Opportunities for Data Exchange

SEVENTH FRAMEWORK PROGRAMME

COMPILATION OF RESULTS ON DRIVERS AND BARRIERS AND NEW OPPORTUNITIES

09 July 2012

Sunje Dallmeier-Tiessen [a], Robert Darby [b, *], Kathrin Gitmans [c], Simon Lambert [b], Jari Suhonen [d], Michael Wilson [b]

[a] CERN. CH1211 Geneva 23. Switzerland
[b] STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX, United Kingdom
[c] Helmholtz Open Access Coordination Office, c/o Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
[d] CSC - IT Center for Science Ltd. P.O. Box 405, FI-02101 Espoo, Finland

* Corresponding author: robert.darby@stfc.ac.uk

Please cite as: Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Suhonen J, Wilson M (2012). Compilation of Results on Drivers and Barriers and New Opportunities. Retrieved from [URL].

EXECUTIVE SUMMARY

Opportunities for Data Exchange (ODE) is a FP7 Project carried out by members of the Alliance for Permanent Access (APA), which is gathering evidence to support strategic investment in the emerging e-Infrastructure for data sharing, re-use and preservation. The ODE Conceptual Model has been developed within the Project to characterise the process of data sharing and the factors which give rise to variations in data sharing for different parties involved. Within the overall Conceptual Model there can be identified models of *process*, of *context*, and of *drivers, barriers and enablers*. The Conceptual Model has been evolved on the basis of existing knowledge and expertise, and draws on research conducted both outside of the ODE Project and in earlier stages of the Project itself (Sections 1–2).

The *process* model describes the functional logic of data sharing in terms of agents, actions and objects. The *context* model 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 (national/regional policy and legislation, infrastructure, funding).

The model of *drivers, barriers and enablers* provides a comprehensive description of the factors that motivate, inhibit and enable the sharing of research data. Drivers, barriers and enablers are 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.

The Conceptual Model was validated, refined and elaborated through a process of consultation and review with expert and interested members of the key stakeholder groups (Section 3). This validation process was conducted in two stages: a workshop on data sharing was held in appendix to the APA Conference in November 2011, in which a group of data sharing experts provided feedback on the Model; and between February and April 2012 telephone interviews based on the Model of drivers and barriers were conducted with 55 individual members of different stakeholder groups, including researchers in all the major disciplinary areas.

Discussions with informed and expert members of different stakeholder groups also served to identify salient issues and converging views in respect of the drivers and barriers that bear on data sharing activities. These have been discussed in thematic sections that provide interpretive summaries of the views and experiences of workshop participants and interviewees (Section 4). The following themes are discussed:

- The role of publishers in data sharing;
- Finance: funding infrastructure and data services;
- Data management: skills training and expert support;
- Standards and interoperability;
- Data citation and description for discovery and use;

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

For each theme a summary of the views and experiences discussed is given, followed by a brief analysis of the most salient drivers and barriers and the enablers that stakeholders can implement to surmount or reduce the operative barriers.

Thematic analysis led to the formulation of a data sharing domain evaluation tool, which might serve to assess the maturity of a data sharing domain by the presence and strength of certain indicators (Annex 2). This is proposed as a high-level domain analysis tool that may be useful in identifying areas that need to be addressed in policy—though it is not part of the Conceptual Model itself.

In conclusion the outputs of this phase of the ODE Project are considered in the context of the European Commission's Horizon 2020 initiative for a global data infrastructure and a digital research area for Europe. The Conceptual Model of data sharing drivers, barriers and enablers and the data sharing domain evaluation are proposed as tools that could have practical value in elucidating the relationships between the Horizon 2020 goals and the conditions needed to bring them about, and could support those charged with formulation and implementation of policy in this area.

1 INTRODUCTION TO THE ODE CONCEPTUAL MODEL

1.1 PURPOSE AND SCOPE

Opportunities for Data Exchange (ODE) is a FP7 Project carried out by members of the Alliance for Permanent Access (APA), which is gathering evidence to support strategic investment in the emerging e-Infrastructure for data sharing, re-use and preservation.[1] The aim of the ODE Project has been to engage in dialogue with relevant stakeholders, in order to collect and document views and opinions on challenges and opportunities for data exchange.

Public funders of research increasingly agree with 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 (OECD, 2007)[2]. In practice data sharing in and among research communities is variable and unevenly distributed. While there are certainly drivers which have encouraged some research communities to share some types of data, there are many barriers preventing some communities from sharing any data, and all researchers from sharing some types of data.

Hodson (2009) summarises the commonly accepted barriers to data sharing:

> *…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.*

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. Examples of successful data sharing can present a persuasive case to stakeholder organisations. But to arrive at the stage where re-use 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 identification of the enablers through which barriers can be overcome.

The ODE Conceptual Model is designed for this purpose and contains analytic representations of the data sharing system under different aspects. Within the overall

[1] http://ode-project.eu.

[2] For example: the US National Institutes of Health (NIH) *Data Sharing Policy and Implementation Guidance* (NIH, 2003); the Wellcome Trust *Policy on Data Management and Sharing* (Wellcome Trust, 2007; 2010); the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) *Proposal Preparation Instructions: Project Proposals* (DFG, 2010; 2012); and the US National Science Foundation (NSF) *Data Sharing Policy* (NSF, 2011).

Conceptual Model there can be identified models of *process*, of *context*, and of *drivers, barriers and enablers*.

The *process* model describes the functional logic of data sharing in terms of agents, actions and objects; the *context* model maps the systemic scholarly communication context in which data sharing occurs; and the model of *drivers, barriers and enablers* provides a comprehensive description of the factors that motivate, inhibit and enable the sharing of research data. These component models taken together constitute the overall Conceptual Model of data sharing.

Different stakeholder groups have partial views of the data sharing process, context, and drivers, barriers and enablers according to their primary interests. The Conceptual Model is designed to provide a rounded representation of the data sharing system that incorporates the views of the different stakeholders – librarians, publishers, data centre service providers, funding bodies, infrastructure providers, researchers, citizen scientists, and organisations in the commercial sector (including software developers, publishers, and providers of citation services).

The Conceptual Model has been evolved on the basis of existing knowledge and expertise, and draws on research conducted both outside of the ODE Project and in earlier stages of the Project itself. It has also been carefully validated through a process of consultation and review with expert and interested members of the key stakeholder groups, described in detail below.

The Conceptual Model is proposed as a robust validated model in-the-round of the data sharing process, context, and drivers, barriers and enablers. It is a solid basis on which to develop an understanding of data sharing today, analyse the factors that motivate, enable and inhibit data sharing, and formulate requirements in order to achieve the mature culture of data sharing anticipated by the OECD recommendation.

1.2 Baseline for the Conceptual Model

The baseline from which the ODE Conceptual Model has been developed was established from existing published knowledge and from prior activity within the ODE Project (Reilly et al., 2011; Schäfer et al., 2011).

1.2.1 Sources consulted

The key published sources consulted in development of the Conceptual Model are listed in the Bibliography. They include studies on the benefits of preservation, barriers to preservation, costing of preservation, data sharing communities, and differences between disciplines in attitudes to data sharing. Many of these studies provided analytical representations of data preservation and sharing systems and processes, and enumerated drivers, barriers and enablers that bear on success and failure in data sharing. They were used to inform development of the data preservation and sharing

process and context models, and to elaborate a comprehensive list of drivers, barriers and enablers in data sharing.

The process and context models of data sharing (Section 2.1 and 2.2) were developed with particular reference to the OAIS Reference Model for long term digital preservation and access (ISO14721:2003; CCSDS, 2009) and a variety of digital preservation lifecycle costing models. The OAIS Reference Model, which has been extended and validated in the SHAMAN[3] and CASPAR[4] digital preservation projects, provides a functional representation of the data preservation process, including ingest, archival storage, data management, access, and dissemination.

There have been a number of studies on costing digital preservation, most based on lifecycle Activity Based Costing (ABC), where the overall process is divided into its component activities, which are then added together to arrive at a total cost for digital preservation. These all provide models for breaking down the data preservation and management life cycle into discrete components. Examples consulted include:

- The Princeton DataSpace Model (Goldstein and Ratliff, 2010), a basic ABC Pay Once Store Forever (POSF) costing model;
- The LIFE[3] digital preservation costing model (Wheatley and Hole, 2009), which describes the following activities for the preservation lifecycle: Acquisition, Ingest, Metadata, Bit storage, Content preservation, and Access;
- The Keeping Research Data Safe (KRDS) Project cost framework for long term data preservation, which can be used to generate local cost models (Beagrie et al., 2008; Beagrie et al., 2010).

These studies largely focused on preservation roles and activities. The scope of the ODE Project embraced data sharing more broadly, to include data discovery, access and re-use in addition to preservation. While the Conceptual Model draws on existing preservation process models to a large extent, it also shifts the emphasis to data activities within the overall scholarly communication system, and models roles and activities related to data discovery, access and re-use.

Two studies proved useful in elaborating the model of drivers, barriers and enablers (Section 2.3): a large-scale survey of researchers, publishers and data managers on barriers to digital preservation and re-use of data conducted by the PARSE.Insight Project (Kuipers and van der Hoeven, 2009); and the KRDS Benefits Framework for long term data preservation.

The PARSE.Insight survey of researchers, data managers and publishers 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.

[3] Sustaining Heritage Access through Multivalent Archiving. http://shaman-ip.eu/
[4] Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. http://www.casparpreserves.eu/

The KRDS Benefits Framework 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.

