Enabling Scientific Data Sharing and Re-use

R. Darby, S. Lambert,
B. Matthews, M. Wilson
Science and Technology
Facilities Council,
Didcot, UK

K. Gitmans
Alfred Wegener Institute
for Polar and Marine Research
Bremerhaven, Germany

S. Dallmeier-Tiessen,
S. Mele
CERN
Geneva, Switzerland

J. Suhonen
CSC - IT Center for Science
Espoo, Finland

Abstract—
Research data sharing is one of the key challenges in the e-science era. IT technologies facilitate an enhanced management and sharing of research data. It is crucial to understand the current status of research data sharing in order to facilitate enhanced data sharing in the future.

In this study, a conceptual model has been developed to characterize the process of data sharing and the factors which give rise to variations in data re-use. The study goes beyond a solely technical analysis and includes also psychological, social, organizational, legal and political components. The model was developed based on the literature and 21 face to face interviews with research, funding, data centre and publishing experts. It was validated by both a vigorous workshop and a further 55 structured telephone interviews. The overall model identifies sub-models of *process*, of *context*, and of *drivers, barriers and enablers*. These provide a comprehensive description of the factors that enable or inhibit the sharing of research data. They affect *whether* data are shared, *how* they are shared, and *how successfully* they are shared. Implementing the enablers will help the research community overcome the barriers to data re-use to facilitate future e-science endeavors.

Keywords- digital libraries, data management, data preservation, cyber-infrastructure

I. INTRODUCTION

Public funders of research increasingly follow [1] guidance from the Organisation for Economic Co-operation and Development (OECD) that publicly-funded research data should as far as possible be openly available to the scientific community in order to maximise the return on the public sector investment [2]. However, to date systematic data sharing is not common practice in the research community.

In order to bring the OECD recommendation into common practice, stakeholder groups need to be persuaded by a value proposition for data sharing which is compelling and appeals to their strategic objectives. Until re-use of digitally preserved data has become customary and its benefits are taken as axiomatic, value propositions can best be supported by compelling examples of successful data sharing, and by developing realistic models of data sharing to show how these successful cases can generalize.

The current study has identified examples of successful data sharing [3], and has developed a conceptual model of data sharing in order to identify the drivers, barriers and enablers to overcome those barriers to data sharing.

This paper briefly outlines the methodology used, then summarizes the conceptual model, before defining the identified drivers, barriers and enablers. The themes which emerged from the study are discussed followed a list of recommendations to each of the six stakeholder groups in order to enable more sharing of scientific data.

II. METHOD

The method used to establish the enablers for data sharing consisted of six stages: 1) the collection of examples of successful data sharing , 2) the drafting of a baseline model; 3) testing this baseline model for completeness via a workshop; 4) updating the model; 5) identifying the most important drivers, barriers and enablers in the model through interviews with experts; 6) deriving recommendations for stakeholders. Each step will be described in more detail.

The baseline from which the conceptual model was developed was established from existing published knowledge and from face to face interviews with 21 research, funding, data centre and publishing experts [4].

The published sources consulted in development of the conceptual model include studies on the benefits of preservation [5], barriers to preservation [6], costing of preservation [7,8], data sharing communities [9,10,11], and differences between disciplines in attitudes to data sharing [12,13,14]. Many of these studies provided analytical representations of data sharing systems and processes, and enumerated drivers and barriers that bear on success and failure in data sharing. They were used to inform development of the data sharing process and context models, and to elaborate a comprehensive list of drivers and barriers to data sharing.

The process and context models of data sharing were developed with particular reference to the Open Archival Information System (OAIS) Reference Model [15] for long term digital preservation and access. These studies largely focused on preservation roles and activities.

The work reported was partially funded by the European Commission FP7 ICT grant 261530 to the Opportunities for Data Exchange (ODE) project.

The conceptual model embraces data sharing more broadly, to include data discovery, access and re-use in addition to preservation. Two studies proved especially useful in elaborating the model of drivers and barriers: a large-scale survey on barriers to data re-use, and a framework of the benefits of long term data preservation.

The PARSE.Insight survey [16] received responses from 1,840 people overall, including 1,389 researchers, 273 data managers, and 178 publishers. It provided evidence across a wide range of disciplines about levels of data sharing, researchers' motivations for data sharing, and the barriers to sharing data that they had encountered.

The KRDS Benefits Framework [17] described a taxonomy of data sharing benefits and provided an analytical tool that could be used to evaluate the benefits in a particular instance of potential digital preservation.

