time after publication or dissemination of results.	-	-	After completion of the project or its major publications.	-
American Psychological Association	APA *Ethical Principles of Psychologists and Code of Conduct* (2002)[114]	After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release.	-	-	This does not preclude psychologists from requiring that such individuals or groups be responsible for costs associated with the provision of such information.	After research results are published.	Psychologists who request data from other psychologists to verify the substantive claims through reanalysis may use shared data only for the declared purpose. Requesting psychologists obtain prior written agreement for all other uses of the data.

[113] American Sociological Association, *Code of Ethics* (1997)
[114] American Psychological Association, *Ethical Principles of Psychologists and Code of Conduct* (2002)

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
Office for Research Integrity	*ORI Introduction to the Responsible Conduct of Research* (2004)[115]	Once a researcher has published the results of an experiment, it is generally expected that all the information about that experiment, including the final data, should be freely available for other researchers to check and use.	-	-	-	Once a researcher has published the results of an experiment.	Freely available. Research data must be made available in response to Freedom of Information Act (FOIA) requests (OMB, Circular A-110).

[115] Office for Research Integrity, *Introduction to the Responsible Conduct of Research* (2004)

In **Australia**, the original ministerial *Accessibility Framework* was proposed as a strategic framework to improve access to research information, outputs and infrastructure. It is an agreed system-wide approach for managing research outputs and infrastructure, so that they are discoverable, accessible and shareable, in order to improve the quality of research outcomes, reduce duplication and better manage research activities and reporting. The broad principles of the Framework and its supporting projects are:[116]

- publicly funded research outputs and data should be managed in ways that maximise public benefit;
- institutions or individuals receiving public money have an obligation to make the results of their research publicly available as soon as possible;
- outputs should be made accessible through an institutional digital repository, a subject matter digital repository and/or open access publication;
- constraints to open access should only apply in a very few cases, for example for reasons of national security, privacy or cultural sensitivity;
- at the research proposal stage, researchers should consider how to ensure readability, durability and re-usability of research outputs and data;
- the provision of open access should include the curation and preservation of digital material including cataloguing, archiving, reproducing, safekeeping and media migration of research outputs.

The Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC) have included an expectation that data would be shared in their funding conditions and have jointly stated their desire to ensure the widest possible dissemination of the research supported by their grants, in the most effective manner and at the earliest opportunity (see the policy statements in Table 5 below). Given that most ARC and NHMRC projects have a three-four year timeframe there will be some delay for these policies to take full effect. Both organisations took part in developing the joint *Australian Code for the Responsible Conduct of Research*. The Code includes a chapter on responsible data management and sharing where it states:[117]

> Policies are required that address the ownership of research materials and data, their storage, their retention beyond the end of the project, and appropriate access to them by the research community. The responsible conduct of research includes the proper management and retention of the research data. Retaining the research data is important because it may be all that remains of the research work at the end of the project. While it may not be practical to keep all the primary material (such as ore, biological material, questionnaires or recordings), durable records derived from them (such as assays, test results, transcripts, and laboratory and field notes) must be retained and accessible. The researcher must decide which data and materials should be retained, although in some cases this is determined by law, funding agency, publisher or by convention in the discipline.

The Code also states explicit requirements on institutional level for data management and retention:

> Each institution must have a policy on the retention of materials and research data. The institutional policy must be consistent with practices in the discipline, relevant legislation, codes and guidelines.
> In general, the minimum recommended period for retention of research data is 5 years from the date of publication. However, in any particular case, the period for which data should be retained should be determined by the specific type of research.
> Institutions must provide facilities for the safe and secure storage of research data and for maintaining records of where research data are stored.
> Each institution must have a policy on the ownership of research materials and data during and following the research project. The policy must guide researchers in the management of research data and primary materials, including storage, access, ownership and confidentiality.

The Accessibility Framework offers the ARC, NHMRC and other funding bodies the possibility of further strengthening their competitive grant funding rules to mandate rather than encourage the deposit of research outputs into repositories, including open access repositories.

[116] Cf. Alexander Cooke, *Research Infrastructure and the Open Access Agenda* (2008)
[117] National Health and Medical Research Council, Australian Research Council, Universities Australia, *Australian Code for the Responsible Conduct of Research* (2007)

Table 5. Data sharing policy statements from Australian funding agencies

Funding Body	Policy Document	Data Sharing Policy Statement	Mandate for Projects	Possible Sanctions	Cost of Data Sharing	Timing of Data Sharing	Accessibility
Australian Research Council	ARC Funding Rules (2008)[118]	The ARC … encourages researchers to consider the benefits of depositing their data and any publications arising from a research project in an appropriate subject and/or institutional repository wherever such a repository available to the researcher(s).	If a researcher is not intending to deposit the data from a project in a repository within a six-month period, he/she should include the reasons in the project's Final Report. Any research outputs that have been or will be deposited in appropriate repositories should be identified in the Final Report.	-	-	Within six months of the completion of the research	-
National Health and Medical Research Council	NHMRC *Project Grants Funding Policy*[119]	The NHMRC encourages researchers to consider the benefits of depositing their data and any publications arising from a research project in an appropriate subject and/or institutional repository wherever such a repository is available to the researcher(s).	Any research outputs that have been or will be deposited in appropriate repositories should be identified in the Final Report.	-	-	-	To maximise the benefits from research, findings need to be disseminated as broadly as possible to allow access by other researchers and the wider community.

[118] ARC, *Discovery Projects: Funding Rules for Funding Commencing in 2008* (2008)
[119] National Health and Medical Research Council, *Project Grants Funding Policy for Funding Commencing in 2008* (2006)

The **Canadian** Institutes of Health Research have subscribed to the open sharing of their research output and in their policy state:[120]

> Recognizing that access to research data promotes the advancement of science and further high-quality and ethical investigation, CIHR explored current best practices and standards related to the deposition of publication-related data in openly accessible databases. As a first step, CIHR will now require grant recipients to deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database, as already required by most journals, immediately upon publication of research results (e.g., deposition of nucleic acid sequences into GenBank).
> CIHR now requires grant recipients to retain original data sets arising from CIHR-funded research for a minimum of five years after the end of the grant. This applies to all data, whether published or not. The grant recipient's institution and research ethics board may have additional policies and practices regarding the preservation, retention, and protection of research data that must be respected.

The **German** Research Foundation (Deutsche Forschungsgemeinschaft, DFG) has tied open access into its funding policy since 2006 when the DFG's Senate and Joint Committee recommended encouraging funded scientists to digitally publish their results and make them available via open access. The DFG expects the research results funded by it to be published and to be made available, where possible, digitally and on the internet via open access. To achieve this, the contributions involved should either be deposited in discipline-specific or institutional electronic archives (repositories) following conventional publication, or should be published in a recognised peer-reviewed open access journal.[121]

The research funding agencies in Germany are gradually following the line set by the DFG. Fraunhofer Gesellschaft's *Open Access Policy* states in the preamble:[122]

> It is an essential requirement that scientific data and research findings should be made openly and immediately available, with the obvious exception of confidential information supplied by customers of the Fraunhofer Institutes for the purposes of a specific project.

Yet the policy then goes on to make policy statements on dissemination of articles and scientific papers alone.

The Max-Planck Society has set up an Open Access Unit as part of the Max Planck Digital Library whose long term aim is to assist in making all knowledge produced at the Max Planck Institutes freely available via the internet, and by this help to achieve the goal set by the Berlin Declaration 'to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access'. While independent open access policies are pursued at individual Max Planck Institutes it is the responsibility of the Max Planck Open Access Unit to support the institutes in these activities and to help formulate a coherent open access policy for Max Planck Society as a whole.[123]

The Helmholtz Association (Helmholtz Gemeinschaft) was actively involved in preparing the Berlin Open Access Declaration, and has followed the open access principles since then. The Association's formal Open Access Policy covers only research publications,[124] but the Association has taken steps to promote open data sharing alongside publications. In 2005 it set up a working group on defining open access principles and services at the Helmholtz Society that defined four directions, one of them dedicated to models for open access to research data.[125] The Open Access Project offers support to individual researchers as well as Helmholtz Centres in realising the open access principles.[126]

A similar Open Access Workgroup was set up at the Leibnitz Gemeinschaft that first published nine theses on open access, and a year later its *Guidelines to Open Access in the Leibnitz Society* (2007).[127] The Guidelines state that researchers must publish their output digitally and make it freely accessible as soon as possible. The choice of channel for publication is left up to the researcher, but repositories and archives within the Leibnitz Society are the recommended practice.

---

[120] CIHR, *Policy on Access to Research Outputs* (2007)
[121] http://www.dfg.de/en/news/information_science_research/other_news/info_wissenschaft_04_06.html
[122] Fraunhofer Gesellschaft, *Open Access Policy* (2008)
[123] Max Planck Open Access Unit, http://www.mpdl.mpg.de/profile/openaccess_en.htm?mp=25
[124] Open Access at the Helmholtz Association, http://www.helmholtz.de/en/research/open_access/
[125] Arbeitsgruppe Open Access in der Helmholtz-Gemeinschaft, *Realisierung des offenen Zugangs zu Publikationen und Daten aus der Helmholtz-Gemeinschaft* (2005)
[126] http://oa.helmholtz.de/index.php?id=137#c1115
[127] *Leitlinie zu Open Access in der Leibniz-Gemeinschaft* (2007)

In **Finland**, a high-level working group was appointed by the Minister of Culture in 2004, to develop a national policy position on open access to scholarly output. The workgroup made 30 recommendations to research funders, universities and public research institutes, individual researchers, scientific journals and societies, libraries and the Ministry of Education.[128] The working group recommended that funding should be made available for further institutional repositories, that research grants should cover also the preparation of open access publications and that the ministry should fund building a supportive infrastructure for the accessibility and preservation of primary research materials.

The **French** Research Agency (Agence nationale de la recherche, ANR) has not developed a data archiving or sharing policy as yet, but does encourage researchers to disseminate publications arising from their research via open access repositories.[129] The Hyper Archives en Ligne (HAL) repository[130] run by the Centre pour la communication scientifique directe (CCSDS)[131] of the French National Centre for Scientific Research (CNRS) caters for researchers from all fields, offers long-term preservation for the deposited content, and is a preferable option for depositing French research articles.

In **Spain**, the Centro de Investigaciones Sociologicas (CIS) has established data access policies based on national legislation and royal decrees:[132]

> The Centro de Investigaciones Sociológicas works under the principles of objectivity and neutrality in its actions, of equal access to its data and of respect for the rights of the citizens and statistical privacy.

The **Japanese** research funding organisations do not explicitly mention data in their funding guides and handbooks, although dissemination of publications is covered.[133]

Open access has prompted research funding organisations, institutions and even departments within research institutions to publish policies that often also cover data as one part of the research output. The recommendation to publish these in open access repositories as soon as the research results are published, is a common requirement. The Registry of Open Access Repository Material Archiving Policies (ROARMAP) is part of the services offered by the eprints.org. Their list of policies from research funding bodies currently stands at 29 policies;[134] countries where research data are part of the open access policy include Australia, Belgium, Canada, Ireland, and Switzerland.

The research funding organisations around the globe are making significant progress in taking a pro-active stance in mandating sharing of research data and are developing policies in relation to data. Incentives for them to develop data sharing and publication policies come from different directions:

- there are international examples of such policies;
- the culture of sharing research outputs is gradually changing in research communities and domains of science towards openness;
- government departments and audit offices are putting pressure on research funders to ensure that outputs from publicly funded research are freely available;
- the open access movement is expanding its scope to cover research data as well as the publications that are based on them.

Models of funding research vary from country to country – large countries have research foundations and associations that comprise dozens of research institutions, smaller countries have fewer fund-making organisations, but they often specialise in specific areas of science and research. This diversity is reflected in the data sharing policies that in some cases are quite detailed and in others necessarily general, leaving the details of what data should be shared how and when to institutional and funding programme policies. This is part of the explanation why not all research

[128] Opetusministeriö, *Recommendations for the Promotion of Open Access in Scientific Publishing in Finland* (2005)
[129] L'ANR incite les chercheurs à intégrer leurs publications dans le système d'archives ouvertes
[130] http://hal.archives-ouvertes.fr/index.php?langue=en&halsid=vjs4o2h09m4lmhfeo4f9cfe3c0
[131] http://www.ccsd.cnrs.fr/
[132] CIS Legislation, http://www.cis.es/cis/opencms/EN/7_cis/legislacion.html
[133] Cf. for example Japan Society for the Promotion of Science, *Handbook for the Use of Kakenhi Research-in-Aid for Scientific Research* (2008)
[134] http://www.eprints.org/openaccess/policysignup/

> funding organisations have made data sharing a condition for receiving funding – these conditions can be imposed on the next level. The other part of the explanation is the services that research funding organisations can offer for data archiving and dissemination – where these have been set up centrally by the funders, or a distributed network of institutional repositories in research institutes of the funding organisation is present, data deposit and dissemination are most often included in the conditions for receiving funding.
>
> Major differences remain in the degree to which funding organisations see making data accessible and acting as its long-term guardians as an appropriate or feasible role for themselves. Co-ordination of data sharing principles among research funding organisations is called for to support availability of data for re-use in all fields of research and to thus facilitate interdisciplinary research. There is much room for development in this area both on national and international levels where legal obstacles have to be analysed and a culture change will be necessary in many domains.

2.4 INSTITUTIONAL POLICIES

Increasingly, the focus of publishing research outcomes and maintaining the underlying data is shifting from national and research domain level funding and services to individual research institutions who see this as a way of promoting their own work. Open access principles are having a profound effect on rapid development of institutional level policies for openly sharing research outputs, but not all of these incorporate research data yet. The ROAR service currently lists only 24 institutional mandates for open access publication,[135] but in reality this number is much larger. In some sciences the so called Mertonian tradition of open science has created a culture where proprietary exploitation of data, as opposed to inventions derived from data, has been discouraged and availability of datasets on which a published work was based has been required as a condition of publication.[136] These principles, where formulated, were the antecendents of open access.

A large number of **inter-institutional policies** can be found that address data sharing either directly or indirectly, most often as part of open access policies.

Many large scale and long-term research projects have been defining their own data sharing policies, to foster exchange and promote the results of the project. An example of this is the US NIH funded Human Brain project that was launched in 1993 to develop and support a new science: neuroinformatics. The principal investigators in the project offer guidelines in support of data sharing, designed to reduce technological and sociological barriers to effective data sharing. A website (datasharing.net) was established by one of the partners, and a statement with principles for data sharing was issued:[137]

- Current NIH policy mandates data sharing for high-direct-cost grantees. We urge broader, voluntary, sharing of data, and adoptions of norms for the proper presentation and use of shareable data.
- Just as results are published freely and openly, without restrictions, so most data should be made available for sharing, consonant with appropriate privacy or proprietary restrictions.
- Just as publication is timely, so data should be made available without delay.
- Just as publications are citable archives, so shared data and its locators should be maintained.
- Just as citation of others' publications is essential to scientific communication, so citation and acknowledgment of shared data should be required.
- Just as publication costs are recognized as appropriate direct costs of research, so expenses of data sharing should be supported.

The Atmospheric Radiation Measurement (ARM) Program in the United States is a multi-laboratory, interagency program, funded by the U.S. Department of Energy and is a key contributor to national and international research efforts related to global climate change. It has an extensive *Data Sharing and Distribution Policy* that supports the open sharing of data principles:[138]

- Free and open sharing of data.
- Timely (e.g., 'near real time' where desired) delivery of processed data from the ARM Data Archive to Science Team members.

[135] http://www.eprints.org/openaccess/policysignup/
[136] http://sciencecommons.org/about/towards/
[137] Human Brain Project, *Principles of Data Sharing* (2002)
[138] ARM, *Data Sharing and Distribution Policy* (2006)

- Timely access to data by the general scientific community through the ARM Data Archive.
- Timely sharing of all data among various participants in ARM-sponsored programs.
- Recognition of data sources either through co-authorship or acknowledgments as appropriate.
- Sharing of data of common interest from external sources when possible, Some sources restrict secondary distribution of data. In these cases, ARM will seek specific allowances to distribute such data to members of the ARM Science Team, but will observe restrictions on further distribution from the ARM Data Archive if required.

Groupings of universities and bodies where university rectors represent their institutions have developed shared positions, codes and memoranda that the participating institutions subscribe to. The Russell Group of universities in the UK approved their Code of Practice on Good Research Governance that states:[139]

> All members of the Russell Group [...] confirm that they either have in place, or are developing and implementing, policies and procedures to [...] ensure, as robustly as possible, that for all research activities:
> - research results and data are recorded, dated and subsequently stored securely, preferably at a level higher than the individual.

In Italy, the Conference of Italian Universities Rectors (CRUI) has acknowledged the importance of full and open access to the research information and data for research and scientific education, and is promoting free dissemination on the web of research achievements developed in Italian universities and research centres.[140]

Universities UK (UUK) has defined a position statement on access to research publications:[141]

> Universities UK support the principle that the outcomes of publicly funded research should be made available as widely as possible with no barriers to access.

The practical implications of such group statements for individual subscribing institutions have most often not been considered, neither is there an explicit mechanism in place for verifying that the principles are, indeed, being followed.

There has been an upsurge in the number of **institutional policies** in recent years. In 2007 RIN surveyed data policies of UK universities and reported that:[142]

> [U]niversities tend to treat the management of datasets produced as a research output as almost exclusively a matter for researchers and their departments. The universities interviewed do not have policies requiring researchers to manage data in any particular ways, except in accordance with funding body requirements or best practice. Although a few universities are seeking actively to gather datasets together in their repositories, those interviewed do not require or even encourage the deposit of data, although they acknowledge that this issue will have to be addressed at some point, given the increasing importance of data and the need for integration with other research outputs.

Policy statements on research data deposit and sharing have started to emerge, but are few, and commonly are attached to institutional repositories that universities are maintaining.[143] University policies on deposit and sharing of research papers and conference presentations via institutional repositories are more common.[144] The JISC commissioned report *Dealing with Data* recommends:[145]

> Institutions need to have a data management, curation, sharing and preservation policy. The institutional policy should recommend that researchers deposit their data in an appropriate public data repository, (which may not necessarily be only in an institutional repository), where these exist.

In the United States, university level policies have existed for some time, but are most often focused on re-iterating or detailing regulations of funding organisations for research projects.[146] Most policies reviewed placed the responsibility for data retention and decisions to disseminate it on the researcher who has primary responsibility for the research project. For example, the University of Kentucky *Data Retention and Ownership Policy* states:[147]

---

[139] Russell Group, *Code of Practice on Good Research Governance* (2005)
[140] Roberto Delle Donne, *CRUI and Open Access in Italy* (2007)
[141] Universities UK, *Access to Research Publications: Position Statement* (2005)
[142] RIN, *Research Funders' Policies for the management of information outputs* (2007), p. 63
[143] For example: Queen Margaret University, http://eresearch.qmu.ac.uk/policies.html; Brunel University, http://bura.brunel.ac.uk/deposit-guide.html
[144] http://www.eprints.org/openaccess/policysignup/
[145] Liz Lyon, *Dealing with Data* (2007) p. 47
[146] See: Gunta Liders, *Comparison of Institutional Data Policies* (2002)
[147] University of Kentucky, *Data Retention & Ownership Policy* (1999; 2006)

> Research data […] must be retained by the principal investigator for a period of five years after publication or submission of the final report on the project for which the data were collected, whichever is longer. If the retention requirements specified in other statutes or external agency's regulations are longer, the agency requirements will apply. During the retention period, data must be immediately provided to University administration, upon request. Data must be available to representatives of external sponsors or designated governmental officials, as appropriate.

The Faculty of Medicine at Harvard University makes the following provision for data:[148]

> Primary data should remain in the laboratory at all times and should be preserved as long as there is any reasonable need to refer to them. The chief of each research unit must decide whether to preserve such primary data for a given number of years or for the life of the unit. In no instance, however, should primary data be destroyed while investigators, colleagues, or readers of published results may raise questions answerable only by reference to such data.

The University of Massachusetts commissioned a report to develop its own policy on data:[149]

> Researchers shall endeavor to make their data publicly available as soon as possible, and to the extent possible. Access may be delayed while the correctness of the data is being verified, until an initial publication based on the data appears, for the minimum period needed to file a patent application, or for any other reasonable need. Data should be released early if benefit to the public is likely.

The University of Rochester has put in place several policies on openness in research, data sharing plans and also an interim policy on data retention:[150]

> The original research data shall be in the custody of the senior investigator on behalf of the University, but must be returned to the University upon request of the Provost. Additionally, such data must be available to representatives of external sponsors of the research or designated governmental officials, when such access is appropriate.

The James Madison University has included a through *Institutional Data Stewardship Model* in its Manual of University Policies and Procedures.[151] The Model details responsibility and rights over data produced by the University staff, and sensitive data should be protected.

In Europe, university level policies are appearing for open access to research outputs and these often already cover, or at least hold a promise for including in the future, research data. The issues of cost of self-archiving of both publications and data into institutional repositories has been dealt with in very few of these policies, but positive examples are appearing (e.g., the University of Minho in Portugal).

A recent survey of professors in humanities, social sciences and behavioural sciences at Finnish universities revealed that 90% of them had no guidelines or policies from their departments on the preservation of digital research data.[152] The University of Helsinki now requires that researchers deposit copies of their research articles in the university's open repository.[153] Concurrently with establishing the open access mandate for publications the rector also appointed a working group with the task to analyse how and to what extent could research data of the university be made accessible to researchers and the general public.[154]

It has also been noted that the tradition of institution level policies will be difficult to introduce in some countries because of existing traditions:[155]

> The ideal green road is a world in which each institution has the responsibility to implement broad institutional mandates to deposit academic output or else a place where patchwork mandates on individual or faculty levels are established. However, discourse in countries such as France and the Netherlands for example suggests that other methods are necessary to encourage the participation of researchers where institutional mandates could be more difficult to enforce. This is where incentives need to play a more significant role.

---

[148] Faculty of Medicine, Harvard University, *Guidelines for Investigators in Scientific Research* (1988)
[149] University of Massachusetts, *Data Ownership, Retention, and Access at the University of Massachusetts Amherst* (2006)
[150] University of Rochester, *Interim Policy on Access to and Retention of Research Data* (n.d.)
[151] James Madison University, *Data Stewardship Model* (2008)
[152] Arja Kuula, Sami Borg, Open Access to and Reuse of Research Data – The State of the Art in Finland (2008)
[153] University of Helsinki, *Open Access to Research Publications in the University of Helsinki* (2008)
[154] Marjut Salokannel, *University of Helsinki Opens its Research Vaults: A few Words on Open Access and the New Research Environment in Finland* (2008)
[155] DRIVER, *A DRIVER's Guide to European Repositories* (2008), p. 50

As discussed above, many research funding organisations require a data management plan to be part of a research project proposal. The **project level** data policy is expected to include high-level statements on how data generated or compiled in the research project will be made available for access and use. Yet, projects' data sharing principles and policies are often limited to reciting the funding agency's principles. A recent survey in Australia discovered that 80% of respondents acknowledged that they do not have a formal data management plan, and it concluded that:[156]

> Research data management would be easier for all concerned if researchers, research units and research organisations all had policies and plans surrounding the creation and management of data.

To respond to this situation, an extensive guidance document has been published by the e-Research and OAK-Law projects[157] to help organisations develop their data management plans, to provide for IPR protection and data sharing. It suggests that a Data Management Plan should also deal with access issues, including:

- which data is to be made accessible (some data may be subject to legal restrictions (such as confidentiality or privacy restrictions) and may not be able to be made openly accessible);
- when the data is to be made accessible (at the conclusion of the research project or before?);
- who may access the data (open or restricted access?);
- how the data will be made accessible (via an online repository?);
- how wide are the access rights to be granted (for example, access to the entire dataset, or will some parts of the dataset require the user to enter a password to access the data?); and
- reuse of the data, namely, what reuse rights are to be granted to end users and how these rights will be granted (for example, through an end user copyright licence such as a Creative Commons licence?).

There is also an expectation that research publishers can play at least as important a role in facilitating data sharing as do research funders, especially through their power to influence practice. Journal policies are usually expressed within 'instruction for authors' statements. A recent analysis of journal policies in the genetic research area[158] showed that of the 70 journal policies found, 53 made some mention of sharing publication-related data within their Instruction to Author statements. Of the 40 policies with a data sharing policy applicable to gene expression microarrays, 17 were classified as weak and 23 as strong (strong policies required an accession number from database submission prior to publication). Existence of a data sharing policy was also associated with the type of journal publisher: 46% of commercial journals had data sharing policy, compared to 82% of journals published by an academic society; all five of the open access journals had a data sharing policy. Most of the analysed policies state that data must be made available in a public database. A few are less specific, stating that sharing via public webpages or supplementary journal information is sufficient, or the policy leaves location unspecified. Thus, the responsibility for retaining and sharing data is left with the researchers. For example, *Nature*'s view on original data is:[159]

> Authors should retain all original data and analyses to:
> - Provide evidence of deposition in recognized repositories at submission
> - Be prepared to provide any additional data that referees and review process may require

A requirement to provide supporting data with articles submitted was introduced to the *Chemical Communications*. Their submission guidelines state that:[160]

> Experimental information must be provided to enable other researchers to reproduce accurately the work.

In Japan, many institutional repositories of universities have declared an 'everything is welcome' policy for voluntary deposit of not only written articles, but also images, datasets and multimedia content, with positive results.[161]

> The emerging institutional policies still remain ad hoc and do not appear to be well coordinated. To develop uniform data sharing policies and put them into practice, the institutions will require significant help and guidance.

[156] Margaret Henty, et al., Investigating Data Management Practices in Australian Universities (2008)
[157] Anne Fitzgerald, Kylie Pappalardo, Anthony Austin, *Practical Data Management: A Legal and Policy Guide* (2008)
[158] Heather Piwowar, Wendy Chapman, *A Review of Journal Policies for Sharing Research Data* (2008)
[159] Maxine Clarke, *Raw Data Policy of Scientific Journals? The Nature Perspective* (2007)
[160] RSC Chemical Communications, *Guidelines for Authors* (2007)
[161] Cf. http://ir.library.osaka-u.ac.jp/metadb/up/DRFIC2008/ChibaUniv.pdf

3. INFRASTRUCTURE AND SERVICES FOR RESEARCH DATA SHARING

In contrast with printed books and articles, sharing and re-using digital data requires considerable infrastructure to be available for researchers. Hence, the world of research data sharing is not yet remotely as organised or secure as the traditional system of libraries. Typical problems encountered by researchers seeking to re-use existing data include:[162]

> In many scholarly fields, the data used in large fractions of published books and articles do not exist in public archives and often cannot be located anywhere. Data sometimes exist on individual researchers' web sites, without professional backups, off-site replication, plans for format conversion and migration, or professional cataloguing. Sometimes URLs are given, but they often do not last for long. Data created more than 5 to 10 years ago may exist in defunct storage media or the inaccessible formats of old statistics or database packages, operating systems, or hardware.
> Data citation practices differ across fields, among archives, and across and even within individual publications. Data are sometimes listed in the references, sometimes in the text, and only occasionally with enough information to guarantee future access to the identical data set. Data referred to may no longer exist, may not be available at all, or may be available only from the author or with his or her approval (which is sometimes forthcoming only if the author thinks he or she won't be criticized). Data listed as available from the author are unlikely to be available for long and will not be accessible after the author retires or dies.

For data to be re-used, they need to be actively curated, adequately described and linked to services that disseminate the data to other researchers and stakeholders. The current methods of storing research data are as diverse as the disciplines that generate it and are necessarily driven by the myriad ways in which researchers need to subsequently access and exploit the information. Institutional repositories, data centres on domain and national levels, international data grids, and all other methods of storing and sharing data have to exist within an infrastructure that enables researchers to access and exploit the data, and variant models for this infrastructure can be conceptualised. The JISC report *Dealing with Data*[163] discussed two high-level data flow models:

- The Domain Data Deposit Model, that is based on a strong integrated community foundation with well-established common standards, policies and practice. It is supported by clear research funder policy, advocacy and investment. The data centre provides training to the community and proactively promotes its services.

