Enabling sharing and reuse of scientific data

This is an Authors' Accepted Manuscript of an Article whose final and definitive form, the Version of Record, has been published in the *New Review of Information Networking* Volume 19 Issue 1 on 18 April 2014, available online at: http://www.tandfonline.com/doi/full/10.1080/13614576.2014.883936.

Enabling Sharing and Reuse of Scientific Data

S. Dallmeier-Tiessen[1], R. Darby[2], K. Gitmans[3], S. Lambert[2], B. Matthews[2], S. Mele[1], J. Suhonen[4], M. Wilson[2]

[1]CERN, Geneva, Switzerland

[2]Science and Technology Facilities Council, Didcot, UK

[3]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

[4]CSC - IT Center for Science Espoo, Finland

Keywords: digital libraries, data management, data preservation, e-infrastructure, data sharing

Abstract

The purpose of this study was to develop an understanding of the current state of scientific data sharing that stakeholders could use to develop and implement effective data sharing strategies and policies. The study developed a conceptual model to describe the process of data sharing, and the drivers, barriers, and enablers that determine stakeholder engagement. The conceptual model was used as a framework to structure discussions and interviews with key members of all stakeholder groups. Analysis of data obtained from interviewees identified a number of themes that highlight key requirements for the development of a mature data sharing culture.

1 Introduction

Public funders of research increasingly follow the lead given in the *OECD Principles and Guidelines for Access to Research Data from Public Funding* 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 (see also [Royal Society 74]). In practice data sharing in and between research communities is variable and unevenly distributed. While there are certainly drivers which have encouraged some research communities to engage in data sharing and reuse, there are numerous barriers that inhibit the development of data sharing cultures within communities and that affect the willingness and ability of individual researchers to share their data.

In a discussion of recent changes in data sharing culture, Hodson summarises a number of the commonly perceived barriers to data sharing among researchers:

> *Of course, not all data can or should be shared. Issues of privacy, commercial potential and intellectual property rights all need to be taken into account. Fundamental characteristics of academic culture also need to be respected – to a point. Academic reputation is built upon publications. And publications are built upon data. Hence there is pressure on researchers not to share their data, at least until they have published, for fear of being pipped at the post.* [Hodson]

There are incentives to share data, not only for funding organisations and researchers, but also for other groups that have a stake in the production and communication of scientific research: these include universities and other research organisations with research management and library functions; data centres and organisations that exist to preserve data and make them available; and publishers. If data sharing is to become established in common research practice, stakeholder groups need to be persuaded by a value proposition which is compelling and appeals to their strategic objectives. Examples of successful data sharing can present a persuasive case to stakeholder organisations. But to arrive at the stage where reuse of digitally preserved data has become customary and its benefits are taken as axiomatic, development of policy and infrastructure needs to be supported by realistic models of data sharing, which afford an understanding of the drivers and barriers that affect the different stakeholders in the system, and promote implementation of the enablers through which barriers can be overcome.

The work described in this paper was undertaken in the project Opportunities for Data Exchange (ODE). It was motivated by recognition of the importance of creating a scientific data e-infrastructure, justified in terms of its positive effects on the process and practice of science: 'It is vital that those responsible for financing, generating and preserving primary data take full advantage of these opportunities and create an environment where datasets act as the pieces of a jigsaw. Multi-disciplinary approaches to understanding how that jigsaw fits together will further enhance the intellectual capital of Europe' (quoted in [Science and Technology Facilities Council]).The project gathered evidence to support strategic investment in the emerging e-Infrastructure for data sharing, reuse and preservation. Its broad approach was to engage in dialogue with relevant stakeholders, in order to collect and document views and opinions on challenges and opportunities for data exchange.

The work discussed here was undertaken within the project in order to analyse current data sharing practice in terms of an analytical model describing the drivers and barriers that affect stakeholder engagement in data sharing, and the enablers through which stakeholders can reinforce drivers and overcome or minimise barriers. This model formed the basis of discussions

and structured interviews with key members of stakeholder groups; further analysis identified a number of themes that recurred in stakeholders' discussions of data sharing opportunities and challenges, and helped the project to identify the principal areas where targeted effort was most likely to realise practical benefits in terms of a maturing data sharing culture.

In this paper we outline the method used in this work, describe the conceptual model including drivers, barriers and enablers that was developed as a framework for inquiry, and consider some of the themes that were identified through discussions and interviews with key stakeholders. The discussion of the themes covers strengths and weaknesses of the present situation, highlights promising developments, and outlines areas for action, and is intended to provide a baseline for understanding aspects of the practice and prospects for data sharing.

2 Method

The method of study consisted of six stages: 1) the collection of examples of successful data sharing; 2) the drafting of a simple conceptual model for representing the processes and contexts of data sharing; 3) testing this model for completeness via a workshop; 4) updating the model; 5) identifying the most important drivers, barriers and enablers in the model through structured interviews with experts; 6) identifying key themes in data sharing and opportunities and challenges for stakeholders. Each step is described in more detail in the following pages.

The baseline from which the conceptual model was developed was established from existing literature and face-to-face interviews with 21 research, funding, data centre and publishing experts [Schäfer et al.].

The published sources consulted in development of the conceptual model include: studies on the benefits of preservation [Blue Ribbon Task Force on Sustainable Digital Preservation and Access]; barriers to preservation [Gardner et al.]; costing of preservation [Wheatley and Hole; Beagrie, Lavoie and Woollard; Beagrie, Chruszcz and Lavoie]; data sharing communities [Birnholtz and Bietz; Parr and Cummings; Tenopir et al.]; and differences between disciplines in attitudes to data sharing [Borgman; Key Perspectives; Consultative Committee for Space Data Systems]. Many of these studies provided analytical representations of data preservation and 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 data preservation and sharing process and context models, and to elaborate a comprehensive list of drivers and barriers to data sharing.

The studies cited above focused largely on preservation roles and activities. Data preservation and reuse are intimately linked, as preservation is intended to support future use of the preserved entities. The conceptual model developed within this study embraces data sharing in the broad sense, to include data discovery, access and reuse in addition to preservation. Two studies proved especially useful in elaborating the model of drivers and barriers: a large-scale survey of researchers on barriers to data reuse undertaken within the PARSE.Insight project [Kuipers and van der Hoeven], and the Keeping Research Data Safe project framework for the benefits of long-term data preservation [Beagrie, Lavoie and Woollard].