Once a baseline conceptual model had been drafted based on these resources, its completeness was initially tested in a workshop of experts who vigorously debated it in November 2011. 20 experts participated in the workshop: 3 STM publishers, 4 managers of data centres in the particle physics, biological sciences, chemistry and archeology, 6 providers of data preservation and storage services, 4 researchers in the humanities, social science, earth sciences, and 3 providers of infrastructure services. Workshop participants had several days before the meeting to review the model, before an afternoon of facilitated debate in which several themes emerged. An analysis of the workshop discussion resulted in elaboration of the content and changes to the presentation of the conceptual model.

Once the revised model was available a second stage of validation was undertaken through telephone interviews with individual experts to identify the most important drivers, barriers and enablers. 55 interviews were conducted between February and April 2012.

Interviewees represented each of the stakeholder groups identified in the model (see Table 1). Although there was a broad international spread of interviewees, nearly 60% of interviewees came from the UK or Germany. 40% were with researchers who customarily produce and use data. Interviewees were selected from a range of academic disciplines: earth and environmental sciences, social sciences and humanities, medical and life sciences, physical sciences, engineering and technology, and computer sciences and mathematics.

Interviews were scheduled to last approximately 30 minutes. Prior to the interview interviewees were sent a document outlining the conceptual model of drivers and barriers. An interview *pro forma* was used, in two slight variations, one for researchers, and one for non-researchers. The pro-forma provided interviewers with a structured set of questions designed to stimulate critical engagement with the conceptual model, and allowing interviewees to elaborate on their views and experiences in data.

Analysis of the collected corpus of interviews informed the validation and qualification of the conceptual model. It also identified the most important themes and converging views on those themes in data sharing which are discussed in section 5 below, after a summary of the conceptual model, and the drivers, barriers and enablers.

Details of the questions, interviews, workshop, and selection of experts are available in the full report of this study [20].

III. Conceptual Model

The conceptual model is divided into three parts, or subsidiary models: the data sharing *process*, the data sharing *context*, and data sharing *drivers, barriers and enablers.*

The *process* model (see Figure 1) combines models of the data preservation process, and the scientific process to describe the overall data sharing process in terms of agents, actions and objects. The purpose of the process is to produce economic and social benefit through scientific discovery, new social policy, or through commercial benefit from new inventions. The preservation planning phase includes the identification of both data and supporting information needed to interpret the data, while the pre-archive phase addresses the collection and preparation of this data and information required to make the data re-usable. The model has been used as a key analytical tool to derive the model of data sharing drivers and barriers.

The *context* model (see Tables 1 and 2) describes the systemic scholarly communication context in which data sharing occurs. This context is described in terms of stakeholder roles (researcher, funder, publisher, etc.), and key variables that qualify the generic model, including research discipline, research sector, and geopolitical context (e.g. national/regional policy and legislation, infrastructure, funding).

The model of *drivers, barriers and enablers (see Table 3)* gives a comprehensive description of the factors that enable or inhibit the sharing of research data.

Functional roles are described in terms of the key stakeholder groups to which they belong. The key variables in data sharing are described below.

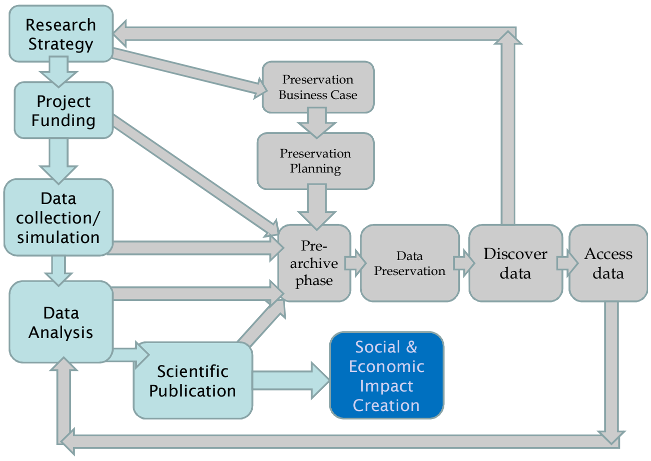

Figure 1: The idealized data sharing process incorporating views of the data preservation process on the right, and the scientific process on the left.