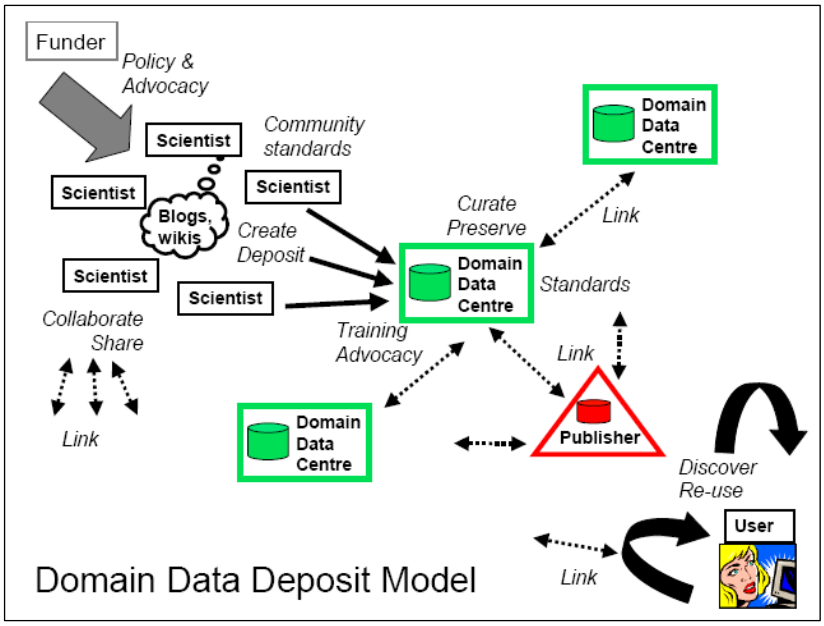

Figure 2. The Domain Data Deposit Model (*Dealing with Data*, 2007, p. 58)

[162] Gary King, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing (2007)
[163] Liz Lyon, *Dealing with Data* (2007), p. 57

- The Federation Data Deposit Model, where groups of repositories have joined together in a federation, which is based on some agreed level of commonality documented in some form of partner agreement, but where there is a broader practice base. Policy, advocacy and training are provided primarily within the federation. The federation is supported by strong institutional buy-in, and there may be little or no investment by research funding organisations. There will be some common technical standards, but the maturity of these may be variable.

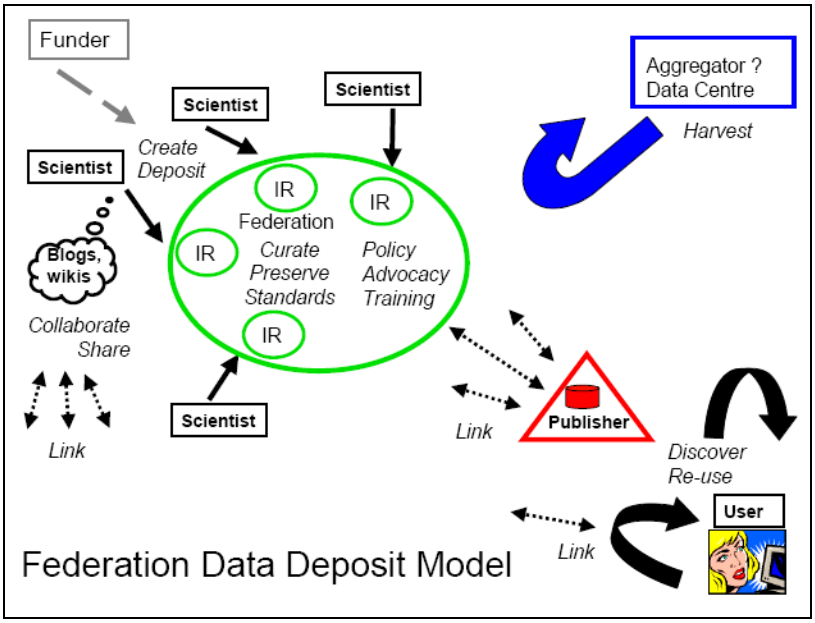

Figure 3. The Federation Data Deposit Model (Liz Lyon, *Dealing with Data, 2007*, p. 59)

The high availability of social networking software such as blogs, wikis, and other web data publication opportunities have been taken up by (predominantly) young scientists who are depositing data straight onto web-based resources. This method has been called the 'Open Science Model'.

In addition, there are many so called 'big science' projects worldwide that generate terabytes of data per year or even per month. These volumes of research data have their own storage requirements, because they cannot be managed in a local or web environment to support the required dynamic collaboration work and complex analyses of the data. Grid infrastructures that can host large and complex datasets and include the computing resources to manage and analyse data are the main environment for hosting and sharing the 'big science' data. Data grids are in essence distributed federated systems of heterogeneous IT environments that appear as a single system to end users. Large (public) investments have been made to set up grid infrastructure for several research communities. National e-science grid services are also available to host big datasets that are too large for storing at local institutions. Multi-disciplinary data can usually be deposited in these grids for free, but with some peer review process to assess the value of the data resource. Grid-based data sharing should be considered as a further model for sharing (e-science) data.

Discussion of effective infrastructures for curating research data is taking place at all levels wherever research is reliant on the long-term stewardship of digital material. Different national agendas are informing variant infrastructure models, but there is also significant variance in different domains of science and research:[164]

> Data represents the front line of [US] cyberinfrastructure development: its main site of operation, its most tangible output, and in some regards the target of its highest ambitions. From this perspective, cyberinfrastructure is principally about data: how to get it, how to share it, how to store it, and how to leverage it into the major downstream products (knowledge, discoveries, learning, applications, etc.) we expect research and sciences to produce. At the same time, there is significant variation (both within and across disciplines) as to what counts as data. For some, data is first and foremost a question of things: samples, specimens, collections. For others, data is what comes out of a model — or perhaps the model itself. Data may be tactile, visual, textual, numeric, tabular, classificatory, or statistical. Data may be an intermediate

[164] Steven Jackson, Paul Edwards, Geoffrey Bowker, Cory Knobel, *Understanding Infrastructure: History, Heuristics and Cyberinfrastructure policy* (2007)

> outcome, a step on the road to higher-order products of science (publications, patents, etc.). Or data may be the product itself. Where a discipline or research project fits within this spectrum will have enormous consequences for its positioning vis-à-vis cyberinfrastructure. This specificity alone guarantees that cyberinfrastructure should and assuredly never will be a singular or unified thing.

This chapter looks at the governmental investments that are being made into developing e-research and data sharing infrastructures; compares the existing services for data sharing and for data discovery; looks at emerging tools that support access to data as part of the research process and finally discusses the development of the skill-base of researchers to make use of the data sharing services.

3.1 GOVERNMENT FUNDING FOR E-RESEARCH INFRASTRUCTURE BUILDING

Infrastructure provision for data sharing carries a significant cost. Most of the funding for this has come from national governments who continue to support directly the (centralised) services and participate in funding research domain level and institutional services through funding research. With new models for sharing research data appearing, the question of whose funds could or should be used for developing and maintaining the services, and the discussion of revenue streams to cover the costs, is frequently raised. The cost of data sharing is also vital for budgetary planning purposes on all levels – from national governments to research projects. Yet real cost figures are hard to obtain as data sharing is 'bundled' with other services, most often with archiving and curating data. A separate JISC project has reported on modelling the costs and benefits of data sharing.[165]

The OECD publishes statistics on its own member states and the total OECD investment in research and development climbed to USD 818 billion in 2006, up from USD 468 billion in 1996. Gross domestic expenditure on R&D grew by 4.6% annually (in real terms) between 1996 and 2001, but growth slowed to less than 2.5% a year between 2001 and 2006.[166] The European Union has set a target to increase research spending to 3% of GDP by 2010.

The European Commission funding for digital preservation and digital repository infrastructures has been on the rise and currently supports linking of digital repositories (2007-08: €50 million), research on digital preservation (2007-08: €25 million) and research on accessibility and usability (2005-08: €10 million).

The UK government science research budget in 2005-6 allocated funds for research areas as follows: Arts and Humanities £68m; Biotechnology and Biological Sciences £321m; Central Laboratories of the Research Councils £301m; Engineering and Physical Sciences £575m; Economic and Social Sciences £126m; Medicine £503m; Natural Environment £370m; Particle Physics and Astronomy £342m; other UK Government departments £701m. Funding for preservation and curation across the research councils varies according to needs and scale of the disciplines involved. However from the above it can be seen that where data centres and services exist they represent approx 1.4-1.5% of total research expenditure (excluding indirect overheads and full economic costs).[167] A recent JISC commissioned report[168] noted that researchers are reportedly still railing against research councils' policies that require to share data, but these requirements are not backed up by enabling budgets and there is a view of a gap between what would be good to do for data sharing and what is actually achievable with the available resources.

In the United States the National Science and Technology Council has identified long-term preservation and the maintenance of and access to long-lived science and engineering data collections and federal records as one of five strategic priorities for R&D by federal agencies in the 2007 Presidential budget. The NSF's Office for Cyberinfrastructure has been created with a budget of USD 127 million, potentially rising to USD 182.42 million per annum. On a mandate from the US Congress the Library of Congress has developed the National Digital Information Infrastructure and Preservation Program (NDIIPP) which received approximately USD 100 million. The US National Archives and Records Administration (NARA) has awarded a consortium led by Lockheed Martin a USD 308 million contract to build a digital archives system for electronic records created by the federal government.

---

[165] Neil Beagrie, Julia Chruszcz, Brian Lavoie, *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities* (2008)
[166] *OECD Science, Technology and Industry Outlook 2008: Highlights* (2008)
[167] Cf. Neil Beagrie, *E-Infrastructure Strategy for Research: Final Report From the OSI Preservation and Curation Working Group* (2007)
[168] Liz Lyon, *Dealing with Data* (2007)

In Australia the government's national strategy *Backing Australia's Ability – Building our Future through Science and Innovation* committed Au$ 8.3 billion funding to develop the country's science and innovation base over a 10 year period 2001 to 2010. The Australian National Data Service (ANDS) has received funding of Au$ 24 million over 4 years.

In Sweden the Swedish National Data Service has a budget of SEK 6 million for 2008, planned to increase to SEK 12 million by the year 2012.

These are but a few examples of the significant funding made available for research data infrastructures in different countries. The next sections explore the services that have been created with this funding.

3.2 DATA CURATION AND SHARING SERVICES

Chapter 2 above discussed the policies on different levels that mandate and guide the sharing of research data. Many funding agencies specify which repositories the data resulting from research they fund should be deposited with; some only state the principle of sharing, leaving the choice of method and channel for sharing up to the researchers. Requirements to maintain data underlying an article or publications are also being set by publishers and universities' institutional repositories. The US National Institutes of Health guidance concedes that:[169]

> The method for sharing that an investigator selects is likely to depend on several factors, including the sensitivity of the data, the size and complexity of the dataset, and the volume of requests anticipated. Investigators sharing under their own auspices may simply mail a CD with the data to the requestor, or post the data on their institutional or personal Website. Although not a condition for data access, some investigators sharing under their own auspices may form collaborations with other investigators seeking their data in order to pursue research of mutual interest. Others may simply share the data by transferring them to a data archives [...] that can be particularly attractive for investigators concerned about a large volume of requests, vetting frivolous or inappropriate requests, or providing technical assistance for users seeking help with analyses.

Evidently one method of data sharing does not preclude the use of other ones, and ultimately it is the researcher who decides which (additional) channels to use for dissemination. The knowledge of existing data sharing services, their nature, scope of their collections, target user groups, security and trustworthiness will all play a role in making this decision. The sub-chapters below will attempt a comparison of data sharing services on some of these criteria.

3.2.1 TRANS-NATIONAL INITIATIVES

The European Commission (EC) has positioned itself as both a policy-making body, and a research funding body that provides funds for digital infrastructure and relevant research and networking activities. The EC supports the development of scientific digital repositories, science data infrastructures and the e-science grid infrastructure.

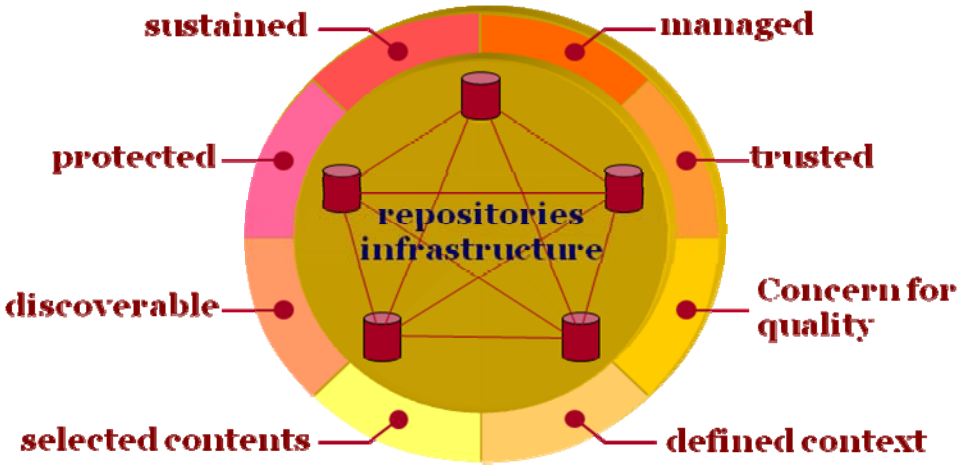

Figure 4. Qualities of data repositories (EC, 2008).

[169] *NIH Data Sharing Policy and Implementation Guidance* (2003)

The Enabling Grids for E-science (EGEE)[170] has been supported with funding from the European Commission since 2001 to develop the European multi-science Grid. The EGEE brings together more than 140 institutions to produce a reliable and scalable computing resource available to the European and global research community. At present, it consists of approximately 300 sites in 50 countries and gives its 10,000 users access to 80,000 CPU cores around the clock. The EGEE services are being used by ca 150 projects in many disciplines,[171] including astronomy and astrophysics, computational chemistry, earth sciences, fusion, high energy physics, life sciences and medical research. Regional and national e-science grid projects are being formed and linked to the EGEE.

The European Commission has co-funded two periods of the DRIVER project[172] as one of the main generic data repository developments for the European research infrastructure development. The primary objective of the project is to create a cohesive, robust and flexible, pan-European infrastructure for digital repositories, offering sophisticated services and functionalities for researchers, administrators and the general public. The support for federated repositories and content harvesting from repositories is provided through the D-Net software.[173] The DRIVER-II project will transform the initial testbed into a fully functional, state-of-the-art service, extending the network to a larger confederation of repositories.

The European Union's Community Research and Development Information Service (CORDIS) has set up a European Strategy Forum on Research Infrastructures (ESFRI) that works on policy issues and a roadmap for developing European research infrastructure. The ESFRI conducted an international peer-review and identified 35 projects for large scale research infrastructures.[174] Several projects in this list are concerned with developing services for data sharing.

The DARIAH project[175] will develop services for facilitating long-term preservation of and access to arts and humanities research data. Among these services is development of integrated middleware for supporting digital access to arts and humanities data and services in Europe. The architecture of this solution is focusing on web and grid services as the foundation for an open semantic service-oriented architecture that allows for distributed preservation services.

A multitude of other EU funded projects are developing subject-specific data repositories and services:

- The NMDB project[176] establishes a digital repository for cosmic-ray data, and develops a real-time database from many neutron monitoring stations.
- The EuroVO-AIDA project[177] unifies digital data collections of astronomy, integrating European data centres into a global Virtual Observatory.
- The IMPACT project[178] unifies data from 10 major databases related to protein families.

PARSE.Insight[179] is a two-year project co-funded by the European Commission under the Seventh Framework Programme and is closely linked to the Alliance for Permanent Access to the Records of Science.[180] The project is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research. The project is developing a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. As a first step, the project is conducting surveys to establish the awareness and needs of different scientific communities for digital preservation. Several online surveys have been launched, including one aimed at researchers in different disciplines; one for publishers; one for research funders; and one for those who manage and curate data. Additional surveys are being conducted in specific disciplines, including high-energy physics, to allow a detailed understanding of their particular perspectives. The surveys shed light not only on the needs for preservation, but also on access and reuse of data. The first survey results are expected to be published in early 2009.

[170] http://www.eu-egee.org/
[171] For a list of projects see: *Third EGEE User Forum* (2008)
[172] http://www.driver-community.eu/
[173] http://www.driver-repository.eu/System-Architecture.html
[174] http://cordis.europa.eu/esfri/large-scale.htm
[175] http://www.dariah.eu/
[176] http://record.oulu.fi/nmdbinfo/
[177] http://www.euro-vo.org/pub/aida/overview.html
[178] http://www.ucc.ie/impact/
[179] http://www.parse-insight.eu/
[180] http://www.alliancepermanentaccess.eu/

3.2.2 NATIONAL LEVEL INITIATIVES

Several countries have initiated national level projects to develop research data repositories and services around them. These are intended to be used by all disciplines and institutions.

In the **UK** a feasibility study for providing national level repository infrastructure for research data is ongoing. The Higher Education Funding Council for England (HEFCE) has funded through its Shared Services programme, and with support from JISC, a UK Research Data Service feasibility study (UKRDS).[181] The UKRDS study is a joint project between RLUK (the Consortium of Research Libraries in the UK and Ireland), and RUGIT (the Russell Group IT Directors Group). The objective of the UKRDS study is to assess the feasibility and costs of developing and maintaining a national shared digital research data service for the UK Higher Education sector. The project has delivered an interim report[182] where preliminary results from a survey it conducted are presented and three scenarios discussed for possible future development. The 'Hybrid/Umbrella organisation' scenario has been selected for further analysis, whereas the 'No Change' and 'Centralised repository' scenarios have been shown to be defective or not acceptable for some stakeholders. The 'umbrella organisation' should be a combination of regional and specialist repositories. Such an organisation would be well-placed to act as a mediator, as a standards-setting body and as a source of information about data archiving and repositories. In time it might become a data repository in its own right or take on other functions as required. This would introduce the shared services model into the environment of grid computing and cloud-based data storage, with an emphasis on distributed shared services, rather than centralised shared services. The UKRDS project has analysed thoroughly the involved stakeholders and their interests and will now proceed to define the organisational roles and responsibilities within the selected scenario.

Large repositories have or are being developed also by a few national bodies in the UK such as the National Archives, the British Library, the Ordnance Survey and the University of London Computing Centre.

In the **United States** a national digital data framework is being developed as part of the cyberinfrastructure. It will be an integral component in the national cyberinfrastructure framework that consists of a range of data collections and managing organisations, networked together in a flexible technical architecture using standard, open protocols and interfaces, and designed to contribute to the emerging global information commons. It will be simultaneously local, regional, national and global in nature, and will evolve as science and engineering research and education needs change and as new science and engineering opportunities arise. To achieve this goal the NSF, through coordination by its Cyberinfrastructure Office, will catalyse the development of a system of science and engineering data collections that is open, extensible, and evolvable; and will support development of a new generation of tools and services for data mining, data discovery, integration, visualisation, analysis and preservation.[183]

The NSF has envisaged that from a technological perspective, the national data framework must provide for reliable preservation, access, analysis, interoperability, and data movement, possibly using a web or grid services distributed environment. The architecture must use standard open protocols and interfaces to enable the broadest use by multiple communities. It must facilitate user access, analysis and visualization of data, addressing issues such as authentication, authorization and other security concerns, and data acquisition, mining, integration, analysis and visualization. It must also support complex workflows enabling data discovery. Such an architecture can be visualized as a number of layers providing different capabilities to the user, including data management, analysis, collaboration tools, and community portals. The connections among these layers must be transparent to the end user, and services must be available as modular units responsive to individual or community needs. The system is likely to be implemented as a series of distributed applications and operations supported by a number of organisations distributed throughout the country.

NSF's Office of Cyberinfrastructure has initiated a Sustainable Digital Data Preservation and Access Network Partners (DataNet) project. The DataNet seeks to foster the development of new types of organisations that integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise. The goal of the DataNet is to provide reliable digital preservation, access, integration and analysis capabilities for science and engineering over a decades-long timeline. Up to USD 1 million plus

[181] http://www.ukrds.ac.uk/
[182] UKRDS, *UKRDS Interim Report* (2008)
[183] NSF, *Cyberinfrastructure Vision for 21st Century Discovery* (2007), pp. 2, 24-25

indirect costs is available in this programme over a five-year period, with the possibility of a five-year renewal. The DataNet programme is currently selecting its partner organisations and a first project – TeraGrid[184] – has received its funding through this scheme.

In **Australia** a project to create the Australian National Data Service (ANDS) has been started as a reaction to the needs raised by the PMSEIC Data for Science Working Group, which discussed the idea of a new National Centre for Data for Science.[185] A detailed scoping report was produced for developing the ANDS that presented both a rationale for such a service and its components with estimated budgets and timelines.[186] The ANDS vision is that the disparate collections of research data around Australia can be transformed into a cohesive corpus of research resources. ANDS should provide common services in support of the corpus of research data collections and provide integration infrastructure that facilitates sharing and reuse of data, so that researchers can more easily discover, access, use, analyse, and combine digital resources as part of their activities. The ANDS should also support and advise researchers and research data managers in appropriate digital preservation strategies. The practical activities planned were:

- sustain the community of interest needed to build a federated solution to research data management and to achieve the necessary community agreements;
- develop frameworks for bridging from a research data federation to other federated data management communities;
- identify and provide the national operational services on which data federation depends; and
- assist existing data federating research communities to successfully migrate into this more systemic federated framework.

The ANDS will not itself host research data and is dependent on research outputs stored in institutionally-supported repositories or discipline-based service providers.

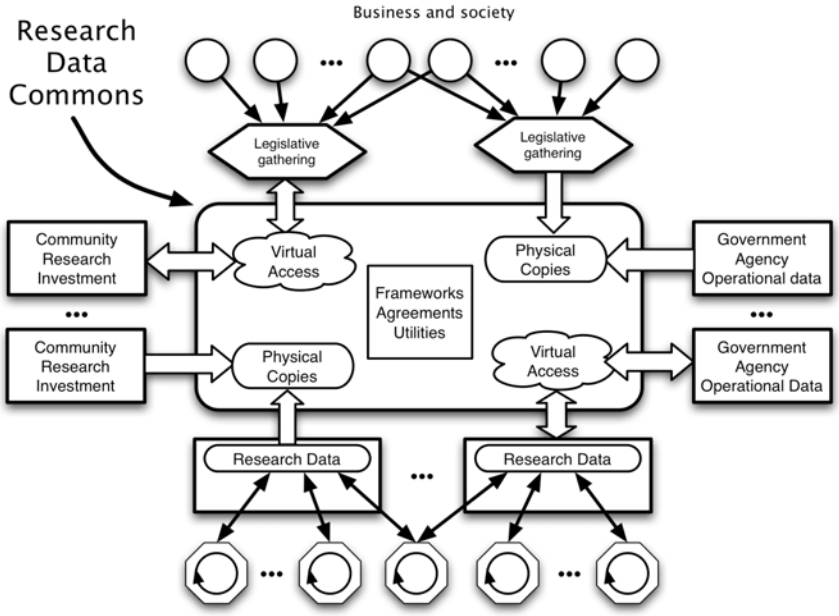

Figure 5. The ANDS solution architecture (*Towards the Australian Data Commons*, 2007)

The ANDS report proposed four large-scale programmes for the project that together encompass the entire data management lifecycle:

- Frameworks: Aimed to establish agreements around policies and responsibilities that will allow a cohesive network of research repositories to be established and to grow into an Australian 'data commons'.
- Utilities: ANDS will deliver a set of utility services that facilitate discovery of, and access to, research data held in Australian repositories.

[184] http://www.teragrid.org/
[185] Working Group on Data for Science, *From Data to Wisdom: Pathways to Successful Data Management for Australian Science. Report to PMSEIC* (2006)
[186] ANDS Technical Working Group, *Towards the Australian Data Commons: A Proposal for an Australian National Data Service* (2007)

- Repositories: To achieve the aim that Australian research data is routinely deposited into stable and sustainable data management and preservation environments, ANDS will improve and/or supplement the available data management options at institutions, noting that ANDS is focussed on data federation and is not established to fund data management itself.
- Researcher Practice: ANDS will help Australian researchers and research data managers to develop the necessary skills to create, manage, and share high-quality research data.

Subsequently, ANDS has found the last two programmes too general in practice and have replaced them with a single programme 'Feeding the Data Commons', which is now the priority of the project. The ANDS project board has also found the initial start-up and challenge of addressing research data issues at a national level to be quite daunting and to facilitate the realisation of the plans have contracted the implementation of the ANDS Establishment Project to Monash University.

The **Canadian** Social Sciences and Humanities Research Council conducted an analysis for a National Data Archive in 2002[187] that favoured a National Research Data Archive Network over a centralised operating agency within the National Archives of Canada.

> All reviewed proposals for a national data service have opted for a distributed, umbrella-type approach where the national service provides the environment for repositories – common principles and standards that data repositories in the country apply, and the development of tools that facilitate interaction between repositories. The main expected outcomes are better data curation and dissemination services that are based on shared tools and principles.

3.2.3 RESEARCH DOMAIN LEVEL INITIATIVES

Research funding agencies have established centralised data repositories and services to cater for research projects they fund, or are supporting repositories in institutions in their domain. Several research disciplines have self-organised networks of data repositories and data sharing services, for example social science data archives have a history of over 40 years. The number of domain-specific repositories in the OECD countries is vast and they cannot all be reviewed in the scope of this report. The overview below includes some examples that represent the best practice of domain data repositories (both in Table 6 and the following description below).