Other sources consulted included a longitudinal study providing in-depth insight into data sharing and the evolution of academic trust networks (Wilson, 2008), and a comparative study of data sharing in different academic disciplines (Key Perspectives Ltd, 2010).

1.2.2 HYPOTHESES ON DRIVERS AND BARRIERS FROM THE ODE SURVEY

Hypotheses on the benefits and barriers to digital data sharing and re-use were derived from the 21 interviews undertaken with experts in the ODE Project and reported in the *Baseline Report on Drivers and Barriers in Data Sharing* (Schäfer et al, 2011). Interviewees were selected from a range of stakeholder groups, including data providers, data users and data infrastructure providers. These hypotheses directly informed the model of data sharing drivers, barriers and enablers described in Section 2.3.

2 THE ODE CONCEPTUAL MODEL

The ODE 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 describes the functional logic of data sharing in terms of agents, actions and objects. The model builds on the Open Archival Information System (OAIS) Reference Model of digital preservation (International Organization for Standardisation, ISO 14721:2003; Consultative Committee on Space Data Systems, CCSDS, 2009). But the process model describes the digital data sharing process as a socio-technical whole consisting of dissemination and use activities in addition to preservation proper. The model has been used as a key analytical tool to derive the model of data sharing drivers and barriers.

The *context* model maps 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 (national/regional policy and legislation, infrastructure, funding).

The model of *drivers, barriers and enablers* is designed to provide 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.

2.1 DATA SHARING PROCESS

2.1.1 PREMISES OF THE DATA SHARING PROCESS

The process model for data sharing assumes that the aim of research is to achieve social and economic impact. This can be achieved in different disciplines in different ways: for example, in the social sciences through changes to social policy; in engineering disciplines by the creation of new technologies which can be exploited commercially; and in the biosciences by the development of new medicines which can be exploited commercially to improve the health of the population.

Research is both cumulative and currency-driven: researchers require access to existing research and underlying data, both in historical archives and in accessible stores of the latest outputs. This in turn implies a requirement on researchers to share their research data as early as possible in the research process. To be shared effectively, data must be meaningful, that is, stored, described and organised in such a way that others can find, access, understand and use them. As Attwood et al. (2009) argue:

> *Merely increasing the amounts of information we collect does not in itself bestow an increase in knowledge. For information to be usable it must be stored and organised in ways that allow us to access it, to analyse it, to annotate it and to relate it to other information; only then can we begin to understand what it means; only with the acquisition of meaning do we acquire knowledge. The real problem is that we have failed to store and organise much of the rapidly accumulating information (whether in databases or documents) in rigorous, principled ways, so that finding what we want and understanding what's already known become exhausting, frustrating, stressful and increasingly costly experiences.*

Van den Eynden et al. (2011) describe various ways to share research data, including:

- depositing them with a specialist data repository, data centre, data archive or data bank;
- submitting them to a journal to support a publication;
- depositing them in an institutional repository;
- making them available online via a project or institutional website;
- making them available informally between researchers on a peer-to-peer basis.

Each of these ways of sharing data has advantages and disadvantages: data centres may not be able to accept all data submitted to them; institutional repositories may not be able to afford long-term maintenance of data or support for more complex research data; and websites are often ephemeral with little sustainability. Consequently, approaches to data sharing may vary according to research environments and disciplines, due to the varying nature of data types and characteristics, and the resources available to the community.

2.1.2 THE DATA SHARING PROCESS MODEL

The data sharing process model is a combination of two component processes:

- the research process, where data is consumed, produced, processed and interpreted; and
- the data preservation process, where data preservation and sharing feed back into other research processes.

The data sharing process as a synthesis of the component research and preservation processes is shown in Figure 1 below.

Different actors are engaged at different stages in these processes: research planners (usually senior research staff), research funders, researchers, publishers (and suppliers of supplementary services such as citation indexes), data centre managers (possibly library managers), data centre staff (possibly library staff), infrastructure providers, and suppliers of supplementary services, such as data discovery.

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 the research process is *scientific publication*, which in turn leads to the indirect outcomes of *social and economic impact*. Although this is not explicit in the process model, it should be noted that the path to social and economic impact need not necessarily pass through formal scientific publication: re-use of exchanged data by industry or policy makers could itself produce socio-economic impact without accompanying scientific publications.

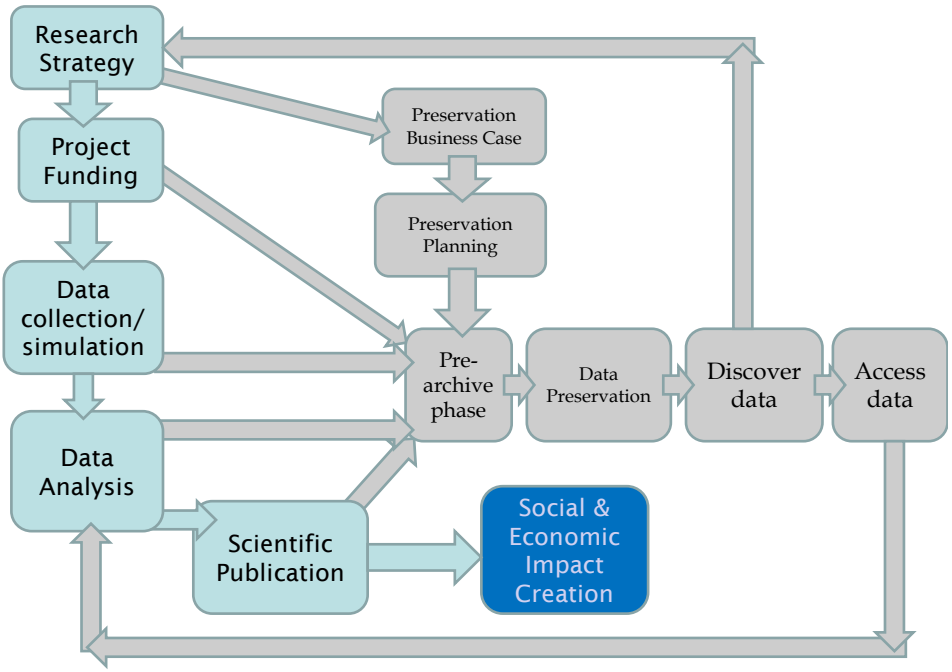

FIGURE 1. THE DATA SHARING PROCESS, COMBINING THE RESEARCH PROCESS (IN LIGHT BLUE) AND THE DATA PRESERVATION PROCESS (IN GREY)

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* generated during the research process.

The division of the research process into *data collection* and *data analysis* raises the issue of which data should be preserved to enable sharing and re-use. In many disciplines *raw data* are collected, then normalised or calibrated; then through the analysis process sets of *derived data* are produced at each stage, before the final *resultant data* are produced. Resultant data are usually the data which are published or archived when data preservation is a requirement of publication. However, in order to replicate results data from earlier stages are often required. Different disciplines treat these data sets differently.

Data analysis may include combining data from multiple sources. Access to each individual data set may become easier, but the convenience of analysing multiple types of data, and being able to cope with large amounts of data, requires automated support, which in turn requires that the appropriate metadata is available. Critical parts of this metadata must be captured during the initial preservation process to form the basis of the ongoing preservation activities. As Van den Eynden et al. (2011), argue:

> *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.*

This underlines the fact that 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 potential role of supplementary services to digital preservation such as data discovery or search engines, which may be integrated across many data archives. These could be generic (e.g. figshare[5], DataCite[6]) or specialised to different disciplines (e.g. DRYAD[7] in Biosciences, PANGAEA[8] in Earth sciences). Discovery services could also link to other supplementary services, for example, linking citation counts on published articles to the data supporting the article or even citation counts on that data[9]. There is scope for novel integrating services to support data discovery, which could be provided by several of the actors in the process.

2.2 DATA SHARING CONTEXT

The context variables apply to situations where data sharing could take place.

2.2.1 DATA SHARING ROLES

Functional roles are described in the following table in terms of the key stakeholder groups to which they belong.

[5] http://figshare.com/
[6] http://search.datacite.org/ui
[7] http://datadryad.org/
[8] http://www.pangaea.de/
[9] See for example the INSPIRE service for high energy physics data citations: http://inspirehep.net/

Stakeholder group	Roles
Policy-makers	National policy-makers International 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

2.2.2 DATA SHARING VARIABLES

The key variables in data sharing are described below.