TABLE I. STAKEHOLDERS IN THE DATA SHARING PROCESS

Stakeholder group	Roles
Policy-makers	National policy makers Regional policy makers
Funders	Research funders Infrastructure funders
Researchers	Data producers Data consumers
Research and education organisations	Research planners and managers Librarians
Data management and infrastructure service providers	Data centre managers and staff Other infrastructure providers
Publishers	Publishers

TABLE II. SIGNIFICANT FACTORS IN THE CONTEXT OF DATA SHARING

Variable	Factors
Academic discipline	Source of data Cost of data collection Possibility to collect data again Complexity of data analysis
Country	Legislation Infrastructure Funding
Age of researcher	Willingness to invest effort for possible long-term benefit
Sector	Non-commercial research Commercial research Education

TABLE III. DRIVERS, BARRIERS AND ENABLERS TO DATA SHARING

Drivers
a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
Barriers
f) Individual Contributor Incentives g) Availability of a Sustainable Preservation Infrastructure h) Trustworthiness of the data, Data Usability, Pre-archive activities i) Data Discovery j) Academic Defensiveness k) Finance l) Subject Anonymity and Personal Data Confidentiality m) Legislation/Regulation
Enablers
n) Individual Contributor barriers o) Availability of a Sustainable Preservation Infrastructure p) Trustworthiness of the data, Data usability, Pre-archive activities q) Data Discovery r) Academic Defensiveness s) Finance t) Subject Anonymity and Personal Data Confidentiality u) Legislation/Regulation

IV. DRIVERS, BARRIERS AND ENABLERS

A set of drivers, barriers and enablers has been derived from the conceptual model. It has been validated successfully through the workshop and interviews. Each driver is shown in Table 4.

TABLE IV. DRIVERS TO DATA SHARING

Societal benefits	Economic/commercial benefits; Continued education; Inspiring the young; Allowing the exploitation of the cognitive surplus in society; Better quality decision making in government and commerce; Citizens being able to hold governments to account.
Academic benefits	The integrity of science as a activity is increased by the availability of data; Increased public understanding of science.
Research benefits	**For the data contributor:** Validation of scientific results by other scientists; recognition of their contribution.
	For the data user: Re-use of data in meta-studies to find hidden effects/trends (e.g. greater geographical spread is obtained by combining datasets; larger sample size from combining data sets increases statistical significance of small factors); To test new theories against past data; To do new science not considered when data was collected without repeating the experiment; To ease discovery of data by searching/mining across large datasets with benefits of scale; To ease discovery and understanding of data across disciplines to promote interdisciplinary studies; To combine with other data (new or archived) in the light of new ideas.
Organisational benefits	**Producer Organisation:** Publication of high quality data enhances organizational profile; Citation of data enhances organisation profile.
	Publisher Organisation: Preserved data linked to published articles adds value to the product.
	Infrastructure Organisation: Data preservation is more business; Reputation of institution as "data holder with expert support" is increased (e.g. universities with medieval library collections, fossil collections etc.).
	Consumer Organisation: Organisational need to combine data from multiple sources to make policy decisions; Re-use of data instead of new data collection reduces time and cost to new research results; Use of data for teaching purposes.

The eight tables below each describe a barrier with the enablers to overcome it and promote the benefits that the barrier blocks. The enablers are described in the same table as the barriers they overcome.

TABLE V. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO INDIVIDUAL CONTRIBUTOR INCENTIVES

Individual Contributor Incentives
Barriers: Journal articles do not describe available data as a publication; Published data is not recognized by the community as a citable publication; There is a lack of specific funding in grants to address the pre-archive activities for data preservation; There is a lack of mandates to deposit of high quality data with appropriate metadata in preservation archives; Journals do not require data to be deposited in a form where it can be re-used as a condition of publication; Data publication and data citation counts are not tracked and used as part of the performance evaluation for career advancement; There is a lack of high status awards to individuals and institutions which contribute data that is re-used.
Enablers: Journal articles describing available data as a publication; Citation of data itself, and the articles describing it; Specific funding in grants to address the pre-archive activities for data preservation; Enforced funding regulation to ensure the depositing of high quality data with appropriate metadata in preservation archives; Journals requiring data to be deposited in a form where it can be re-used as a condition of publication (e.g. Nature, but see [18], and [19] on poor conformance rates); Track data publication and data citation counts, and then use them as part of the performance evaluation for career advancement; High status awards to individuals and institutions which contribute data that is re-used

TABLE VI. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO THE AVAILABILITY OF A SUSTAINABLE PRESERVATION INFRASTRUCTURE