[187] SSHRC, *National Data Archive Consultation. Building Infrastructure for Access to and Preservation of Research Data* (2002)

Table 6. Examples of research domain level data sharing infrastructure solutions

Name	Data Domains	Scale of Services	Sponsors	Year Founded	Data Collection Principles	Licence to Share Data	Accessibility	Added Value
United Kingdom								
Antarctic Environmental Data Centre (AEDC)[188]	Data collected in Antarctica and the Southern Ocean	Domain / National	NERC, British Antarctic Survey	n/a	Data collected by UK funded scientists in Antarctica and the Southern Ocean	Mandated by the funding agency	On permission from the data owner	n/a
Archaeology Data Service (ADS)[189]	Archaeology	Domain / National	AHRC, University of York	1996	Mandatory for AHRC funded projects. Recomm-ended deposit by: British Academy, Carnegie Trust, Council for British Archaeology, ESRC, Leverhulme Trust, NERC, Wellcome Trust's History of Medicine Programme	Depositors are asked to specify that the ADS may have a non-exclusive license to distribute their datasets	Open after accepting a licence agreement on-line	Guides to good practice learning and teaching programme
British Atmospheric Data Centre (BADC)[190]	Atmospheric data and associated data	Domain / National	NERC	1985	Mandatory for NERC funded projects	An individual by signing the NERC grant award allows NERC the right to use and share the data but it remains their intellectual property	Both open access and with user registration, complying with licence agreement, and sometimes permission from data owner	Community-based value-added services and tools.
European Bio-informatics Institute (EMBL-EBI)[191]	DNA sequences, protein sequences, microarray expression data, macro-molecular structures, protein interactions and biological pathways	Domain / International	European Molecular Biology Laboratory, Wellcome Trust, EU, UK Research Councils, NIH	1995	BBSCR funded projects submitted by researchers	n/a	Open	Data submission tools Guidance on data use and tools for data analysis User training

[188] http://www.antarctica.ac.uk/about_bas/our_organisation/eid/aedc.php
[189] http://ads.ahds.ac.uk/
[190] http://badc.nerc.ac.uk/home/index.html
[191] http://www.ebi.ac.uk/

Name	Data Domains	Scale of Services	Sponsors	Year Founded	Data Collection Principles	Licence to Share Data	Accessibility	Added Value
UK Data Archive (UKDA)[192]	Social science, economics, politics, sociology, history, geography, public records of government	Domain / National	ESRC, JISC, University of Essex	1967	Mandatory for ESRC funded projects TNA deposited records Active collection policy	The Deposit Licence assigns to the University of Essex the right to disseminate and publish the data	Requires registration and complying with licence agreement	Guidance on data management. Participation in content federation projects. Analytical on-line tools for accessing data
NERC Earth Observation Data Acquisition and Analysis Service (NEODAAS)[193]	Remote sensing data from polar-orbiting sensors and geostationary satellites	Domain / National	NERC, Plymouth Marine Laboratory, University of Dundee	1981 / 1995	Direct satellite data acquisition	Mandated by the funding agency	Limited free access; Low resolution with simple web registration; High resolution with full registration	Data analysis and processing
EDINA[194]	Geospatial and other data	Cross-domain / National	JISC, University of Edinburgh	1996	Individual data services	Mandated by the individual data services	Free at the point of use for staff and students in learning, teaching and research through institutional subscription	Services and software for building and running online services
NERC Environmental Bioinformatics Centre (NEBC)[195]	Sequence data for genomes, genome and proteome data	Domain / National	NERC	2002	Mandatory for NERC funded projects in the subject area	Mandated by the funding agency	Public after an exclusive use period but requires permission from the data owner	Data analysis and processing
National Geoscience Data Centre (NGDC)[196]	Geological and environmental information	Domain / National	NERC	n/a	Mandatory for NERC funded projects	Copyright is vested with the clients that commissioned the data collection work	On permission from the data owner	Guidance to data owners

[192] http://www.data-archive.ac.uk/
[193] http://www.neodaas.ac.uk/index.php
[194] http://edina.ac.uk/
[195] http://nebc.nox.ac.uk/
[196] http://www.bgs.ac.uk/services/ngdc/

Name	Data Domains	Scale of Services	Sponsors	Year Founded	Data Collection Principles	Licence to Share Data	Accessibility	Added Value
Manchester Information Datasets and Associated Services (MIMAS)[197]	Census, satellite image, chemical and other data	Cross-domain / National	JISC, ESRC, University of Manchester	1994	Individual data services	Mandated by the individual data services	Staff, students, researchers, teachers or instructors at UK Higher and Further Education institutions free of charge as end-users of authorised institutions	Federated access
British Oceanographic Data Centre (BODC)[198]	Marine data	Domain / National	NERC, Proudman Oceanographic Laboratory	1969 / 1989	Mandatory for NERC funded projects	Mandated by the funding agency	Open after accepting a licence agreement on-line	Data and software services
United States								
Inter-University Consortium for Political and Social Research ICPSR[199]	Social science data	Domain / National and International	University of Michigan	1962	Funding agency mandates, serial collection, active collections policy researcher deposits,	With the Deposit Licence the owner grants ICPSR the right to re-disseminate copies of the data	Open access. Some data open to users from member organisations only	Training and guidance Qualitative research Data federation services
National Center for Biotechno-logy Information NCBI[200]	Biomedical information	Domain / National	NIH, NLM	1988	Funding agency mandates	n/a	Open	Data federation services Research User training courses
The Henry A. Murray Research Archive[201]	Quantitative social science	Domain / International	Ford Foundation, Harvard University	1976	On a voluntary basis	Voluntary	Free	Part of the DataVerse Network
Roper Center for Public Opinion Research[202]	Social science data, survey data	Domain / International	Funded by the membership fees and fees for non-members for acquiring data	1947	Collecting from commercial and media survey organisations	n/a	Open to members, non-members have access according to the selected package of services	User training

[197] http://www.mimas.ac.uk/
[198] http://www.bodc.ac.uk/
[199] http://www.icpsr.umich.edu/
[200] http://www.ncbi.nlm.nih.gov/
[201] http://www.murray.harvard.edu/
[202] http://www.ropercenter.uconn.edu/about_roper.html

Name	Data Domains	Scale of Services	Sponsors	Year Founded	Data Collection Principles	Licence to Share Data	Accessibility	Added Value
National Archive of Criminal Justice Data[203]	Criminal justice data	Domain / National	Department of Justice, John D. and Catherine T. MacArthur Foundation	1978	On a voluntary basis	Copyright is vested with the clients that commissioned the data collection work	Open. Restricted Access Data Archive contains sensitive data and has special procedure for acquiring data	Educational activities
Canada								
Geoscience Data Repository (GDR)[204]	Geoscience	Domain / National	Natural Resources Canada	2003	Produced and/or compiled by Natural Resources Canada	Copyright of Her Majesty the Queen in Right of Canada.	Free for non-commercial purposes	n/a
Germany								
Scientific Drilling Database (SDDB)[205]	Scientific drilling data	Domain / International	GeoForschungs Zentrum, ICDP, Helmholtz Association	n/a	Data resulting from ICDP projects	Creative Commons license	Open access for research purposes	n/a
Zentrum für Psychologische Information und Dokumentation (ZPID)[206]	Psychology	Domain / National	DFG, University of Trier	2002	Voluntary	Based on the deposit agreement	Depends on the use agreement signed for each dataset	Content harvested via OAI-PMH
Netherlands								
Data Archiving and Networked Services DANS[207]	History Archaeology Socio-cultural Sciences Social Sciences Statistics Netherlands Data	Domain / National	KNAW, NWO	2005	Both individual researchers and research institutes may deposit their data sets with DANS	The Depositor grants the Repository a non-exclusive licence to make the dataset (or substantial parts thereof) available to third parties by means of on-line transmission.[208]	Requires registration and compliance with licence agreement	Semi-automated self-archiving tools for data. Participation in data federation projects. Support and guidance for data management, digitisation, feasibility studies. Data Seal of Approval set of quality criteria.

[203] http://www.icpsr.umich.edu/NACJD/
[204] http://gdr.nrcan.gc.ca/index_e.php
[205] http://www.icdp-online.org/contenido/lakedb/front_content.php
[206] http://www.zpid.de/index.php?lang=EN
[207] http://www.dans.knaw.nl/en/
[208] *DANS Licence Agreement* (2008)

Name	Data Domains	Scale of Services	Sponsors	Year Founded	Data Collection Principles	Licence to Share Data	Accessibility	Added Value
France								
Centre de Données Socio-Politiques (CDSP)[209]	Social science Political science Survey data	Domain / National	Fondation Nationale des Sciences Politiques; Centre National de la Recherche Scientifique	2005	Voluntary	The owner assigns to the CDSP the right to disseminate the data	Requires signing an on-line user licence	Data management guidance to researchers International federation of data
Spain								
Centro de Investigaciones Sociológicas (CIS)[210]	Survey data	Domain / National	Government	1977	Created by the CIS	CIS owns the data	Open	Provides support for training and research in the area of social sciences
Denmark								
Danish Data Archive (DDA)[211]	Social sciences Health Historical population	Domain / National	National Archives	1973	Voluntary	The owner assigns to the DDA the right to disseminate the data	Open and on permission from the data owner for research and statistical purposes	Advice on data documentation practices
Sweden								
Swedish National Data Service (SND)[212]	Social sciences Humanities Medicine	Cross-domain / National	DISC	2006	Mandated by the funding agency	The owner assigns to the SND the right to disseminate the data	Free for students and researchers from academic institutions	Professional advice on documenting and archiving data
Norway								
Norwegian Social Science Data Services (NSD)[213]	Social sciences	Domain / National	Research Council of Norway, Ministry of Education and Research	1971	Mandated by the funding agency	The owner assigns to the NSD the right to disseminate the data	Free Requires registration	On-line data analysis tools Data protection ombudsman Guidance on data management

[209] http://cdsp.sciences-po.fr/index.php?&idRubrique=cdsp&lang=ANG
[210] http://www.cis.es/cis/opencms/EN/index.html
[211] http://www.sa.dk/content/us/about_us/danish_data_archive/
[212] http://www.snd.gu.se/
[213] http:/www.nsd.no/

Archiving and sharing of electronic research data has a long history in the **United Kingdom** – the UK Data Archive, for example, was first established as the SSRC Data Bank in 1967.[214] In the 1970s the requirement to archive the data created by research projects came to be connected with ESRC's funding for research. Until the emergence of Internet-based data sharing tools, the model where research data was deposited with an academic data centre as a requirement by the research funders or as best practice was working relatively successfully. In contrast in the government sector the National Digital Archive of Datasets[215] was first set up only in 1998 and has had to make a considerable investment into making accessible again older datasets going back to the 1960s. While there has been a conscious centralised development of national level policies, standards and infrastructure for e-government initiatives to cope with the Internet era challenges, the research domain in the UK has largely continued to rely on the established framework of research councils' services for data curation and sharing.

The data centres of UK research councils (see Table 6 above) have been surveyed and analysed in a number of recent reports.[216] Findings of these studies will not be repeated here; only significant examples of data sharing services are discussed.

The UK Data Archive (UKDA) is a centre of expertise in data acquisition, preservation, dissemination and promotion and is curator of the largest collection of digital data in the social sciences and humanities in the UK. It is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and the University of Essex. It houses several thousand datasets of interest to researchers in all sectors and from many different disciplines. It provides resource discovery and support for secondary use of quantitative and qualitative data in research, teaching and learning. As a lead partner of the Economic and Social Data Service (ESDS), the UKDA is working with MIMAS and the Cathie Marsh Centre for Census and Survey Research (CCSR)) at the University of Manchester. The UKDA also provides preservation services for other data organisations, supports the National Centre for e-Social Science (NCeSS)) and facilitates international data exchange through agreements with other national archives. The UKDA hosts the History Data Service and Census.ac.uk, facilitating access to the census data resources for UK higher and further education.

The UKDA together with ESRC is moving towards a life-cycle oriented approach where ESRC notifies UKDA within the first 3 months of award of grant to enable evaluation of the potential of long term preservation of the outputs. The award holder would then be contacted early on in the process before data has been collected, to advise on the preservation schedule and to assist in identifying and overcoming issues of consent, ethics, confidentiality, copyright, formats and metadata, so that these issues can be resolved before they become a problem.

A network of seven data centres provides support and guidance in data management to the researchers funded by NERC. These centres are responsible for the long-term curation of data and provide access to NERC's data holdings. The centres provide focus for the following disciplines: Atmospheric science; Earth sciences; Earth observation; Marine science; Polar science; Science-based archaeology; Terrestrial & freshwater science; Hydrology; Bioinformatics - data management for environmental genomics research.

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

The Arts and Humanities Research Council and the Biotechnology and Biological Sciences Research Council (BBSRC) take a devolved approach to data repositories. The Medical Research Council (MRC) has no data centres. The AHRC who had a long track record in centralised data centres has now changed its stance and no longer funds the Arts and Humanities Data Services. Its position now is:

> Council believes that long term storage of digital materials and sustainability is best dealt with by an active engagement with HEIs rather than through a centralised service.

[214] http://www.data-archive.ac.uk/ukda40/about/origins.asp
[215] http://www.ndad.nationalarchives.gov.uk/
[216] Rachel Heery, Andy Powell, *Digital Repositories Roadmap: Looking Forward* (2006); Liz Lyon, *Dealing with Data* (2007)

The International Council for Science (ICSU) created a network of **scientific World Data Centres** (WDC)[217] during the International Geophysical Year of 1957-1958. Originally consisting of 27 centres in the US, Russia, Europe and Japan, there are now 50 WDCs in 12 countries and their holdings include a range of solar, geophysical, environmental and human dimensions data. Individual data centres are hosted and funded nationally in a variety of different institutions and they serve mainly a national mandate, collecting data in a specific domain from research institutions in their country. A secondary function is to be part of the WDC network and provide information on data in other WDCs. All data held in WDCs are available on a full and open access basis, either directly online or for no more than the cost of copying and sending the requested information. Of the 50 WDCs, nine are in China, one in India and all the others are in OECD countries:[218]

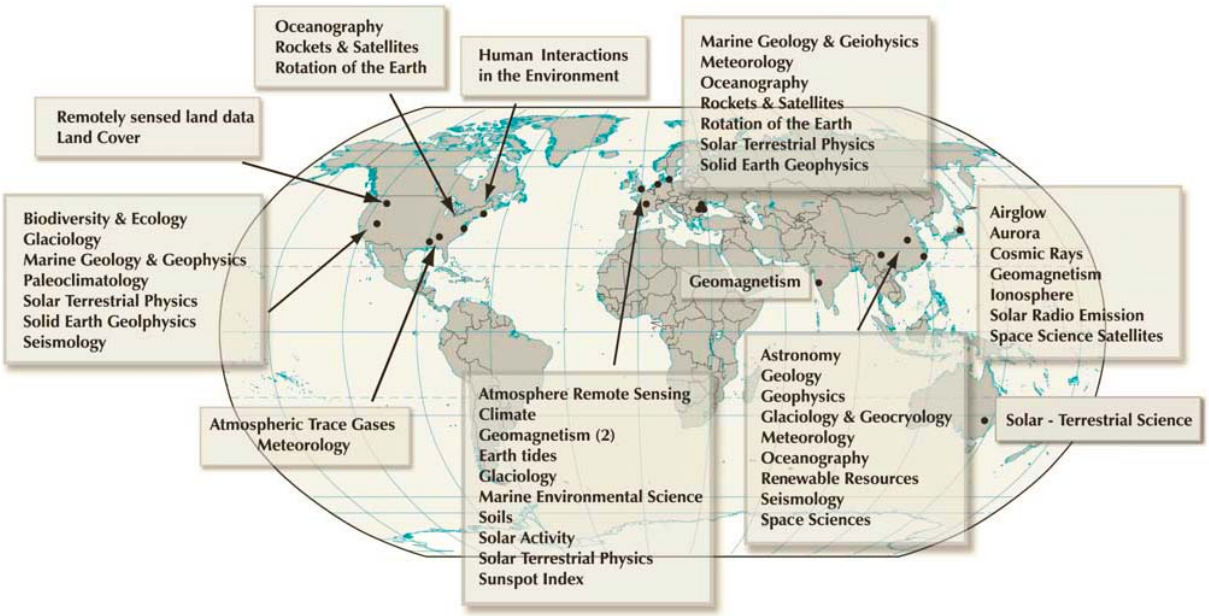

Figure 6. Network of ICSU science World Data Centres.

Examples of significant domain-specific data repositories among the WDCs include:

France:

- Bass2000 Solar Survey Archive at the Observatoire de Paris[219]

Germany:

- Global Geodynamic Project Information System and Data Center at the Helmholtz-Zentrum Potsdam[220]
- WDC for Climate at the Max-Planck-Institute for Meteorology[221]
- WDC for Marine Environmental Sciences, maintained jointly by the Alfred Wegener Institute for Polar and Marine Research, a centre of the Helmholtz Association, and the Center for Marine Environmental Sciences at the University of Bremen[222]
- WDC for Remote Sensing of the Atmosphere at the German Aerospace Center[223]

Japan:

- Data Analysis Center for Geomagnetism and Space Magnetism at Kyoto University[224]
- Data Archives and Transmission System is an archive of the data acquired by Japanese scientific satellites and spacecraft, operated by the Japanese Aerospace Exploration Agency[225]

[217] http://www.wdc.rl.ac.uk/wdcmain/
[218] ICSU Ad hoc Strategic Committee on Information and Data, *Final Report to the ICSU Committee on Scientific Planning and Review* (2008), p. 27
[219] http://bass2000.obspm.fr/home.php
[220] http://ggp.gfz-potsdam.de/
[221] http://www.mad.zmaw.de/wdc-for-climate/
[222] http://www.wdc-mare.org/
[223] http://wdc.dlr.de/
[224] http://swdcwww.kugi.kyoto-u.ac.jp/index.html
[225] http://www.darts.isas.jaxa.jp/

Netherlands:

- ISRIC World Soil Information is funded by the Netherlands Government and has a strategic association with Wageningen University and Research Centre[226]

UK:

- WDC for Glaciology at the Scott Polar Research Institute, University of Cambridge[227]
- WDC for Geomagnetism at the British Geological Survey in Edinburgh[228]
- The UK Solar System Data Centre at the Rutherford Appleton Laboratory[229]

US:

- WDC for Human Interactions in the Environment hosted by NASA's Socioeconomic Data and Applications Center[230]
- WDC for Meteorology in Asheville, hosted by the National Climatic Data Center[231]

These data centres are mostly hosted by institutions that collect the data themselves, but all have global significance as vital data collections. Hence, the WDCs have a strong focus on developing tools for data access and analysis. In most cases the data are openly accessible, but a requirement to quote the data source is added.

The tradition of **social science data archives** also goes back over 40 years in Europe. The European data archives have united into the Council of European Social Science Data Archives (CESSDA)[232] that acts as an umbrella organisation. Since the 1970s its members have worked together to improve access to data for researchers and students. Collectively the 20 constituent CESSDA member organisations serve some 200,000+ social science and humanities researchers each year, providing access to and delivering over 70,000 data collections per annum and acquiring a further 3,500 data collections each year.

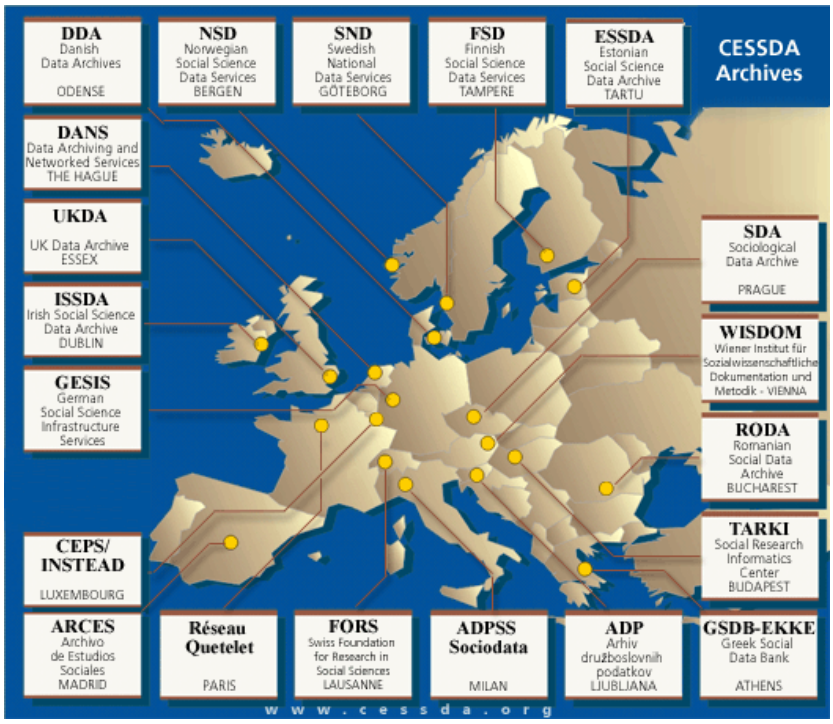

Figure 7. Map showing CESSDA member organisations.[233]

[226] http://www.isric.org/
[227] http://www.wdcgc.spri.cam.ac.uk/
[228] http://www.wdc.bgs.ac.uk/
[229] http://www.ukssdc.ac.uk/wdcc1/data_menu.html
[230] http://www.gateway.ciesin.org/wdc
[231] http://www.ncdc.noaa.gov/oa/wmo/wdcamet.html
[232] http://www.cessda.org/
[233] http://www.cessda.org/about/members/

A similar umbrella organisation that covers social science data archives from around the world is the International Federation of Data Archives for the Social Science (IFDO)[234] with 29 member organisations (10 of these are also CESSDA members). IFDO facilitates worldwide cooperation and exchange of data and social science information between national data archives.

The funding basis for social science data archives varies – individual university, funding agency, academy of science and national archives feature among the sponsors of these data repositories. The range of additional disciplines and the scope of data that are accepted to the data archives varies from country to country, sometimes including also humanities, economics, political science, and other domains. Many social science data archives also collect data directly from government organisations and the private sector. Access to data is usually provided after a user licence has been accepted. Some examples of well-established social science data archives from OECD countries include:

- ASSDA
  The social science data archives can be funded by individual universities, like the Australian Social Science Data Archive (ASSDA)[235] which was established in 1981 at the Australian National University. It now holds of over 3000 datasets not only from academic institutions but also from government and private organisations. Access to and analysis of data is provided through the Nesstar system interface.

- DDA
  The Danish Data Archive (DDA)[236] is a national data bank dedicated to the acquisition, preservation and dissemination of machine-readable data created by researchers from the social science and the health science communities. DDA also holds quantitative historical data materials and transcribed historical censuses. DDA is an independent unit within the group of Danish National Archives.

- CDSP
  The French Centre de Données Socio-Politiques (CDSP)[237] was founded in 2005 mainly for collecting survey data that its hosts, the Fondation Nationale des Sciences Politiques (FNSP) and the Centre National de la Recherche Scientifique (CNRS), collect and study. CDSP has joined with other similar data centres in France – the socio-demographic data collections of the Institut National d'Études Démographiques (INED)[238] and the micro-data archive from official statistics at the Centre Maurice Halbwachs[239] – into a Quetelet Network.[240] This is a national organisation that facilitates access for researchers to statistical data in the social sciences and offers a shared access and retrieval service to the three data collections.

- GSDB
  The Greek Social Data Bank (GSDB)[241] was founded in 1998 to support and promote social empirical research in Greece and to disseminate its results. The National Centre for Social Research (EKKE) supports the GSDB as a trusted archive for social data in Greece, covering not only the needs of EKKE but those of the entire research community. GSDB publishes information about its collections through the National Data Inventory System (SIB).

- SSJDA
  The Social Science Japan Data Archive (SSJDA)[242] is a unit within the Information Centre for Social Research on Japan at the Institute of Social Science of the University of Tokyo. Established in 1998, the SSJDA collects data from those organisations and researchers that are unable to distribute their data individually, but are willing to deposit them. The collection holds mostly micro level data and the archive is collaborating with the national Statistics Bureau.

[234] http://www.ifdo.org/network/index.html
[235] http://assda.anu.edu.au/index.html
[236] http://www.sa.dk/content/us/about_us/danish_data_archive/
[237] http://cdsp.sciences-po.fr/index.php?&idRubrique=cdsp&lang=ANG
[238] http://www-enquetes.ined.fr/index.html
[239] http://www.cmh.acsdm2.ens.fr/adisp_eng.php
[240] http://www.centre.quetelet.cnrs.fr/
[241] http://www.gsdb.gr/databank_role_en.html
[242] https://ssjda.iss.u-tokyo.ac.jp/en/index.html

- DANS
  The Dutch Data Archiving and Networked Services (DANS)[243] is an institute under the auspices of the Royal Netherlands Academy of Arts and Sciences and is also supported by the Netherlands Organisation for Scientific Research. DANS was formed in 2005 on the basis of the socio-scientific Steinmetz Archive and the former Netherlands Historic Data Archive. DANS stores and makes accessible research data in social sciences but also in the arts and humanities. DANS has developed a data depositing and archiving system EASY that is made available to all researchers. DANS also offers various support and consultancy services to researchers.

- NSD
  The Norwegian Social Science Data Services (NSD)[244] has been developed with the support of the Research Council of Norway since 1971. In 2003 NSD became a Limited Company owned by the Ministry of Education and Research but it still receives its main funding from the National Research Council. NSD data collections have been built up over the years and cover a variety of areas on individual and regional levels, on political and administrative systems, national statistics, population and demographic data, higher education statistics and data resulting from research at Norwegian institutions.
  NSD is the privacy ombudsman for research and student projects carried out at all universities, the state, specialised and private university colleges, as well as in many health enterprises and other research institutions. This means that researchers and students have to report the use of personal information to NSD who will recommend proper procedures in accordance with the Personal Data Act and the Personal Health Data Filing System Act.
  NSD has developed software for accessing and analysing their data. The Nesstar system[245] is being used by many social science data archives around the world.

- CIS
  The Centro de Investigaciones Sociológicas (CIS) is an autonomous state agency attached to the Office of the President, whose purpose is to study Spanish society, primarily through survey-based research. CIS holds survey results data that it has collected, but also provides access to other datasets through its Archivo de Estudios Sociales (ARCES) service where data created by other institutions are deposited and CIS is given the licence to preserve and disseminate the data.[246]

- SND
  The Swedish National Data Service (SND)[247] is a new service formed on the basis of the current Swedish Social Science Data Service, and will be hosted at the University of Gothenburg with funding from the national Database Infrastructure Committee. It is to be a data archive for social science and epidemiological research datasets. It will also develop a common information portal on all data holdings in Sweden to facilitate access to data in a distributed database environment. As a national resource for the entire research community it will cooperate closely with the national statistical authorities.

- ICPSR
  The US Inter-University Consortium for Political and Social Research (ICPSR)[248] – a membership-based organisation with over 500 member colleges and universities around the world – maintains and provides access to an archive of social science data. ICPSR tailors itself as a content management organisation, curating social science data and providing user support services. ICPSR is acting as a data archiving service provider to several funding agencies, and is providing its infrastructure and data analysis tools for several data collections or databases that otherwise maintain their own brand, for example:

[243] http://www.dans.knaw.nl/en/over_dans/
[244] http://www.nsd.uib.no/nsd/english/index.html
[245] http://www.nesstar.com/
[246] http://www.cis.es/cis/opencms/EN/6_arces/donaciones.html
[247] http://www.snd.gu.se/
[248] http://www.icpsr.umich.edu/

- The National Archive of Criminal Justice Data (NACJD)[249] where projects funded by the National Institute of Justice and the Office of Juvenile Justice and Delinquency Prevention have to deposit their data.
- The Substance Abuse and Mental Health Data Archive (SAMHDA)[250] that is maintained by the Substance Abuse and Mental Health Services Administration and is collecting data on mental health and abuse of substances.
- The National Archive of Computerized Data on Aging (NACDA)[251] that is funded by the National Institute on Aging and collects data relevant to gerontological research.
- The International Archive of Education Data (IAED)[252] that is collecting and analysing data related to education in the United States and other countries.

ICPSR has prepared a comprehensive set of guidelines for preparing data for archiving. While these guidelines were written with social science data in mind, they are broadly applicable.

The European network of social science data archives will be further developed through the CESSDA PPP project (see Chapter 3.3.1 below) that also has plans to apply for fixed national mandates for these archives.

Data repositories also exist in a variety of other domains, notably in biological research areas, medicine and health research. Some of these are briefly described below.

The US National Institutes of Health National Center for Biotechnology Information (NCBI)[253] plays an important role in the management of genome data at the national level, supporting public databases, developing software tools for analysing data, and disseminating biomedical information. NCBI has developed a number of data mining tools for their databases that support researchers in their work and runs a range of courses to train their end-users. The NCBI's internationally renown GenBank[254] is a genetic sequence database with an annotated collection of all publicly available DNA sequences. Submitters to GenBank currently contribute over 3 million new DNA sequences per month to the database and research funding agencies in several countries require their fundees to deposit data with Genbank. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI that all exchange data on a daily basis.

In Australia, the emphasis is on developing an infrastructure where institutional repositories are federated through the use of shared services (see Chapter 3.2.4 below), but some domain specific data centres also exist. Several science data centres and databases are hosted by the Commonwealth Scientific and Industrial Research Organisation (CSIRO).[255] These provide not only data collection and storage services, but an extensive range of access and analysis tools as well:

- Australian Marine and Atmospheric Research Data Centre[256]
- Australian e-Health Research Centre[257]
- Australia Telescope Online Archive[258]
- Australian National Wildlife Collection Sound Archive[259]

The Canadian Geoscience Data Repository (GDR)[260] is a collection of earth sciences sector geoscience databases that is managed by the Natural Resources Canada. The data offered free of charge from the repository remains in the GDR's copyright and a permission has to be sought for commercial exploitation of the data.

[249] http://www.icpsr.umich.edu/NACJD/
[250] http://www.icpsr.umich.edu/SAMHDA/about/about-samhda.html
[251] http://www.icpsr.umich.edu/NACDA/
[252] http://www.icpsr.umich.edu/IAED/
[253] http://www.ncbi.nlm.nih.gov/
[254] http://www.ncbi.nlm.nih.gov/Genbank/index.html
[255] http://www.csiro.au/
[256] http://www.marine.csiro.au/datacentre/
[257] http://aehrc.com/hie/index.html
[258] http://atoa.atnf.csiro.au/
[259] http://www.csiro.au/places/ANWCSoundArchive.html
[260] http://gdr.nrcan.gc.ca/index_e.php

The international Scientific Drilling Database (SDDB)[261] is managed by the GeoForschungs Zentrum Potsdam in **Germany**. It holds data resulting from International Continental Scientific Drilling Program (ICDP) projects and other projects supported by the ICDP. Most data in the SDDB are published under the Creative Commons license under the condition that the authors of the dataset are cited.