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

2.3 DATA SHARING DRIVERS, BARRIERS AND ENABLERS

DRIVERS

a) Societal benefits
b) Academic benefits
c) Research benefits
d) Organisational incentives
e) Individual contributor incentives

BARRIERS

f) Individual contributor barriers
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

Each driver and barrier is described below with enablers to overcome the barriers and promote the benefits of data sharing. For ease of reference the enablers are described following the barriers they overcome, rather than altogether in a list following all the barriers as above.

a) Driver: 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.

b) Driver: Academic benefits

1) The integrity of science as a activity is increased by the availability of data;

2) Increased public understanding of science.

c) Driver: Research benefits

Usual reasons given to preserve research data for sharing are benefits for the re-user, not the contributor:

1) Validation of scientific results by other scientists;
2) 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);
3) To test new theories against past data;
4) To do new science not considered when data was collected without repeating the experiment;
5) To ease discovery of data by searching/mining across large datasets with benefits of scale;
6) To ease discovery and understanding of data across disciplines to promote interdisciplinary studies;
7) To combine with other data (new or archived) in the light of new ideas.

d) Driver: Organisational incentives

Producer Organisation:

1) Publication of high quality data enhances organisational 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: institutions hosting data services and professional data expertise can build profiles within disciplinary communities;

Consumer Organisation:

6) Organisational need to combine data from multiple sources to make policy decisions;
7) Re-use of data instead of new data collection reduces time and cost to new research results;
8) Use of data for teaching purposes.

e) Driver: Individual contributor incentives

Research data contributors perceive their rewards as:

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.

f) Barrier: Individual contributor barriers

Barriers to contributing data may include:

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 of 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 re-used 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 re-used.

n) Enabler: Individual contributor barriers

The barrier to contributing data for publication can be overcome by several proposed solutions:

1) Journal articles describing available data as a publication;
2) Citation of data itself, and the articles describing it;
3) Specific funding in grants to address the pre-archive activities for data preservation;
4) Enforced funding regulation to ensure the depositing of high quality data with appropriate metadata in preservation archives;
5) 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 Piwowar and Chapman, 2008, and Alsheikh-Ali et al., 2011 on poor conformance rates);
6) Tracking data publication and data usage and citation counts, and using them as part of the performance evaluation for career advancement;

7) High status awards to individuals and institutions which contribute data that is re-used.

g) Barrier: Availability of a sustainable preservation infrastructure

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:

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

o) Enabler: Availability of a sustainable preservation infrastructure

This barrier can be overcome by several proposed solutions for the publication of data:

1) In archives supported by journal publishers (e.g. Nature) sustained by a business model;
2) In archives supported by learned societies (e.g. the CAS Registry[10] of the American Chemical Society) sustained by a business model;
3) In archives funded by funding bodies (e.g. UK Economic and Social Data Service[11]);
4) In institutional archives (e.g. ESO archive of astronomical images, university archives proposed by NSF and UK Research Councils).
5) 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 e.g. the SCIDIP-ES project.[12]

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.

h) Barrier: 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.

[10] http://www.cas.org/expertise/cascontent/registry/index.html
[11] http://www.esds.ac.uk/
[12] http://www.scidip-es.eu/

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:

1) Not 'feeling safe' in dealing with unfamiliar data;
2) Impossibility of data centre staff having detailed technical knowledge of all data (e.g. museum curators);
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 co-ordinate system);
5) Datasets not meaningful in themselves; need algorithms and software to interpret them;
6) Lack of clear definition of the metadata that the potential data users will require to interpret the data;
7) Lack of a process to ensure quality standards and ensure acquisition of metadata;
8) Lack of data management training for staff;
9) Cost of providing the effort to ensure the quality standards are enforced, and the metadata gathered.

p) Enabler: Trustworthiness of the data, data usability and pre-archive activities

These barriers can be overcome by a combination of:

1) Agreeing auditable standards for publishable data quality and metadata within disciplines;
2) Certification of data centres for data quality and usability by a trustworthy body;
3) Peer reviewing of data supporting academic research publications to certify its quality;
4) The development of education and training materials for these data quality standards;
5) The training of data producers with these materials;
6) Implementation of automated data quality and metadata content tools to test pre-archive data;
7) Providing the rewards to lead to the contribution of producer effort required (see the incentives barrier below);
8) Inclusion of a mandatory data management/preservation preparation stage in research project proposals;
9) Introducing 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;
10) Overcoming the financial barrier to pre-archive activities (see the finance barrier below).

i) Barrier: Data discovery

There is no infrastructure to support international, cross-disciplinary data discovery.

q) Enabler: Data discovery

This barrier can be overcome by the following suggestions:

1) Open Linked Data initiative lead by the founder of the World Wide Web, Tim Berners-Lee;
2) Persistent, unique data identifiers with search engines (e.g. DataCite);
3) Interoperating Data Centres in specific disciplines (e.g. CESSDA in Social Science[13]).

j) Barrier: Academic defensiveness

Data producers may be defensive about publishing data for a variety of reasons:

1) Security concerns over the danger of 'being hacked' and not being preserved as it is;
2) Fear of failure to validate their results;
3) Fear that others will gain benefit from their data;
4) Fear of misuse of data for purposes for which it is not suited will harm the data contributor;
5) Fear of misuse of data to justify arguments which the contributor would find unacceptable will harm the data contributor.

r) Enabler: Academic defensiveness

Scientific claims must be subject to validation or correction, and it is incumbent on scientists to substantiate their claims with relevant supporting data. Given that proper preservation can establish data provenance and integrity, and put appropriate commercial, confidentiality, and security safeguards in place, individual anxieties about releasing data on the grounds that others may invalidate, misinterpret or otherwise exploit them should have no place in academic practice and are to be strongly deprecated. It is in the nature of science to advance through exploitation of existing knowledge and in this sense all data in the long term reverts to the public good.

1) Data centres meeting minimum standards of data curation must be available to scientists in all disciplines, so that they have confidence their data will be correctly attributed, its integrity will be maintained, and any restrictions such as embargos and protection of commercial confidence will be properly applied;
2) Scientists should be trained and assessed not only in disciplinary knowledge but in disciplinary norms and professional ethics.
3) Legitimate short-term professional and commercial advantage may be secured through embargo periods on the publication of data after they have been

[13] http://www.cessda.org/accessing/catalogue/

collected, analysed and/or contributed. Acceptable embargo periods vary by discipline, e.g. raw data collected by large neutron and synchrotron facilities may enjoy embargo periods of up to 3 years; whereas in genomics immediate publication of gene sequences is a professional requirement.

k) Barrier: Finance

Archiving costs alone are argued to be small in studies of preservation costing (Beagrie et al., 2010). 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 (Beagrie et al., 2010). 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 or 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.

s) Enabler: Finance

This barrier can be overcome by:

1) 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;
2) 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 or national 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 Publishing Group).

l) Barrier: Subject anonymity and personal data confidentiality

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.

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 de-anonymisation.

t) Enabler: Subject anonymity and personal data confidentiality

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.

m) Barrier: Legislation/regulation

There are *perceived* to be conflicts:

1) Between the data protection and freedom of information legislation;
2) Between international and national legislation;
3) Between the legislation of different countries;
4) Between national and regional legislation;
5) In the enforcement of legislation by different agencies;
6) In the understanding on legislation by different stakeholders;
7) Between the regulations of different stakeholders designed to implement legislation.

u) Enabler: Legislation/regulation

These barriers can be overcome by:

1) Unifying legislation at the European level;
2) Unifying the implementation of European directives in national legislation, and the enforcement of the European directives;
3) Greater education as to the exact entailments of the legislation for research data sharing.

3 VALIDATION OF THE CONCEPTUAL MODEL

The Conceptual Model was evolved on the basis of existing validated and published models, including the OAIS Reference Model (ISO 14721:2003) and the KRDS Benefits Taxonomy (Beagrie et al., 2008). The model of drivers, barriers and enablers was developed from the quantitative analysis of barriers to digital preservation undertaken in the PARSE.Insight survey (Kuipers et al., 2009), and the ODE Project's *Baseline Report on Drivers and Barriers in Data Sharing* (Schäfer et al., 2011), which was based on interviews with 21 key stakeholders. To a large extent most of the components of the ODE Conceptual Model were pre-validated.

The Conceptual Model was further validated and qualified through structured discussion and dialogue with informed and interested stakeholders. This validation was carried out in two complementary stages:

- A workshop at the 2011 APA Annual Conference, at which a group of 11 delegates from different stakeholder groups was invited to respond to the Conceptual Model through guided discussion;
- Telephone interviews based on the model of drivers and barriers with 55 members of different stakeholder groups, including researchers in all the major disciplinary areas.

These validation activities are described in the following section.

3.1 DATA SHARING WORKSHOP

3.1.1 PURPOSE OF THE WORKSHOP

Following initial elaboration of the ODE Conceptual Model, it was tested against an expert peer group in a Data Sharing Workshop held at the APA 2011 Annual Conference. The purpose of this workshop was to gather considered feedback on the Model through guided discussion, which could be used to further elaborate and refine the Model and to inform further analysis of drivers and barriers in data sharing.

3.1.2 WORKSHOP REPORT

The Workshop took place at BMA House in London on Monday 7 November 2011. APA Conference delegates were invited to join members of the ODE Project for guided discussion. Eleven APA delegates participated in the workshop, including 3 scientific, technical and medical (STM) publishers, 6 providers of data preservation and storage services, and 2 providers of infrastructure services. Participants had expertise or extensive experience in a range of subject areas, from broad-level STM services, to specific disciplinary expertise in medical sciences, biological sciences, and history.

Workshop participants were provided in advance of the conference with the latest version of the ODE Conceptual Model, incorporating the process, context, and drivers, barriers and enablers models. They were asked to consider the following questions:

- Is the list of drivers and barriers complete?
- Do the drivers and barriers listed require further elaboration?
- Which drivers and barriers have been most important in their experience?
- Can they provide examples of successful data re-use, that is, where the drivers have been strong enough to overcome the barriers?
- Can they provide examples of instances where barriers have not been overcome and a project has failed?

Guided discussion sought to elicit the participants' views and experiences in respect of data sharing, with reference the Conceptual Model and the questions listed above.

An audio recording of the workshop was made and a transcript prepared for the Project record. This was used to inform a revision of the Conceptual Model. The experiences shared and views expressed by the workshop participants have also been anonymised and incorporated into the 'Themes in data sharing' section of this report (Section 4).

3.2 INTERVIEWS

3.2.1 PURPOSE OF INTERVIEWS

Interviews with stakeholders served a twofold purpose:

First, they provided an enhanced peer review of the Conceptual Model of drivers, barriers and enablers on the part of experts and informed individuals across all stakeholder groups. Collectively, their views served to validate the general Conceptual Model, while by virtue of their own particular domain expertise they were able to qualify and elaborate parts of the model with greater clarity and precision. In this respect, the interviews extended the process undertaken in the workshop.