Availability of a Sustainable Preservation Infrastructure.
Barriers: Until there is an infrastructure for data preservation which has credible sustainability and credible chances of data discovery and re-use, then data producers will not make the effort to prepare data for publication and re-use. Specific barriers have been identified: Absence of data preservation infrastructure; Charges for access to infrastructure (e.g. professional bodies); Journals are not necessarily good at holding data associated with articles; Lack of data reviewers in infrastructure to assure data quality; Risk that data holders cease to operate, and archive is lost.
Enablers: This barrier can be overcome by several proposed solutions for the publication of data: in archives supported by journal publishers (e.g. Nature) sustained by a business model; in archives supported by learned societies (e.g. the CAS Registry of the American Chemical Society) sustained by a business model; in archives funded by funding bodies (e.g. UK Economic and Social Data Service); in institutional archives (e.g. ESO archive of astronomical images, university archives proposed by NSF and UK Research Councils). via e-infrastructure to support/share the effort of creating the metadata needed to enable the re-use and combination of data from multiple sources. In order to address not only the elite institutions (which may be able to sustain themselves into the long term future, and their own archives), but also the long tail of less well endowed and less productive research institutions, institutional archives alone will not be a credible sustainable solution. If there is a combination of archives, then there is a clear need for an integration infrastructure to facilitate data discovery – inter-disciplinary, international, and across classes of organisation.

TABLE VII. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO TRUSTWORTHINESS OF THE DATA, DATA USABILITY AND PRE-ARCHIVE ACTIVITIES

Trustworthiness of the data, Data Usability and Pre-archive activities
The pre-archive phase of data preservation is where the data quality is checked, and the metadata is gathered and linked to the data to make it usable. When preparing data for publication and re-use, ensuring the appropriate quality of data and provision of sufficient metadata to ensure that the designated community can use the data raises significant problems for data producers: Not “feeling safe” in dealing with unfamiliar data; Impossibility of data centre staff having detailed technical knowledge of all data (e.g. museum curators); Lack of clear definition of the level of data quality that the potential data users will require; Interdisciplinary data requires a unifying factor for data to make reuse easier (e.g. data maps to a common geographical coordinate system);
Enablers: These barriers can be overcome by a combination of: agreeing auditable standards for publishable data quality and metadata within disciplines; certification of data centres for data quality and usability by a trustworthy body; peer reviewing of data supporting academic research publications to certify its quality; customer reviewing of data and metadata to ensure their usability; the development of education & training materials for these data quality standards; the training of data producers with these materials; implement automated data quality and metadata content tools to test pre-archive data; providing the reward to lead to the contribution of producer effort required (see the incentives barrier below); inclusion of a mandatory data management/preservation preparation stage in research project proposals; introduce specific job profiles with career paths for data preparation and quality assurance staff – such staff may be embedded in research groups or hosted in data centres; overcoming the financial barrier to pre-archive activities (see the finance barrier below).

TABLE VIII. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO DATA DISCOVERY

Data Discovery
Barrier: There is no single infrastructure to support international, cross disciplinary data discovery.
Enabler: This barrier can be overcome by the following suggestions: Open Linked Data initiative; Persistent, unique data identifiers with search engines (e.g. DataCite); Interoperating Data Centres in specific disciplines (e.g. CESSDA in Social Science).

TABLE IX. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO ACADEMIC DEFENSIVENESS

Academic Defensiveness
Barrier: Data producers may be defensive about publishing data for a variety of reasons: Security concerns over the danger of "being hacked" and not being preserved as it is; Fear of failure to validate their results; Fear that others will gain benefit from their data; Fear of misuse of data for purposes for which it is not suited will harm the data contributor; Fear of misuse of data to justify arguments which the contributor would find unacceptable will harm the data contributor.
Enabler: The first problem should not be supported on grounds of professional ethics. The second barrier is addressed in some disciplines by incorporating embargo periods on the publication of data after it has been collected, analyzed and/or contributed – e.g. 3 years after collection of raw data by large neutron and synchrotron facilities. This approach is completely unacceptable in other disciplines where immediate publication is the norm (e.g. publication of gene sequences in bioscience). The third and fourth barriers need to be addressed.