The psychology data archive PsychData[262] at the Institute for Psychology Information (ZPID) is the long-term archive of primary research data from all areas of psychology and the human sciences that provides the data for research purposes only. The service is supported by the German Research Association and hosted by the University of Trier. The ZPID has developed a web-based data deposit service and guidance tools for new and starting projects. The metadata in PshychData is open for harvesting via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH),[263] but the actual data can only be accessed according to the rules agreed on in both in the deposit agreement and the user agreement.

Data repositories for specific research domains are currently the predominant type of data repositories. In some disciplines there is a long tradition of depositing research data in national repositories so they can be re-used. Many of these data repositories have historically started out as projects, but have succeeded in institutionalising the project databases into data archives with more secure funding. Recently data archives are being set up consciously, using federated repository models and having close partnerships with other data archives in the same country and internationally.

A recent study in Finland showed that academics consider domain level data archives the best option for curating and disseminating their data:[264]

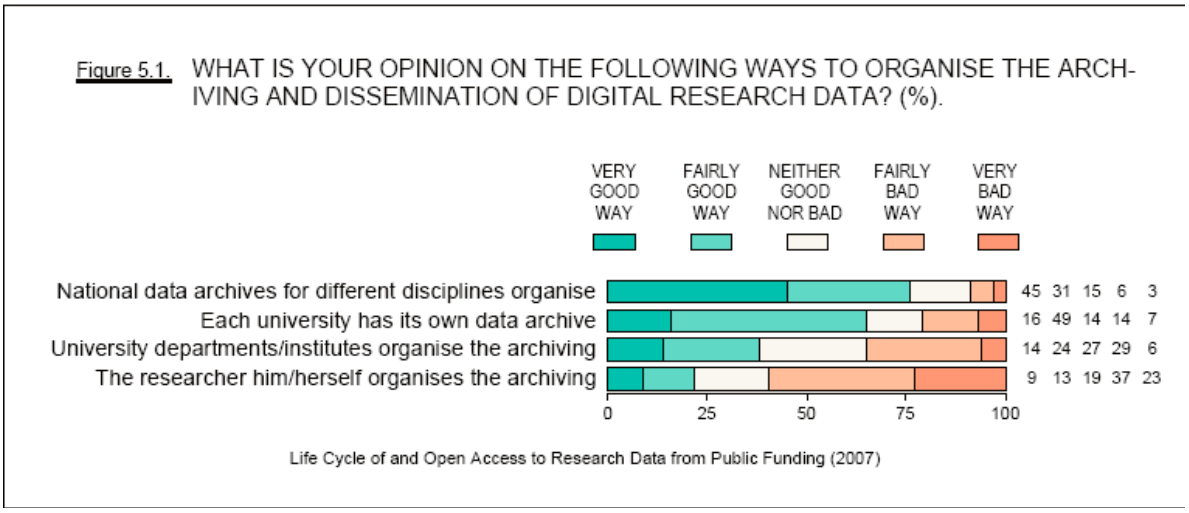

Figure 8. Support for research domain data repositories.

The funding base for research domain level data repositories is varied – most frequently it is the research funding agency that maintains the repository, but it can also be individual institutions or direct government funding. The established data archives (e.g. ICPSR, UKDA) have also succeeded in offering their services to several funding agencies and other organisations in need of digital preservation and data dissemination services.

[261] http://www.icdp-online.org/contenido/lakedb/front_content.php
[262] http://www.zpid.de/index.php?wahl=dienste&uwahl=psychdatainfo&lang=EN
[263] http://www.openarchives.org/pmh/
[264] Arja Kuula, Sami Borg, *Open Access to and Reuse of Research Data – The State of the Art in Finland* (2008), p. 19

3.2.4 INSTITUTIONAL AND PROJECT LEVEL INITIATIVES

Whilst centrally funded research domain-specific data centres provide expert curatorial facilities for the deposit, management and re-use of research data, institutional repositories are developing as an alternative location to deposit research outputs. It is fair to say that to date, most emphasis has been on the deposit of textual outputs; however, some organisations are beginning to explore the practical benefits and value of institutional digital repositories for the storage of primary research data. Many university libraries provide access to research datasets, but these are data licensed from research domain data archives. Data repositories for storing and curating a university's own research data are being set up separate from the licensed data access services. Several research projects that have started to develop a data repository are being institutionalised and federated with other repositories. Some examples of these repositories are described in Table 7 below.

Table 7. Examples of institutional data sharing infrastructure solutions

Name	Data Domains	Scale of Services	Core Partners	Year Founded	Architecture	Licence to Share Data	Accessibility
United Kingdom							
DataShare[265]	Social sciences	Domain / National	EDINA, LSE	2007	Federated repositories	n/a	Open
eBank[266]	Crystallographic data	Domain / National	UKOLN, University of Southampton	2003	Federated repositories	n/a	Open
eCRYSTALS[267]	Crystallographic data	Domain / National	UKOLN, DCC, University of Southampton, University of Manchester	2007	Federated repositories	At present there are 'gentleman's agreements' and no formal agreements in place	Open
GRADE[268]	Geo-spatial data	Domain / National	EDINA, University of Edinburgh, University of Strathclyde, Kingston Centre for GIS, University of Southampton	n/a	Centralised	n/a	Users must agree to the Terms & Conditions of the repository
Ensembl[269]	Genome and biological data	Domain / National	EBI, Wellcome Trust	1999	Centralised	Data generated by Ensembl members	Freely available for one's own use
United States							
Reciprocal Net[270]	Molecular structures	Domain / International	Indiana University, Los Alamos National Lab, MIT, and 16 other universities	1995	Federated repositories	n/a	Much of the data is freely available to the general public
UC DATA[271]	Computerized social science data	Domain / International	University of California, Berkeley	1962 / 1991	Central, storage from ICPSR	Data generated by UC DATA	Unless restricted by licensing and use agreements, data are available to the general public
CPANDA[272]	Arts and cultural policy data	Domain / International	Princeton University, Pew Charitable Trusts	2003	n/a	Licence is held by the Princeton University	Free for non-commercial purposes after signing the user agreement

[265] http://www.disc-uk.org/datashare.html
[266] http://www.ukoln.ac.uk/projects/ebank-uk/
[267] http://ecrystals.chem.soton.ac.uk/
[268] http://edina.ac.uk/projects/grade/
[269] http://www.ensembl.org/index.html
[270] http://www.reciprocalnet.org/index.html
[271] http://ucdata.berkeley.edu:7101/
[272] http://www.cpanda.org/

Name	Data Domains	Scale of Services	Core Partners	Year Founded	Architecture	Licence to Share Data	Accessibility
SIO Explorer[273]	Ocean engineering data	Domain / National-International	Geological Data Center, San Diego Supercomputer Center, UCSD Library, IGPP, Birch Aquarium	2007	Centralised	Free access	n/a
DDCN[274]	Transition and merging markets data	Domain / International	William Davidson Institute, NSF	n/a	Federated repositories	n/a	Free of charge
GeoData Center[275]	Geophysical data	Domain/ National	University of Alaska Fairbanks	n/a	Collected by the Institute	Copyrighted	Free for non-commercial use after filling out a permission form
IDEALS[276]	Digital research outputs	National	University of Illinois at Urbana-Champaign	n/a	Collected by the university	IDEALS requires non-exclusive rights to distribute the data to be assigned to the repository	Open
Australia							
UQ eSpace[277]	The University of Queensland's institutional digital repository for publications, research, and teaching materials	National	Funded by Department of Education, Employment and Workplace Relations. Testbed for the Australian Partnership for Sustainable Repositories.	2008	Collected by the University of Queensland	Unless otherwise indicated, copyright is the property of the University of Queensland.	Open

[273] http://nsdl.sdsc.edu/
[274] http://ddcn.prowebis.com/
[275] http://www.gi.alaska.edu/services/geodata/
[276] http://www.ideals.uiuc.edu/
[277] https://espace.library.uq.edu.au/about.php

The Directory of Open Access Repositories (DOAR)[278] currently lists 44 open access institutional data repositories from around the world. Only 5 of these are from the UK. An analysis of institutional repositories at Australian universities[279] conducted in March 2007 revealed that out of 19 existing repositories 12 accepted research data and datasets for archiving, but all deposits were voluntary and mostly intended as datasets accompanying a research paper.

In the **United Kingdom**, the JISC Digital Repositories Programme has funded several projects that are looking into setting up institutional data repositories and providing new services for repositories. Several of these (e.g. GRADE, StORe, SPECTRa, eCHRYSTALS) have already been reviewed in detail in other reports.[280] As a result of these efforts the first multi-disciplinary institutional data repositories are beginning to appear, and several institutionally supported data repositories limited to one research discipline are also appearing.

The DISC-UK DataShare project[281] led by EDINA explores new pathways to assist academics in sharing their research data via institutional repositories. The project partners already have data repositories: the Edinburgh DataShare[282] is a pilot digital repository of multi-disciplinary research datasets produced at the University of Edinburgh; the Edinburgh DataShare is hosted by the University's Data Library, the first university data library in the UK that was established in 1983; Oxford University's Nuffield College manages a Data Library[283] with a small number of datasets; the LSE Research Laboratory Data Service[284] collects primary data created by LSE researchers; the University of Southampton has a large e-prints repository and its library is offering access to a large number of licensed datasets from other data centres.

A project at the University of Oxford is scoping the requirements for repository services to manage research data.[285] Led by the Oxford e-Research Centre, the project has interviewed a number of academics about their data management practices and requirements for services. The project will also look into federating the existing repositories in the University of Oxford to a single set of services, including repositories for data.

York St. John University has recently opened its institutional repository with a collection of digital video and audio resulting from research projects.[286]

The eBank-UK project[287] has established an institutional repository that supports, manages and disseminates metadata relating to crystal structure data. As part of the larger landscape eBank-UK is investigating the role of aggregator services in linking data-sets from grid-enabled experiments to open data archives and through to peer-reviewed articles. One of the outcomes of the eBank-UK project is eCrystals,[288] the archive for crystal structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service.

Ensembl[289] is a joint project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI) to develop a repository system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is collecting and releasing its data and analysis into the public domain immediately and imposes no restrictions on access to, or use of, the data provided and the software used to analyse and present it.

In the **United States**, compared with the number of institutional repositories for textual materials, the number of institutionally supported data repositories is not very large and most of them appear to be specific to individual research disciplines. Many university repositories offer licensed data from other, subject-based data archives to their researchers and hold very little or no data themselves. Similar to the JISC digital

[278] http://www.opendoar.org/find.php?search=&clID=&ctID=9&rtID=2&cID=&lID=&rSoftWareName=&submit=Search&format=summary&step=20&sort=r.rName&rID=&ctrl=new&p=1
[279] Kylie Pappalardo et al., *A Guide to Developing Open Access Through Your Digital Repository* (2007)
[280] Cf: Liz Lyon, *Dealing with Data* (2007), Ch. 5.3
[281] http://www.disc-uk.org/datashare.html
[282] http://datashare.edina.ac.uk/dspace/
[283] http://www.nuff.ox.ac.uk/projects/datalibrary/index.html
[284] http://rlab.lse.ac.uk/data/default.asp
[285] http://www.ict.ox.ac.uk/odit/projects/digitalrepository/
[286] http://www.yorksj.ac.uk/library/resources/archive/index.aspx
[287] http://www.ukoln.ac.uk/projects/ebank-uk/
[288] http://ecrystals.chem.soton.ac.uk/
[289] http://www.ensembl.org/index.html

repositories programmes, the NSF has funded a National Science Digital Library (NSDL) programme[290] that brings together best examples of repositories.

Reciprocal Net[291] is constructing a distributed, open, extensible digital collection of molecular structures. Associated with the collection will be software tools for visualizing, interacting with, and rendering printable images of the contents; software for the automated conversion of local database representations into standard formats which can be globally shared; and tools and components for constructing educational modules based on the collection.

SIO Explorer[292] is making data, documents and images from 822 expeditions by the Scripps Institution of Oceanography (SIO) since 1903 accessible as part of the NSDL.

The Davidson Data Center and Network (DDCN)[293] is an integrated, fully searchable database on transition and emerging markets. DDCN archives and provides free access to socio-economic micro and macro data on transition economies.

The Geophysical Institute at the University of Alaska Fairbanks is maintaining, among other national data collections, a GeoData Center[294] that provides data management and archive services for the Geophysical Institute and maintains a variety of geophysical data collections in support of scientific research.

The Henry A. Murray Research Archive at Harvard University[295] is a permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, and provides physical storage for the entire IQSS DataVerse Network. The collection comprises over 125 terabytes of data, audio, and video. The archive accepts and preserves in perpetuity all types of data of interest to the research community, including numerical, video, audio, interview notes, and other data.

The University of Illinois at Urbana-Champaign has set up the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS)[296] as a service to preserve and provide persistent and reliable access to the digital research outputs, including datasets, of faculty, staff, and students on the UIUC campus in order to give these works the greatest possible recognition and distribution. IDEALS has published a comprehensive set of policies and requires non-exclusive rights to distribute the data to be assigned to the repository.

The Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill maintains an archive of social science and health data.[297] Data archived with the Institute will be made available on the Institute's Web site at no cost to depositors or users and the ODUM collection is linked with the DataVerse network.

In **Australia**, investigation into appropriate repository infrastructure at the institutional level has progressed through three research projects – ARROW, DART and ARCHER.

The Australian Research Repositories Online to the World (ARROW) project[298] came into existence in response to a call for proposals in 2003 by the Australian Commonwealth Department of Education, Science and Training (DEST). DEST was interested in furthering the discovery, creation, management and dissemination of Australian research information in a digital environment. Specifically, it wanted to build the Australian research information infrastructure through the development of distributed digital repositories and the common technical services supporting access and authorisation to them. The design of the ARROW repository solution was informed by the desire to:

- Use a common underlying repository for a range of content types.
- Provide content management modules for different use cases.
- Expose the content as widely as possible using a number of different technologies.

[290] http://nsdl.org/
[291] http://www.reciprocalnet.org/index.html
[292] http://nsdl.sdsc.edu/
[293] http://ddcn.prowebis.com/
[294] http://www.gi.alaska.edu/services/geodata/
[295] http://www.murray.harvard.edu/
[296] http://www.ideals.uiuc.edu/
[297] http://www.irss.unc.edu/odum/jsp/content_node.jsp?nodeid=7
[298] http://arrow.edu.au/

The general architecture of the ARROW project's original infrastructure solution is the following:[299]

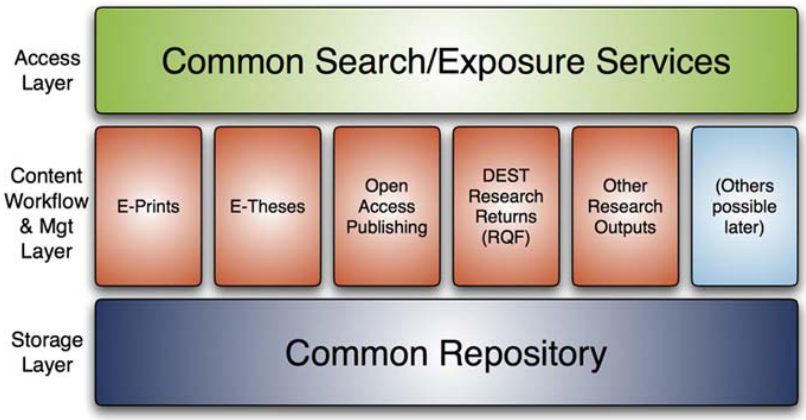

Figure 9. High-level ARROW solution architecture.

The shared storage layer is used by different repository management and workflow solutions that cater for different types of content and repositories. A shared discovery service can be built on top of the different repository management systems (see also Chapter 3.3 below). The Fedora-based repository management solution VITAL developed by the ARROW project and its commercial partner VTLS is currently in use at 14 university repositories, in some cases also for managing research data.

The following Dataset Acquisition, Accessibility and Annotations e-Research Technologies (DART) project[300] is built on the design proposed by the ARROW project to develop a proof of concept system for a federated repository solution and collaborative research work environment. The high-level architecture of the DART solutions shown in Figure 10 continues to use the Fedora repository management system and to support working with large datasets will make use of the Storage Resource Broker (SRB).

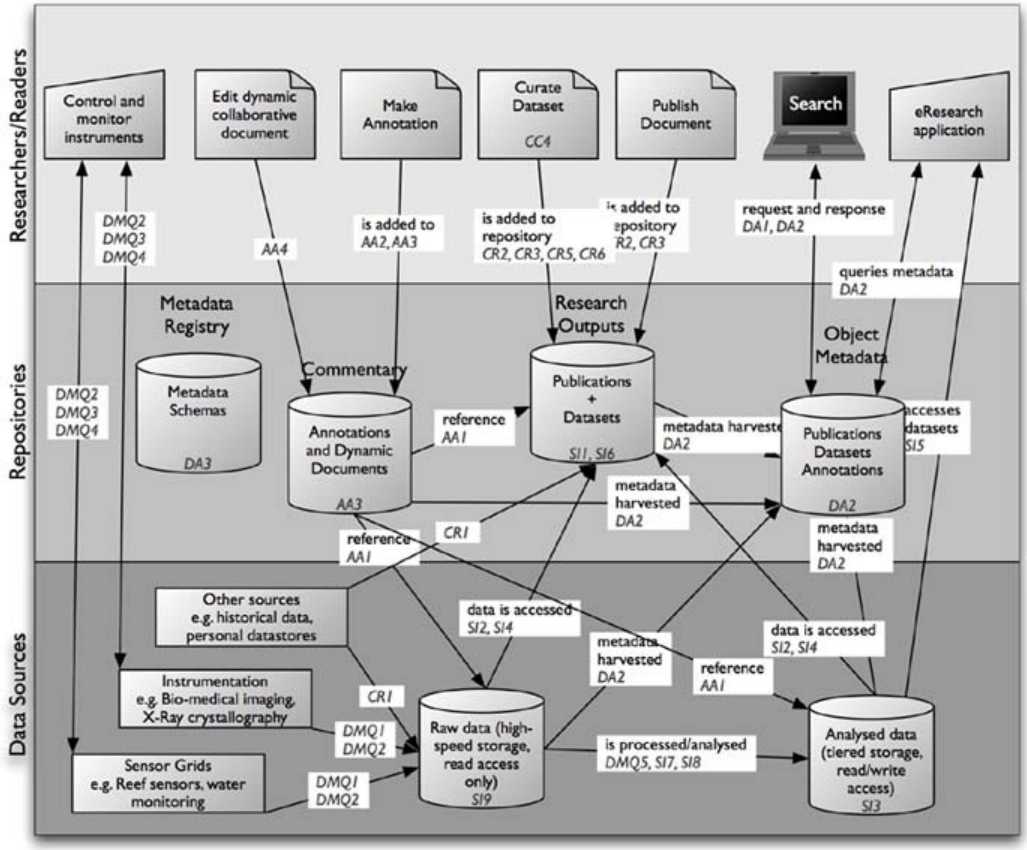

Figure 10. High-level DART architecture.

[299] Andrew Treloar, David Groenewegen, *ARROW, DART and ARCHER: A Quiver Full of Research Repository and Related Projects* (2007)

[300] http://dart.edu.au/

In Figure 10[301] the uppermost layer shows researchers, readers and computer programs. The middle layer shows the proposed repositories (including traditional publications as research outputs, and raw data) and the data flows between them and the datasets in the lowest layer. The lowest layer shows the data sources and their associated storage.

DART has been working with researchers in three different domains: x-ray crystallography, digital history and climate research. Of these, the x-ray crystallography demonstrator is the one that has progressed to the greatest extent and a Gridsphere portal has been created.[302]

The DART proof-of-concept design has been taken up by the Australian Research Enabling Environment (ARCHER) project[303] to integrate it with other open source components and to provide a robust and comprehensive end-to-end set of data acquisition and management tools. The requirements of the national ANDS project are being integrated into the design of the ARCHER tools to support large datasets using the SRB technology. The ARCHER research repository provides tools for managing datasets through the research process and federated access to data from web, desktop, or standard file access protocols (e.g. GridFTP and SRB).[304]

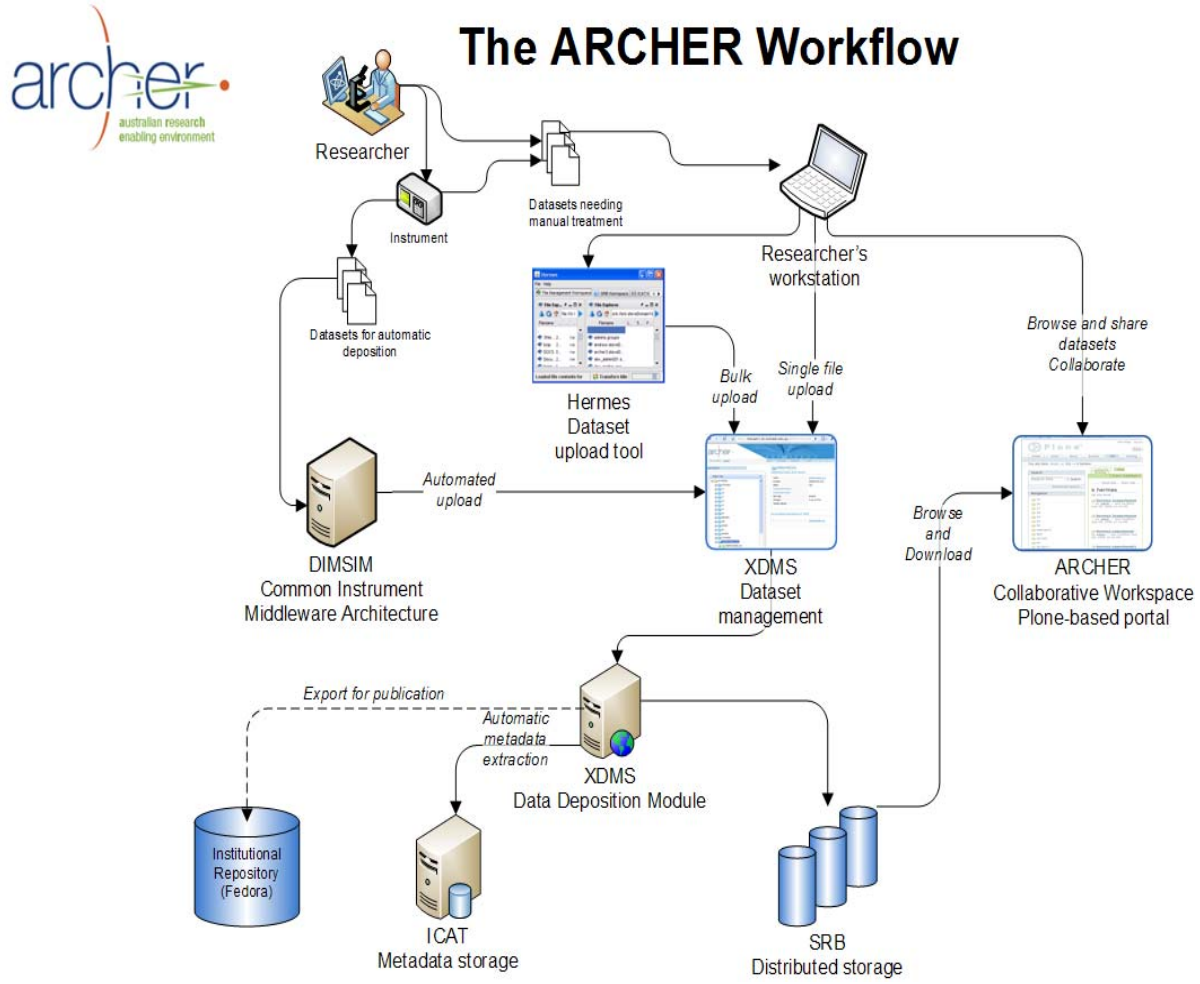

Figure 11. The ARCHER toolset components and workflow.

The concept of data repository that in the original ARROW project was envisaged as a single underlying repository underpinning all the research outputs of a university has developed in the DART and ARCHER projects to suggest an alternative model. This is based on two different kinds of repository: one optimised for collaboration and one for publication of data. This new concept will be implemented in the ongoing Australian

[301] Andrew Treloar, *The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: Building the New Collaborative E-research Infrastructure* (2006)
[302] For a demo see: http://dart.edu.au/assets/images/workpackages/ultra-short-800x600/ultra-short-800x600.html
[303] http://archer.edu.au/
[304] Cf. http://archer.edu.au/products/index.html

Repositories for Diffraction Images (TARDIS) project.[305] TARDIS is a multi-institutional collaborative venture that aims to facilitate the archiving and sharing of raw X-ray diffraction images from the Australian protein crystallography community. It is based on a model of federated repositories that share dataset descriptions and will develop a staging repository for refining a dataset collection's metadata and for holding it before publication.

In **Germany,** the Helmholtz Open Access Project supports scientists as well as the respective Helmholtz Centres in the realisation of open access. The project newsletter and the Helmholtz Open Access Workshops are two examples of how the project informs and advises researchers. Most Helmholtz Centres run institutional repositories that contain a significant and steadily growing share of the scientific output, including data, of the Helmholtz Association.

The Germany Science Association is funding the establishment of a German Network of Certified Repositories, certified by Deutsche Initiative für Netzwerkinformation (DINI).[306] The network is intended as the German contribution to interconnecting repositories in the European Research Area via the DRIVER project tools. The network focuses on universities and their institutional repositories, but also includes the Max-Planck Society, Fraunhofer and Leibniz Societies' repositories, in order to document the whole of the German research output.[307] A search for datasets in German open access repositories returns almost a thousand hits.

In most OECD countries institutional level data repositories are only at the exploratory stage and the complex issues of copyright, data protection, responsibilities and funding the infrastructure are being investigated. The institutional repositories that already hold a few datasets use standard repository management software and harvest metadata via OAI-PMH.

In **Spain**, a consortium of seven universities in Madrid[308] is following up the successful establishment of an e-prints repository with development of a data repository.[309]

The **Japanese** National Institute of Informatics (NII) has launched its Institutional Repositories Program[310] with funding for over ten collaborative projects on institutional repository federation and exploration of including other types of digital objects in addition to text. The Japanese national access portal to institutional repositories currently shows availability of 457 datasets in 7 university repositories.[311]

> The institutional infrastructure for data sharing is still emerging. Most institutional repositories aspire to be more than a stockpile of e-prints and provide safe harbours for a more inclusive range of the intellectual output of local research and teaching. These repositories retain institutional identity, but may choose to limit access to specific sub-communities within the institution instead of public, web-wide open access. Institutional repositories are more interested in collecting research materials near the end of the research life cycle, employing an acquisition process that is simple and a preservation process that is not designed to support complex content.[312] These limitations may decrease over time as the institution-based data repository movement matures.

[305] http://www.tardis.edu.au/
[306] http://www.dini.de/service/dini-zertifikat/
[307] http://www.aepic.it/conf/viewpaper.php?id=271&cf=10
[308] http://www.consorciomadrono.es/
[309] Cf: Juan Corrales Correyero, Alicia López Medina, *La red de repositorios institucionales del Consorcio Madroño ante el 'diluvio de los DATA'* (2008) http://www.consorciomadrono.es/noticias_eventos/2008/AliciaLopez.pdf
[310] http://www.nii.ac.jp/irp/en/
[311] http://jairo.nii.ac.jp/en/
[312] Cf. Ann Green, Myron Gutmann, *Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives* (2006), p. 4

3.3 DATA FEDERATION AND ACCESS SERVICES

Research has become international and interdisciplinary. Locating data in disparate repositories in different countries, gaining access to them through a maze of licence agreements in different languages and permissions, and re-using them in a multitude of file formats can be a daunting task. These barriers are not easy to overcome – the sheer diversity of data makes it difficult to design tools with the range and ability to accommodate and translate between the distinctly different data needs of the various domain communities:[313]

> Even where technical solutions can be devised, how can participants from one disciplinary community make sense of data produced under the very different procedures and understandings of another? As work in the field of science and technology studies (STS) has demonstrated, data are the product of working epistemologies that are very often particular to disciplinary, geographic, or institutional locations. Data oriented to the needs, practices, and cultures of the ocean sciences might not be easily or automatically translatable into the idiom and usages of atmospheric science. Beyond such issues of recognition and fit, questions of trust loom large. Can I trust those I share my data with to make reasonable and appropriate use of it, and on a timeline which doesn't jeopardize my own interests around publication, credit, and priority? Or conversely: can I trust the data I'm getting, particularly as collaborative webs widen and my first–hand knowledge of the data and its producers recedes?