Secondly, these interviews provided an opportunity for the respondents to discuss their own experiences of data sharing and the barriers they and others had encountered. The respondents were encouraged to expound on their knowledge and experience of strategies, practices and projects by means of which barriers had been reduced or surmounted. The interviews also sought to elicit respondents' views of new opportunities and possibilities for future development in data sharing systems and practices.

3.2.2 METHOD

Project partners used their peer networks to collect a list of over 350 possible interviewees across all the major stakeholder groups. From this list approximately 220 people were invited to participate in an ODE telephone interview. The selected subset was randomly chosen from the full list, and corrected for a balanced distribution across stakeholder groups, roles, subject areas and countries. Initial invitations yielded replies

from approximately 70 people willing to participate in an interview. It was not possible to interview all of these people within the constraints of the Work Package. In all 55 interviews were conducted by Project partners between February and April 2012.

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 as per Section 2.3. An interview pro forma was evolved, in two slight variations, one for researchers, and one for non-researchers (see Annex 1). 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 sharing.

Initial analysis of the collected corpus of interviews identified a number of salient themes and converging views on key issues, such as finance and funding, the role of publishers and data description and citation. These themes were used as organising principles for more detailed analysis, and through a process of refinement yielded the Themes in Data Sharing discussed in Section 4. Interview analysis also informed the validation and qualification of the Conceptual Model.

3.2.3 INTERVIEW DISTRIBUTION AND ANALYSIS

The net for interviewees was cast among the Project peer networks, and mostly embraced stakeholders in the ERA, the United States and Australia. This is reflected in the distribution of interviewees by country (Figure 2), which is mostly among the countries of Western Europe, with strong bias towards UK and Germany: together these two countries supplied 32 out of 55 interviewees, or nearly 60% of the interview total.

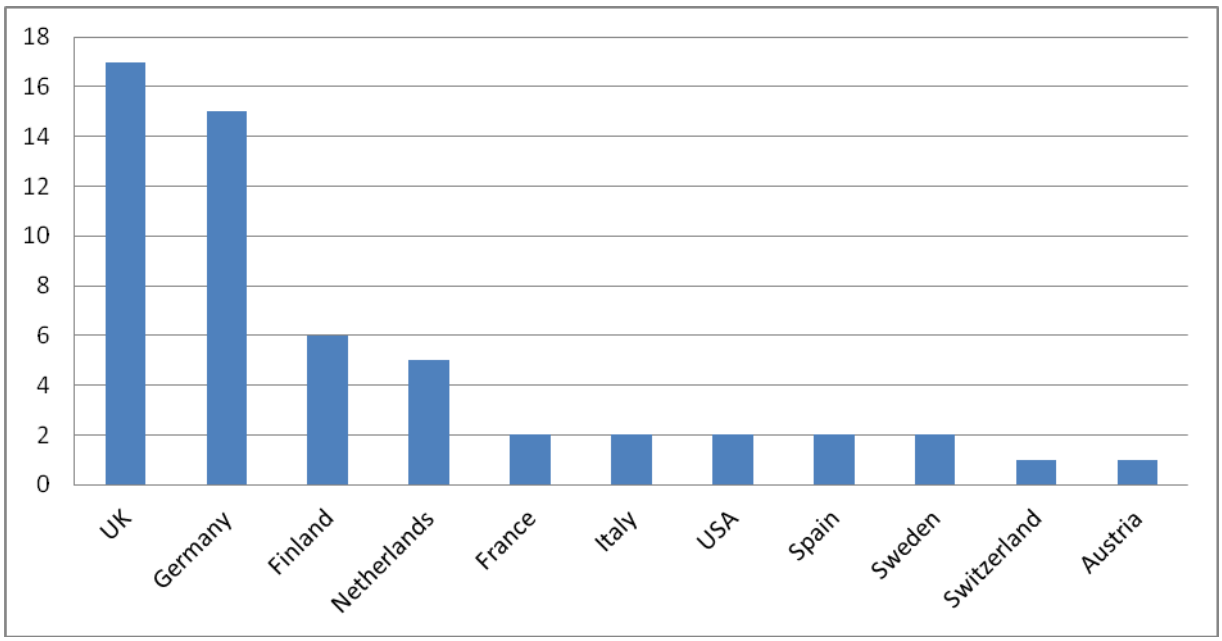

FIGURE 2. WP5 INTERVIEWS BY COUNTRY

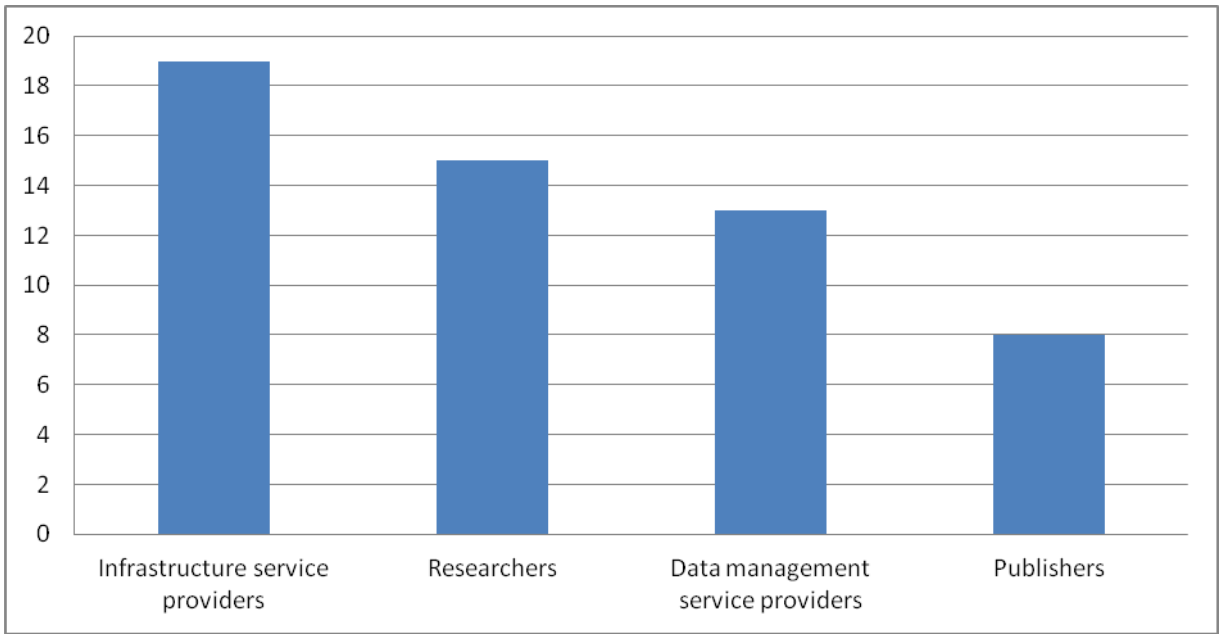

FIGURE 3. INTERVIEWS BY STAKEHOLDER GROUP

Broadly interviewees fell into one of four stakeholder groups: infrastructure service providers, researchers, data management service providers, and publishers (Figure 3). This distribution should be further qualified:

- The stakeholder categories are very broad; further analysis of individual roles identifies a wide mix of researchers, research managers, policy-makers, funders, data centre staff, librarians, infrastructure providers, publishers and other service providers;
- Many interviewees fulfilled more than one role in their professional activities and identified with more than one stakeholder group; the distribution in Figure 3 is an approximate reflection of the interviewee's primary role and stakeholder identity as presented in the interview;
- Funders and policy-makers have not been included in the distribution. A number of interviewees were based in organisations that had policy-making and funding functions, but which also might operate facilities, provide infrastructure and services, and undertake primary research. Those interviewees based in funding organisations were primarily involved in providing infrastructure or data management services, so they have been enrolled in one of these stakeholder groups as appropriate.

It was important for the Project to elicit substantive comment on the Conceptual Model from people who create and use research data. For this purpose interviewees were classified as either researchers, i.e. users and producers of data, or non-researchers, i.e. providers of services and resources.

In practice this was often too simplistic a distinction: many interviewees fulfilled multiple roles, and might provide research support or data services as well as conducting primary research their own right. For this reason a significant number of interviewees

fall into the hybrid 'Non-researcher/researcher' category. Taking 'pure' and hybrid researchers together, we can say that of the 55 interviews conducted, 22 or 40% were with researchers who customarily produce and use data (Figure 4).