The baseline conceptual model developed by the project was tested in a workshop of experts held in November 2011. 20 experts participated in the workshop: three STM publishers, four managers of data centres in the fields of particle physics, biological sciences, chemistry and archaeology, six providers of data preservation and storage services, four researchers in the humanities, social

science and earth sciences fields, and three providers of infrastructure services. The conceptual model was validated and revised based on analysis of the workshop discussion.

The revised model informed development of an interview script designed to allow structured information collection from interviewees about their experiences and understanding of the most important drivers, barriers and enablers in data sharing. Project members carried out 55 telephone interviews with individual experts and key members of stakeholder groups between February and April 2012. The distribution of interviewees was selected to be representative of key stakeholder groups, and included researchers, infrastructure service providers, data management service providers and publishers. There was a broad spread of interviewees across the European Research Area, the United States and Australia, although with almost 60% of them based in either the UK or Germany. 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. Many interviewees fulfilled multiple stakeholder roles in their professional activities and identified with more than one stakeholder group, for example as active researchers who might also be involved in providing data centre or infrastructure services for data sharing, or who worked within funding and policy-making organisations. Particular effort was made to ensure the interview sample included a representative proportion of individuals engaged in primary research who produce and consume different kinds of data in a range of disciplinary fields; 40% of those interviewed were researchers who customarily produce and use data.

Interviewees were provided with a structured list of drivers and barriers ahead of their interview, which was scheduled to last approximately 30 minutes. An interview pro forma was used to provide interviewers with a structured set of questions designed to stimulate critical engagement with the conceptual model, allowing interviewees to identify the most important data sharing drivers and barriers in their field, and to elaborate on their views and experiences in data sharing.

Analysis of the collected corpus of interviews identified a number of themes that highlight key requirements for the development of a mature data sharing culture. A selection of these is discussed in detail in section 5 below, following a summary of the conceptual model of data sharing drivers, barriers and enablers.

3 Conceptual model: process and context

The conceptual model of data sharing is based on a general model of the *processes* underlying the scientific endeavour as it relates to the generation, analysis, preservation and reuse of data. This model is a fundamental framework for positioning the activities related to data sharing. In addition, a model of *context* allows the characterisation of the variables that impinge on the practice of data sharing. These models together allow the derivation of a set of drivers and barriers that are the subject of section 4.

3.1 Process model

The *process* model maps out the key stages in the research cycle where activities that enable data sharing can occur, and where the influence of drivers and barriers will affect the ultimate success or failure of data sharing. This model is illustrated in Figure 1, giving the sequence of activities in the research and preservation process, and the information flows between them.

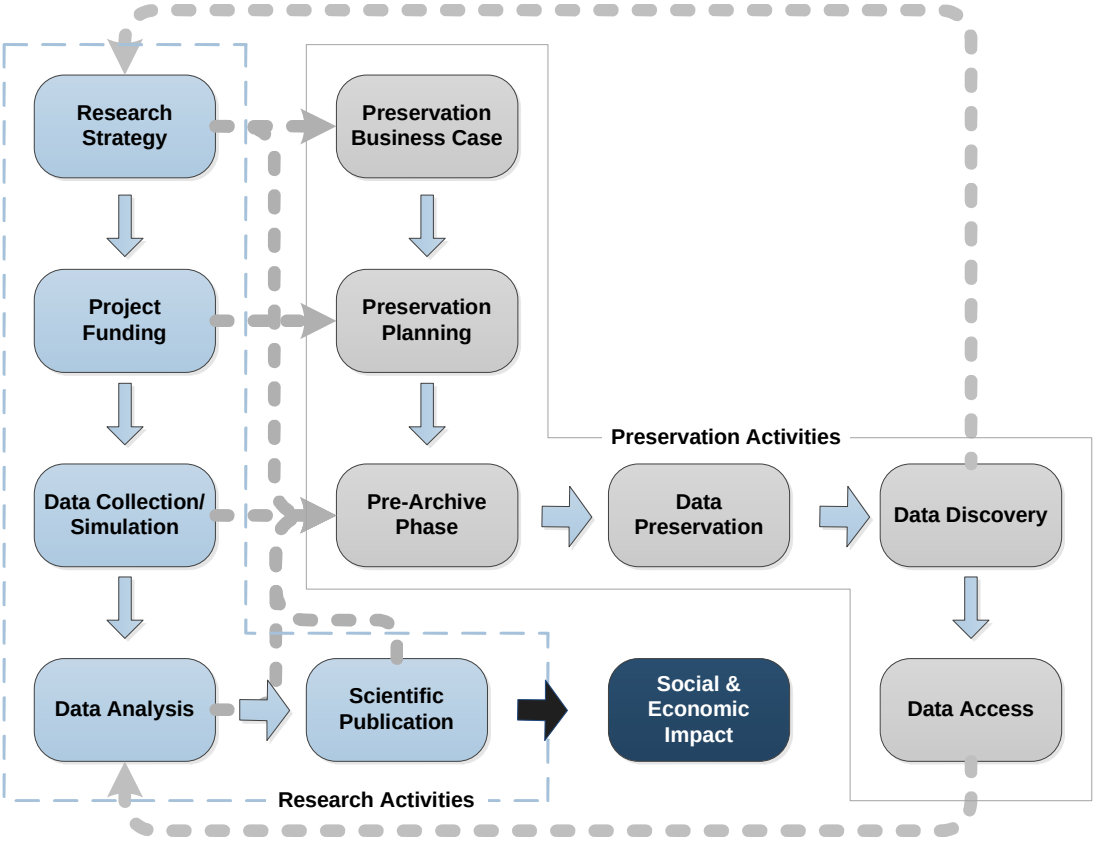

Figure 1: The idealised data sharing process

Different actors and stakeholders are engaged at different stages in these processes (see Table 1 below). Specific research activities are undertaken within the scope of *research strategies*, which at their broadest are formulated at national and international levels, but will also be articulated by funders of research, and research and education organisations. These strategies will implicitly or explicitly address requirements for preservation and sharing of data, and should in the particular research instance initiate the preservation process through the requirement for a *preservation business case* and *planning for preservation of data* produced during the research process.