TABLE X. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO FINANCE

Finance
Barrier: Archiving costs alone are argued to be small in studies of preservation costing. Pre-archive collection of metadata and quality checking of data must be undertaken by the data provider (perhaps with guidance from the preservation service staff) but they need to perceive sufficient benefit to justify this effort from their own costs, or have them explicitly funded. Data discovery costs can be high if data archives are to be linked to promote data discovery as part of a large data infrastructure. Lack of pre-archive funding by contributor; Lack of archiving funding by infrastructure; Lack of data discovery and access funding; Risk of lack or return on long term investment in preservation infrastructure; Risk of high costs in answering questions about projects or data after their funding has expired.
Enabler: Only investing in archiving services as sustained infrastructure, leaving the investment in pre-archive (by the producer project) and data access (by the consumer project) activities to be included in research project costs funded at project review; Publicising case studies of successful data sharing and re-use which have achieved significant impact. There is perceived to be a need for central funding for discovery integration costs as part of a discipline based/interdisciplinary national/international data infrastructure. Possible sources of funding to overcome this barrier include publishers, who can sell data discovery services, or EU, national or state public funding for infrastructure. Commercial business models for publishers to provide data discovery services need to be tested, although they have been established in some disciplines (e.g. American Chemical Society), and by the most prestigious journal publishers (e.g. Nature).

TABLE XI. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO SUBJECT ANONYMITY AND PERSONAL DATA CONFIDENTIALITY.

Subject Anonymity and Personal Data Confidentiality
Barrier: There is a genuine need/desire among researchers in medical and social science research disciplines to preserve the anonymity of subjects who contribute data to their studies, not least to ensure that they will be willing to contribute data again in the future. The research is dependent on subjects contributing data, so this is a strong driver to preserve anonymity. Specific barriers identified are: Lack of funding for anonymising data, which is costly; Lack of agreed standards for anonymising data; Lack of trust in the preservation infrastructure to prevent de-anonymisation.
Enabler: This barrier is usually overcome by only publishing data through a "Data Enclave" which is a secure environment that allows for remote access to confidential micro-data where the combination of data sets which may reveal the identity of subjects is prevented. This issue is not a binary one of data which can identify individuals or anonymous data, but a spectrum where different classes of data require different levels of security.

TABLE XII. BARRIERS AND ENABLERS TO DATA SHARING RELATED TO LEGISLATION AND REGULATION.

Legislation/Regulation
Barrier: There are *perceived* to be conflicts between the data protection and freedom of information legislation; between international and national legislation; between the legislation of different countries; between national and regional legislation; in the enforcement of legislation by different agencies; in the understanding on legislation by different stakeholders, and between the regulations of different stakeholders designed to implement legislation. All of these conflicts create barriers to data sharing.
Enabler: These barriers can be overcome by unifying legislation at the highest level, but more importantly, unifying the national implementation of European directives and international agreements as national legislation, and the national enforcement of the those directives, and by greater education as to the exact entailments of the legislation for research data sharing.

The interviews highlighted that the drivers to data sharing are frequently blocked by the barriers identified. Many interviewees reported experiences or initiatives which facilitated overcoming the barriers. The enacting of the enablers will involve most of the stakeholders identified in Table 1, and cannot be undertaken by those who fulfill a single role alone. A simple recommendation from the study would be that all the enablers should be enacted, but this naive view does not prioritize those enablers which will have most effect on data sharing.

The interviews conducted in the fifth phase of the study identified common themes among the barriers and enablers, which can be used to prioritize those actions which will have most effect.

V. Emergent themes

Ten themes emerged from the collated evidence supported by converging views which have been used to organize and interpret the evidence in this section. These have been grouped below under the three overall topics of *digital publication*, *data management infrastructure*, and *culture and policy*. The themes highlight what has been achieved so far to promote data sharing in e-science, and present the challenges and opportunities the next decade will bring.

A. Data Publication Practice Themes

1) The role of publishers in data sharing

The role of publishers in data publication and sharing is widely discussed and excites a range of opinions. By and large publishers appear to be open to the ideas of supplementary data publication, standard data citation in publications, reciprocal linking of publications and datasets, and facilitating access to data both through appropriate licensing and through provision of tools that allow users to discover and interrogate data linked to publications. There are positive examples of publishers engaging in all these areas and of a willingness to engage further where suitable collaboration partners can be found. Other views expressed by some publishers, data centre managers and researchers indicate a perception that as a whole the publishing community has not gone far enough or fast enough in areas such as: implementing best practice in data citation; developing industry standards for data citation or using existing standards, such as DataCite DOIs; incorporating quality assurance and peer review of data into editorial processes; and bringing standalone data journals to market.