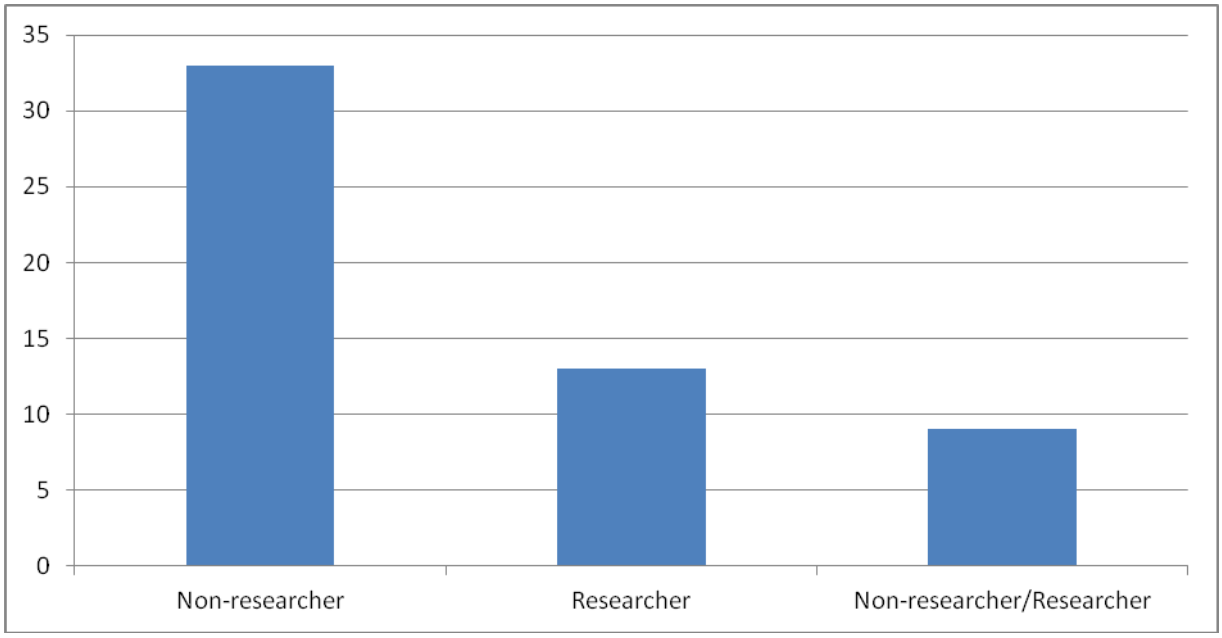

FIGURE 4. INTERVIEWS BY ROLE

Interviews with researchers were intended to capture a range of responses from researchers across a variety of disciplines. While it was impossible within the scope of the Work Package to achieve a comprehensive representation of different academic disciplines, an effort was made to obtain input from researchers handling different kinds of data with specific management requirements. Thus interviewees were able to speak from experience of their challenges in handling data in earth and environmental sciences, social sciences and humanities, medical and life sciences, physical sciences, engineering and technology, and computer sciences and mathematics (Figure 5). This allowed a number of issues specific to different data types to be discussed.

Inevitably, with a small sample self-selected through peer networks, there is a risk of response bias, and it has not been possible to control for this risk. It should be borne in mind, however, that the aim of the interview process was not to collect a large statistical sample of views, but to recruit a selection of 'peer reviewers' for the Conceptual Model, who were expert or informed in different aspects of data sharing, and represented a range of geopolitical contexts (primarily within the ERA), stakeholder groups, roles and academic disciplines.

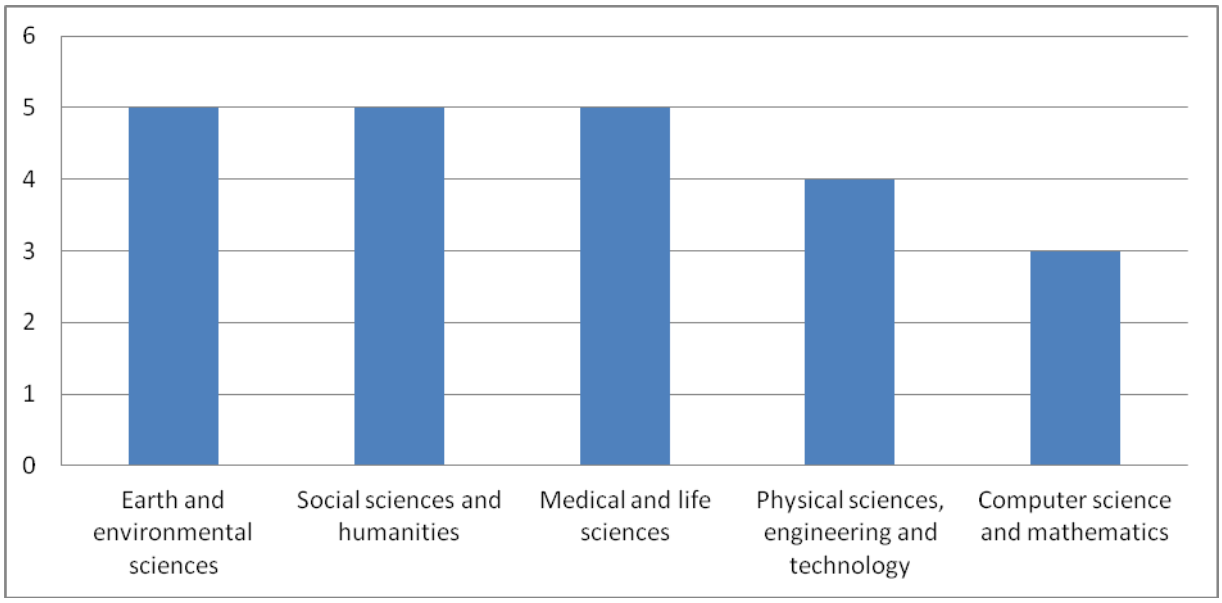

FIGURE 5. INTERVIEWS BY SUBJECT AREA

4 THEMES IN DATA SHARING

This section presents a number of thematic expositions based on the evidence collected in the workshops and telephone interviews. A number of strong themes and converging views emerged from the collated evidence, and these have been used to organise and interpret the evidence in such a way as to illuminate different aspects of the Conceptual Model. Themes have been selected for their relevance to the key concerns of the Project.

The following thematic studies are presented:

- The role of publishers in data sharing;
- Finance: funding infrastructure and data services;
- Data management: skills training and expert support;
- Standards and interoperability;
- Data citation and description for discovery and use;
- Public visibility of research data;
- Data sharing culture;
- National and regional policy and legal frameworks;
- Incentives in the academic reward system for good data practice;
- Quality assurance of data.

These treatments synthesise and organise the information and views offered by the interviewees. They are evidence-based and are substantiated by the Project interview transcripts. For reasons of confidentiality, the evidence has been aggregated and presented anonymously; where it is considered particularly relevant, a view may be attributed by the role of the interviewee, e.g. publisher or researcher.

The thematic expositions that follow are organised into three parts: a summary of the theme; a discussion of the evidence pertaining to the theme provided by the workshop participants and interviewees; and a brief analysis of the most salient drivers and barriers and the enablers that stakeholders can implement to surmount or reduce the operative barriers.

4.1 THE ROLE OF PUBLISHERS IN DATA SHARING

4.1.1 SUMMARY

Publishers have a major role to play in creating and supporting the infrastructure and services that allow data to be shared and discovered. Key areas where the industry can lead are: linking data and publications, establishing standards in data citation and description (e.g. machine-readable ontologies), developing data journals devoted to the publication and validation of data, and building services that allow users to discover and interrogate data. There is a strong argument for the benefits of collaboration among publishers and with other stakeholders providing infrastructure and data services. 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.

4.1.2 DISCUSSION

THE PUBLISHER PERSPECTIVE

Many publishers (including Elsevier, IOPP, Sage, Springer and Wiley) support Principle 7 of the STM Brussels Declaration, which states:

> *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.*[14]

Accordingly, most publishers express willingness to provide at least basic supplementary data citation and linking services to data held in external repositories. Journal publishers' data hosting services are limited in scope and use, and do not 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 at least this may be due to the fact that these services are not actively promoted. One major publisher indicated that while individual journal editors have the freedom to actively promote data publication in their journals, this is a matter of editorial choice and not general publisher policy.

[14] http://www.stm-assoc.org/brussels-declaration/

THE CRITICAL PERSPECTIVE

Many respondents were critical of the current state of data publishing, linking and citation. The following points were made:

- Supplementary data may be presented in a highly processed state, suitable for publication (e.g. in graphs or charts), but not suitable 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 respondents 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;
- Publishers may bar or restrict access to data and publications for automated data-mining.

POSITIVE EXAMPLES AND NEW OPPORTUNITIES

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.

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

- DataCite and the CODATA Data Citation Standards and Practices Task Group[15] are working to develop best practices for data citation. First results will be released in October 2012. The goal is to release recommendation guidelines for the use of DOIs. DataCite is talking to STM about citation practice, and has also entered into agreement with CrossRef to implement interoperability of their DOIs[16];

[15] http://www.codata.org/taskgroups/TGdatacitation/index.html
[16] http://www.crossref.org/10quarterly/quarterly.html#dois_in_use

- The JISC-funded REWARD project[17] brings together the UCL Institute of Archaeology, UCL Library Services and Ubiquity Press to encourage the archiving of research data using the UCL Discovery institutional repository. Researchers are asked to manage their data using the Digital Curation Centre's *DMP Online* tool,[18] and then to make the data openly available in the institutional repository via publishing a data paper in the *Journal of Open Archaeology Data*. This will make the data citable and reuse trackable, important factors for the 2014 national research assessment exercise, the Research Excellence Framework (REF). Five case studies will be followed during the course of the project in order to assess the effectiveness of the systems involved.
- 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;[19]
- Elsevier collaborates with the PANGAEA earth and environmental sciences data library for reciprocal linking[20]. This is a model that other institutions and disciplines are becoming interested in;
- One publisher spoke of exploring more flexible file formats for supplementary data, mentioning Wolfram Alpha's Computable Document Format (CDF)[21]. 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;
- The Bodleian Libraries are working with Oxford University Press (OUP) on data linking models;
- CrossRef is currently piloting CrossMark[22], 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;
- JISC's Managing Research Data Programme 2011-2013[23] focuses on data publishing, in particular data journals.