The key activities in the research process are *data collection/simulation* and *data analysis*, which will generate the data that is fed into the preservation process. The direct output of a given research process is *scientific publication*, which in turn may lead to *social and economic impact*. But *data preservation* can also support the same end by making available for *discovery* and *access* the data outputs of a given research activity, which may in turn become data inputs into further research activities. Clearly the data preservation stage is critical to the success with which research-generated data can be recycled through future research activities: 'A crucial part of making data user-friendly, shareable and with long-lasting usability is to ensure they can be understood and interpreted by any user. This requires clear and detailed data description, annotation and contextual information' [Van den Eynden et al. 9]. It is in these activities that the expert support of data and information professionals based in research organisations and in data centres can have a determinate effect on the quality of data preservation and the usability of a given dataset.

Although research and data preservation are conceptually distinct processes, they are in practice not easily separable, and in fact may be advanced by the same activities. Hence *data collection* proceeds hand-in-hand with *data preservation*, as data and the transformations they undergo are

recorded and described. As raw data are transformed through the research process they are also travelling towards the definitive form in which they will be preserved.

The division of the post-data preservation stage between *data discovery* and *data access* highlights the role of discovery functions within data repositories as well as generic services, such as the DataCite Metadata Store, which allow users to search across an aggregation of repositories. Discovery services might also link to supplementary services that incentivise researchers to take ownership of their data, for example by linking citation counts on published articles to data supporting the article, or by providing measures of data use and citation.

3.2 Context model

The *context* model describes the systemic scholarly communication context in which data sharing occurs. This model aims to capture the different perspectives that give rise to differences between one scientific discipline and another, or one country's scientific community and another's, or the different points of view of actors within the scientific process. This context is described in terms of stakeholder roles (Table 1), and key variables that qualify the generic model, including research discipline, research sector, and geopolitical context (Table 2).

Table 1: Stakeholders and roles 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 2: Significant variables and factors in the context of data sharing

Variable	Factors
Academic discipline	Source of data Cost of data collection Possibility of collecting 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

4 Drivers, barriers and enablers of data sharing

The assemblage of *drivers, barriers and enablers* provides a comprehensive description of the factors that motivate, inhibit and enable the sharing of research data. These may be variously defined in terms of individual-psychological, social, organisational, technical, legal and political components. They affect *whether* data are shared, *how* they are shared, and *how successfully* they are shared. These factors were developed by examination of the process and context models, extracting and structuring the essential mechanisms that could encourage or inhibit data sharing, and were further refined and completed using the results of workshop discussion of the model involving data experts.

Table 3 lists the primary categories in which drivers and barriers and enablers are described; the rest of this section analyses these primary categories in terms of their key constituent factors.

Table 3: 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 and Enablers
Related to … 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

4.1 Drivers to data sharing

General benefits for data sharing are classified according to the category of beneficiary: society at large, academics and researchers, and different stakeholder organisations (Table 4).

Table 4: Benefits accruing from data sharing

Societal benefits	1. Economic/commercial benefits; 2. Continued education; 3. Inspiring the young; 4. Allowing the exploitation of the cognitive surplus in society; 5. Better quality decision making in government and commerce; 6. Citizens being able to hold governments to account.

Academic benefits	1. The integrity of science as an activity is increased by the availability of data; 2. Increased public understanding of science.
Research benefits	**For the data contributor:** 1. Validation of scientific results by other scientists; 2. Recognition of their contribution.
	For the data user: 3. Reuse 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); 4. To test new theories against past data; 5. To do new science not considered when data was collected without repeating the experiment; 6. To ease discovery of data by searching/mining across large datasets with benefits of scale; 7. To ease discovery and understanding of data across disciplines to promote interdisciplinary studies; 8. To combine with other data (new or archived) in the light of new ideas.
Organisational benefits	**Producer Organisation:** 1. Publication of high quality data enhances organizational profile; 2. Citation of data enhances organisation profile.
	Publisher Organisation: 3. Preserved data linked to published articles adds value to the product.
	Infrastructure Organisation: 4. Data preservation is more business; 5. Reputation of institution as "data holder with expert support" is increased.
	Consumer Organisation: 6. Meeting organisational need to combine data from multiple sources to make policy decisions; 7. Reuse of data instead of new data collection reduces time and cost to new research results; 8. Use of data for teaching purposes.
Individual contributor benefits	1. Preserving data for the contributor to access later - sharing with your future self; 2. Peer visibility and increased respect achieved through publications and citation; 3. Increased research funding; 4. When more established in their careers through increased control of organisational resources; 5. The socio-economic impact of their research (e.g. spin-out companies, patent licenses, inspiring legislation); 6. Status, promotion and pay increase with career advancement; 7. Status conferring awards and honours.

4.2 Barriers

The sections below list the key barriers identified within the conceptual model. The barriers are expressed in terms of disincentives related to particular areas: for example, the first set of barriers are counters to the individual contributor incentives already identified as drivers. It is worth noting that drivers tend to be in terms of overall aims, the barriers in term of specific concrete factors. The corresponding enablers are not enumerated here, for reasons of concision; however some of them emerge in the discussion of themes in section 5, and they may be seen in full in the ODE project reports available on the website. The full final report of the work undertaken in the project which is summarised in this paper can be found in [Dallmeier-Tiessen et al.].

4.2.1 Individual contributor incentives

Individual researchers may not be motivated to take the effort to contribute their research data, for reasons including the following.

1. Journal articles do not describe available data as a publication;
2. Published data is not recognized by the community as a citable publication;
3. There is a lack of specific funding in grants to address the pre-archive activities for data preservation;
4. There is a lack of mandates to deposit high quality data with appropriate metadata in preservation archives;
5. Journals do not require data to be deposited in a form where it can be reused as a condition of publication;
6. Data publication and data citation counts are not tracked and used as part of the performance evaluation for career advancement;
7. There is a lack of high status awards to individuals and institutions which contribute data that is reused.

4.2.2 Availability of a sustainable preservation infrastructure

Until there is a sustainable infrastructure for data preservation which facilitates data discovery and reuse, data producers will not make the effort to prepare data for publication. Specific barriers identified include:

1. Absence of data preservation infrastructure;
2. Charges for access to infrastructure;
3. Journals are not necessarily good at holding data associated with articles;
4. Lack of data reviewers to assure data quality;
5. Risk that data holders cease to operate, and archive is lost.

4.2.3 Trustworthiness of the data, data usability and pre-archive activities

In the pre-archive phase of data preservation data quality is checked, and the metadata gathered and linked to the data to make them usable. Data producers may struggle to validate the integrity of their data and create usable metadata for a number of reasons:

1. Not “feeling safe” in dealing with unfamiliar data;

2. Lack of expert support: data centre cannot have detailed technical knowledge of all data they handle;

3. Lack of clear definition of the level of data quality that the potential data users will require;

4. Interdisciplinary data requires a unifying factor for data to make reuse easier (e.g. data maps to a common geographical coordinate system).