Most publishers consulted believed they could play a larger role in enabling people to publish data and make it discoverable and usable. By acting in collaboration with community stakeholders they could promote the adoption of common data formats and standards of data referencing and description. Such collaborative approaches might embrace publishers, researchers and libraries, in much the same way as electronic article preservation is being tackled collaboratively. Initiatives such as ORCID and DataCite are examples of cross-industry approaches to developing standards and solutions for the scholarly communication field, which could provide a positive model for the development and embedding of data standards, e.g. machine-readable taxonomies.

2) Data citation and description for discovery and use

Over 75% of interviewees expressed views on good data citation. Data citation practice is not yet customary after the manner of citing publications such as journal articles. But the importance of citation to the recognition of data as a primary research output, rather than a by-product of research, is now starting to be recognized. Routine citation of data sets will enhance their status as research outputs, and increase the potential impact of research, to the benefit of both the data creator and the research itself. But citation is most effective when applied according to established universal standards, as regards both metadata formats and semantics. The granularity of data to be cited is often unclear - a complete database is often too much, but the way to cite subsets of data is not agreed. Since a piece of research will be based on a subset of data, it is the relationship of this subset to a resulting publication, an initiating proposal , and other data sets which needs to be captured and maintained. The citation and linking mechanisms must support the creation and evolution of these relationships so that they can support citation logging, and impact analysis. Good data citation leads to better impact for the research, and ultimately benefits the researcher.

B. Data Management Infrastructure Themes

1) Finance: funding infrastructure and data services;

For the long-term viability of data-sharing, it is essential that protected funding be dedicated both in research grants for data management activities, and at national and regional level to sustain the preservation and sharing infrastructure, and maintain data centres providing services across all academic disciplines. These objectives can be most effectively achieved by co-ordination between stakeholders.

A coordinated, systemic approach to financing data preservation and sharing is widely agreed to be a worthy ideal, but one very difficult to achieve in practice, as there are many different kinds activity, service and infrastructure that need to be financed by different stakeholders at many different levels, ranging from individual funded research projects to large supranational infrastructures. But certainly without mechanisms for stakeholders to co-ordinate their spending, there will be avoidable gaps and redundancies in provision, and inefficiencies in the use of public money. Coordinated activity, though in practice it may be difficult to achieve, will tend towards greater cost-effectiveness across the entire data ecosystem, and will distribute service provision more efficiently so as to reduce gaps and redundancies.

2) Data management: skills training and expert support;

Research data management planning is a foundation for good research. Training programmes aim to equip researchers and data custodians with the skills they need to share and preserve data effectively. Service providers such as data centres and libraries play a central role in aiding researchers to perform data management. It is essential for the institutions of higher education to include discipline-focussed training programmes in curricula for students and researchers, so that emerging researchers learn at early stage to take ownership of their data and acquire the proper data management skills.

Improving the skills and understanding of researchers in data management is essential. Training should begin in the institutions that train researchers, at the outset of postgraduate study at the latest, possibly even earlier. Education and research institutions should also offer ongoing support, especially to early-career researchers; while it is not realistic to expect research institutions to meet the highly specialized data management needs of all their researchers, they should at the least serve an intermediary function and guide researchers to appropriate sources of specialist support.

Institutional libraries are ideally positioned to offer basic skills support and signposting services, as it is the traditional skill of the librarian to know where the required information can be found. If this is a role librarians will increasingly be required to undertake, it should be reflected in the training delivered through professional qualification courses for librarians.

Specialist data management services tailored to specific discipline and data requirements can best be provided by experts based in data centres and specialist data service providers. There may be scope for such providers to become more proactive in delivering skills training to students and early-career researchers.

A coordinated national approach to training researchers in data management may be most effective. This might involve, for example, a mandate from the national government for all higher education providers to include a data skills training module in all postgraduate course; funders could also include data skills training requirements in postgraduate study grant conditions; accredited courses could be delivered by institutions or by specialists based in data centres.

There is also a demand for professional training and defined career paths for data librarians, and this may need to be reflected in professional librarian training courses.

3) Quality assurance of research data.

There are two aspects of quality of data: fitness for purpose and trustworthiness. This theme is more concerned with the second of these, but the first is related, and potential re-users of data will want to know about both. The need is particularly acute for cross-disciplinary reuse, when the potential re-user might not have in-depth expertise and ability to evaluate the data being considered. There are deep problems of anticipating what might in future be done with data gathered for a particular purpose, and establishing provenance to ensure that datasets that have been combined or processed or migrated retain their value. Some of these are current research problems, though a professional and reliable infrastructure of repositories can help by ensuring that basic requirements are met when data is accepted. There is a separate issue of the quality of repositories themselves; that is, their trustworthiness to preserve data for the long term.