In some domains with a long history of collaboration, norms of data sharing exist and are highly structured. In others, the collaboration has not been a tradition, and rules and procedures for sharing data are relatively undefined. Social science data archives have quite a long history of cross-repository access services – resource discovery in the social sciences now extends far beyond consulting a stand-alone research aid or search tool. Several alliances across repositories exist and the emergence of a new set of tools allows researchers to do complex and innovative searches to locate and explore data. The possibility of cross-database and cross-site searching has been made much easier by the emergence of a standard XML-based mark-up language for social science metadata called the Data Documentation Initiative (DDI),[314] now in its third version. The existence of common metadata mark-up has enhanced the development of software tools for exploratory data search and analysis.

Technical solutions for data federation from different repositories in one research domain and across domains are beginning to bridge the gaps between disciplines. Search and retrieval services based on OAI-PMH have made finding e-prints and other scholarly works from repositories around the world very easy. Similar portal services that harvest metadata from disparate data repositories are emerging that allow the retrieval of entire cross-sections of national- or research-domain level research output.

3.3.1 INTERNATIONAL INITIATIVES

The OpenDOAR service[315] provides a quality-assured listing of open access repositories around the world that contain research information. The OpenDOAR project team is undertaking a survey of the available repositories to determine the scope and scale of the developing repository network. OpenDOAR staff harvest and assign metadata to allow categorisation and analysis to assist the wider use and exploitation of repositories. OpenDOAR is maintained by SHERPA, and the project is a partnership between the University of Nottingham, UK and Lund University, Sweden. The service is supported by the Open Society Institute, JISC, CURL and SPARC Europe.

In 2007, the European organisation for social science data archives, CESSDA, received EU funding for CESSDA PPP – Preparatory Phase Project for a Major Upgrade of the Council of European Social Science Data Archives Research Infrastructure.[316] The project is focussing on tackling and resolving a number of strategic, financial and legal issues, including work on developing a data portal to allow seamless access to data holdings across Europe, developing common authentication and access middleware tools, developing metadata standards, creating thesauri management tools, extending the coverage of CESSDA, investigating the potential of grid technologies, and improving data harmonisation tools. In developing the next version of the common, multilingual data access portal, the project will also investigate data storage, discovery and retrieval to allow

[313] Steven Jackson, et al., *Understanding infrastructure: History, Heuristics and Cyberinfrastructure Policy* (2007)
[314] http://www.icpsr.org/DDI/
[315] http://www.opendoar.org/
[316] http://www.cessda.org/project/

extensive access to several types of complex data. The portal will have a sign-on procedure to authenticate and authorize legitimate users.[317]

Another EU FP7 funded project is the Ground European Network for Earth Science Interoperations - Digital Repositories (GENSI-DR).[318] The project will establish an open earth science (data from space, airborne, in-situ sensors) digital repository and integrated access portal.

Pan-European and international access to biobanks and biomolecular resources is being developed by the BBMRI project for biomedical and biological research.[319] The BBMRI services format will be a distributed hub structure in which the hubs coordinate activities, including collection, exchange and analysis of samples and data for the major domains. The biobanks, biomolecular resources and technology centres that are members of BBMRI are associated with their specific domain hub. This structure provides flexibility for new members and partners to be connected at any time and the structure can easily be adapted to emerging needs in biomedical research. The IT infrastructure employs federated database architecture and grid computing technology integrates the network of hubs, members and partners into a single virtual infrastructure.

The PANGAEA project[320] hosted by the Alfred Wegener Institute for Polar and Marine Research and Centre for Marine Environmental Sciences at the University of Bremen in Germany operates an open access library aimed at archiving, publishing and distributing georeferenced data from earth system research. Data are archived as supplements to publications or as citable data collections. The system guarantees long-term availability of its content through a commitment of the operating institutions. The data and any associated material in PANGAEA are made available under the Creative Commons Attribution license, although some datasets have restrictions on their use. PANGAEA is offering services as a designated archive for WDC Mare, Ocean Drilling Program and the journal *Earth System Science Data*. The PANGAEA search engine is powered by the open-source software panFMP (PANGAEA Framework for Metadata Portals). panFMP is a flexible framework for building geoscientific metadata portals independent of content standards for metadata and protocols. Data providers can be harvested with commonly used protocols (e.g. OAI-PMH).

The iRODS project[321] at the San Diego Supercomputing Center in the US has developed an integrated rule-based data system that organizes distributed data into shared collections. iRODS allows users to access files seamlessly across a distributed environment, based on their attributes rather than just their names or physical locations. It replicates, synchronises and archives data, connecting heterogeneous resources in a logical and abstracted manner and supports a distributed workflow system.

The US NCBI has developed a life sciences search portal Entrez[322] that can query both published articles and datasets concurrently from hundreds of international libraries and data collections.

> Cross-domain repository search tools (e.g. OpenDOAR, OAIster) allow searching for datasets in institutional repository catalogues using OAI-PMH, but because the practice of accepting datasets into institutional repositories and describing them there vary, these search portals are not (yet) the most efficient means of accessing data. Considerably more effort is being channelled into developing domain-specific data access tools that rely on metadata standards used in individual disciplines. Agreement on metadata schemas and protocols for information exchange is one of the key outcomes of the international services in development.

3.3.2 NATIONAL INITIATIVES

Portals offering access to all types of scholarly output on a national level are being developed in many countries. These services rely primarily on institutional repositories and on OAI-PMH for harvesting metadata; some services are also looking into linking publications and their underlying data.

In the **United Kingdom**, the JISC has funded several projects that develop tools for cross-repository search. The Intute Repository Search[323] is a MIMAS project, which been commissioned by JISC in partnership with

[317] Cf. *CESSDA PPP Presented at IASSIST Conference* (2008)
[318] http://www.genesi-dr.eu/
[319] http://bbmri.eu/bbmri/
[320] http://www.pangaea.de/
[321] http://irods.sdsc.edu
[322] http://www.ncbi.nlm.nih.gov/sites/gquery

UKOLN and SHERPA. The project will provide a ubiquitous middle layer operating between academic, research and institutional repositories, which can feed into personalised learning and research spaces. Intute is exploring the development of full-text searching and discovery with the help of text mining tools; adding personalisation features to the search functions; enhancing and augmentation of metadata creation via automatic classification and meaning-based taxonomy extraction. Supporting data repositories search is not explicitly in the project scope description but Intute does allow searching for datasets.

The Federated Access to Repositories (FAR)[324] led by the LSE had a focus on the access management requirements needed for the storage and sharing of data sets in institutional repositories. The project contributed to the DISC-UK DataShare project to help develop an interface to the repository for dataset management for specifying a set of access conditions for the deposited datasets and sharing them with other users registered in the repository. The project developed specifications for extending the functionality of existing institutional repository management software packages to allow, among other things, for more flexible searching.

JISC has also funded a Data Audit Framework (DAF)[325] to provide research institutions with the means to identify, locate, describe, and assess how they are managing their research data assets. DAF combines a set of methods with an online tool to enable data auditors to gather information on data assets. DAF will help ensure that research data produced is preserved and remains accessible in the long term.

In the **United States**, NDIIPP has funded development of the Dataverse Network.[326] Developed at Harvard University by the Institute for Quantitative Social Science, this is open-source software for digital library management, dissemination, exchange, and citation of virtual collections (dataverses) of quantitative data. Dataverses can be used or administered through web-based clients that communicate with a host dataverse network. Individual dataverses are self-contained virtual data archives, served out by a Dataverse Network, appearing on the web sites of their owners (e.g. individuals, departments, projects, or universities). Dataverses are branded in the style of the owning entity, but are easy to set up, require no local software installations, and offer the services of a modern archive controlled by the dataverse owner. Data is displayed in a hierarchy; descriptive information can be searched. Dataverses and Dataverse Networks can federate with each other, and with other systems through open protocols (OAI-PMH and Z39.50).

In **Australia**, a number of projects have been funded by the government in support of the Australian Accessibility Framework. The Australian Access Federation (AAF) project[327] is focussing on policy frameworks and developing an overarching set of governance policies for the trust federation. A shared authorisation framework for access to research resources and network infrastructure to allow researchers and research resources to connect at the required bandwidth will be established.

Macquarie University has led the Meta Access Management System (MAMS) Project.[328] This project allows for the integration of multiple solutions to managing authentication, authorisation and identities, together with common services for digital rights, search services and metadata management. The project provides an essential middleware component to increase the efficiency and effectiveness of research infrastructure. The MAMS new conceptual architecture is capable of supporting multiple, independent models implemented locally within organisations, but with the potential for inter-institutional communication.

The Persistent Identifier and Linking Infrastructure (PILIN) project[329] was a subset of the ARROW project to develop an abstract model for identifiers and their management and a persistent identifier management infrastructure. The project prepared for a sustainable shared identifier infrastructure in Australia. The Identifier infrastructure enables management of the full lifecycle of a data resource from creation to archiving.

Using these facilities, the ARROW project (see Chapter 3.3.2 above) developed the ARROW Discovery Service[330] for searching all of Australia's research repositories. The service is provided by the National Library

[323] http://www.intute.ac.uk/irs/
[324] https://gabriel.lse.ac.uk/twiki/bin/view/Projects/FAR/WebHome
[325] http://www.data-audit.eu
[326] http://thedata.org/
[327] http://www.aaf.edu.au/
[328] http://www.melcoe.mq.edu.au/projects/MAMS/
[329] http://www.pilin.net.au/
[330] http://search.arrow.edu.au/

of Australia and information is harvested from the hosting repositories via OAI-PMH. Each record found through this service links back to a record in the original repository. The contents of the ARROW Discovery Service are also indexed by Google, increasing the visibility of Australian research throughout the world.

In **Germany** the Publication and Citation of Scientific Primary Data (STD-DOI) project,[331] funded by the German Science Foundation, aims to make primary scientific data citeable as publications. In this system, a data set would be attributed to its investigators as authors would be for a work in the conventional scientific literature. Thus, scientific primary data should not be exclusively understood as part of a scientific publication, but may have its own identity. The service connects repositories or 'data publications agents' where each dataset is assigned a persistent identifier and is linked with search facilities. The persistent identifiers (both DOI and URN) identify datasets and are resolved to the valid location (URL) where the datasets can be found. In addition, the data publications are included in the catalogue of the German National Library of Science and Technology (TIB). The TIB acts as a registration agency for persistent identifiers.

In **the Netherlands**, a national show-case portal for scholarly output has been developed in several stages. The Dutch Academy of Arts and Sciences-funded SURF foundation is a collaborative organisation for higher education institutions and research institutes aimed at encouraging innovations in ICT. The Digital Academic Repositories (DARE) programme is an initiative to develop institutional repositories at universities and making the research results accessible. The DAREnet portal was developed as part of the project, presenting the national 'Cream of Science' works, a national site for doctoral theses ('Promise of Science') and a knowledge bank of the universities of applied sciences.

The completion of the DARE programme has provided the foundation for the SURFshare programme[332] 2007-2010. The SURFshare programme focuses on four main areas: Digital Author Identifier (DAI), persistent identifiers, complex objects and metadata. Activities are undertaken around the following themes:

- Innovation of the scientific publication cycle, including enhanced publications.
- Collaboratories: virtual research groups in which scientists can share their sources and can collaborate from various locations.
- Quality assessment, dissemination and impact measurements within an Open Access environment.
- Registration of research data and facilities for their sustainable accessibility.

Part of this work has been the development of the original DAREnet portal into the NARCIS service.[333] NARCIS provides access to all Dutch scholarly information – full-text publications from all Dutch universities, KNAW, NWO, and a number of research institutes; datasets from DANS; and descriptions of research projects, institutes and researchers. At the moment, DANS is the only contributor of datasets to the NARCIS service.

The **Canadian** Association of Research Libraries has developed a national search portal[334] linking together fourteen institutional repositories and collections that can also be searched for datasets. Current holdings contain 33 datasets.

In **Japan**, the national Institute of Informatics (NII) launched JAIRO (Japanese Institutional Repositories Online) in 2008. JAIRO succeeded JuNii+ (test version) in which academic information (journal articles, theses or dissertations, departmental bulletin papers, research papers, etc.) accumulated in Japanese institutional repositories can be searched for cross-sectionally. As of October 2008, JAIRO allows about 540,000 objects in 84 institutional repositories to be searched for, including 457 datasets.

In **France**, the EducNet, a service supported by the Ministry of Education and Research, has commissioned a study on legal requirements on dissemination of educational resources and research data. It covers legal, economic and social drivers for making research data available and offers guidance on complying with the legal requirements.[335]

[331] http://www.icdp-online.org/contenido/std-doi/front_content.php
[332] SURFshare programme 2007-2010. *SURF Platform ICT and Research* (2007)
[333] http://www.narcis.info/index
[334] http://carl-abrc-oai.lib.sfu.ca/index.php/search
[335] Jean-Michel Bruguiere, *Droit des données publiques* (2008)

Most countries either already have or are in the process of developing a national access portal to e-prints and other content held in institutional repositories. Research data is sometimes part of these plans, but the added complexities of different metadata requirements and lack of data citation rules do not hold promise for the e-prints portals becoming main access gateway to data as well.

3.3.3 DOMAIN LEVEL INITIATIVES IN INDIVIDUAL COUNTRIES

The majority of data repositories are organised around research disciplines and funding agencies in each discipline. These repositories provide access to their own collections and in almost all cases also federated access to other domain data repositories through metadata harvesting. The number of domain-based data access services is large; some have been already been discussed under international initiatives (section 3.3.1) above. A few examples are discussed in this chapter.

In the **United Kingdom**, JISC has funded a series of portal development projects through its Portals Programme. The eight research subject hubs that were created[336] now rely on the Intute system for search and access management.

The UK Citation, Location and Deposition in Discipline & Institutional Repositories (CLADDIER) project[337] produced a demonstration system to cross-search and link publications in institutional repositories to data held in the British Atmospheric Data Centre. The work of the project includes exploration of data citation and dataset publishing and concrete recommendations for data and publications linking.[338]

PerX[339] is a subject-based cross-repository search tool for resource discovery in engineering. PerX is currently offered as a pilot service.

The Go-Geo service[340] maintained by EDINA provides access to geospatial information and services. Users can access data through a map-based interface or via an advanced search and query builder. A large proportion of the data offered through this service is harvested from ADS, UKDA and NSRI.

In the **United States**, several social sciences data archives are collaborating in the Data Preservation Alliance for the Social Sciences (Data-PASS) partnership,[341] funded by the NDIIPP, and devoted to identifying, acquiring and preserving data at risk of being lost to the social science research community. As part of this collaboration a shared catalogue has been developed,[342] using the Dataverse Network services. The metadata schema used for searching in the Dataverse Network is a downsized version of DDI[343] that is more data-specific than the OAI metadata schema based on the main Dublin Core metadata elements. The Data-PASS shared catalogue gives links to data at partner site and the catalogue server may cache a copy of data for faster performance. Data analysis tools (subset extraction, descriptive statistics, crosstabs) are offered in the catalogue system and data can be downloaded in multiple formats.

Several domain-specific data aggregation services are being developed in **Australia**. Collaborative services for researchers in geosciences will be based on grid technology to federate content from individual repositories in Australia. The AuScope project will develop an AuScope Exploratorium[344] that is designed to overcome the issues with non-standardised data in different institutions by developing an interoperable middle layer between the institutional collections and the user interface. The AuScope Exploratorium will build on the experiences from the Solid Earth and Environment GRID (SEEGRID)[345] collaboration project.

BioGrid Australia[346] is a demonstrator system applying the Service Oriented Architecture to create a single logical database spanning data sources at multiple institutions. It is a clinical, laboratory and genetic data

[336] http://www.jisc.org.uk/whatwedo/programmes/portals/spp2.aspx; http://www.jisc.org.uk/whatwedo/programmes/portals.aspx
[337] http://claddier.badc.ac.uk/trac
[338] Brian Matthews, Katherine Portwin, Catherine Jones, Bryan Lawrence, *Recommendations for Data/Publication Linkage* (2007)
[339] http://www.engineering.ac.uk/index.html
[340] http://www.gogeo.ac.uk/cgi-bin/index.cgi
[341] http://www.icpsr.umich.edu/DATAPASS/
[342] http://dvn.iq.harvard.edu/dvn/dv/datapass/
[343] See: *Data-PASS Metadata Requirements* (2007)
[344] http://www.auscope.org/home_frame.htm
[345] https://www.seegrid.csiro.au/twiki/bin/view/Main/WebHome
[346] http://www.biogrid.org.au/pages/index.php

federation service that allows researchers to combine data from multiple sources to produce larger, more statistically significant sets for clinical research. The data is physically located across jurisdictions within independent hospitals and research organisations, and with authorization can be combined, searched and queried over the Internet in a de-identified format. The federation of repositories is a network without centralised control where every member is a trusted peer and where an agreed communication protocol is used. The challenges of ethics, privacy, and data ownership have been addressed with solutions for protecting patient confidentiality an important part of the BioGrid network. BioGrid builds on the earlier successful pilot Bio21 Molecular Medicine Informatics Model (Bio21:MMIM) which was a virtual repository of clinical and genetic data sets that were physically located at various organisations, but were federated to be integrated, searched and queried seamlessly via a federated data integrator.[347]

BlueNet is the Australian Marine Science Data Network,[348] which links data repositories and marine resources that currently reside in individual academic and government institutions. BlueNet is established at the University of Tasmania and its BlueNet MEST tool provides searching for metadata records describing marine data sets submitted for curation by the BlueNet project.

In **Canada**, the Ontario Data Documentation, Extraction Service and Infrastructure Initiative (ODESI)[349] is using a web-based data extraction system to provide access to Statistics Canada datasets, data files from Gallup Canada and other polling companies, public-domain files such as the Canadian National Election Surveys, and selected files from the ICPSR. The data are marked up using DDI to allow data resource discovery, distributed access, extraction and analysis. The project is funded by the Ontario Council of University Libraries and OntarioBuys.

The University of Copenhagen in **Denmark** is supporting the Global Biodiversity Information Facility (GBIF).[350] This is a global distributed network of interoperable databases that contain primary biodiversity data – data associated with specimens in biological collections, as well as documented observations of plants and animals in nature. The core description of such collections has been cast into common data exchange formats (e.g. Darwin Core) that can be shared using internet protocols. A federated data access portal has been developed[351] that allows searching from ca 150 million data records around the world.

> Significant resources are being invested in developing domain-specific data access, federation and analysis tools. Despite being frequently developed by short-term projects, institutions are taking on the commitment for supporting these services as valuable resources for the academic community. The domain specific data access portals are also aimed at harvesting metadata and content from other, international sources, to make it available to researchers at the host institution or within the country. The ability to link to international resources depends crucially on the interoperability of metadata schemas used within and across domains. Access to both data and publications through the same search is increasingly becoming a standard service.

3.4 DATA SHARING AND VIRTUAL RESEARCH ENVIRONMENTS

Many sciences are increasingly becoming collaborative sciences. Collaboration involves sharing and re-use of data between several research groups within disparate disciplines that use different tools for collecting and processing data and also describe data differently. Frequently the data need to be shared between small and medium-size research centres and institutes that often have very different computing environments and levels of IT expertise. Data re-use in this setting has required much effort for manual data conversion and has been error-prone. There is a clear need for automation of the research process that is linked directly to data sources.

[347] Marienne Hibbert, *Australian Cancer Grid* (2005)
[348] http://www.bluenet.org.au/index.html
[349] http://search1.odesi.ca/home.xqy
[350] http://www.gbif.org/
[351] http://secretariat.mirror.gbif.org/welcome.htm

The emerging Virtual Research Environment (VRE) (also called Reseachers' Toolbox and Researchers' Desktop) services help researchers in all disciplines manage the increasingly complex range of tasks involved in carrying out research. A VRE provides a framework of resources to support the underlying processes of research on both small and large scales and allow researchers not only to re-use existing data, but to do this in collaboration with other researchers. Use of VREs allows for a faster learning process and helps to build structured knowledge environments.

Virtual Research Environments are seen as vital components of the e-research infrastructure. In the **United Kingdom**, JISC is funding a second phase of VRE projects.[352] The first phase included fourteen projects to explore the definition of and technological solutions for VREs. In the second phase (2007-2009) four integrating pilot projects are funded.

myExperiment[353] is a collaborative environment where scientists can safely publish their workflows, share them with groups and find the workflows of others. Workflows, other digital objects and collections can be swapped, sorted and searched like photos and videos on the web. Sharing, reusing and repurposing research workflows is reducing time to experiment, supports sharing of expertise and helps to avoid reinvention. The workflows being shared with this service include plugins, mashups and links with specific data collections and centres.

The myGrid[354] network, hosted by the University of Manchester, has developed e-science tools for e-laboratories. Among the myGrid tools is the Taverna Workbench that allows users to construct complex analysis workflows from components located on both remote and local computers, run these workflows on their own data and visualise the results. Taverna is an open-source workflow tool which provides a workflow language and graphical interface to facilitate the easy building, running and editing of workflows over distributed computer resources.

All personal, organisational and experimental data in myGrid are stored in a central repository known as the myGrid Information Repository (mIR). The mIR is based on the myGrid Information Model which defines the basic concepts through which different aspects of an e-science process can be represented and linked.

Code Analysis, Repository and Modelling for e-Neuroscience (CARMEN)[355] is an e-science pilot project funded by the UK Engineering and Physical Sciences Research Council. It will deliver a virtual laboratory for neurophysiology, enabling sharing and collaborative exploitation of data, analysis code and expertise. The project involves 20 scientific investigators and practising researchers – neurophysiologists, neuroinformaticists and computer scientists at 11 UK universities who are addressing the complete life-cycle of neurophysiology data.

The aim of the project is to provide a web-based computing infrastructure to enable integration of data, software and knowledge from distributed neuroscientists. These innovations embody a virtual neuroscience laboratory, linking experimental and analytical neuroscientists in a translational pipeline which challenges contemporary neuroscience, offering potential for rapid and expedient advancement.[356] The primary advantage of CARMEN is to reduce the requirement for expensive and often ethically contentious experimentation, by allowing maximum benefit to be derived from experimental data and analysis methods.

The Repository for the Laboratory (R4L)[357] project aims to take aspects of the work of the eBank project a step further by setting up systems that will enable machine generated data to be semi-automatically deposited in institutional laboratory repositories. The scope of the project extends beyond crystallography to other areas of chemistry. Links will be made between datasets and the resultant publications and a report-editing tool allows automatic production of statistics, tables and graphs to assist in the preparation of publications.

[352] http://www.jisc.ac.uk/whatwedo/programmes/vre2.aspx
[353] http://www.myexperiment.org/
[354] http://www.mygrid.org.uk/
[355] http://www.carmen.org.uk/
[356] Martyn Fletcher, et al., *Neural Network Based Pattern Matching and Spike Detection Tools and Services – in the CARMEN Neuroinformatics Project* (2008)
[357] http://r4l.eprints.org/

The omixed[358] project has developed a suite of tools for supporting experimental data management in multi-site, multi-omics studies. It is a collaborative project between the NERC Environmental Bioinformatics Centre (NEBC) and the University of Manchester. omixed lets experimental data, annotation and associated files be stored on a secure server and provides controlled access via a standard web browser. omixed offers a highly customisable data model and specific facilities for collaborative working. It can be easily integrated with existing data analysis pipelines as all of its functionality is exposed via a web service interface.
In the **United States**, the NSF has established the five-year initiative Cyber-Enabled Discovery and Innovation (CDI)[359] to create revolutionary science and engineering research outcomes made possible by innovations and advances in computational thinking. Thematic areas of the project are:

- From Data to Knowledge: enhancing human cognition and generating new knowledge from a wealth of heterogeneous digital data.
- Understanding Complexity in Natural, Built, and Social Systems: deriving fundamental insights into systems comprising multiple interacting elements.
- Building Virtual Organizations: enhancing discovery and innovation by bringing people and resources together across institutional, geographical and cultural boundaries.

The first round of projects, funded in Autumn 2008, aim to connect data repositories with research workflows and publication processes in Virtual Research Environments.[360]

OpenWetWare[361] is a collaborative web resource that provides biological researchers with an online platform for storing, managing, and sharing primary and preliminary research data and know-how. It has a focus on information not typically published in scientific literature, such as control experiments and negative results – knowledge that is otherwise lost in offline lab notebooks or shared only within small communities, but can now be easily disseminated and analysed. OWW has applied for NSF funding to extend the services it is already offering.

In **Australia**, a number of VRE developments are under way. The AustLit project,[362] for example, supports Australian literary and print culture researchers. It offers comprehensive bibliographical records, access to federated data sources, specialist dataset creation, scholarly editing and publishing services and also community engagement and knowledge transfer via social and scholarly networking. The data analysis tools of the portal can present information on a map interface, user-tagging of query results and annotation of sources, and even astrological analysis of data contributors to the portal.

In Germany a number of VRE projects are ongoing. WIKI Next Generation Enhanced Repository (WIKINGER)[363] is an interdisciplinary research project where IT, engineering and humanities specialists are working together to develop new tools for knowledge creation and organisation. The pilot system is being developed for the studies of the Catholic Church in the 19th and 20th century, but the final tool set should be independent of discipline and geographic location. The WIKINGER tools use semantic networking to connect to a research discipline's core data repositories and textual databases.

The eSciDoc[364] is a joint project of the Max Planck Society and FIZ Karlsruhe, funded by the Federal Ministry of Education and Research, and aimed at building an e-research platform for multi-disciplinary research organizations. eSciDoc consists of a set of infrastructure services that provide common functionality, and discipline- or task-specific applications. Each eSciDoc Solution is built on the eSciDoc Infrastructure. The infrastructure services ensure sustainability while services connect, disseminate, visualize, publish, manage, and work with data. The eSciDoc Infrastructure is implemented as a set of loosely coupled services that allows for a flexible configuration of the infrastructure in which services may be omitted or replaced with ones that are more adequate for individual disciplines. The eSciDoc Infrastructure provides services for object storage, search and indexing, statistics and reporting, persistent identification, workflows, validation, and

[358] http://www.omixed.org/
[359] http://www.nsf.gov/crssprgm/cdi/
[360] http://www.nsf.gov/awardsearch/progSearch.do?SearchType=progSearch&page=2&QueryText=&Prog Organization=&ProgOfficer=&ProgEleCode=&BooleanElement=true&ProgRefCode=&BooleanRef=true&ProgProgram=CDI-VIRTUAL+ORGANIZATIONS&ProgFoaCode=&RestrictActive=on&Search=Search#results
[361] http://openwetware.org/
[362] http://www.austlit.edu.au/
[363] http://www.iais.fraunhofer.de/704.html
[364] http://www.escidoc-project.de/JSPWiki/en/Startpage

transformation. As an example, the eSciDoc has developed a Scholarly Workbench that provides a generic solution for communities in the arts and humanities, to store their digital objects and make them processable and reusable within a collaborative environment.

Other similar projects include Im Wissensnetz[365] to develop eScience Semantic Desktops and StemNet[366] that is developing a knowledge management system for hematopoietic stem cell transplantation research. Both make use of ontologies and text mining tools to allow semantic analysis of literature and data sources.