4.2.4 Data Discovery

The principal barrier to data discovery may be expressed in terms of the fragmentation of discovery services and the lack of an integrated infrastructure to support international, cross-disciplinary data discovery.

4.2.5 Academic Defensiveness

Many researchers have concerns about the potentially negative consequences of making their data publically available. Data producers may be defensive about publishing data for a variety of reasons:

1. Concerns at the danger of data “being hacked” and not being preserved as it is;

2. Fear that analysis of data by others will invalidate their results;

3. Fear that others will gain benefit from their data;

4. Fear that misuse of data for unsuitable purposes will harm the data contributor;

5. Fear that misuse of data to justify arguments which the contributor would find unacceptable will harm the data contributor.

4.2.6 Finance

Archiving costs alone are argued to be small in studies of preservation costing [Beagrie, Lavoie and Woollard 31-52]. 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. The data ecosystem is composed of many stakeholders in relationships of mutual dependence and there are consequently numerous points where lack of financing can compound structural weaknesses:

1. Lack of pre-archive funding by contributor;

2. Lack of archiving funding by infrastructure;

3. Lack of data discovery and access funding;

4. Risk of lack of return on long-term investment in preservation infrastructure;

5. Risk of high costs in answering questions about projects or data after their funding has expired.

4.2.7 Subject anonymity and personal data confidentiality

There is a genuine need (as well as regulatory/statutory requirements) for researchers in medical and social science disciplines to preserve the anonymity of subjects who contribute data to their studies. Specific barriers include:

1. Lack of funding for anonymising data, which is costly;

2. Lack of agreed standards for anonymising data;

3. Lack of trust in the preservation infrastructure to prevent deanonymisation.

4.2.8 Legislation/regulation

There is often a perception, which might not reflect reality, of conflicts:

- between 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 (actual and perceived) create barriers to data sharing.

4.3 Discussion

The assemblage of drivers, barriers and enablers and affords a generic analysis of the factors that can promote or inhibit data sharing, but the degree to which factors affect a given case will depend on:

- the stage in the research cycle at which they apply;
- the stakeholders involved; and
- the specific context variables, including the discipline in which the data is produced and consumed, the research sector, the geopolitical context, and the disposition of the researchers involved.

For this reason, it is no easy matter for stakeholders to implement policies and practices that will have optimum overall results in promoting effective data sharing. Nevertheless, the study was able to identify from interviews conducted with reference to this model a number of areas where

strategic investment of resource and effort is likely to have a positive effect in the development of data sharing infrastructure, culture and practice. These are discussed in the following section.

5 Themes in data sharing

A number of salient themes emerged from analysing the collated evidence of the workshop and interviews conducted with reference to the model of drivers and barriers, which identify opportunities and challenges for stakeholders working to develop a more effective data sharing culture. The themes are partly aligned with elements of the process model, which is to be expected inasmuch as they relate to particular activities within the research cycle. Some themes however, such as 'standards and interoperability', cut across several elements of the model. These themes highlight areas where strategic effort and investment is required on the part of stakeholders in order generate positive effects in the development of data sharing. The themes that emerged were classified as follows:

Data publication themes

- The role of publishers in data sharing;
- Data citation and description for discovery and use;

Data management infrastructure themes

- Funding infrastructure and data services;
- Data management: skills training and expert support;
- Quality assurance of data.
- Standards and interoperability;

Culture and policy themes

- Data sharing culture;
- Public visibility of research data;
- National and regional policy and legal frameworks;
- Incentives in the academic reward system for good data practice.

In what follows, a selection of four themes has been made for discussion, reflecting a wide coverage of the research data sharing process and a variety of perspectives. This is not to imply that the other themes are of less interest or significance, but simply that for the purpose of this paper a degree of focus was desirable. A more comprehensive discussion of all themes may be found in the project reports available from the Opportunities for Data Exchange website.

5.1 The role of publishers in data sharing

From the evidence of interviews across all stakeholder groups, there was broad consensus that publishers have a major role to play in creating and supporting the infrastructure and services that allow data to be shared and discovered. This view was strongly espoused by those publishers who were interviewed, and on the evidence of publishers' policies and practices, is one widely held across the industry.

Many publishers (including Elsevier, the Institute of Physics, Sage, Springer and Wiley) support Principle 7 of the STM "Brussels Declaration": 'Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars.' Accordingly, most publishers express willingness to provide at least basic supplementary data citation and linking to data held in external repositories. All publishers consulted expressed interest in developing data services, both those based around supplementary datasets on their own platforms, and tools for discovering, linking, and using datasets held by external databases. Journal publishers' data hosting services are limited in scope and use, and publishers do not expect to assume a role in long-term preservation.

Although usage of publishers' supplementary data publishing services is growing, this is from a very low base. Partly this may be because these services are not actively promoted. One major publisher indicated that while individual journal editors have the freedom to promote data publication in their journals, this is a matter of editorial choice and not general publisher policy. And while some 'data journals' have emerged dedicated to the publication and description of datasets that are preserved and described according to defined standards, they have tended to come from community-based initiatives in well-defined disciplines, and have excited little interest on the part of larger commercial publishers. (Examples of data journals mentioned in interviews include *Earth System Science Data*, published by Copernicus Publications, and *Journal of Open Archaeology Data*, published by Ubiquity Press.)

Many interview respondents in other stakeholder groups were critical of the current state of data publishing, linking and citation, and felt that publishers could do more to enable data usage. The following points were made:

- Supplementary data may be presented with articles in a highly processed state, suitable for publication (e.g. in graphs or charts), but not for detailed analysis, data mining, or repurposing;
- Peer review processes or quality standards for supplementary data are rarely rigorous or transparent. Data may be submitted as part of an article peer review, and may be reviewed to some extent (often undefined), or may be submitted post-review. Supplementary data may be quality-assured only by minimal file integrity checks. This makes it very hard to establish a level of trust in the reliability and provenance of supplementary data made available with articles;
- Supplementary data citation may not meet user requirements. One major publisher declared a general policy of citing supplementary data by the article, and not separately, for the simple reason that there is an added cost to creating DOIs for datasets as separate entities;
- Data citation methods are various: citations may be formatted and placed inconsistently in articles, and can be difficult to locate or identify;
- Publishers can fail to identify data citation in submitted papers. Two interviewees cited instances of prominent journals removing or failing to include DataCite DOIs in article reference lists because they were not identified in editing as valid citations.