Several respondents supported the idea of publishing datasets as standalone entities in dedicated data journals. One respondent observed that in some areas many articles are more or less de facto data publications anyway, being 'just some numbers plus some text

17 http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata/planning/reward.aspx
18 http://www.dcc.ac.uk/dmponline
19 http://datadryad.org/jdap
20 http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01434
21 http://www.wolfram.com/cdf/
22 http://www.crossref.org/crossmark/
23 http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx

around without any context'. Arguably many papers reporting experiments could be more effectively produced as standalone data publications or published through documented data sharing. Advantages cited for data publication in dedicated data journals include:

- Datasets are consistently assigned DOIs, ensuring long-term accessibility;
- Datasets are persistently linked to associated publications;
- Datasets are subject to formal quality checks and peer review;
- Appropriate Creative Commons or other licences are assigned;
- The data publisher can support publication in a wide variety of structures and formats (i.e. also as tables, maps, graphs, animations);
- Data publications are citable entities for which usage metrics can be provided, so that impact can be measured. Arguably this would raise the profile of data publication as a valid research output in its own right, and this might be reflected in research assessment exercises and in community recognition, for example through rewards for good data practice, along the lines of the BioMed Central Research Awards[24].

Data publications might be either published by commercial publishers, e.g. *Journal of Open Archaeology Data* (Ubiquity Press)[25], or *Earth System Science Data* (Copernicus Publications)[26]; or published as extensions of publicly-funded data centres, e.g. the journals *Data Supplements* and *Scientific Technical Report Data*, published by the German Research Centre for Geosciences[27].

CHALLENGES FOR PUBLISHERS AND OTHERS

There are exemplary initiatives, such as the successful incorporation of DataCite DOIs into publication citations, or the reciprocal linking relationship between Elsevier and PANGAEA. But are these exceptions? 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;
- Data repositories do not exist 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

[24] http://www.biomedcentral.com/researchawards
[25] http://openarchaeologydata.metajnl.com/
[26] http://www.earth-system-science-data.net/
[27] http://www.gfz-potsdam.de/portal/cms/Bibliothek/Publizieren/Daten

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.

CONCLUSION

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.

Arguably there are valid viewpoints from both sides of the issue, and some of the disagreements about the overall picture may reflect gaps in perception and expectation between publishers and other actors in data sharing.

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 through the LOCKSS and Dutch KB initiatives. Initiatives such as ORCID[28] and DOIs 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.

[28] http://orcid.org/

4.1.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Organisational incentives
    - Business opportunities for publishers to develop supplementary data discovery and analysis services and to bring data journals to market.

BARRIERS

- Availability of a sustainable preservation infrastructure
    - publishers are not suitable repositories for long-term data preservation;
- Trustworthiness of data, data usability and pre-archive activities
    - supplementary data may be presented in highly processed formats not amenable to detailed analysis or reuse, may not be subject to explicit quality assurance and peer review processes, and may lack relevant provenance and context information;
- Data discovery
    - data citation formats and standards are not embedded in the scholarly communication system.

ENABLERS

Stakeholders	**Action points**
Researchers	Insist on good data citation practice when publishing work
Research and education organisations	Include data publications in research evaluation and reward activities
Funders	Include data publications in research evaluation and reward activities
Policy-makers (national and regional)	
Service providers (infrastructure and data management)	Infrastructure providers, such as citation registries and linking services, should work with publishers to embed citation and linking standards. Data centres should be models of best practice in data citation and develop reciprocal linking relationships with publishers.. Data centres should publish clear data citations guidelines and recommendations and provide examples and support to their stakeholders. Data centres should work with publishers to facilitate the development of data discovery and

	interrogation tools on publisher platforms.
Publishers	Work to develop best practices in supplementary data publication, including rigorous and transparent peer review and quality assurance processes. Collaborate with other publishers and with other stakeholders on standard data citation in publications, reciprocal linking of publications and datasets, and facilitating access to data through appropriate licensing models. Adopt innovative approaches to developing data formats and tools that allow users to discover and interrogate data linked to publications. Develop market in standalone data journals.

4.2 FUNDING INFRASTRUCTURE AND DATA SERVICES

4.2.1 SUMMARY

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.

4.2.2 DISCUSSION

FINANCING DATA MANAGEMENT THROUGH RESEARCH FUNDING

While many funders do now provide dedicated funding in their research grants for data management and sharing activities, other funders have been slow to do so, and overall there is a need for funders to make these funds and the data management requirements tied to them more visible to researchers.

Most respondents acknowledged that funder mandates for data management are becoming standard and are beginning to have an effect, with data management being more often costed and built into the project at an early stage. But there was concern that in areas infrastructure and services may not exist or may be inadequate to meet funder requirements for long-term preservation of accessible data. One researcher respondent felt strongly that funding for pre-archive activities and archive management was critically low, and that there was a serious imbalance in the system, with funding flows

going disproportionately to publishers, and not into development of infrastructures and tools.

There was a view expressed that on the whole funders could be more co-ordinated and proactive in making funds and the data management requirements tied to them more visible to researchers and research organisations. This would help to establish data management as a standard project cost, rather than an optional extra, and the internal systems to support data management activities would be progressively integrated into research organisations as sustainably-funded elements of their infrastructure, alongside laboratories, IT equipment and library services.

DIFFERENT DATA, DIFFERENT FUNDING REQUIREMENTS

Some respondents felt that funding was for the most part not a problem, and the cost of data preservation could largely be accommodated within the existing system. The degree to which this is the case may vary depending on the subject area and the nature of the typical data output. There is a clear difference between on the one hand the extremely large 'big science' data sets, such as those emerging from the Large Hadron Collider (LHC), which often have their own dedicated funding and management infrastructure, and on the other hand the long tail of small data sets that come out of many small research projects.

The long tail of smaller data sets may not necessarily present a big funding challenge, as their management is more easily absorbed into library funding models and the existing repository infrastructure. Libraries can underpin a sustainable model, because they are an established part of a long-term network infrastructure that spreads risk of asset loss even in the case of organisation failure.

But in between the few extremely large and the many very small data sets there is an intermediate zone of data that does not fit easily into existing library infrastructure and lacks the highly-resourced management given to 'big science' data. Many respondents felt that data management resources within organisations and systemic architecture were insufficiently funded. One university library-based data services provider observed that the library was in fact dealing with more demand for data services than it could satisfy. The funds are being provided to researchers and are being used for data preparation; the challenge for organisations is to provide the resources and technical solutions to deliver the required capacity, infrastructure and standards. In the long-term there is a risk of tension in resource allocation between the bodies that fund the research and the institutions that have custodianship of the data and must put in place policies, systems, resources and staff to maintain the data. How cost-sharing is structured for the long-term is a matter of ongoing negotiation between stakeholder groups.

FUNDING INFRASTRUCTURE AND SERVICES

Long-term sustainability of data preservation and sharing requires sustainable infrastructure, data centres and organisations, and a sharing of costs among stakeholders in the system, including national governments and regional powers such as

the EC for the large-scale infrastructure, data centres and discovery and access services, and research organisations.

Some felt that the burden of cost was not being effectively distributed throughout the data sharing system. It is all very well for funders to pay for and encourage data sharing, as many nowadays do; but if – as is currently the case - parts of the system that are to enable data sharing do not exist, or do not function well for lack of investment and development, or are not sustainably funded, then there are risks of system breakdown, system inefficiency, and stakeholder disengagement. So the urgent questions are: who pays, and for what?

Respondents variously indicated that there are not sufficient data repositories in all disciplines; that too much funding was flowing to new data, and not enough being dedicated to preserving old data and making it usable; that the preservation and sharing infrastructure is in some areas not available, or is inadequate; or that there is variable provision in the service layer.

Perceptions of overall service provision and resourcing were mixed, with contrasting views of static or deteriorating provision and of progressive improvement in services. One respondent from the publishing industry highlighted the fact that a number of data repositories in the UK had closed in recent years due to withdrawal of funding – most notably the Arts and Humanities Data Service (AHDS). Another respondent involved in UK service provision noted that funded programmes to breach some of the service gaps do exist, such as JISC's Managing Research Data Programme[29], which is promoting data management and funding projects in HEIs.

A SYSTEMIC APPROACH

National programmes aimed at developing infrastructure and broad service provision can take a systemic perspective, and identify synergies between different services and economies of scale. Organisations managing such programmes can work with allied organisations in other countries and feed into policy agendas and funding frameworks.

Other respondents highlighted systemic risks, and stressed the need for actors to take a systemic perspective, which is able to accommodate the 'total cost of ownership' throughout the data life cycle, from the inception of research to preservation of the research assets for long-term use. Cost is incurred not only in data preparation within the scope of the research project, but also in curation and storage for long-term use, and in developing and maintaining basic infrastructure, including services, platforms and portals for discovery and access, and tools for the manipulation of data. Data citation, for example, is a critical but also highly expensive component of a data sharing system. The lead for this systemic approach must be taken at a high level, by funding agencies and policymakers.

A systemic approach may be more able to assess where funding should enter the system and how it can be used for maximum efficiency and value. For example, there are

[29] http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

synergies and economies of scale to be had in the development of common services, such as data portals that are enriched by high-value metadata, technical expertise and support, and other services. Collaborative approaches provide very good examples for data preservation and sharing activities: e.g. the European Molecular Biology Laboratory[30], which is a non-profit organisation and a basic research institute funded by public research monies from 20 member states, and the National Center for Biotechnology Information[31].