The Dutch NARCIS portal[367] is combining the open access repositories and open datasets from DANS with the national Current Research Information System (NOD). NOD is a database with information on current research projects, researchers and research institutes covering all disciplines and giving access to university and non-university research information. Using a system of Digital Author Identifiers (DAI) assigned to every researcher in the Netherlands and persistent identifiers on objects in repositories, the datasets and publications in the various repositories are linked with NOD. This allows for fast connection between researchers in different domains and provides information about their involvement in different research projects.

A different angle on reasons for repositories to develop new services in support of research models and workflow is to maintain the momentum of content collection that often dwindles after the setting up of a repository. Academics often find a wide variety of reasons for delaying the archiving of their research output and institutional repositories have very little leverage to coerce them to do so. In order to sustain the growth of content in repositories, a consortium of universities and institutional repositories in Japan has started a project 'Integration and Presentation of Diverse Information Resources'[368] that will develop a range of tools for making more efficient use of digital repositories, and support academics in their conduct of research and production of research papers. The tools under development include: topic maps and tag clouds of institutional repository contents, co-evolutional academic research and education through activity-support layers between information stores, plug-in-based data conversion framework and integrated search tools.

> It is likely that the concept of Virtual Research Environments will keep evolving as new projects in this area are funded. The existing examples already automatically link data repositories and research workflows into researchers' workbench tools. It will be important to ensure that the future VREs are based on architectures that are extensible and support, but do not restrict, the resources needed by individual research teams.

3.5 DEVELOPING RESEARCHER SKILLS FOR SHARING DATA

Promoting data publishing and sharing is particularly challenging in a context where researchers tend not to think much about such issues as a matter of course. Data publishing to a standard that facilitates re-use requires effective planning and management of data through the life-cycle of a project. Research funding agencies around the world tend towards encouragement rather than enforcement of good data management practices. There is concern that progress toward effective data management and sharing is too slow.[369]

In the setting of scarce resources and time, and often lack of previous experience or knowledge of issues around good data management, researchers will need outside help to ensure that their project's data outputs can be disseminated. This help and guidance is available and, indeed, recommended to be used by funding agencies. Some funding agencies (e.g. NERC, ESRC) recommend that research projects get in touch with the designated data repository at the start of the project to begin a dialogue on good data management and documentation practices. In practice, such services are not used very often – the experience of RELU's programme-specific Data Support Service[370] shows that even when expert support is available, researchers

[365] http://www.im-wissensnetz.de/
[366] http://www.stemnet.de/?site=Projektziel&language=en
[367] http://www.narcis.info/?wicket:interface=:1::::
[368] Cf. Daisuke Ikeda, Sozo Inoue, *A New, Sustainable Model for the Institutional Repository: A CSI Project 'Integration and Presentation of Diverse Information Resources'* (2008)
[369] RIN, *To Share or not to Share* (2008) p. 29
[370] http://www.data-archive.ac.uk/relu/

will not necessarily avail themselves of it.[371] While established data archives and centres have the knowledge and know-how of data management best practice and can be incentivised by their funders to share this with data creators, institutional repositories at the moment have very little knowledge or guidance to offer, especially discipline-specific data management advice.

Best practice guidance has been accumulated into published manuals in some domains. For example, the ESDS has produced a number of guides on data processing, archiving, data consent issues and data documentation;[372] the Rural Economy and Land Use Programme has its *Guidance on Data Management*;[373] the ICPSR has published a *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*;[374] The Australian National University has published a *Data Management Manual*[375] for the Australian Partnership for Sustainable Repositories (APSR) programme, which comes with templates and checklists. Another guide published in Australia looks at the intellectual property rights and other legal issues when developing data management plans: *Practical Data Management: A Legal and Policy Guide.*[376]

The research project data management plans that have been made a condition for receiving funding from some research funding agencies in the US, Australia and the UK hold significant potential for making researchers more aware of data sharing issues, and ensuring that research projects have the necessary data management skills available. The requirements for data management plans would have to be supplemented with templates and concrete examples of how to make the requirements operational. The work on data management plans currently being published in Australia is leading the way in this area.

The Dutch social science data archive DANS has started a programme to advise and acknowledge partners who meet the quality guidelines for data by issuing a 'Data Seal of Approval'.[377] The Seal will show that the organisation is 'intending to ensure that in the future, research data can still be processed in a high-quality and reliable manner, without this entailing new thresholds, regulations or high costs'. The list of criteria contain 17 guidelines for the application and verification of quality aspects with regard to creation, storage and (re-)use of digital research data. The first requirement states:

> The data producer deposits the research data in a data repository qualified according to these guidelines.

The Digital Curation Centre (DCC) and the DigitalPreservationEurope (DPE) project have jointly developed a digital repository assessment method based on risk analysis. The DRAMBORA toolkit[378] has been piloted in over 20 digital repository settings, among them data centres. The outcome of a DRAMBORA assessment is a register of risks associated with management of digital objects; the risk register can be used to continuously manage the digital collection and mitigate the risks. The assessment methodology lends itself easily for use in everyday data management of a research project, or could be used in combination with the Data Audit Framework (DAF) toolkit by a university department or institute. JISC has funded the development of the Digital Asset Assessment Tool (DAAT)[379] which is more focused on digital preservation issues.

A number of recent reports have looked at the personal skills, training opportunities and career perspectives of researchers and data scientists.[380] All of these analyses recognise that there is currently a gap between the existing and needed levels of skills to ensure adequate data management for the long term. Researchers face a big and continuing challenge in remaining properly skilled because technology and data management tools are moving very quickly and they need to stay abreast of general developments and developments specific to their field. The discussion of best methods to teach these skills to researchers is ongoing.

The role of repository and library staff is increasing in supporting researchers in their data management. Most institutional repositories are located in libraries that have very little experience with research data and it is

[371] RIN, *To Share or not to Share* (2008) p. 27
[372] http://www.esds.ac.uk/support/onlineguides.asp
[373] RELU DSS, *Guidance on Data Management* (2006)
[374] ICPSR *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (2005)
[375] ANU, *Data Management Manual: Managing Digital Research Data at the Australian National University* (2008)
[376] Anne Fitzgerald, Kylie Pappalardo, Anthony Austin, *Practical Data Management: A Legal and Policy Guide* (2008)
[377] http://www.datasealofapproval.org/
[378] http://www.repositoryaudit.eu/
[379] http://www.ulcc.ac.uk/daat
[380] Alma Swan, Sheridan Brown, T*he Skills, Role and Career Structure of Data Scientists and Curators* (2008); RIN, *Mind the Skills Gap: Information-handling Training for Researchers* (2008); Mark Thorley, *Providing Appropriate Skills for Data Curation* (2008); Margaret Henty, Belinda Weaver, Stephanie Bradbury, Simon Porter, *Investigating Data Management Practices in Australian Universities* (2008); Margaret Henty, Developing the Capability and Skills to Support eResearch (2008)

clear that repository staff are required to advance their own skills set, too. The SHERPA project has collected main requirements for repository staff into a statement *Institutional Repositories: Staff and Skills Set*[381] that it intends to revise annually. From the researchers' point of view, the main expectations for repository staff are that they understand the research process in different domains, can make the business case for long-term archiving and curation of data to researchers, and can help to appraise and identify data of long-term value. In the United States the DigCCurr project,[382] although not focussed specifically on research data, is identifying and developing responses to the digital curation skills needed in the future.

> Awareness raising of data management, curation and sharing requirements, issues and available best practice is needed for both researchers and repository staff to ensure effective management of data through the life-cycle of a project and in a data repository.

[381] SHERPA, *Institutional Repositories: Staff and Skills Set* (2008)
[382] http://ils.unc.edu/digccurr/index.html

4. ANALYSIS

This report has analysed the top-down drivers for establishing data sharing infrastructures – policies, strategies and development plans – and charts examples of typical data sharing infrastructure provision in OECD countries. Over 300 initiatives, projects and repositories were looked at in the course of the analysis. The report, however, does not aim to provide a full inventory of all data sharing initiatives across all the OECD countries, but presents best practice examples that cover the range of approaches taken in different research domains.

The report has focused primarily on analysing the **enablers** of research data sharing, and less on the various barriers (see Table 8), but the instances of best practice referenced in this report demonstrate the approaches used to overcome the barriers.

Table 8. Typical enablers and barriers to research data sharing.

Enablers	Barriers
Funding provision	Lack of funding, resources and time
Policies	Restrictive policies
Professional codes of ethics	Lack of adequate data sharing infrastructure
e-Research strategies and planning	Incompatible data formats and description standards
Open Access principles	Lack of data management skills
Data repositories	Data and privacy protection regulations
Services in support of data sharing	Intellectual property rights
Good data management skills	Differences between disciplines and research domains
Cost efficiency calculations	Researchers' reluctance to share their research data

Sharing and reusing data are a form of **collaboration**, and it is through collaboration that the examples of effective enablers to data sharing have been established. Whether it is research agencies agreeing on a set of principles in a policy statement, or a project developing tools for sharing data within a specific research area, it is the declared vested interest and active collaboration that has delivered the result.

The number of stakeholders with interests in establishing the enablers for research data sharing is large, but the boundaries of their **roles** are often still unclear:[383]

> Effective coordination at both national and international levels will not happen by accident. It will require a strengthening of procedures and mechanisms and close collaboration between all the key agencies to ensure that there is clarity as to roles and responsibilities and awareness of new developments and opportunities, in order to avoid both wasteful duplication and damaging gaps in provision.

This analysis has structured the taxonomy data sharing initiatives into five **levels**:

- International and trans-national initiatives.
- National and government-initiated initiatives.
- Research domain and funding agency initiatives.
- Institutional initiatives.
- Individual project initiatives.

The distinctions between these levels are not always unequivocal: a research council may fund the development of services for several interdisciplinary research themes; many research projects nowadays are in some way international; projects that receive funding from international organisations, like the EU, are doing their work and developing services in individual countries in the context of their own institutions. The classification of initiatives proposed in this report follows two main categories: 1. degree of remoteness from the actual research data (e.g. a national level policy statement is further removed from the specific details of data management than an institutional or project level policy); 2. source of core funding for an initiative or the main target group of the developed services (e.g. a project developing an access portal to research output that harvests all repositories in one country is classified as a national level initiative).

This top-down, or hierarchical, taxonomy of stakeholders in data sharing infrastructure provision was used for comparison of both the policies and the services in different countries. Despite the differences that exist

[383] OSI e-Infrastructure Working Group, *Developing the UK's E-infrastructure for Science and Innovation* (2007), p. 13

between countries in terms of the models used for research funding, as well as the levels at which decisions are taken, there is agreement on the expected strata of responsibility for applying the instruments of data sharing. This supports the structure of stakeholder taxonomy used in the study.

Further stakeholder groups can be identified when research data sharing is analysed for other objectives, but have not been included in the current study for the following reasons:

- The general public, who, as taxpayers, are also investors in public research, have a strong interest in seeing that the results their investments yield are effectively managed and used. Issues of access to publicly funded information have been taken up on international level,[384] and are also being pushed by self-organised action groups and initiatives.[385] Direct involvement of the general public in establishing data sharing infrastructures is currently possible only through lobbying for better access provision and open access policies, and as end-users accessing data from repositories and through various access portals.
- Industry and private sector organisations are frequently users that benefit from re-use of research data, but have traditionally kept their own data outputs proprietary and inaccessible. However, private research institutions are increasingly outsourcing their research activities to universities and specialist research institutes. So far, the public-private research partnerships have tended to add to the complication of management of the resulting data, because of rights allocation and conditions for re-use. In this sense the participation of private organisations in research data sharing has been rather negligible, mainly through their funding for infrastructure building in research institutions that can also be used for sharing results from other research. The calls for open access to all research output and better tools for controlling (delayed) access to data hold promise to bring a change to this situation.

The study of data sharing initiatives in the OECD countries confirmed the traditional perception that the policy instruments are clustered more in the upper end of the stakeholder taxonomy – i.e. at the level of national and research funding organisations – whereas the services and practical tools are being developed by organisations at the lower end of the taxonomy.

4.1 POLICY SUPPORT FOR DATA SHARING

Chapter 2 of the report discussed data sharing policies that are being developed by stakeholders on all levels, from individual research projects to international organisations. The lack of a universal model for data sharing policies appears to be a fundamental consequence of research funding models differing between individual countries. This study found no evidence of either a universal model or agreement on what a data sharing policy should include.

Significant effort is being put internationally into developing policies and guidance by national level bodies, by research funding agencies, individual research institutions and key stakeholders in the international community. The long absence of coherent, accessible, and transparent data access policies has been seen as creating artificial barriers to interdisciplinary research and to effective data collections management:[386]

> Researchers working at the interface between disciplines can find themselves subject to conflicting data release policies and deposition requirements. Collections managers who work with multiple communities are often faced with differing rules for deposition, conflicting technical standards, and varying access restrictions. Development of a comprehensive set of policy statements for data access and release that provides for consistency and coherence across disciplines while meeting the distinct needs of individual disciplines and communities, that are transparent and readily accessible to the community, and that prevent unnecessary proliferation and duplication of standards could greatly facilitate progress in research, education, and collections management.

This situation is rapidly improving, especially since international organisations like the European Commission, OECD, CODATA and others have started to develop principles for open data sharing and agendas for implementing them. Research funding agencies have over years developed policies in support of their data centres and infrastructure, but are now in the process of formulating general data sharing policy statements.

[384] E.g., OECD, see: http://www.oecd.org/document/55/0,2340,en_2649_201185_35397879_1_1_1_1,00.html
[385] Cf. http://www.taxpayeraccess.org/
[386] National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (2005), p. 37

Data sharing principles are provided using similar models to those the funding agencies have used for their guidance for research ethics and protecting the privacy of research subjects.

On an **international level**, the key players (organisations like OECD, UNESCO, and the EU and interest groups like CODATA and ESFRI) have concentrated their policy statements around the principle of open access to publicly funded research outputs. While OECD, UNESCO and CODATA have policies explicitly for data sharing, the European Commission is looking at data sharing issues in the broader context of open access to public domain information.

No **national level** policies or strategic documents that explicitly mandate the sharing of research data were found. Nevertheless, the provision of access to research data is seen as a vital element of the general research infrastructure, and all research infrastructure development strategies acknowledge the need to develop the means for accessing data. Applying Open Access principles to data is discussed at the national level in Germany.

Because of differences in data collection, use and management practices in different domains of research, national level policies remain too general to be useful in practice. Typically, the national level strategic documents draw on the following incentives and deal with these aspects:

Table 9. Typical characteristics of national level strategy documents for data sharing.

Policy Enablers	Aspects Covered
International level examples and mandates	Vision for data sharing infrastructure
Open Access principles	Priority development areas
Government funding for research infrastructure development	Distribution of funding for infrastructure development
Shared principles agreed by research funding agencies	Roles and responsibilities of stakeholders
	Shared policy statements (e.g., Open Access)

National level strategy documents are being developed to set priorities for government spending on research infrastructure development. The significant rise in collaboration among funding agencies to agree on common principles and standards for access to research data is also becoming a trigger; this is supported by statements from international organisations and interest groups and the Open Access movement who are increasingly embracing primary research data. The collaborative effort of developing such policies has usually been led by an umbrella organisation (e.g. national research foundation, academy of sciences) or in some cases as a bottom-up process by initiatives from research communities (e.g. RIN in the UK). The main benefit of national level strategic documents is the identification of roles for developing further, more specific policies and data sharing services.

The main burden of developing and implementing data sharing policies is currently being carried by **research funding agencies**, with an expectation (but not a mandate) that individual research institutions and departments will follow these up with their own policy statements. Measures to motivate researchers to share their data incorporate conditions being attached to funding schemes or the provision of data sharing policies backed up by services offered to recipients of funding. The prospect of a more pro-active stance in mandating the sharing of data is evidenced in the recent initiatives of funding agencies to agree on common principles for data sharing.

Typically, but not in all cases, the funding agency policies draw on the following incentives and enablers:

Table 10. Typical characteristics of national level policies for data sharing.

Policy Enablers	Aspects Covered
International level examples and statements	General policy statements
National strategic planning documents and mandates	Obligation / mandate to share data
Research associations' statements and codes of ethics	Division of responsibilities between stakeholders
Open Access principles	What data sharing channels should be used
Government funding for research infrastructure	How can the costs involved in data sharing be covered
Government audit and watchdog offices' reports and requirements	What sanctions can be applied if the data sharing requirements are not being met
	Data access principles and protection of data subjects' rights
	Conditions of exclusive use of data

Funding agencies are well positioned to follow up on how research projects fulfil their policy requirements, but the practice is variable. Researchers in disciplines that have large centralised data centres benefit from the availability of expertise and resources for data curation, whereas other funding agencies often do not have efficient mechanisms in place for ensuring that their policies are being followed. A significant agreement of common principles and standards amongst the funding agencies for widening access to research data is being stimulated by statements from international groups including the Open Access movement.

A natural focal point where higher level policy requirements and incentives for researchers to share their data meet is at the **institutional and departmental level**. Therefore, creating data management and sharing policies on an institutional and/or departmental level would be the rational choice. These policies could still follow the broad requirements of national agencies and the research domain, but they would be designed for operation in the context of individual research projects. Institutions themselves have a vested interest in sharing data, and in some jurisdictions may share or own the intellectual property rights to the data, but institutional data sharing policies are not yet very common. Whilst growing awareness of the open access principles is increasing interest in methods for data sharing, most of the existing institutional level policies for openly sharing research outputs do not yet incorporate research data.

Table 11. Typical characteristics of institutional level policies for data sharing.

Policy Enablers	Barriers / Aspects Covered
International level examples and statements	General policy statements
National strategic planning documents and mandates	Obligation / mandate to share data
Funding agency rules and policies	Responsibilities of stakeholders
Research associations' statements and codes of ethics	What data sharing channels should be used
Open Access principles	How can the costs involved in data sharing be covered
Publishing and publicising the institution's research output	Data access principles and protection of data subjects' rights
Traditions and agreements in research domains	Conditions of exclusive use of data

The emerging institutional policies still remain *ad hoc* and do not appear to be well coordinated. To develop uniform data sharing policies and put them into practice, the institutions will currently require significant help and guidance. Services are being developed (e.g., Data Audit Framework) that facilitate this work, and the number of published data management guides is on the rise.

The analysis of data sharing policies on different levels demonstrates that for the policies to be effective and operational, they should be as close to the actual research and researchers as possible. Whether this means an institutional level policy or a funding agency policy will depend on the research funding model used in a given country and existing services for data sharing. Ideally, the policy should be accompanied with guidance and case study examples, helping researchers to comply with the policies.

Several calls for more uniform data sharing policies have been made to facilitate shared principles across interdisciplinary research boundaries. The Government Accountability Office in the US that looked at data management and sharing policies and practice in four funding agency domains concluded that:[387]

> While the policies generally underscore the importance of making data openly available at minimal cost, we found that they vary among and within agencies because they are often tailored to the needs of different research programs or projects within the same agency. Policies must take into account their applicability to specific research projects, relevant legal and regulatory restrictions, the existence of appropriate archives, and the characteristics of particular research fields.

Suggestions for what a data sharing policy should include have also been discussed in several countries. The RIN report *Stewardship of Digital Research Data* sets out five principles for good data management:[388]

[387] GAO, *Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research* (2007)
[388] RIN, *Stewardship of Digital Research Data - Principles & Guidelines* (2008)

> I. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with these principles.
> II. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.
> III. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.
> IV. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.
> V. Digital research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations.

These principles are expanded in the report with guidance on how policy and practice may need to be changed to ensure that they are to comply with the principles. The report elaborates upon how these five principles provide a broad framework for developing good practice. It calls for a coordinated approach to provide a cohesive framework of policies and procedures for key agents and stakeholders in order to maximise the potential benefits of digital research data.

The quoted GAO report in the US suggests that:

> The scientific community generally believes that data-sharing policies should address how data are to be made available and that, when the appropriate infrastructure exists, data acquired in federally funded research should be made accessible through unrestricted archives.

Hence, data sharing policies should, as a minimum, address:

- The cost of making data available.
- When data are to be shared.
- How data are to be shared.
- What data are to be shared.

A suggestion for a research funding agency's open access policy is to:[389]

- Establish clear policy to mandate their researchers to deposit their research outputs.
- Establish clear policy to provide the funding for open access publishing – make them part of research costs.
- Support and/or create repositories.
- Provide clear advice to researchers and provide it again.

The open access discussion has also reiterated the conclusion that both RIN in the UK and GAO in the US have made, that policy is the easy bit – resourcing, implementation and enforcement is more difficult.[390]

4.2 COSTS OF DATA SHARING INFRASTRUCTURE

Data management and sharing needs long-term vision and long-term support that individual institutions and projects alone cannot provide. The significant costs of data sharing infrastructure provision have mostly been borne by national governments who continue to support directly the (centralised) services and participate in funding the research domain level and institutional data curation and dissemination services. With new models for sharing research data appearing, the question arises of whose funds could or should be used for developing and maintaining the services on institutional, project and individual researcher levels. Cost figures of data sharing are also vital for budgetary planning purposes on all levels. Yet real cost figures are hard to obtain as data sharing is 'bundled' with other services, most often with archiving and the preservation of data.

[389] Robert Terry, *Open Access and the Wellcome Trust* (2006)
[390] Peter Murray, *Open Access to Research Data: Surmountable Challenges* (2007)

To estimate the cost of curating and making data available for re-use institutions first need to take stock of their data resources. Tools like the DCC's Data Audit Framework (DAF) help with the identification of data assets, and ULCC's DAAT and DCC/DPE DRAMBORA are of value in assessing what risks are being faced in managing and curating them. There is a host of existing data stored in institutions that potentially have a considerably larger need for collection, curation and dissemination. This cannot be achieved without significant cost and effort – as the NDAD's collection and archiving of legacy datasets, and the Data-Pass project led by ICPSR have shown, the retro-curation or digital archaeology of data can be extremely costly.

4.3 DATA SHARING INFRASTRUCTURE PROVISION

The infrastructure that researchers can use to make their data available for re-use exists on several levels:

- Individual researcher's computing facilities, including blogs, wikis and other social networking software
- Computing facilities of the project, department or institute where the researcher is working
- Institutional repositories where research outputs are deposited
- Discipline-specific national or international data repositories
- Virtual Research Environment tools and services
- National e-science grid services
- International grids and e-research infrastructures

There is growing support on both international and national level for using grid technology not only to process and store research data, but also for sharing and curating data. Middleware tools, description standards and protocols are being developed that allow interacting with data stored in e-science grids as part of the research process. The great majority of these services are still experimental and project-based.

The proposals for national data services have opted for a distributed, umbrella-type approach where a national service provides the environment for repositories – common principles and standards that data repositories in the country apply. The main expected outcomes are better data curation and dissemination services that are based on shared tools and principles. A systematic and comprehensive approach to develop national research data management services has been taken in Australia with the planned Australian National Data Service (ANDS). A similar analysis and planning exercise has been initiated in Canada.

Data repositories for specific research domains are currently the predominant type of data repositories. In some disciplines there is a long tradition of depositing research data in national repositories. Many of these data repositories have historically started out as projects, but have succeeded in institutionalising the project databases into data archives with a more secure funding. Most frequently it is the research funding agencies that now maintain the data repositories, but it can also be individual institutions or direct government funding. Some long-established data archives (e.g. ICPSR, UKDA) have succeeded in offering their services to several funding agencies and other organisations in need of digital preservation and data dissemination services.

There is no obvious case, as yet, for replacing the existing data curation and sharing infrastructure based on data centres supported by research funding organisations. Institutionally distributed local data storage and sharing may provide a more agile approach, with the advantage of closeness to researchers, but a key disadvantage is the current shortage of expertise and resources at institutional level. On the other hand, domain-specific data repositories hold materials grouped by subject, data types or purpose commonly used in the discipline and thus intrinsically support domain- or discipline-oriented research needs. Domain-specific repositories act and speak on behalf of their designated communities and strive to provide support throughout the data life-cycle. As part of their key missions, they seek to know what the community wants and expects in terms of content, format, delivery options, support, and training.

The institutional infrastructure for data sharing is only emerging. The existing institutional repositories aspire to be more than a stockpile of e-prints and provide safe harbours for a more inclusive range of the intellectual output of local research and teaching. These repositories retain institutional identity, but may choose to limit access to specific sub-communities within the institution instead of public, web-wide open access. Institutional repositories are more interested in collecting research materials near the end of the research life cycle, employing an acquisition process that is simple and a preservation process that is not designed to

support complex content.[391] These limitations may decrease over time as the institution-based data repository movement matures.

In the main, universities and research institutes are either not ready, or it is not appropriate for them, to take the task of data sharing from centralised domain-specific data services. Clear policies, more resources and more skills are needed to allow universities to enter the data management and curation realm, but with the choice of data sharing channels expanding, the physical location of data becomes less and less relevant: access and dissemination services can harvest data from a variety of repository environments. Given that one method of data sharing does not preclude the use of other ones, and ultimately it is the researcher who decides which (additional) channels to use for dissemination, the university and institutional data repositories and social networking web services may become more popular in the future. It would still be in the interest of research funding agencies to ensure that data created with their funding are released to the public domain, are adequately described, curated over time, and the necessary data security rules are effectively applied. The centralised data repositories will continue to provide such a data control regime, but they should consider implementing mechanisms that allow institutional repositories to harvest metadata and link to actual data in their repositories, giving the institutions an opportunity to disseminate the data as linked resources.

The IT infrastructure offered to the researchers by the project, department or institute is already being used for sharing of preliminary data, subsets of larger datasets and results on individual experiments during the course of a research project. The centrally funded data repositories do not cater for this stage of the research process. The social networking software solutions like blogs and wikis are a popular choice for this type of data sharing and the emerging Virtual Research Environment tools are taking advantage of their widespread use to fill the gap in the data sharing cycle.

An example of how different scenarios for building data sharing infrastructures and an assessment of their value to different stakeholders has been developed in the medical sector[392] (see Figure 12 on the next page). This matrix could be further developed into a generic template that can be used by individual institutions and research domains or funding agencies to analyse their own requirements, the pros and cons of possible scenarios for data sharing and adequate division of roles and responsibilities.

[391] Cf. Ann Green, Myron Gutmann, *Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives* (2006), p. 4

[392] Heather Piwowar, Michael Becich, Howard Bilofsky, Rebecca Crowley, *Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers* (2008)

Figure 12. An example of analysis of data sharing scenarios.

Questions	Attribute	Description	Example	Impact for Data Producers	Impact for Data Consumers	Impact for Other Stakeholders
Where are the data stored? How are the data integrated with other data sets?	Centralized	Multiple datasets hosted at a single location in a common format	The Cancer Genome Atlas (TCGA) Data Portal (http://cancergenome.nih.gov/dataportal/)	Sharing often facilitated by well-developed interfaces.	High visibility, easy retrieval, easy aggregation within repository.	Requires funding of centralized repository development and maintenance, often limited to common data types.
	Federated	Physically separate datasets that use information technology to provide a virtual common dataset	Cancer Biomedical Informatics Grid (caBIG) (https://cabig.nci.nih.gov/)	Limited to federation participants. Often requires strict data standards.	Relatively easy retrieval and aggregation for federation participants.	Requires funding of relatively complex infrastructure and participant adoption.
	Distributed	Physically and virtually separate datasets	Data posted on Web site, as supplementary information, or emailed on request	Control retained over location, format, and data elements.	Low visibility, often difficult retrieval, interpretation, aggregation, consistency, and sustainability.	Requires no centralized funding. Allows only ad-hoc access control. Rarely maintained long term.
What control is placed on access to the data?	Open	All data can be viewed and reused by anyone	Single Nucleotide Polymorphism database (dbSNP) (http://www.ncbi.nlm.nih.gov/projects/SNP)	Open sharing of all data, no opportunities for decreasing security risks.	Easy and open participation for all investigators and project types.	Maximizes potential benefits of reuse. Appropriate for non-sensitive datasets.
	Hybrid	A subset of the data is provided openly, while other data are available only to permitted individuals through access or reuse limitations	Database of Genotype and Phenotype (dbGaP) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap)	Allows efficient and appropriate reuse of all data, provides opportunity to limit risks for sensitive subsets.	Easy and open participation for low-risk data; additional steps and qualifications required for complete data access.	Maximizes reuse while providing mechanism to protect sensitive data subsets. Requires ongoing access-granting role.
	Controlled	Only permitted individuals can access the data	National Institute of Mental Health (NIMH) Human Genetics Initiative (http://nimhgenetics.org/)	Allows appropriate sharing of very sensitive data; risks are minimized.	Data available for appropriate reuse; access permission is relatively time-consuming and complex.	Necessary wherever privacy and security of the data are a major consideration (e.g., de-identification can not be guaranteed).
When access is controlled, who determines permissions?[a]	Local	Access decisions for external investigators made by local data stewards on a study-by-study basis	The Cancer Text Information Extraction System (caTIES) (http://caties.cabig.upmc.edu/)	Local data producers are comfortable because they retain control, requires ongoing access-decision role.	Equity depends on local adherence to formal guidelines, otherwise decision-making may appear ad-hoc and opaque.	Facilitates gradual transition from sharing within a community to sharing more openly, as organizations gain comfort with risks and benefits.
	Central	Access decisions made by a usage committee or central source of authority	Shared Pathology Informatics Network (SPIN) (http://spin.nci.nih.gov/)	Data providers and custodians surrender control decisions, must trust central authority.	Equity depends on central adherence to formal guidelines.	Enables binding decisions across a diverse community.