Nevertheless, several positive examples of collaboration involving publishers and other service providers and publicly-funded stakeholders were given:

- The Dryad biosciences data repository links data to published articles through standard DOI citation, agreed with its partner journals through a Joint Data Archiving Policy;

- Elsevier collaborates with the PANGAEA earth and environmental sciences data library for reciprocal linking [Elsevier]. This is a model that other institutions and disciplines are becoming interested in;
- DataCite and the CODATA Data Citation Standards and Practices Task Group have been working since 2010 to develop best practices for data citation and guidelines for the use of DOIs. DataCite has held discussions with STM, the International Association of Scientific, Technical and Medical Publishers, about citation practice, and has entered into agreement with the CrossRef DOI resolver service to implement interoperability of their DOIs;
- One publisher had been exploring more flexible file formats for supplementary data, notably Wolfram's Computable Document Format (CDF). This is a data representation format that builds algorithms into a portable document so that data can be both presented in a strong visual form and processed interactively;
- CrossRef has been developing CrossMark, a version control service that allows publishers to update DOI citations to publications that have been altered or withdrawn and alert citing sources to the change or withdrawal. Such a service could be valuable applied to datasets also, allowing for control of flawed datasets and research that potentially builds on flawed data or data that has since been corrected;
- In the UK the Managing Research Data Programme 2011-2013 run by the Joint Information Systems Committee (JISC) includes a focus on data publishing. In the REWARD Project university College London (UCL) and Ubiquity Press sought to encourage the archiving of research data using the UCL Discovery institutional repository linked to a data paper in the *Journal of Open Archaeology Data*. The objective of the project was to make the data citable and reuse trackable, important factors for the 2014 UK research assessment exercise, the Research Excellence Framework (REF). The project followed five case studies in order to assess the effectiveness of the systems involved.

But even where publishers are open to greater collaboration with key stakeholders, it is not necessarily a simple matter to establish viable partnerships. This may be for several reasons:

- There can be a lack of trust between commercial publishers and data centres and other publicly-funded service providers, which inhibits collaboration;
- There are no suitable data repositories in some disciplines, in particular in the humanities;
- Repositories may not follow best practice, e.g. in metadata standards, use of persistent identification;
- There can be a mismatch between the technological capabilities of publishers, e.g. in data management technologies and discovery tools, and those of potential partners;
- There are unresolved differences between stakeholders over issues of intellectual property and data rights. While publishers may argue that their use of copyright serves to protect intellectual property and guarantee its integrity, there is a widespread perception that copyright is used to restrict sharing and exploit data for commercial advantage. It will take a lot of engagement on the part of publishers to change perceptions.
- Publishers may see no commercial rationale for providing the services that other stakeholders ask for. There are very few data journals, and it may be that larger publishers do not see a viable market for such publications until there is general recognition in the academic system for data papers as research outputs commensurate with articles or conference papers.

There is a strong argument for the benefits of collaboration among publishers and with other stakeholders providing infrastructure and data services. By and large publishers appear to be open

to 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 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. But 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 critical areas such as implementing industry standards and best practice in data citation; incorporating robust data quality assurance and peer review in editorial processes; and bringing standalone data journals to market. Three points clearly emerge:

- There is a demand for the publishing industry to provide more data publication and data usage services than are currently provided, and there are sure to be business opportunities for publishers to exploit;
- Some of the best examples of the industry contributing to the growth of a rich data sharing culture are those where publishers have collaborated with publicly-funded organisations providing other data services, whether infrastructure services such as DataCite or data centres such as PANGAEA. For such collaborations to be successful may require open-mindedness on both sides;
- There is scope for publishers to collaborate among themselves in order to embed industry standards and best practice in data citation and description.

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 can 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 publishing standards.

5.2 Quality assurance of research data

Data centres exist not only to preserve data and make them available for reuse, but also to provide quality assurance, so that datasets can be used with confidence in their integrity and validity.

Any potential data user needs to assess a dataset for its relevance to their requirements and its validity as accurate, trustworthy data. The need is particularly acute for cross-disciplinary reuse, when the potential user might not have in-depth expertise and the ability to evaluate the data being considered. Data centres have a role to play in the application of rigorous quality assurance standards for any data that they accept, to validate both the data as such, through expert analysis, internal consistency checks and file integrity assurance, and the metadata by which datasets are described and made referenceable.

Dataset metadata and accompanying documentation should be sufficient for any researcher with reasonable subject expertise to assess the fitness of a particular dataset for an intended research purpose. There must also be a sufficient audit record to provide confidence in the dataset, which requires documentation of the origins of data and successive steps taken their curation, and also, in the case of processed data, tracking the methods, software and calibrations used to generate the dataset. This is particularly necessary for data with potentially a long lifetime of usefulness,

such as earth observation data, or where complex processing has been carried out. Data collection and processing can be highly specialised, and in some areas may depend on unique instruments, software and collection methods. Not only is considerable expense involved in training data experts able to undertake meaningful quality assurance of highly specialised datasets, but the actual processes of quality assurance and documentation can be labour-intensive and time-consuming.

In order for the quality assurance performed by data centres to have validity, there must in turn be quality assurance of data centres themselves, in respect of their authority to assure data quality, and of their trustworthiness to preserve data for the long term. These are distinct but interlinked capabilities; for example, a trustworthy repository should clearly associate certain information ('metadata') with the ingested content, and monitor the needs of the designated community, and these are both aspects of assurance of data quality. Data centres can be certified by standards such as those issued by the Data Seal of Approval Board, the Deutsche Initiative fűr Netzwerkinformation (DINI) Certificate, the Center for Research Libraries' *Trustworthy Repositories Audit and Certification* standard (TRAC) and ISO 16363:2012, *Space Data and Information Transfer Systems: Audit and Certification of Trustworthy Digital Repositories*. It is reasonable to expect that any viable data centre should strive to be certified to the highest possible relevant standard.

Granted that the guarantors of data over the long term have a responsibility to assure the quality of their holdings, quality assurance of data is not simply something that happens at the end of the research process when the final outputs are deposited for preservation. Quality assurance should begin with the data management plan for a research project, and should proceed hand in hand with data collection and processing through the lifetime of the project. Keeping full records of data collection and processing as they happen is both good research practice and invaluable preparation for the ultimate deposit of data with the preservation service and their formal quality assurance during the ingest process.