If data reuse is to be encouraged then it is vital that the potential re-users should have confidence that the data is fit for their purpose and trustworthy. There are of course intrinsic problems, including necessarily not knowing what use might be made of one's data by an unknown person at an unknown time in the future; and the difficulties of recording provenance of datasets.

At one level, a general professional approach involving repositories, libraries and/or publishers can help to capture and maintain at least some metadata that is relevant for quality assurance. In some disciplines, data journals with peer review have a role to play.

4) Standards and interoperability;

Many respondents highlighted the challenges of developing and embedding standards for describing and formatting data: these are the bases on which interoperability is established, which in turn allows data to be shared across e-infrastructures and interpreted by end users. The frequency with which these issues were mentioned indicates how central they are to the whole system of data sharing.

Two distinct domains of interoperability were discussed:

Firstly, data description: metadata standards are essential to the process of discovering and identifying relevant data that are distributed throughout multiple databases and data repositories. The emphasis of respondents was often on descriptive standards specific to disciplinary communities, but clearly for cross-disciplinary sharing to become possible, generic standards, ontological mappings and semantic techniques for creating the knowledge context of data will be necessary.

Secondly, data formatting: assuming that distributed data sets have been discovered and their relevance established, in order for them to be usable in aggregate, and in particular for them to usable as machine-readable corpora, data need to be structured and formatted according to consistent standards. As the volume of data grows exponentially, the importance of machine processing becomes greater.

Establishing standards for data discovery and use both present enormous technical and intellectual challenges – but without workable solutions the whole data sharing system is less efficient, and the incentives to share data are less apparent to the researcher. The more visible and accessible data are to other users, the greater their impact and productivity: so standards go to the heart of data sharing. But these are also areas in which some standards have already become well-established (such as DataCite DOIs), and where ongoing initiatives and projects are working to promote the adoption of standards.

C. Culture and Policy Themes

1) Data sharing culture;

There is a social dimension to data sharing, and in large part this is determined by the practices that have become established over many years in different research communities. Discipline is the primary determinant in this respect: some disciplines, such as the bio-molecular sciences, or high energy physics, have well established cultures of collaboration and data sharing; whereas others have a traditionally closed or proprietorial approach to data, and do not have a widespread culture of openness.

The advent of internet-based technologies has introduced other demographic distinctions, as between older and younger researchers, the latter being generally seen as more willing to embrace the data-sharing potential of new technologies. These distinctions are likely to level off in due course. It is also the case that as the internet has facilitated greater interdisciplinary communication and the emergence of distinct new disciplines, such as bioinformatics, traditional data sharing cultures are being challenged.

2) Public visibility of research data

The public visibility of academic practice and outputs, in large part enabled by the internet, has had an irreversible effect on academic data policy and practice. There are benign aspects to this, such as the rise of citizen science, open data campaigners holding those in power to account, and the exposure of academic error and fraud; but also more worrying phenomena: aggressive use of Freedom of Information (FoI) requests to universities from agenda-led campaign groups or commercial interests whose purposes are clearly at odds with the public benefit objectives of academic researchers. Whether this visibility is perceived on balance as a good thing or a bad thing, academic policy and practice must change to reflect the new reality – there is no hiding place for data. Data policies need to take account of this new data transparency to ensure that research data is made available responsibly and securely.

3) National and regional policy and legal frameworks

To enable the great leap forward in data sharing, national and regional policies and laws must create frameworks to manage the negotiation between multiple different interests in a co-ordinated, consistent and equitable way. Clearly there are differences in national custom and other factors that militate against a one-size-fits-all approach to legal and policy frameworks, but researchers should encourage the establishment of at the least a coherent framework and minimum regulatory basis on which national policies and legal structures can be established. This should not only reduce friction in the movement of data across borders; it should actually help to accelerate the development of a global data sharing culture.

4) Incentives in the academic reward system for good data practice

There is great scope for data to form part of the system of academic recognition and reward, just as publications do now. There are benefits in broadening the basis for recognition and reward, whether through highly formalized measures based on citations, or less formal types of peer recognition. However some barriers stand in the way. If data citation is to be taken up on a par with conventional citation of papers, then equivalent formalization of citation and of evaluation is required.

VI. CONCLUSIONS

This study showed research data sharing as one of the key challenges in this e-science decade.

Researchers, policy-makers, funders, service providers and publishers have made advances across many disciplines in the last 10 years to facilitate the sharing and re-use of research data. However, this study clearly identified remaining barriers to maximising the re-use of data, and the study has described important enablers which each of these groups can put into place to further increase the benefits of data sharing and re-use to all those involved with research data over the next 10 years. In order to facilitate enhanced data sharing in the future, proper curation of data will be needed to ensure that the data retains its value over the long term, and therefore continues to reflect well on its originators. The effort in preparing data for sharing must be balanced by the rewards. At present the rewards do not always appear sufficiently concrete.

ACKNOWLEDGMENT

The authors acknowledge the contribution of their partners in the ODE project and the cooperation of all those who were interviewed in the study.

REFERENCES

[1] Royal Society, "Science as an Open Enterprise", Royal Society: London, June 2012.

[2] OECD, "Principles and Guidelines for Access to Research Data from Public Funding", OECD Publications: Paris, 2007.

[3] Opportunities for Data Exchange, "10 tales of Drivers and Barriers in Data Sharing", Alliance for Permanent Access: London, 2011.

[4] A Schäfer, H Pampel, H Pfeiffenberger, S Dallmeier-Tiessen, et al., "Baseline Report on Drivers and Barriers in Data Sharing, Opportunities for Data Exchange", Deliverable to the European Commission D3.1, 2011.

[5] Berman, F., Lavoie, E. et al., "Sustainable economics for a digital planet: ensuring long-term access to digital information", Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010, NSF:USA.

[6] Gardner D, Toga AW, Ascoli GA, Beatty JT, et al., "Towards Effective and Rewarding Data Sharing", Neuroinformatics, 2003, **1**(3) pp.289-95.

[7] P. Wheatley, B. Hole, "LIFE3*:* Predicting Long Term Digital Preservation Costs", in iPRES 2009: the Sixth International Conference on Preservation of Digital Objects, California Digital Library, 2009.

[8] N. Beagrie, J. Chruszcz, and B. Lavoie, "Keeping Research Data Safe: a cost model and guidance for UK universities", 2008, JISC:London.

[9] M.D. Wilson "A case study in the rewards of long term data sharing - 26 years of the MRC Psycholinguistic Database", UK e-Science All Hands Meeting '08 (AHM '08), Edinburgh, Sep 2008.

[10] J. P. Birnholtz, M. J. Bietz (2003) "Data at work: supporting sharing in science and engineering", in GROUP '03, Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, ACM:New York, NY.

[11] C. Parr, M. Cummings, "Data Sharing in Ecology and Evolution: Why Not?" *Trends in Ecology and Evolution*, July 2005. HCIL-2005-06, CS-TR-4708, UMIACS-TR-2005-16.

[12] C. Tenopir, S. Allard, K. Douglass, AU. Aydinoglu, L. Wu, et al.. (2011) *Data Sharing by Scientists: Practices and Perceptions.* PLoS ONE, 2011, 6(6): e21101.

[13] C. L. Borgman, "Research Data: *Who will share what, with whom, when, and why?"* Working Paper Series of the German Data Forum #161, German Federal Ministry of Education and Research, Oct. 2010.

[14] Key Perspectives Ltd, "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability", SCARP project synthesis report, 2010, Digital Curation Centre: Edinburgh, UK.

[15] CCSDS, The Reference Model for Open Archival Information Systems (OAIS), Draft Recommendation for Space Data System Standards, August 2009, (ISO 14721).

[16] T Kuipers, and J van der Hoeven, "D3.4 Survey report, PARSE.Insight: INSIGHT into issues of Permanent Ac-cess to the Records of Science in Europe", 2009, Report to the European Commission: Brussels.

[17] N. Beagrie, B. Lavoie, and M. Woollard, "Keeping Research Data Safe 2: Final Report", 2010, JISC: London.

[18] H. A. Piwowar and W. Chapman, "A review of journal policies for sharing research data", International Conference on Electronic Publishing, 2008, Nature Precedings : hdl:10101/npre.2008.1700.1

[19] A. Alsheikh-Ali, W. Qureshi, M.H. Al-Mallah, et al., "Public availability of published research data in high-impact journals", 2011, *PLoS ONE* 6(9): e24357

[20] ODE "D5.1 Report on Drivers and Barriers, and new opportunities for data sharing", June 2012, Report to the European Commission: Brussels.