In the longer term there is a need to produce and adopt universal rules for data description, to define minimum data curation services, and to identify rules for data security that are designed for use across different disciplines and repository types. The implication here is for more collaboration and the provision of more practical tools for use at the institutional level, with lessons learned from the experiences of established data curation institutions and centres.

4.4 DATA FEDERATION AND ACCESS SERVICES

To support interdisciplinary research researchers need to have access to data from a variety of disciplines and sources. Access tools are the last, critical step in supporting participation in data-driven research. These tools can be for individual dataset or bulk level finding, downloading, analysing, associating, tagging and annotating of data. While the domain data repositories have tailored their access tools for the specific needs of their disciplines, the institutional repository search tools make it difficult or in many cases impossible to query the institutional repository explicitly for datasets. Even the advanced search fields of the repository management systems have rarely been configured to allow specifying the type of the searched object.

Most OECD countries either already have or are in the process of developing a national access portal to e-prints and other content held in institutional repositories. Research data is sometimes part of these plans, but the added complexities of different metadata requirements and the lack of data citation rules do not hold promise for the e-prints portal becoming the main access gateway to data. The cross-domain repository search tools (e.g. OpenDOAR, OAIster) allow searching for datasets in institutional repository catalogues using OAI-PMH, but suffer from the limitations resulting from the varying practices of accepting datasets into institutional repositories and describing them.

Significant resources are being invested in developing domain-specific data access, federation and analysis tools. Despite being frequently developed by short-term projects, institutions are taking the commitment for supporting these services as valuable resources for the academic community. The domain specific data access portals are also aimed at harvesting metadata and content from other, international sources, to make it available to researchers at the host institution or within the country. Access to both data and publications through the same search is increasingly becoming a standard service. The ability to link to international repositories and different types of resources depends crucially on the interoperability of metadata schemas used within and across domains. Agreement on metadata schemas and protocols for information exchange is one of the key outcomes of the international services in development.

Different research domains use different metadata schemas to describe their data. Domain-specific metadata schemas, mark-up languages and adaptations or extensions of widely used metadata standards continue to be developed and are effective in supporting data exchange within a single domain. A few of these metadata standards are used in interdisciplinary data exchange, e.g. DDI, OAI-ORE,[393] eBank UK XML schema,[394] but to achieve interdisciplinary interoperability, more effort needs to be invested into developing tools for semantic interoperability between domains of science and metadata broker technology needs to be advanced.

As repository projects have matured, it has become clear that persistent identifiers are crucial for managing large numbers of digital objects over time. Persistent identifiers (PID) help manage repositories by separating the identification of resources from their physical location, overcoming the problems of changing and disappearing URLs, researchers moving between institutions and projects, and different concurrent versions of a dataset in a repository. Persistent identifiers are also a significant component in enabling data citation that plays a vital role in motivating researchers to share their data via repositories. The existing PID and referencing solutions (e.g. URN, DOI, Handle.net) are being used in data repositories, but the number of projects exploring the potential alternatives for specifically data repositories is confoundingly small.

4.5 DATA SHARING AND VIRTUAL RESEARCH ENVIRONMENTS

In the past, data repositories have not succeeded in building tools for efficiently feeding into the research process. Researchers perceive the cumbersome process of finding data from a number of different repositories a barrier that could be solved with additional services. On the other side there is sometimes an unfortunate and artificial distinction drawn between scholarly publication and the research process. In the Internet age scholarly communication increasingly permeates the processes of research and scholarship and services of data repositories are intertwined with scholarly communication.

A growing number of Virtual Research Environment projects are funded that automatically link data repositories and research workflows into ‘researchers’ workbench’ tools. The potential for automating the research workflow and harvesting data directly from repositories is huge in most disciplines. In the long-term perspective, the embedding of data repositories within the research workflow is of critical importance to the success of data repositories. However, the VREs need to be based on architectures that are extensible and support, but do not restrict, the resources needed by individual research teams. There is a clear need for well-defined data deposit and harvesting APIs for repositories on both research domain and institutional level.

[393] http://www.openarchives.org/ore/
[394] http://www.ukoln.ac.uk/projects/ebank-uk/schemas/

4.6 RESEARCHERS' ROLE IN DATA SHARING

Whether and how a research project's data will be shared in practice depends in large part upon the prevailing attitudes and cultures in the research domains. A barrage of obligations, requirements, incentives and suggestions for researchers to share their data is coming from different directions:

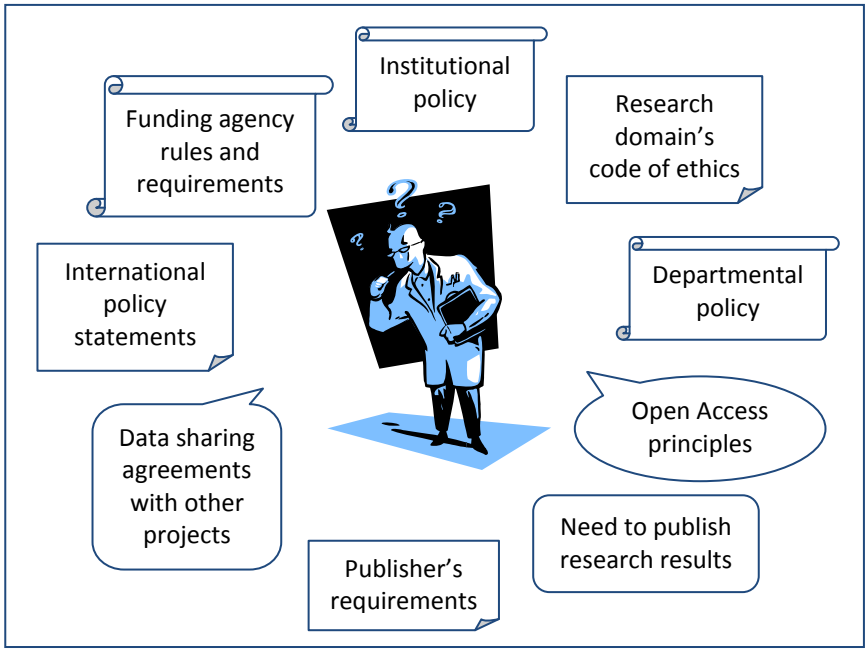

Figure 13. Incentives for researchers to share data.

The policies and conditions of grant produced by the funding agencies are but one among the many reasons for researchers to decide on sharing their research data. Researchers' awareness of these data sharing policies has been reported to be low (a UKRDS survey returned only 66% positive responses[395]) and should be improved, yet researchers have other incentives and requirements to share their data: principally, they want to publicise the results of their research, which in some cases includes data; some publishers require data underlying an article to be made available on request to other researchers; codes of ethics and agreements may require data sharing with other research projects in the same area; adherence to open access principles, which increasingly are applied not only to printed materials, but all other types of research outputs, and other informal or internal agreements can motivate data sharing.

Andrew Treloar, director of the Australian National Data Service project has made a graphical presentation of the general transitions research data makes when it is shared (see Figure 14).[396] For the researcher, migrating data from one domain to the next is associated with real costs, both in terms of time and funding. The private domain is typically characterised by having less metadata, more items, larger objects that are often continually updated, researcher management of the items, less concern for preservation, mostly closed access and little exposure of data. In order to open up a subset of research results to other researchers to access and analyse, systems that support greater structuring of the data collections, as well as more sophisticated access controls are required. Virtual Research Environments are emerging that allow this level of interaction, but in the meantime collaboration environments like wikis and blogs are frequently used for data sharing within a close group of collaborating researchers. The final transition is to a public data repository which is characterised by having more metadata than the collaboration domain, fewer items, smaller objects that are static or derived snapshots, and for which the repository offers organisational management, preservation services, open access and exposure of metadata for harvesting.

[395] UKRDS, *UKRDS Interim Report* (2008), p. 49
[396] Andrew Treloar, Cathrine Harboe-Ree, *Data Management and the Curation Continuum: How the Monash Experience is Informing Repository Relationships* (2008)

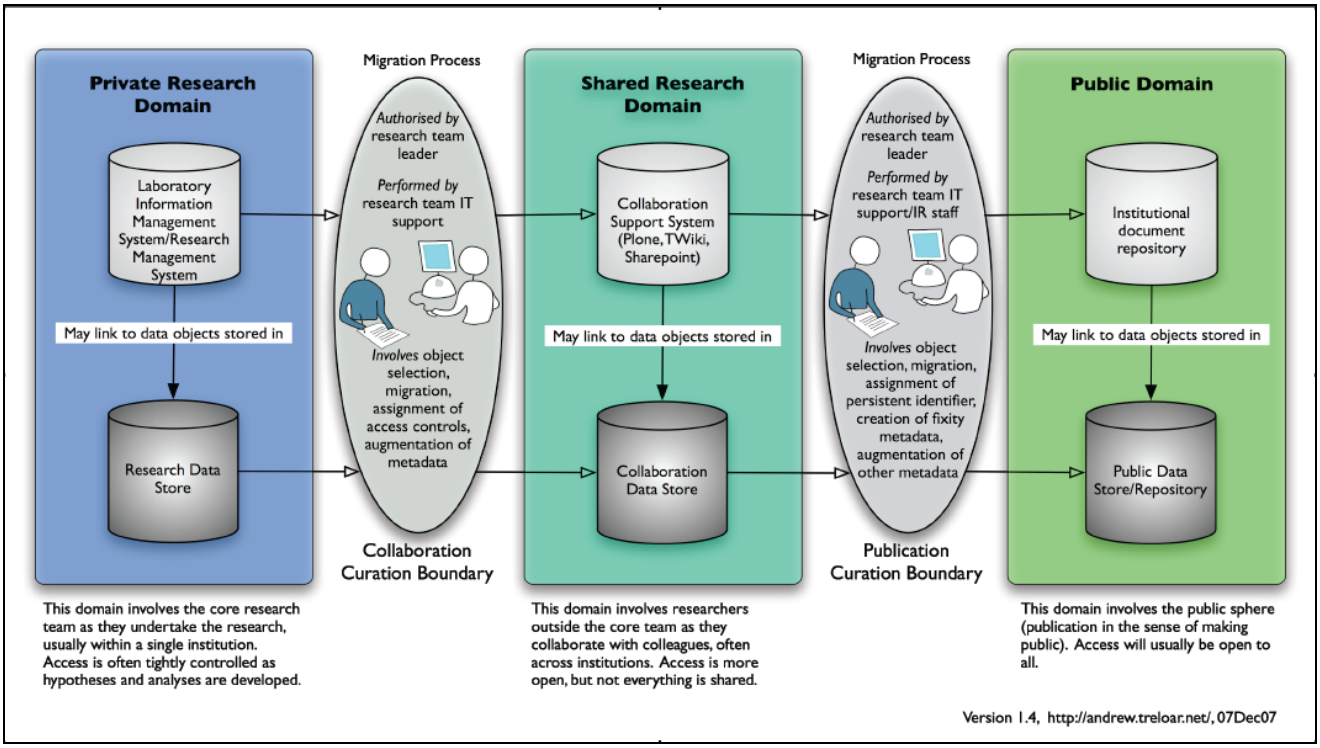

Figure 14. Data transitions between private, shared and public domains.

The incentives for researchers to manage their data in the private research domain so that they can later be shared, are mostly linked with the need to keep the research process on schedule, and also concrete requirements from some funding agencies (i.e. data management plan requirements). The enablers for data management throughout the research process are the data management skills the researchers possess, project data management plans, guidance from domain data archives and data curation specialists, and the emerging virtual research environments in various domains. The infrastructure for data management in the first domain is predominantly provided by the research institutes and in some cases national and international solutions, e.g. grid technologies, are used.

The motivation for opening their data up for collaboration with other research teams would come from project funding rules, research requirements and the wish or need to share data, for example for verification purposes. The enablers for releasing data from the private domain to the shared domain are currently not that many and rely either on institutional infrastructure and services or use the Virtual Research Environments where these exist. A few subject specific data centres have the facilities for uploading data into researcher's private area and giving access to it there to selected other users of the same system. The introduction of this report alluded to the inconsistency in the use of the term 'data sharing', and the interviews carried out as part of this analysis clearly demonstrated that for some researchers making their data accessible to a few colleagues epitomises all the data sharing that is necessary. This is very often done as requested and on an *ad hoc* basis using tools that seem appropriate for both parties at the time. The data documentation is usually provided on the level that is understandable to fellow researchers, but is insufficient for the general public.

The obligation to place their data in the public domain is reiterated to researchers through funding agencies requirements, international interest groups' statements, professional codes of ethics and institutional policies. Yet the incentives to do this are not always so clearly spelled out. In some sciences, publishing results and data as soon as possible is a matter of competition with other research teams and has been likened to adding a new brick in the common wall or a new link to a chain that is being built internationally. The type of international and discipline-specific science data centres and collections where the data is deposited are usually intended to be used by other researchers in the same or neighbouring disciplines. Publishing data in open data repositories would normally require more effort from the researcher, but the institutional repositories often resort to minimum description of archived objects

Policies and requirements alone will not result in a higher use of research data. Optimum accessibility and usability of data presuppose a trajectory of proper organisation and curation of data throughout its life-cycle, with access services and analysis tools that provide the researchers with added value to raw data. But in order

to make data repositories an effective means of data sharing, researchers need to be convinced that preparing their data for online publication is a worthwhile effort. The research world is strongly focussed on publication as an outcome – publications, citation rates and impact factors are the traditional research assessment indicators. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to reputation and ranking among peers. First attempts are being made (e.g. in Germany) to change the research assessment rules to also include data publications. If general agreement was to be reached on this then researchers would have an increased incentive to share data and adhere to various policy requirements, although mechanisms for ensuring the quality of data publications would first have to be put in place. To achieve the rank of a publication, a data publication needs to meet two main criteria: persistence and quality.[397]

Decisions affecting the practice of good data management at the level of an individual research project are influenced by many factors in addition to data sharing policies. The act of creating data management plans (required increasingly to accompany new project funding proposals) has the potential for incorporating structured guidance on how research data should be managed throughout their lifecycle. Examples of data management plans have been published in Australia and the US. A further step could be to link the data management requirements with data curation models such as the DCC Curation Lifecycle Model (see Figure 15)[398] and other models that are being produced in several countries (see an example from US in Chapter 2.2 above).

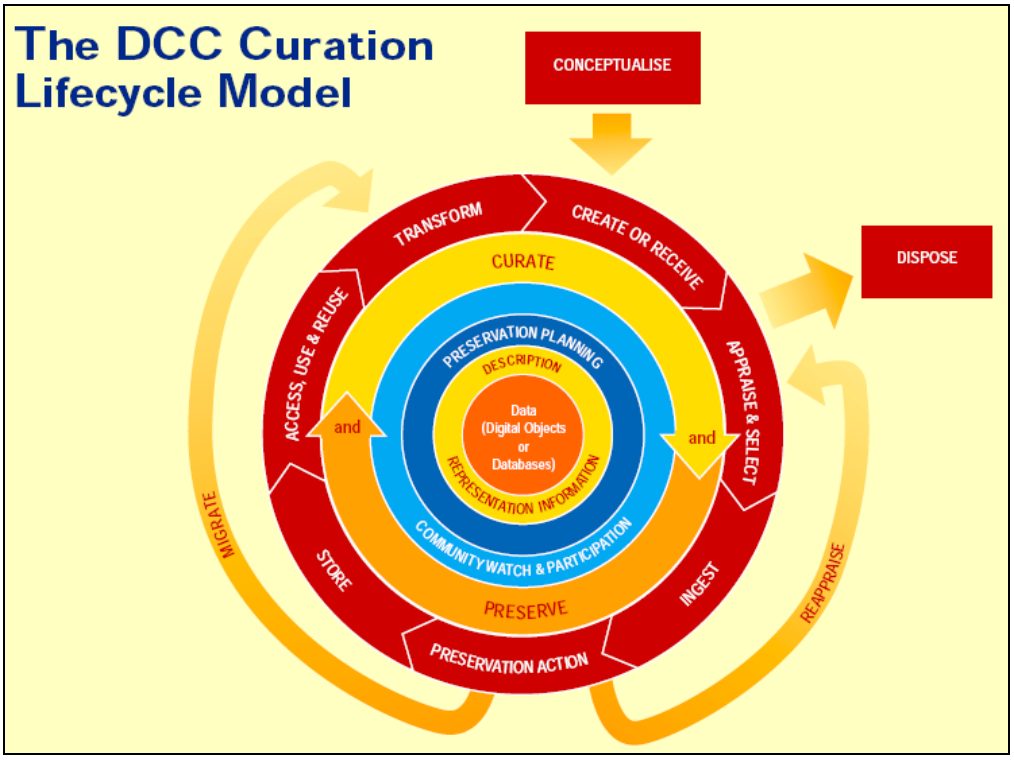

Figure 15. The DCC curation lifecycle model.

Different research disciplines have developed generic models of their research process workflow:

[397] Jens Klump, et al., *Data Publication in the Open Access Initiative* (2006)
[398] http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

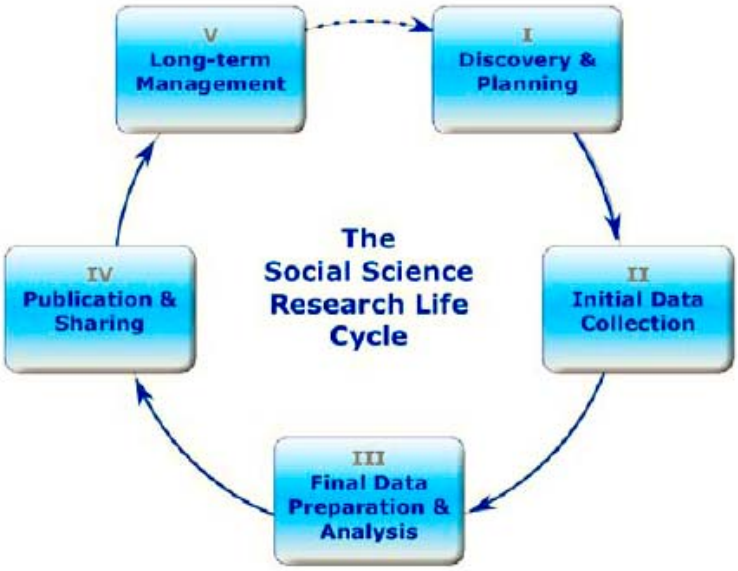

Figure 16. Social science research life cycle.[399]

And first attempts have been made to link the models of research process and data management tasks associated with the stages in the research workflow:[400]

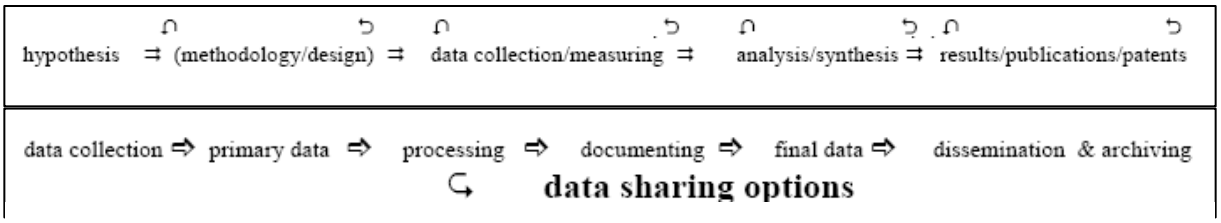

Figure 17. The research and data trajectories (Uhlir & Schröder 2007).

These models demonstrate vividly that dissemination-ready data are created through partnerships among data producers, data analysts, data archivists, and technicians, and that these processes take place over several stages in the research life cycle, and not merely at the end:

[399] Ann Green, Myron Gutmann, *Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives* (2006), p. 4
[400] Paul Uhlir, Peter Schröder, *Open Data for Global Science* (2007) pp. 38-39

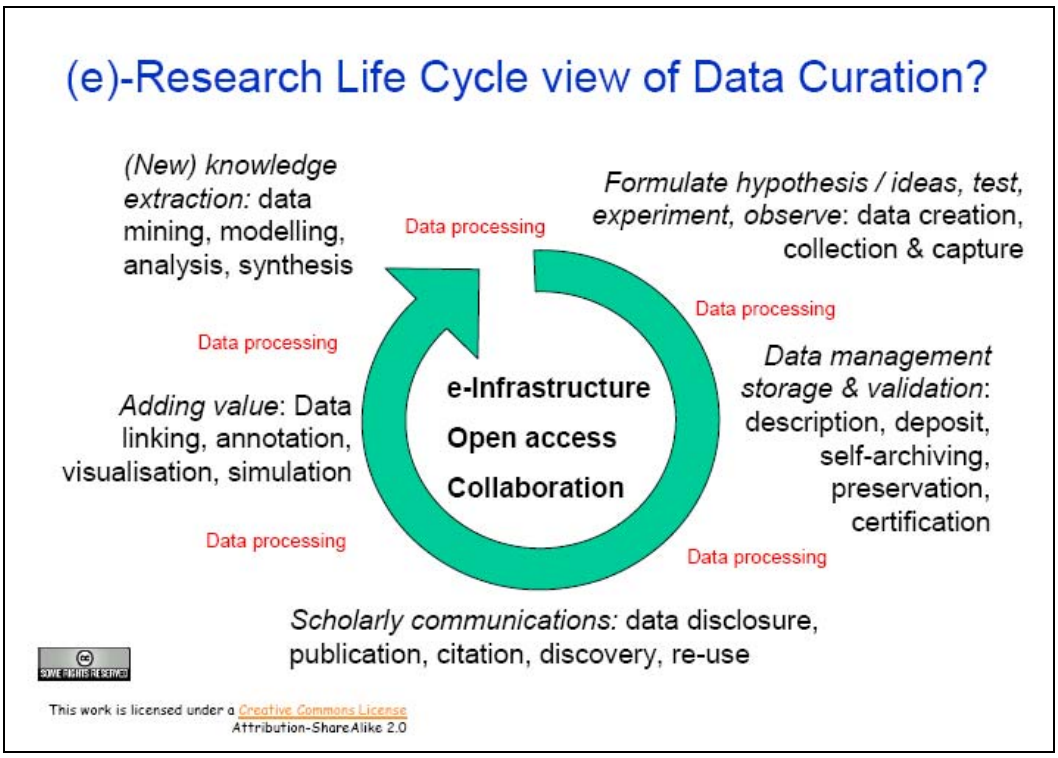

Figure 18. e-Research Life Cycle and data curation (*Dealing with Data*, 2007).

These parallel data lifecycle and research workflow models should be explicitly linked to indicate the data management and sharing tasks and responsibilities of the parties involved at each stage of the data life-cycle. Such a model should form the basis for good data management practice guides to researchers, and for awareness raising of data sharing issues and benefits.

In the near future the preferences of researchers will very likely be for tools and repositories that can support the research process, offer advice and are pro-actively involved with the researchers' needs. Repository and library staff need to live up to these tasks and reconceptualise their role in the scholarly communication system from publishers to active collectors and advisers.

4.7 FUTURE DATA SHARING INFRASTRUCTURE MODELS

What would be the ideal architecture of the data sharing infrastructure in the future? In an ideal world all data would be described using standard metadata, be stored in openly accessible repositories that take care of data curation, and be accessed with simple search and retrieval tools in any language. As long as automatic PID assigning and metadata extraction tools, and data curation services are freely available in the global data sharing infrastructure, the physical location of datasets is no longer significant and researchers can freely connect to them using automated research workflow components as 'building blocks'. The multilingual tools and services available to researchers have been certified as trusted, provenance of datasets can be traced using Current Research Information Systems (CRIS), and researchers receive due credit for having produced and published their datasets.

The data sharing infrastructure landscape of today falls short of the fully automated data re-use environment. Examples where domain level data centres provide infrastructure and services (storage, curation) to institutional data centres and other domains already exist (e.g. ICPSR, UKDA, the proposed collaboration between SHERPA-DP and AHDS that, however, was not realised after the AHDS funding was withdrawn). Several projects (e.g. ARROW,[401] SURFshare) have explored the models for interaction between institutional repositories, research management systems and researcher identity databases. Networks of repositories that provide their metadata for harvesting and content for federated use have been identified in several countries as the favoured model for building national data sharing infrastructure.

[401] See: ARROW HERDC Working Group, *Interim Report* (2008)

Putting the concerns of active data lifecycle management at the forefront of the argument, Gutman and Green[402] propose a layered approach where researchers, institutional repositories and centralised domain level data centres all have a role. A dialogue among the three should start from the planning stage of a research project and continue throughout its course. Documentation and ensuing articles from the project should be deposited with the institutional repositories, preliminary data could be shared through domain-specific tools supporting the research process, and the final datasets should be deposited with the domain data centre which takes the responsibility for its long-term curation and dissemination, while being connected with the contextual information (publications) that continue to be available from institutional repositories. This model would be a good compromise for the next stage of development of the research data sharing infrastructure. To realise it, all stakeholders will have to intensify their collaboration with each other.

[402] Ann Green, Myron Gutmann, *Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives* (2006), p. 9

5. RECOMMENDATIONS

The analysis in this report suggests that in order to improve the data sharing infrastructure provision in the UK, JISC and the wider stakeholder community should focus their future activities in the following areas:

CO-ORDINATION AND POLICY

- UK research funding organisations should jointly agree on data sharing principles and develop a set of common criteria for their data sharing policies.
- UK research funding organisations should each publish and impose a data management policy that is applicable to all grant holders.
- UK research funding organisations' data sharing policies should recommend that universities and research institutions develop their own data sharing policies.
- JISC, through the DCC, should develop, publish and promote a model institutional data sharing framework.
- Data sharing policies should recommend data deposit in an appropriate open access data repository and/or data centre where these exist.
- The Digital Curation Centre (DCC) should provide templates and assistance to institutions for the construction of data management and sharing plans that meet the requirements of the funding organisations.
- JISC should analyse the results from the PARSE.Insight survey results (due to be published in early 2009) to draw further conclusions for development of data sharing policies.
- JISC should monitor the development of European Union recommendations on open access to research outputs and public domain data, and produce guidance analysing the impact of these positions on the UK research community.
- JISC should explore the possibilities of institutionalising the current UKRDS project[403] into an office tasked with co-ordination of data sharing policy and infrastructure development.
- The DCC Research Data Management Forum,[404] the UK Data Forum[405] and other similar fora should collaborate in the identification and promulgation of key data sharing principles and practices.
- The research assessment rules need to be changed to include also data publications and data citations as criteria.
- JISC should commission a study to estimate the volume of legacy and orphaned data assets that are in need of curation and could be made accessible through existing data sharing services.

INFRASTRUCTURE DEVELOPMENT

- Policies alone will not result in a higher use of research data – to ensure optimum accessibility and usability of data, a coherent set of services for collecting, curating and accessing data needs to be defined and implemented as data management infrastructure.
- JISC should continue to support its repositories programme and more specifically enhance its support to the development of data repositories.
- JISC, through the DCC, should develop a template matrix for analysing possible scenarios for data sharing as a guide to development of the UK research funding organisations' data policies and services.