5.3 Standards and interoperability

Internet technologies have facilitated the growth of interdisciplinary data sharing in forms and at volumes previously inconceivable. Untold potential has been released, for example, by the combination of massive data sets from different disciplines and sources in the social and medical sciences, or by using common reference data systems, such as Geographical Information Systems (GIS), to bring disparate sets of data together. But in many respects interdisciplinary data sharing is at a very early stage. Even where infrastructure and standards within disciplines may be well-developed, these often do not interoperate effectively with infrastructure and standards outside of the disciplinary domain. This can frustrate researchers' efforts to discover, understand and use data from outside of their own disciplinary zone.

Efficient machine discovery and interpretation processes are becoming more and more important to the researcher. Many interviewees 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.

Data sharing requires common infrastructures, rules and semantics. These common requirements are expressed and realized as standards among the communities in which data sharing takes place. Standards may exist at different levels of community, from the global to the highly localized. Two distinct domains of interoperability can be identified:

- *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 interviewees 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.

- *Data formatting*: assuming that distributed datasets 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.

The absence of well-established standards in both data description and data formatting are significant barriers to data discovery and use. Establishing those standards presents 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. Some standards have already become well-established (for example, DataCite DOIs), and there are ongoing initiatives and projects working to promote the adoption of standards, such as the European INSPIRE directive to establish an infrastructure for spatial information [European Commission, "INSPIRE"].

Key areas in which there is a need for greater effort towards the establishment of standards include:

- Common identifiers and resolvers for basic entities such as datasets and authors should be applied by data centres and publishers: the DOI system and the ORCID universal researcher ID initiative were both mentioned as examples of cross-industry approaches to developing standards and solutions for a global scholarly communications infrastructure.

- Publishers should establish industry standards for data citation, just as there are currently well-established standards for citing other forms of publication.

- The presumption should in all cases be in favour of open standards: this applies to both description and data formatting. The proprietary PDF is widely used by publishers as a format for supplementary data, but it is essentially data-unfriendly: it cannot easily accommodate large datasets, it does not have the capacity to create well-structured data sets, and in aggregate PDF files are not amenable to machine processing. This is a prime example of the adverse consequences that arise from applying a proprietary format in a system that must be open to be effective; there are counter-examples of the successful adoption of open file formats, such as the Crystallographic Information File (CIF), a standard format for representing crystallographic information, promulgated by the International Union of Crystallography (IUCr).

- Semantic Web/Linked Data approaches are crucial for creating data discovery pathways for users navigating unfamiliar metadata schemas and ontologies. It is relatively easy to construct domain-specific data preservation tools, but synthesising the metadata into integrated discovery services is a challenge on a different scale. One researcher in the bioinformatics field stated that there are some 250 metadata standards for various kinds of

data. It is a challenge for researchers to select the most appropriate standard; and it is a challenge for discovery service providers to integrate the domain. Different kinds of data require different standards; but managing multiple domain-specific data repositories soon becomes a problem of scalability and effective management.

While peer group inside knowledge may to some extent compensate for lack of standards at a local level within disciplinary communities, this tacit knowledge is not available to disciplinary outsiders, and so the capacity of data to travel across disciplinary boundaries can be seriously compromised by poor description and formatting. But it is only as the culture of interdisciplinary data use grows that solutions will start to be applied. Technical solutions to mapping metadata, data formats and systems and synthesizing them into integrated discovery services can help researchers communicate across disciplinary barriers, and so build bridges between different domains and forge new and creative combinations of data.

The critical stage in the lifetime of a given output of data is that phase where it goes from being working research data to a defined dataset, which is ingested into a store and acquires its storage format and structure and descriptive wrapper. How standards are applied to the dataset at this stage will determine the destiny of that dataset: it will affect how easily the data can be shared, discovered and used, and will ultimately affect the impact of the data set. Getting it right at this stage requires skills and effort on the part of researchers, data specialists and service providers to ensure that data are described correctly and appropriately and are stored in a manner consistent with their anticipated use.

All of this implies a need both for the standards and infrastructure, and for training of researchers and data managers in how to understand and apply them. In practice researchers are often careless and ill-informed about standards when it comes to preservation of their data. Indeed, if the motivation to share data is weak or does not exist, there is little or no incentive to invest the time and effort required to help other researchers find and use the data. A large part of the problem of standards is making the researchers themselves understand the fundamental importance of sharing data at all, before they begin to consider how data can be shared most efficiently.

A key requirement here will be enhanced status for data preservation and publication within the academic system. When these activities are recognized as research outputs in their own right of a standing commensurate with formal publications, and are subject to the same degree of scrutiny by peer-reviewers, funders and bodies that evaluate research for quality and impact, then researchers will be incentivized to take the steps that will maximize the visibility and impact of these outputs.

Developing standards is necessarily a community effort that engages multiple stakeholders. For example, the INSPIRE Directive established a European infrastructure for spatial information to support Community environmental policies and activities with environmental impact. It is ‘based on the infrastructures for spatial information established and operated by the 27 Member States of the European Union’, and is both a technical specification that has led to open interfaces and greater interoperability, and a model regional policy framework for participating member states.

In another example, the European Data Infrastructure (EUDAT) initiative is seeking ‘to support a Collaborative Data Infrastructure which will allow researchers to share data within and between communities and enable them to carry out their research effectively’. To that end it is working to

build the common service architecture and trust framework that will enable communication between data systems and the development of a rich service layer.

Standards evolve in unpredictable ways: they often emerge from communities of specialized knowledge and practice, and grow by gradual extension and acceptance across different areas of a community, by absorption of other standards, or by migration to other communities and adaptation to new contexts. Moreover there are hardly any universal or fixed standards: rather, there is a variety of competing and constantly evolving standards, promulgated and championed by different stakeholders in the scholarly communication system: researchers, publishers, computer scientists, information scientists and others. Which standards become dominant or widely accepted may depend on a number of often accidental factors.

Nevertheless with robust frameworks, community engagement, and education in basic good practice, it should be possible to improve systems interoperability and foster the adoption of high-quality standards in data description and formatting. Semantic ontologies and ontology mapping, well-structured open data formats, and intelligent discovery interfaces that can parse user requests are all technical requirements. But standards also go to the heart of basic data management on the part of researchers and data centres. Data centres can play a role by ensuring standards are applied to data that are submitted to them, and by working to embed the knowledge of those standards among their research communities through outreach and proactive engagement with research organisations.