COSTS AND BENEFITS

Some respondents emphasised the need for robust cost and effort modelling in funding data management activities, which can be extremely intensive. It is not simply a question of whether data is made available or not; but rather, of how available, and how useful the data will be made. Archiving costs can be high, and there must be a trade-off at some point between potential value of data and cost of preparation and preservation. There is often a gap between the long-term preservation objective and the funding business case, which is tied to the research project lifetime and is necessarily more short-term in focus. There is a need for funders to develop understanding and business processes to support long-term preservation. NERC, for example, which funds a lot of earth observation research, recognises this in providing for indefinite use of data in business case submissions.

Some respondents expressed the view that there was nothing necessarily wrong with recovering the cost of preserving and sharing data through charging others to use either the data or value-added services. Researchers are provided through their funders and organisations with the means to purchase use of equipment and publications for the purposes of research; why should they, or their organisation or funder on their behalf, not also pay for access to data? This may in fact be a driver to increase the quality and efficiency of data preservation and make preservation more sustainable in the long term.

CONCLUSION

A co-ordinated, 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 use of public money. Co-ordinated 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.

[30] http://www.embl.de/aboutus/index.html
[31] http://www.ncbi.nlm.nih.gov/

4.2.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Societal benefits
    - more efficient funding of data management activities and infrastructure yields greater sharing and re-use of data, and ultimately greater impact for each unit of public money invested.

BARRIERS

- Availability of a sustainable preservation infrastructure
    - there are gaps and redundancies in data service and infrastructure provision, and instabilities where parts of the system are not sustainably funded. These need to be addressed through high-level systemic funding policies that develop synergies and build economies of scale;
- Finance
    - Finance enters the data sharing system at many different points, on many different levels, and from many different sources. This causes inefficient and inequitable distribution. Only a co-ordinated, high-level national/regional approach can begin to address these inefficiencies.

ENABLERS

Stakeholders	**Action points**
Researchers	
Research and education organisations	Develop budgeted internal data management services and infrastructure, building on library and research support services. Develop shared infrastructure and services with other organisations.
Funders	Promote explicit funding for data management in research grants and tie it into explicit data management requirements. Co-ordinate policies with other national funders; seek collaborative synergies.
Policy-makers (national and regional)	Ensure funding for co-ordinated development and maintenance of basic infrastructure. Fund service development.
Service providers (infrastructure and data management)	Look for synergies and economies of scale with other service providers.

	Develop business models for recovery of costs through value-added services. Promote integration of data repositories and stored data in discovery services. Examples: the Registry of Research Data Repositories[32] and the DataCite Metadata Store[33]. Promote standardisation of data repositories.
Publishers	

[32] http://www.re3data.org/
[33] https://mds.datacite.org/

4.3 DATA MANAGEMENT SKILLS TRAINING AND ONGOING SUPPORT

4.3.1 SUMMARY

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.

4.3.2 DISCUSSION

A number of respondents commented on the fact that the skills required of researchers to enable effective data can be difficult to acquire. Researchers need to be taught both specific skills for data management (which will vary according to discipline), and, just as importantly, to take ownership of their data, so that it is preserved and made available to others as effectively as possible. Neither the skills nor data-responsible attitude can be acquired without training. Education in data management best practice should be incorporated into student and researcher training at an early stage.

Universities and research institutions should integrate basic training in data management into their curricula, at postgraduate level at the very least, and possibly earlier if undergraduates are generating data as part of their research. As researchers advanced in training and specialize further there is likely to be a need for them to access special training and expert domain-specific support, which may be available either at their home institutions, or by arrangement with data specialists based at other institutions or in data centres.

Universities are beginning to adopt a more proactive approach to the development of good data management skills. Various examples were given, including the introduction of mandatory training programmes for students and provision of a data librarian support service for students.

While institution-based training in data management basics is clearly necessary, at a more advanced level data management training and support can only reasonably be offered by specialist service providers, typically data centres. There are two main reasons for this:

Data management and data documentation are time consuming: they require specialist knowledge and a considerable amount of intellectual effort. Researchers need ongoing expert support from service providers to ensure data are made available, and to get assistance in data formatting and metadata description.

Secondly respondents felt that data management is nothing they can get reward for, e.g. additional funding or impact for their professional careers.

Views on the role of libraries in data management training were mixed. Some supported the view that libraries could and should play an increased role as data managers and experts on the basis of their traditional role as providers of information management professionals. Others were more inclined to see the role of the library as at the most one of intermediary between researchers and the data centres that would be able to provide the highly specialized support researchers would need.

Some respondents expressed that funders could be more proactive in data management requirements for research grants. This would help that research projects have a data management plan from the beginning. Extra financing would be provided/included for explicit training programmes and ongoing expert support.

CONCLUSION

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 co-ordinated 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.

4.3.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Societal benefits
    - more and better data is made available to be shared and re-used by researchers;
- Research benefits/Individual contributor incentives
    - researchers acquire data skills at an early stage in their careers and benefit from ongoing training and support, making them ultimately better researchers and enhancing the impact of their work;
- Organisational incentives
    - data centres and other service providers can market their specialist skills and training to HEIs and researchers as part of their business model.

BARRIERS

- Individual contributor incentives
    - data skills training is not a mandatory part of researcher training and there is in general a lack of incentives in the academic reward system for good data practice;
- Availability of a sustainable preservation infrastructure
    - there may a lack of data centres and data specialists to meet the needs of researchers, or they may not be a means for researchers to access the training and support they require;
- Finance
    - national funding to HEIs does not require institutions to provide data skills training;
    - there may be funding gaps for data centres and the training of data professionals who can provide specialist services.

ENABLERS

Stakeholders	Action points
Researchers	Participate in training programmes to acquire and develop the skills they need to share and preserve data effectively.
Research and education organisations	Make data management skills training and ongoing support available to early-career researchers. Foster the role of the library in delivering training and support. Make basic data skills training a postgraduate course requirement. More training and career opportunities for data librarians.
Funders	Promote data management plans in research grants to finance professional training programmes and ongoing expert support in the area of data management from the beginning.
Policy-makers (national and regional)	Mandate higher education providers to include data skills training in postgraduate courses.
Service providers (infrastructure and data management)	Data centres to develop specialist data skills training and outreach services, and work with HEIs to target specialist support at early-career researchers. Data skills directories: searchable portals through which researchers can define the data skills they need and find matching sources of training and support.
Publishers	Standardisation of data management methods within publication tools.

4.4 STANDARDS AND INTEROPERABILITY

4.4.1 SUMMARY

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:

- 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.
- 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, such as the INSPIRE[34] directive.

4.4.2 DISCUSSION

The absence of well-established standards in both data description and data formatting are significant barriers to data discovery and use. Concerns were raised about both high-level infrastructural barriers and discipline-specific issues with regard to standards and metadata. This highlights the points that metadata and interoperability issues might be perceived differently from the perspective of different stakeholder groups, and that the nature of the challenges and the existence of solutions varied according to disciplinary area.

Data sharing, in order to take place, requires common infrastructures, rules and semantics. These common requirements are expressed and realized as standards among

[34] http://inspire.jrc.ec.europa.eu/

the communities in which data sharing takes place. Standards may exist at different levels of community, from the global to the highly localized. Many respondents mentioned the need for various standards to support these machine processes, including:

- Common identifiers and resolvers for basic entities such as data sets 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 data sets, 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;
- 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.

CONTEXT KNOWLEDGE AND INTERDISCIPLINARY DATA

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.

A standards-based approach may also hold the key for creating the less well-defined context in which a data set exists – a sort of explicit representation of the tacit knowledge a researcher familiar with the research might bring. Linked Data approaches can recreate the network of information sources that establish the value of the data; these might include statistical data about use and impact, references to publications based on the data, others' opinions of the data. This is less about discovering and

interpreting the data, than it is about being able to make value judgements about the authority and usefulness of data.

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 in a very early stage. Even where infrastructure and standards within disciplines may be well-developed, these often do not interoperate effectively with other infrastructure and standards outside of the disciplinary domain. This can frustrate researchers' efforts to discover, understand and use effectively data from outside of their own disciplinary zone.

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.

THIS IS NOT JUST A TECHNICAL ISSUE

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 data set, which is ingested into a store or collection and acquires its storage format and structure and descriptive wrapper. How standards are applied to the data set at this stage will determine the destiny of that data set: 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.

This latter point is echoed in the concern that efficient machine discovery and interpretation processes are becoming more and more important to the researcher. This is the case not only for researchers who need to master the volumes of data within their own disciplines, but also to enable new high-volume approaches to data processing. Distributed processing and large data infrastructures are providing new interdisciplinary possibilities that have not existed before: e.g. numerical techniques developed by physicists are being applied to massive data sets in other disciplines, such as biology and economics.

Machine-based approaches are important in data creation as well. A lot of data needs a lot of context to interpret: not only basic information about what it is, but capture of other relevant variables: when it was created, using what instruments, how they were calibrated, how the data were subsequently processed. Established standards and processes for generating this data in real time and storing it with the data objects can greatly assist discovery and use. Staff at STFC for example have developed a Core Scientific Metadata Model to describe the transformation data undergoes as it is produced using large scientific instruments[35]. For such models to scale they must be widely adopted.

POSITIVE EXAMPLES AND NEW OPPORTUNITIES

One respondent made an explicit connection between a high-level standards-based data infrastructure and European research competitiveness, and cited the example of the INSPIRE Directive[36]. This 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.

The EUDAT[37] 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 service rich service layer. Another initiative in this context is re3data.org[38], which has as its goal the creation of a global Registry of Research Data Repositories. This will promulgate standards and help to professionalise the data repository landscape.