[403] http://www.ukrds.ac.uk/
[404] http://www.dcc.ac.uk/data-forum/
[405] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/nds/ukdf/default.aspx

- JISC should analyse current practice and develop new services for linking data objects and published research articles in repositories.
- JISC should commission a study to investigate current practice, assess future potential and evaluate the practical and legal issues associated with sharing research data through social networking software.

SERVICES DEVELOPMENT

- JISC should endorse the use of the DCC Data Audit Framework[406] to enable higher education institutions to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation.
- JISC should continue to support its Virtual Research Environment programme and extend it to investigate different mechanisms and services for managing data sharing across research disciplines.
- JISC should fund experimental case studies of archiving and making accessible legacy and orphaned research data assets to determine the cost and models for curating such data resources.
- JISC should develop an observatory of data sharing and dissemination tools that are available for use in different disciplines.

DATA MANAGEMENT PRACTICE

- Support should be given to services that facilitate work on data management plans and develop guidance and case study examples that help researchers to comply with data sharing policies.
- The JISC-commissioned DCC Data Audit Framework should be extended to cover data quality aspects and allow for assessment of the quality of research data assets.
- The DCC's expertise should be made available to aid researchers in developing effective data management and curation plans and practices.
- JISC should produce guidelines for good practice in data citation.

AWARENESS RAISING AND SKILLS DEVELOPMENT

- The DCC resources should be further employed to raise awareness of both data sharing policies and data management issues among researchers.
- The DCC should deliver co-ordinated training programmes and supporting materials, targeted at researchers in specific disciplines, to build data sharing skills and capacity within the sector.

The following recommendations made in the *Dealing with Data* report[407] can be endorsed and reiterated as a result of the analysis in this report:

- All relevant stakeholders should identify and promote incentives to encourage the routine deposit of research data by researchers in an appropriate open access data repository.
- Each funded research project, should submit a structured Data Management Plan for peer-review as an integral part of the application for funding.
- Each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy, which recommends data deposit in an appropriate open access data repository and/or data centre where these exist.

[406] http://www.data-audit.eu/
[407] Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007)

- There is a need to identify and promote scalable and sustainable operational models for data deposit, which are based on co-operative partnerships with researchers and common standards.
- JISCLegal should provide enhanced advice and guidance to the research community on all aspects of IPR and other rights issues relating to data sets.
- Work by JISC and the research councils, on developing model licences for data, should be co-ordinated so that a minimum set of standard licences are adopted more widely.
- More work is needed to identify integrated information architectures, which link institutional repository and data centre software platforms.
- The JISC should fund technical development projects seeking to enhance data discovery services, which operate across the entire data and information environment.
- The JISC should commission work to construct new economic models for preservation and data sharing infrastructure, to develop sustainable solutions.

APPENDIX. REFERENCES

Note: all URLs quoted in the report have been checked to be correct on January 9, 2009.

AHRC, *Research Funding Guide* (2008) http://www.ahrc.ac.uk/FundingOpportunities/Documents/Research%20Funding%20Guide.pdf

Alliance of German Science Organisations, *Priority Initiative 'Digital Information'* (2008) http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/allianz_initiative_digital_information_en.pdf

American Psychological Association, *Ethical Principles of Psychologists and Code of Conduct* (2002) http://www.apa.org/ethics/code2002.html

American Sociological Association, *Code of Ethics* (1997) http://www.asanet.org/cs/root/leftnav/ethics/code_of_ethics_standards

ANDS Technical Working Group, *Towards the Australian Data Commons: A Proposal for an Australian National Data Service* (2007) http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf

ANU, *Data Management Manual: Managing Digital Research Data at the Australian National University* (2008) http://ilp.anu.edu.au/dm/ANU_DM_Manual_v1.03.pdf

Arbeitsgruppe Open Access in der Helmholtz-Gemeinschaft, 'Realisierung des offenen Zugangs zu Publikationen und Daten aus der Helmholtz-Gemeinschaft' (2005) http://oa.helmholtz.de/fileadmin/Links/Plan_Open_Access_Realisierung__2005-02-03.pdf

ARM, *Data Sharing and Distribution Policy* (2006) http://www.arm.gov/data/policy.stm

ARROW HERDC Working Group, *Interim Report* (2008) http://arrow.edu.au/docs/files/arrow-herdc-interimreport-june08.pdf

Australian Research Council, *Discovery Projects: Funding Rules for Funding Commencing in 2008* http://www.arc.gov.au/pdf/DP08_FundingRules.pdf

BBSRC, *Data Sharing Policy* (2007) http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf

Neil Beagrie, *E-infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group* (2007) http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Neil Beagrie, Julia Chruszcz, Brian Lavoie, *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities* (2008) http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003) http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

BMBF, *Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik* (2001) http://www.bmbf.de/_media/press/A--FIN4_.pdf

Jean-Michel Bruguiere, *Droit des données publiques* (2008) http://www2.educnet.education.fr/sections/legamedia/droit-donne/

CESSDA PPP Presented at IASSIST Conference (2008) http://www.cessda.org/project/iassist2008.html

CIHR, *Policy on Access to Research Outputs* (2007) http://www.cihr-irsc.gc.ca/e/34846.html#5.1.2

Maxine Clarke, *Raw Data Policy of Scientific Journals? The Nature Perspective* (2007) http://www.esf.org/activities/science-policy/corporate-science-policy-initiatives/sharing-research-data-2007.html

Alexander Cooke, *Research Infrastructure and the Open Access Agenda*. Presentation at the Open Access and Research Conference (2008) http://www.oaklaw.qut.edu.au/files/Cooke.ppt

Juan Corrales Correyero, Alicia López Medina, *La red de repositorios institucionales del Consorcio Madroño ante el 'diluvio de los DATA'* (2008) http://www.consorciomadrono.es/noticias_eventos/2008/AliciaLopez.pdf

Council of European Union, *Council Conclusions on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (2007) http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf

Council on Governmental Relations, *Access to and Retention of Research Data: Rights and Responsibilities* (2006) http://206.151.87.67/docs/CompleteDRBooklet.htm

DANS, *Data Seal of Approval* (2008) http://www.datasealofapproval.org/

Data-PASS Metadata Requirements (2007) http://www.digitalpreservation.gov/partners/datapass/high/data-pass_metadata_requirements2007.pdf

Roberto Delle Donne, *CRUI and Open Access in Italy* (2007) http://www.aepic.it/conf/viewpaper.php?id=311&cf=10

Department for Cancer Epidemiology and Genetics, *Data Sharing Policy* (2008) http://dceg.cancer.gov/research/datapolicy

Department of Health and Human Services, *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule* (2004) http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf

Digital Curation Centre, DigitalPreservationEurope, *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)* (2007) http://www.repositoryaudit.eu/

DISC-UK, *DataShare: State-of-the-Art Review* (2007) http://www.disc-uk.org/docs/state-of-the-art-review.pdf

DRIVER, *A DRIVER's Guide to European Repositories* (2008) http://dare.uva.nl/document/93898

EPSRC, *Funding Guide* (2008) http://www.epsrc.ac.uk/CMSWeb/Downloads/Publications/Other/FundingGuideApril2008.pdf

ESFRI Working Group About Digital Repositories, *ESFRI Position Paper* (2007) ftp://ftp.cordis.europa.eu/pub/esfri/docs/digital_repositories_working_group.pdf

ESRC, *Data Policy* (2000) http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

ESRC, *Research Funding Guide* (2008) http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/ESRC_Research_Funding_Guide_June_2008_tcm6-9734.pdf

European Commission, *Communication on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (2007) http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf

European Commission, *The European Research Area: New Perspectives. Green Paper* (2007) http://ec.europa.eu/research/era/pdf/era_gp_final_en.pdf

European Commission, *Scientific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age*. Stakeholder conference 15-16 February 2007 http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1720&CFID=10805786&CFTOKEN=923bfcc24f508329-EB66FF6E-B7BB-326C-50432D6EF6A5CADB

European Commission, *Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe* (2006) http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

Stephen E. Fienberg, Margaret E. Martin, Miron L. Straf, (Eds.), *Sharing Research Data* (National Research Council, Committee on National Statistics, 1985)

Anne Fitzgerald, Kylie Pappalardo, Anthony Austin, *Practical Data Management: A Legal and Policy Guide* (2008) http://eprints.qut.edu.au/archive/00014923/01/Microsoft_Word_-_Practical_Data_Management_-_A_Legal_and_Policy_Guide_doc.pdf

Martyn Fletcher, Bojian Liang, Leslie Smith, Alastair Knowles, Tom Jackson, Mark Jessop, Jim Austin, Neural Network Based Pattern Matching and Spike Detection Tools and Services – in the CARMEN Neuroinformatics Project // *Neural Networks*, Special Issue on Neuroinformatics, Vol. 21, No. 8 (2008) http://www.carmen.org.uk/publications/CARMEN-Neural-Network-SI-Neuroinformatics.pdf

Fraunhofer Gesellschaft, *Open Access Policy* (2008) http://www.fraunhofer.de/fhg/Images/OpenAccessPolicy_tcm6-101804.pdf

Yukiko Fukasaku, *International Initiatives in Data Sharing: OECD, CODATA and GICSI* (2007) http://www.esf.org/activities/science-policy/corporate-science-policy-initiatives/sharing-research-data-2007.html

Cita Furlani, Chuck Romine, Chris Greer, *Briefing: Interagency Working Group on Digital Data* (2008) http://www7.nationalacademies.org/data/Data_Furlani.ppt

Cita Furlani, Chuck Romine, Chris Greer, *Interagency Working Group on Digital Data* (2008) http://iwg.cfa.harvard.edu/twiki4/pub/IWGDD/IwgddPresentationsOfInterest/IWGDD_Presentation_for_COS_Jan_25-08_FINAL.ppt

Government Accountability Office, *Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research* (2007) http://republicans.energycommerce.house.gov/Media/File/News/10.22.07_GAO_Report_Data_Sharing_Climate_Research.pdf

Ann Green, Myron Gutmann, *Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives* (2006) http://hdl.handle.net/2027.42/41214

Harvard University, Faculty of Medicine, *Guidelines for Investigators in Scientific Research* (1988) http://www.hms.harvard.edu/integrity/scientif.html

Rachel Heery, Andy Powell, *Digital Repositories Roadmap: Looking Forward* (2006) http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/rep-roadmap-v15.pdf

Margaret Henty, Developing the Capability and Skills to Support eResearch // *Ariadne*, Issue 55 (2008) http://www.ariadne.ac.uk/issue55/henty/

Margaret Henty, Belinda Weaver, Stephanie Bradbury, Simon Porter, *Investigating Data Management Practices in Australian Universities* (2008) http://www.apsr.edu.au/orca/investigating_data_management.pdf

Marienne Hibbert, *Australian Cancer Grid* (2005) http://www.apac.edu.au/communication_media/RDC_workshop/Aust_Cancer_Grid-Hibbert.pdf

High Level Group on Digital Libraries, *Position Paper on Digital Research Data Access and Preservation* (2008) http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/minutes/scientific_information/research_data_preservation.pdf

Human Brain Project, *Principles of Data Sharing* (2002) http://datasharing.net/Principles.html

ICPSR *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (2005) http://www.icpsr.umich.edu/access/dataprep.pdf

ICSU Ad hoc Strategic Committee on Information and Data, *Final Report to the ICSU Committee on Scientific Planning and Review* (2008) http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf

Daisuke Ikeda, Sozo Inoue, *A New, Sustainable Model for the Institutional Repository: A CSI Project 'Integration and Presentation of Diverse Information Resources'* (2008) http://ir.library.osaka-u.ac.jp/metadb/up/DRFIC2008/Ikeda.pdf

Steven Jackson, Paul Edwards, Geoffrey Bowker, Cory Knobel, Understanding infrastructure: History, heuristics and cyberinfrastructure policy // *First Monday*, Vol. 12, No. 6 (2007) http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1904/1786

James Madison University, *Data Stewardship Model* (2008) http://www.jmu.edu/JMUpolicy/12052.shtml

Japan Society for the Promotion of Science, *Handbook for the Use of Kakenhi Research-in-Aid for Scientific Research* (2008) http://www.jsps.go.jp/english/e-grants/data/handbook.pdf

Joint Task Force on Library Support for E-Science, *Agenda for Developing E-Science in Research Libraries* (2007) http://www.arl.org/bm~doc/ARL_EScience_final.pdf

Sarah Jones, Seamus Ross, Raivo Ruusalepp, *Data Audit Framework Methodology* (2008) http://www.data-audit.eu/DAF_Methodology.pdf

Gary King, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing // *Sociological Methods & Research*, Vol. 36, No. 2 (2007), pp. 173-199 http://gking.harvard.edu/files/dvn.pdf

Jens Klump, Roland Bertelmann, Jan Brase, Michael Diepenbroek, Hannes Grobe, Heinke Höck, Michael Lautenschlager, Uwe Schindler, Irina Sens, Joachim Wächter, Data Publication in the Open Access Initiative // *Data Science Journal*, Vol. 5 (2006) http://open-access.net/fileadmin/partner/helmholtz/klump_et_al_2006.pdf

Arja Kuula, Sami Borg, Open Access to and Reuse of Research Data – The State of the Art in Finland // *Finnish Social Science Data Archive Reports*, no. 7 (2008) http://www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs07_OECD_en.pdf

Leibniz-Gemeinschaft*, Leitlinie zu Open Access in der Leibniz-Gemeinschaft* (2007). http://www.leibniz-gemeinschaft.de/?nid=akroa&nidap=&print=0

Library and Archives Canada, *Canadian Digital Information Strategy. Draft* (2007) http://www.collectionscanada.gc.ca/cdis/012033-1006-e.html

Gunta Liders, *Comparison of Institutional Data Policies* (2002) http://206.151.87.67/docs/ComparisonInstitutionalDP.htm

Philip Lord, Alison MacDonald, Richard Sinnott, Denise Ecklund, Martin Westhead, Andy Jones, *Large-scale Data Sharing in the Life Sciences: Data Standards, Incentives, Barriers and Funding Models* (2005) http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007) http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Brian Matthews, Katherine Portwin, Catherine Jones, Bryan Lawrence, *Recommendations for Data/Publication Linkage* (2007) http://claddier.badc.ac.uk/trac/raw-attachment/wiki/WikiStart/Report_III_RecommendationsForDataLinking-final.doc

Medical Research Council, *Policy on Data Sharing and Preservation* (2007) http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/MRC002551

Medical Research Council, *Principles for Access to, and Use of, MRC Funded Research Data* (2007) http://www.mrc.ac.uk/consumption/groups/public/documents/content/mrc003759.pdf

Peter Murray, *Open Access to Research Data: Surmountable Challenges* (2007) http://www.esf.org/activities/science-policy/corporate-science-policy-initiatives/sharing-research-data-2007.html

National Health and Medical Research Council, *Project Grants Funding Policy for Funding Commencing in 2008* (2006) see: http://www.nhmrc.gov.au/grants/apply/projects/

National Health and Medical Research Council, Australian Research Council, Universities Australia, *Australian Code for the Responsible Conduct of Research* (2007) http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

National Heart, Lung and Blood Institute, *Data Set Policy* (2005) http://www.nhlbi.nih.gov/resources/deca/policy_new.htm

National Human Subjects Protection Advisory Committee, *Public Use of Data Files* (2002) http://www.hhs.gov/ohrp/nhrpac/documents/dataltr.pdf

National Institute for Allergy and Infectious Diseases, *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (2007) http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html

National Institutes of Health, *NIH Data Sharing Policy and Implementation Guidance* (2003) http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Institutes of Health, *Statement on Sharing Research Data* (2003) http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (2005) http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

National Science Foundation, *Cyberinfrastructure Vision for 21st Century Discovery* (2007) http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

National Science Foundation, *Division of Earth Sciences Data Policy* (2002) http://www.nsf.gov/geo/ear/EAR_data_policy_204.pdf

National Science Foundation, *Division of Ocean Sciences Data and Sample Policy*, NSF 04-004 (2003) http://www.nsf.gov/pubs/2004/nsf04004/nsf04004.pdf

National Science Foundation, *General Grant Conditions* (2007) http://www.nsf.gov/pubs/policydocs/gc1_607.pdf

National Science Foundation SES/SBE*, Data Archiving Policy* (2008) http://www.nsf.gov/sbe/ses/common/archive.jsp

National Strategy for Data Resources for the Social Sciences (2006) http://www2.warwick.ac.uk/fac/soc/nds/National_Data_Strategy_Final_tcm6-16581.pdf

National Strategy for the Development of Data Resources for the Social Sciences. Forward Plan, 2008-2010 (2008) http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/forward%20plan_tcm6-26817.pdf

NERC, *Data Policy Handbook* (2002) http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf

Norges Forskningsråd, *Verktøy for forskning. Nasjonal strategi for forskningsinfrastruktur (2008 – 2017)* (Research Council of Norway, *Tools for research. National research infrastructure strategy 2008-2017*) (2008) http://www.forskningsradet.no/servlet/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition%3A&blobheadervalue1=+attachment%3B+filename%3D17233UtstyrsStrategi9788212025219%5B1%5D.pdf&blobkey=id&blobtable=MungoBlobs&blobwhere=1203712678896&ssbinary=true

NSTC COS, *Terms of Reference of the Interagency Working Group on Digital Data* (2006) http://iwg.cfa.harvard.edu/twiki4/pub/IWGDD/IwgddTermsOfReference/COS_TOR_Digital_Data_IWG_12-06.pdf

OECD, *Declaration on Access to Research Data from Public Funding* (2004) http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

OECD, *The Global Research Village Conference* (2000) http://www.oecd.org/document/37/0,3343,en_2649_34293_1880805_1_1_1_1,00.html

OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007) http://www.oecd.org/dataoecd/9/61/38500813.pdf

OECD, *OECD Science, Technology and Industry Outlook 2008: Highlights* (2008) http://www.oecd.org/dataoecd/18/32/41551978.pdf

OECD, *Science, Technology and Innovation for the 21st Century*. Final Communiqué of the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 http://www.oecd.org/document/15/0,3343,en_2649_34487_25998799_1_1_1_1,00.html

OECD, *Shaping Policies for the Future of the Internet Economy* (2008) http://www.oecd.org/dataoecd/1/29/40821707.pdf

Office for Research Integrity, *Introduction to the Responsible Conduct of Research* (2004) http://www.mtu.edu/research/administration/integrity-compliance/pdf/rcrintro.pdf

Office of Science and Innovation e-Infrastructure Working Group, *Developing the UK's E-infrastructure for Science and Innovation* (2007) http://www.nesc.ac.uk/documents/OSI/report.pdf

Opetusministeriö, *Recommendations for the Promotion of Open Access in Scientific Publishing in Finland* (2005) http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2005/liitteet/opm_250_tr16.pdf?lang=en

OSI Data and Information Creation Working Group, *20/20 Vision: An E-Infrastructure for the Next Decade. Report of the Data and Information Creation Working Group to the e-Infrastructure Steering Group* (2006) http://www.nesc.ac.uk/documents/OSI/data.pdf

OSI e-Infrastructure Working Group, *Developing the UK's E-infrastructure for Science and Innovation* (2007) http://www.nesc.ac.uk/documents/OSI/report.pdf

Kylie M. Pappalardo, , Anne M.Fitzgerald, Brian F. Fitzgerald, Scott D. Kiel-Chisholm, Damien O'Brien, Anthony Auston, *A Guide to Developing Open Access Through Your Digital Repository* (2007) http://www.oaklaw.qut.edu.au/node/32

PARSE.Insight, *Digital Preservation and Scientific Data Infrastructure: A Preliminary Summary of Evidence on Scientific Community Needs* (November 2008)

Dirk Pilat, Yukiko Fukasaku, *OECD Principles and Guidelines for Access to Research Data from Public Funding* // *Data Science Journal*, Vol. 6, Open Data Issue (2007) http://www.jstage.jst.go.jp/article/dsj/6/0/OD4/_pdf

Heather Piwowar, Michael Becich, Howard Bilofsky, Rebecca Crowley, *Towards a Data Sharing Culture:*

Recommendations for Leadership from Academic Health Centers (2008) http://medicine.plosjournals.org/archive/1549-1676/5/9/pdf/10.1371_journal.pmed.0050183-L.pdf

Heather Piwowar, Wendy Chapman, A review of journal policies for sharing research data. In: *Proceedings of the ELPUB 2008 Conference on Electronic Publishing. Toronto, Canada* (2008) http://elpub.scix.net/data/works/att/001_elpub2008.content.pdf

Productivity Commission, *Public Support for Science and Innovation: Research Report* (2007) http://www.pc.gov.au/__data/assets/pdf_file/0016/37123/science.pdf

Graham Pryor, *Research Council Data Policy Statements*. DCC internal document (2007)

RELU DSS, *Guidance on Data Management* (2006) http://www.data-archive.ac.uk/relu/reluaug2006.pdf

Research Councils UK, *Draft Position Statement on Access to Research Outputs* (2005) http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2005statement.pdf

Research Councils UK, *Updated Position Statement on Access to Research Outputs* (2006) http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf

RIN, *Mind the Skills Gap: Information-handling Training for Researchers* (2008) http://www.rin.ac.uk/training-research-info

RIN, *Research Funders' Policies for the Management of Information Outputs* (2007) http://www.rin.ac.uk/policy-information-outputs

RIN, *Stewardship of Digital Research Data: A Framework of Principles and Guidelines: Responsibilities of Research Institutions and Funders, Data Managers, Learned Societies and Publishers* (2008) http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf

RIN, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* (2008) http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf

Royal Statistical Society, UK Data Archive, *Preserving and Sharing Statistical Material* (2002) http://www.data-archive.ac.uk/news/publications/PreservingSharing.pdf

RSC Chemical Communications, *Guidelines for Authors* (2007) http://www.rsc.org/Publishing/ReSourCe/AuthorGuidelines/JournalPolicy/CC/sect3.asp

Russell Group, *Code of Practice on Good Research Governance* (2005) http://www.russellgroup.ac.uk/news/2005/code-of-practice-on-good-research-governance.html

Marjut Salokannel, University of Helsinki Opens its Research Vaults: A few Words on Open Access and the New Research Environment in Finland // *Sciecom Info*, No. 2 (2008) http://www.sciecom.org/ojs/index.php/sciecominfo/article/viewFile/653/447

Science Commons, *Towards a Science Commons* (2007) http://sciencecommons.org/about/towards/

Shared Responsibilities in Sharing Research Data: Policies and Partnerships at the 'Berlin 5 Open Access Conference "From Practice to Impact: Consequences of Knowledge Dissemination"' (2007) http://www.aepic.it/conf/program.php?cf=10#fri1

SHERPA, *Institutional Repositories: Staff and Skills Set* (2008) http://www.sherpa.ac.uk/documents/staffandskills2008.pdf

SSHRC, *National Data Archive Consultation. Building Infrastructure for Access to and Preservation of Research Data* (2002) http://sshrc.ca/web/about/publications/da_finalreport_e.pdf

STFC, *Research Grants Handbook* (2008) http://www.stfc.ac.uk/rgh/PDFs/rgh8.pdf

David F. Strong, Peter B. Leach, *National Consultation on Access to Scientific Research Data* (2005) http://data-donnees.gc.ca/docs/NCASRDReport.pdf

SURFshare programme 2007-2010. *SURF Platform ICT and Research* (2007) http://www.surffoundation.nl/download/SURFshare%20programme%202007-2010%20Condensed%20version%20website.pdf

Alma Swan, Sheridan Brown, T*he Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs* (2008) http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

Robert Terry, *Open Access and the Wellcome Trust* (2006) http://berlin4.aei.mpg.de/presentations/Terry_OA06.pdf

Third EGEE User Forum (2008) http://cdsweb.cern.ch/record/1103667/files/egee-pub-2008-001.pdf

Mark Thorley, *Providing Appropriate Skills for Data Curation*. The Research Data Management Forum (2008) http://www.dcc.ac.uk/data-forum/docs/workshop1/19&20-03-08Forum_Thorley.ppt

To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe (2006) http://www.arl.org/bm~doc/digdatarpt.pdf

Treasury, Department of Trade and Industry, Department for Education and Skills, *Science and Innovation Investment Framework 2004-2014* (HMSO, 2004), http://www.hm-treasury.gov.uk/spending_sr04_science.htm

Andrew Treloar, *The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: Building the New Collaborative E-research Infrastructure* (2006) http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/paper.html

Andrew Treloar, David Groenewegen, ARROW, DART and ARCHER: A Quiver Full of Research Repository and Related Projects // *Ariadne*, Issue 51 (2007) http://www.ariadne.ac.uk/issue51/treloar-groenewegen/

Andrew Treloar, Cathrine Harboe-Ree, *Data Management and the Curation Continuum: How the Monash Experience is Informing Repository Relationships* (2008) http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Paul Uhlir, *Board on Research Data and Information. Draft Prospectus* (2007) http://www7.nationalacademies.org/usnc-codata/BRDI_Prospectus_19_November_2007.pdf

Paul Uhlir, *Draft Policy Guidelines for the Development and Promotion of Public Domain Information* (2003) http://unesdoc.unesco.org/images/0012/001297/129725E.pdf

Paul Uhlir, *Policy Guidelines for the Development and Promotion of Governmental Public Domain Information* (2004) http://unesdoc.unesco.org/images/0013/001373/137363eo.pdf

Paul Uhlir, Peter Schröder, Open Data for Global Science // *Data Science Journal*, Vol. 6 (2007) http://www.jstage.jst.go.jp/article/dsj/6/0/OD36/_pdf

The UK Data Forum, *Progress on Plans to Implement the National Data Strategy* (2006) http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/UKDF-06-19_tcm6-25732.pdf

UKRDS, *UKRDS Interim Report* Version v0.1a.030708 (July, 2008) http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20(2).doc

ULCC, AHDS, *Digital Asset Assessment Tool (DAAT)* (2007) http://www.ulcc.ac.uk/daat

UNESCO, *Draft Charter on the Preservation of the Digital Heritage* (2003) http://unesdoc.unesco.org/images/0013/001311/131178e.pdf

Universities UK, *Access to Research Publications: Position Statement* (2005) http://www.universitiesuk.ac.uk/PolicyAndResearch/PolicyAreas/Documents/Research/OpenAccessUUKPolicyStatementSept2005.pdf

University of Helsinki, *Open Access to Research Publications in the University of Helsinki* (2008) http://www.helsinki.fi/ejulkaiseminen/rinnakkaistallennus/pdf-tiedostot/decision260508_eng.pdf

University of Kentucky, *Data Retention & Ownership Policy* (1999; 2006) http://www.rgs.uky.edu/ori/data.htm

University of Massachusetts, Research Council, *Data Ownership, Retention, and Access at the University of Massachusetts Amherst* (2006) http://www.umass.edu/research/ora/data1.html

University of Rochester, *Interim Policy on Access to and Retention of Research Data* (no date) http://www.rochester.edu/ORPA/policies/retent.pdf

Johannes Velterop, *Analysis of the Responses to the Knowledge Sharing Questions in the Online Public Consultation on the Future of the European Research Area* (2007) http://ec.europa.eu/research/science-society/document_library/pdf_06/summary-reportq21-24092007_en.pdf

Wellcome Trust, *Open and Unrestricted Access to the Outputs of Published Research* http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/index.htm

Wellcome Trust, *Policy on Data Management and Sharing* (2007) http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Wellcome Trust, *Position Statement in Support of Open and Unrestricted Access to Published Research* (2008) http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm

Wellcome Trust, *Q&A: Wellcome Trust Policy on Data Management and Sharing*, http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-management-and-sharing/WTX035045.htm

Caroline Whitbeck, *The Responsible Collection, Retention, Sharing, and Interpretation of Data*. Online Ethics Centre (2006) http://www.onlineethics.org/cms/research/modindex/moddata.aspx

Working Group on Data for Science, *From Data to Wisdom: Pathways to Successful Data Management for Australian Science. Report to PMSEIC* (2006) http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm