



JISC

INFRASTRUCTURE PLANNING AND DATA CURATION

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

SUMMARY AND MAIN FINDINGS

20. NOVEMBER 2008

Prepared by:
Raivo Ruusalepp
Estonian Business Archives Consultancy

BACKGROUND

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.

Through an 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.

This summary report was prepared by Raivo Ruusalepp (EBAC), with contributions from Graham Pryor (DCC). The evidence upon which it is based may be consulted in the full report on the study, which was undertaken between May and November 2008.

STUDY APPROACH

This study provides an analysis of 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. The project timeline did not allow the creation of a full inventory of all data sharing initiatives across all the OECD countries, but instead, the focus is on best practice examples that cover the range of approaches taken.

The report has focused primarily on analysing the **enablers** of research data sharing, and less on the various barriers, but the instances of best practice referenced in the full report demonstrate the approaches used to overcome the barriers. 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:¹

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.

¹ OSI e-Infrastructure Working Group, *Developing the UK's e-infrastructure for science and innovation* (2007), p. 13

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).

KEY FINDINGS

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

Main conclusions from the more detailed analysis in the full report are presented below.

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).

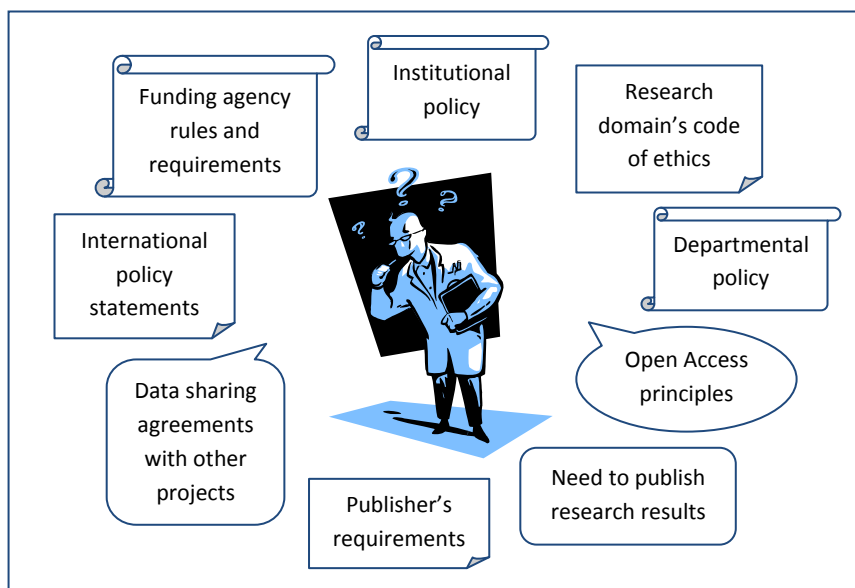
DATA SHARING INFRASTRUCTURE PROVISION

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. A barrage of obligations, requirements, incentives and suggestions for researchers to share their data is coming from different directions:



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 focused 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

The analysis in the full 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 project² into an office tasked with co-ordination of data sharing policy and infrastructure development.
- The DCC Research Data Management Forum,³ the UK Data Forum⁴ 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.
- 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.

² <http://www.ukrds.ac.uk/>

³ <http://www.dcc.ac.uk/data-forum/>

⁴ <http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/nds/ukdf/default.aspx>

SERVICES DEVELOPMENT

- JISC should endorse the use of the DCC Data Audit Framework⁵ 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* report⁶ 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.
- 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.

⁵ <http://www.data-audit.eu/>

⁶ Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (2007)

- 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.

COUNTRY COMPARISON

The table below presents main aspects of research data sharing support and infrastructure provision in five OECD countries (UK, US, Australia, Canada and Germany) that in this study were found to be most active in discussing data sharing.

Category	UK	US	Australia	Canada	Germany
Policies					
Main data sharing policy level	Funding agency	Funding agency and institutional	Funding agency	Funding agency	National and funding agency
Funding agency data sharing policy	Majority of the funding agencies	Majority of the funding agencies	-		Majority of the funding agencies
Data sharing a condition of funding	Majority of the funding agencies	Frequently	Recommended	Frequently	Recommended
Mandate for data deposit	Funding agency's repository; community resources; unspecified	Researcher's choice; funding agency's repository	Subject and/or institutional repository	Appropriate public database	Discipline-specific or institutional repositories
Data management plan required	Some funding agencies	Some funding agencies	All funding agencies	-	-
Data sharing costs part of the funding	Half of the funding agencies	Majority of the funding agencies	-	-	-
Typical data sharing timing condition	3 months of grant ending; when research results are published	2 years; as soon as possible; when research results are published	6 months of grant ending	When research results are published	When research results are published
Data repositories					
Predominant data deposit model	Domain data deposit model	Domain data deposit model	Federation data deposit model	Domain data deposit model	Federation data deposit model
Main data repository type	Centralised in research domains	Centralised in research domains	Institutional	Centralised in research domains	Institutional
Funding for repositories	Funding agencies	Funding agencies	Institutions, project funding		Institutional, funding agencies
National level data repository project / feasibility study	Yes: UKRDS	Yes: DataNet	Yes: ANDS	Yes: NCASRD	No / partially in: Digital Information Initiative
Services for data access					
Access portals to data repositories	National: OpenDOAR	Domain level	National: ARROW, AAF	Domain level	National services: STD-DOI
Guidance for data management					
Guidance to good data management	UKDA, DCC	ICPSR	OAK-Law and e-Research	Research Data Strategy Working Group	Some funding agencies

Other countries that the full report mentions include: Denmark, Finland, France, Greece, Japan, Netherlands, Norway, Spain, and Sweden.