5.4 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. Cultures are governed by behavioural norms, which may be expressed as rules and codes of practice, although for the large part they are absorbed into customary practice as simply the way things are done.

As a general rule, where research units are more distinctly defined within a given community, and where the data processing requirement does not exceed the capacity of the typical research unit to process the data, the tendency to share data is less marked. For example, in bio-molecular sciences, astronomy and areas of earth sciences the size of the datasets and the amount of processing they require necessitates a culture of collaboration and open data sharing. In other areas, such as medical sciences or chemistry, highly-focused research projects may be conducted by small teams and produce small datasets that require minimal processing. The production and use of these data are much more closely allied to professional benefits for the individual researchers, leading to a more competitive culture that does not support data sharing.

There may be other factors that inhibit the growth of a data-sharing culture: where large amounts of confidential personal data are used, as in the medical and social sciences, there are strict legal constraints controlling the publication of such data, and in many cases the cost of compliance is prohibitive. In other areas, including medical sciences, chemistry and engineering, research may be in whole or in part funded from commercial sources, and may be subject to commercial confidentiality requirements. Where data have actual or potential commercial value, there may be a presumption against sharing.

Technological development is in itself a driver for cultural change, and interviews conducted as part of the study found as a general rule that younger researchers and those with greater technological literacy are more open in their attitudes to data sharing. Certainly, as technologies advance the benefits of data sharing might be expected to become more apparent and easier to obtain. Improved data collection, description, deposit, citation, and discovery technologies will allow data to travel faster and further, and researchers will perceive the benefits of accelerated and enhanced impact for their research. Researchers will become more active in data sharing as data becomes more citable and linkable, and are recognised in assessment and evaluation.

An interesting potential accelerant of change in data sharing cultures is the growth of interdisciplinary data use, itself a result of the possibilities unleashed by the internet. A multiplier effect is at work: as more and more data becomes available, including data not originally obtained for scientific purposes, such as public sector data from government bodies, the opportunities become more numerous and the expectations higher. There is a reinforcement of the perception of data as a public good, produced at the taxpayer's expense and held on trust for the wider community, rather than as the private property of the researcher, and it makes the uses to which data are put transparent and accountable. The theme of public visibility of research data was one which was explored in detail, but is beyond the scope of this paper.

It may take a long time for attitudes to change in areas where there is no deep-rooted culture of data sharing. But clearly there are changes to policies and systems that can be made to encourage the development of such a culture. Policy-makers and research funders have a role to play in mandating data sharing and enforcing compliance. Where personal or commercially-sensitive data are involved, data centres can find improved ways to manage these issues and provide guidance and support to researchers. It is of course critical for researchers to receive training in data sharing at an early stage and to continue to benefit from expert support throughout their research careers.

6 Conclusions

This study confirms that the sharing of research data is one of the key challenges and opportunities in this era of e-science. Researchers, policy-makers, funders, service providers and publishers have made advances across many disciplines in recent years to facilitate the sharing and reuse of research data. But there are many barriers to maximising the reuse of data, for all stakeholders and at all points in the data sharing system. Of course it is hoped that these barriers can be overcome, and the thematic discussions include some hints and directions for the future. All stakeholders will need to have some engagement with the issues raised, and to act in concert and with understanding of the impact of their decisions and behaviour on others in the overall data lifecycle. The challenges are complex, and it is impossible to reduce the findings too far without risk of over-simplification. But it is possible to enumerate a number of key topics, recurring across the themes and linked to multiple drivers and barriers, that can be summarised as follows.

- **Data citation:** establishing principles, infrastructure and practices for clear and precise citing of others' data.
- **Evaluation of data in research assessment:** ensuring that openness with data and quality of data are rewarded, thereby providing a powerful incentive for researchers to share their data.
- **Data discovery and reuse:** making it easier for researchers to find the data they need.

- **Review and quality assurance of data:** giving confidence in the quality of data and its fitness for reuse.
- **Infrastructure for data management and curation:** establishing division of responsibilities, technical bases and sustainable funding.
- **Training:** equipping researchers with the required knowledge and skills for making their data effectively shareable.
- **Standards for different purposes:** developing or taking advantage of standards to facilitate the representation of provenance, interoperability, discovery and other uses built on aggregation.
- **Reuse linked to preservation:** understanding that openness and preservation of data are intimately linked, and that techniques and approaches for one will facilitate the other.

The challenges concern all those who hold a stake in data sharing. The conceptual model of data sharing developed as part of this study can help researchers, policy-makers, funders, and data service providers identify the barriers that need to be overcome to enable data sharing. For those stakeholders who provide and maintain data sharing systems the model can be used in developing strategies that will address gaps in infrastructure, service provision and funding; and it can enable researchers to identify the barriers to data sharing they encounter, and to create strategies and data plans to overcome those barriers. The thematic analyses of different aspects of data sharing, along with the proposed key actions stakeholders could take to reinforce data sharing, demonstrate how the model can help stakeholders to structure the development of strategies and practices that promote data sharing.

It was not in the scope of the ODE project to produce concrete recommendations. The areas for action proposed in the thematic analyses have emerged from discussions with stakeholder groups and represent the thinking of interested and expert participants in the field of data sharing about the present state of the field and how it could or should evolve. In this light it is hoped that this work might represent a foundation on which others can build.

References

Beagrie, N., Lavoie, B. and Woollard, M. *Keeping Research Data Safe 2: Final Report*. JISC, 2010. Web. 5 August 2013. <http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>.

Beagrie, N. Chruszcz, J. and Lavoie, B. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. JISC, 2008. Web. 5 August 2013. <http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>.

Birnholtz, J. P. and Bietz, M. J. "Data at Work: Supporting Sharing in Science and Engineering". *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, FL, USA, November 9-12 2003*. New York: ACM, 2003. Print.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. *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*. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010. Web. 5 August 2013. <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf>.

Borgman, C. L. "Research Data: Who will share what, with whom, when, and why?" RatSWD Working Papers 161 (2010). Web. 5 August 2013. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714427>.

The Center for Research Libraries. *Trustworthy Repositories Audit and Certification: Criteria and Checklist*. OCLC and CRL, 2007. Web. 5 August 2013. <http://www.crl.edu/PDF/trac.pdf>.

CODATA. "Data Citation Standards and Practices Task Group". International Council for Science: Committee on Data for Science and Technology, 2010. Web. 5 August 2013. <http://www.codata.org/taskgroups/TGdatacitation/index.html>.

Consultative Committee for Space Data Systems. *Reference Model for Open Archival Information Systems (OAIS): Draft Recommendation for Space Data System Standards*. *(ISO 14721).* CCSDS, 2009. Web. 5 August 2013. <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf>.

CrossRef. Web. 5 August 2013. <http://www.crossref.org/>.

CrossRef. "CrossMark". CrossRef, 2012. Web.5 August 2013. <http://www.crossref.org/crossmark/>.

Dallmeier-Tiessen, S. et al. *Compilation of Results on Drivers and Barriers and New Opportunities*. Alliance for Permanent Access, 2012. Web. 5 August 2013. <http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf>.

DataCite. Web. 5 August 2013. <http://www.datacite.org/>.

DataCite Metadata Store. Web. 5 August 2013. <https://mds.datacite.org/>.

Data Seal of Approval Board. "Data Seal of Approval". Data Seal of Approval, 2013. Web. 5 August 2013. <http://datasealofapproval.org/>.

Deutsche Initiative für Netwerkinformation. "DINI Certificate 2010 for Document and Publication Services". DINI, 2013. Web. 5 August 2013. <http://www.dini.de/dini-zertifikat/english/>.

Dryad. "Joint Data Archiving Policy". Dryad, 2013. Web. 5 August 2013. <http://datadryad.org/pages/jdap>.

Earth System Science Data. Copernicus Publications. 2009 to date. Web. 5 August 2013. <http://www.earth-system-science-data.net/>.

Elsevier. "Elsevier and PANGAEA link contents for easier access to full earth system research". Elsevier, 2010. Web. 5 August 2013. <http://www.elsevier.com/about/press-releases/science-and-technology/elsevier-and-pangaea-link-contents-for-easier-access-to-full-earth-system-research>.

EUDAT. Web. 5 August 2013. <http://www.eudat.eu/>.

European Commission. "INSPIRE. Infrastructure for Spatial Information in the European Community". European Commission, 2007. Web. 5 August 2013. <http://inspire.jrc.ec.europa.eu/>.

Gardner D. et al. "Towards Effective and Rewarding Data Sharing." *Neuroinformatics* 1.3 (2003): 289-95. Print.

Hodson, S. “Data-sharing culture has changed”. *Research Information* Dec. 2009/Jan. 2010. *Research* Information. Web. 5 August 2013. <http://www.researchinformation.info/features/feature.php?feature_id=243>

ISO. *ISO 16363:2012. Space Data and Information Transfer Systems: Audit and Certification of Trustworthy Digital Repositories*. ISO, 2012. Web. 5 August 2013. <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510>.

JISC. “Managing Research Data programme 2011-2013”. JISC, 2011. Web. 5 August 2013. <http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx>.

Journal of Open Archaeology Data. Ubiquity Press. 2012 to date. Web. 5 August 2013. <http://openarchaeologydata.metajnl.com/>.

Key Perspectives. “Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. A Comparative Review Based on Sixteen Case Studies". Digital Curation Centre, 2010. Web. 5 August 2013. <http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf>.

Kuipers, T. and van der Hoeven, J. *D3.4 Survey report, PARSE.Insight: INSIGHT into Issues of Permanent Access to the Records of Science in Europe.* PARSE.Insight, 2009. Web. 5 August 2013. <http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf>.

OECD. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD, 2007. *OECD.* Web. 5 August 2013.

Opportunities for Data Exchange. *Opportunities for Data Exchange*. Alliance for Permanent Access, n.d. Web. 5 August 2013. <http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/>.

ORCID. Web. 5 August 2013. <http://orcid.org/>.

PANGAEA. Web. 5 August 2013. <http://www.pangaea.de/>.

Parr, C. and Cummings, M. "Data Sharing in Ecology and Evolution." *Trends in Ecology and Evolution* 20.7 (2005): 362-63. Print. HCIL-2005-06, CS-TR-4708, UMIACS-TR-2005-16.

The Royal Society. *Science as an Open Enterprise*. London: The Royal Society, 2012. *The Royal Society*. Web. 5 August 2013.

Schäfer, A. et al. *Baseline Report on Drivers and Barriers in Data Sharing*. Alliance for Permanent Access, 2011. Web. 5 August 2013. <http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf>.

Science and Technology Facilities Council. “ODE-Opportunities for Data Exchange.” STFC e-Science. Author 2014. Web. 25 March 2014. <http://www.stfc.ac.uk/e-Science/projects/long-term/39189.aspx>.

STM. “Brussels Declaration”. International Association of Scientific, Technical and Medical Publishers, 2007. Web. 5 August 2013. <http://www.stm-assoc.org/brussels-declaration/>.

Tenopir, C. et al. “Data Sharing by Scientists: Practices and Perceptions”. *PLoS ONE* 6.6 (2011): e21101. Web. 5 August 2013. <doi:10.1371/journal.pone.0021101>.

UCL and Ubiquity Press. “The REWARD Project”. UCL, 2011. Web. 5 August 2013. <http://www.ucl.ac.uk/reward/>.

Van den Eynden, V. et al. *Managing and Sharing Data: Best Practice for Researchers. 3rd edition*. UK Data Archive, 2011. Web. 5 August 2013. <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>.

Wheatley, P. and Hole, B. "LIFE3*:* Predicting Long Term Digital Preservation Costs". *iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*. California Digital Library, 2009. Web. 5 August 2013. <http://escholarship.org/uc/item/23b3225n>.

Wolfram. “Introducing the Computable Document Format (CDF)”. Wolfram, 2013. Web. 5 August 2013. <http://www.wolfram.com/cdf/>.

Acknowledgments

The work reported was partially funded by the European Commission FP7 ICT grant 261530 to the Opportunities for Data Exchange (ODE) project, 2010-12 within the European Commission’s Research Infrastructures Work Programme, funded under the FP7 Capacities programme (European Commission, e-Infrastructure).

The authors acknowledge the contribution of their partners in the ODE project and the cooperation of all those who were interviewed in the study. We dedicate this paper to Professor Michael Wilson who suddenly passed away during its preparation. He was closely involved with the work of the ODE project, and we are indebted to his ideas and insights.