In areas where data sharing practices are well-established, there are likely also to be highly-embedded open standards: in crystallography, for example, the Crystallographic Information File (CIF) is a standard text file format for representing crystallographic

[35] http://epubs.stfc.ac.uk/work-details?w=53953
[36] http://inspire.jrc.ec.europa.eu/
[37] http://www.eudat.eu/
[38] http://www.re3data.org/

information, promulgated by the International Union of Crystallography (IUCr). This also applies to Meteorology for example, where some standard values are easy to define, on a global scale.

CONCLUSION

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, 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 repositories. For example, there is a clear role for data repositories to take a role in ensuring standards are applied to data that are submitted to them, and in working to embed the knowledge of those standards among their research communities through outreach and proactive engagement with research organisations.

4.4.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Research benefits
    - standardisation and interoperability increase effectiveness of data discovery and understanding, facilitate automated processing, and enable interdisciplinary connections and meta-analysis of data sets.

BARRIERS

- Availability of a sustainable preservation infrastructure;
- Trustworthiness of the data, data usability, pre-archive activities;
- Data discovery.

ENABLERS

Stakeholders	Action points
Researchers	Take ownership of data: includes data description and data formatting. Participate in community work to define and establish standards.
Research and education organisations	Provide training on data description and formatting as part of researcher data skills education.
Funders	Mandate data management plans in research grants.
Policy-makers (national and regional)	Streamline policies and initiatives among stakeholders and across disciplines.
Service providers (infrastructure and data management)	Define data description and formatting standards for repositories, and educate researchers about them. Support researchers in submission; provide training in data skills. Work with other data repositories in the community to establish common standards and interoperability, e.g. in ontologies. Build frameworks of trust between data communities and systems. Build integrated infrastructures driven by common architectures and best practices. Promote standardisation and networking of data repositories, to allow metadata exchange and value-added services.
Publishers	Work with data repositories and other stakeholders to improve interoperability. Develop technologies and infrastructures to improve discoverability. Use well-structured open formats for data publishing.

4.5 DATA CITATION AND DESCRIPTION FOR DISCOVERY AND USE

4.5.1 SUMMARY

Efficient discovery and use of data depends on effective citation and description. This requires, at its most basic:

- Standard citation rules followed by researchers and publishers – just as citation of published work follows customary practices;
- Ascription to data sets of persistent URIs;
- Accurate, standards-based description and provenance of data sets, allowing users to easily assess relevance and judge value.

Researchers must in the first instance create and describe citable data sets using appropriate disciplinary metadata; and data centres have a role to play in validation and quality assurance of metadata. But publishers are essential if universal standards of effective citation are to be embedded in the system as whole.

4.5.2 DISCUSSION

Interviewees were asked for their views on good data citation, and over 75% of them expressed a view. There was widespread agreement on the minimum requirements for effective data citation:

- Persistent resolvable identifiers, such as DataCite DOIs, and a stable architecture for resolving them;
- Consistent citation formats;
- Universal data citation rules applied by publishers ;
- Appropriate descriptive metadata associated with the data set, so that users can understand the data and assess their relevance;
- Provenance metadata (creator(s), source organisation, holding organisation), so that users can assess the value of the data set, its authority and trustworthiness.

A number of other respondents also expressed the view that basic citation alone is not sufficient. A key question is the unit of data to cite. In many cases it is not useful to cite the whole of a large database, where a piece of research may be based on a subset of data. There may be requirements for data citation to be able to express different levels of granularity in data sets, and to describe relationships, such as those of subset to container set, or of subset to other subsets. Can citations be constructed to express such relationships?

At a broader level, this leads into the point that any data set exists in numerous relationships to other data sets and to publications: citation and linking mechanisms must be able to support the generation and maintenance of these relationships and allow the user to navigate easily through them. As data publication is established as a research output in its own right, it will become increasingly important to link data sets to associated research outputs, such as publication and other data sets, and to track and log citation and usage for bibliometric and impact analysis.

Effective data citation is not just about making data discoverable: descriptive metadata is also necessary for users to be able to assess the relevance and value of a given data set. This in turn requires standard metadata formats and semantics, so that metadata are consistent and easily readable by the community of users, and amenable to machine processing.

The nuts and bolts of data citation rely on providers of infrastructure services (DOI registries and resolvers), data centres and publishers. Publishers especially must play a central role in establishing standards of data referencing and description and incorporating them into their editorial policies.

It is also important for data centres to support correct data citation to appropriate standards as part of data validation processes, and this in turn requires researchers to learn good data citation practice, just as they are required to understand how to cite published sources in their papers. Both data centres and higher education institutions should support education of researchers in good data citation, and encourage researchers to take ownership of their data through correct citation. Good data citation leads to better impact for the research and ultimately benefits the researcher[39].

CONCLUSION

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.

4.5.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Research benefits
    - Research impact increases by citing data in publications.
- Individual Contributor Incentives
    - Peer visibility and increased respect achieved through publications and citation;
    - Status, promotion and pay increase with career development;

BARRIERS

- Individual contributor incentives:
    - Lack of motivation for data citation;

[39] Data citation is treated in greater detail in ODE D4.2, 'Best practices for citability of data and on evolving roles in scholarly communication'.

- Availability of a sustainable preservation infrastructure:
  - journals are not necessarily good at holding data associated with articles;
- Trustworthiness of the data, data usability:
  - lack of clear definition of the metadata that the potential data users will require to interpret the data;
  - lack of a process to ensure quality standards and ensure acquisition of metadata;
- Finance:
  - lack of scalable cost-effective methods for creating semantically rich data description.

ENABLERS

Stakeholders	**Action points**
Researchers	Take ownership of data citation.
Research and education organisations	Train researchers in data citation basics and best practice.
Funders	Mandate deposit of citable data sets in data management plans.
Policy-makers (national and regional)	
Service providers (infrastructure and data management)	Ensure sustainable registries and architecture citation URIs. Allow data citation to express relationships between related data sets, and facilitate reciprocal linking with related entities, e.g. publications.
Publishers	Establish and embed editorial data citation rules. Apply consistent data citation standards.

4.6 PUBLIC VISIBILITY OF RESEARCH DATA

4.6.1 SUMMARY

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 projects (GalaxyZoo etc.), 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.

4.6.2 DISCUSSION

The public visibility of research makes issues of data openness critical, especially in politically-sensitive research areas, such as climate science, and areas where research intersects with strong public interest, e.g. medical research. Academic defensiveness and legitimate concerns about how research data might be used by those outside the research community can be major barriers to the public sharing of data. Researchers may be concerned about accidental or deliberate misrepresentation or misinterpretation of data, and use of publicly-funded data for the benefit of commercial or politically-motivated interests.

Freedom of Information (FoI) legislation is symptomatic of the new climate, in which the right of access to data held by public bodies is presumed. In respect of academic research data such legislation has proven controversial and caused considerable anxiety among researchers. In the UK in recent years there have been two widely reported controversies:

- The so-called 'Climategate' scandal arose from repeated FoI requests on the part of climate change sceptics for access to climate research data held at the University. The defensive response of scientists at the University arose largely because they perceived these requests to be vexatious and made by agenda-driven campaign groups, not disinterested scientists.
- In 2011 the tobacco company Philip Morris entered a FoI request for access to research data on smoking behaviours and perceptions in the young held by Stirling University. The University initially declined this request and it is subject to ongoing dispute[40].

[40] http://www.bbc.co.uk/news/uk-scotland-tayside-central-14744240

Many within the research community will be sympathetic with the motivations that led researchers in the above cases to resist attempts to access their data. Good data sharing takes time and effort, so it is not difficult for researchers to be obstructive if they are forced to share data against their will. Even where data is openly shared, it is easy to hand over a dump of data with no explanation of what it means or how to use it; or to make it available in theory but presented in a form that effectively limits its usability.

But the fact remains that researchers do not own data produced in the course of research funded by the public purse, and they cannot arbitrarily control access to their data or how they are used by others. It is not a useful response for researchers to bury data in the hope that they won't be asked to produce them; far better to plan for managed data sharing at an early stage in research so that the data can be responsibly shared, and production and preservation work is properly undertaken and funded. Where valuable or politically sensitive data is made publicly available, elementary safeguards against abuse can be put in place – for example, requiring registration for access to data, so that there is an audit trail of usage.

Public controversies such as those above have encouraged greater institutional ownership of the data produced by their researchers or under their stewardship. Historically, for example, climate science has not been open in its data practices, but under pressure of circumstances its approach to data management is changing, and this is being reflected in institutional data management policies.

Aside from reasons of academic defensiveness, researchers can be daunted by the prospect of preparing data for sharing, in terms of both effort and cost. Here again there is a role for institutions and service providers to give support to ensure that data is made available in a suitable manner – taking account of confidentiality, commercial and other legal issues, and providing adequate data description to make it discoverable and usable.

SUPPORTING ACCESS TO DATA

Researchers may regard requests for access to their data from interests outside the research community as a necessary evil, but there is reciprocal potential, especially in areas bearing on public policy that can work to the researcher's advantage. Public interest and advocacy organisations can use their resources and high profile to help researchers get access to the data they need.

For example, the UK organisation Doctor Foster is an independent body that aims to improve the quality and efficiency of health and social care and other public services through better use of information[41]. It is a research information unit based at Imperial College London and to some extent bridges academic research and public advocacy roles. One advantage of its high public profile is that it has a degree of political agency when it comes to mobilizing data sources inside large organisations such as the National Health Service, and so can support researchers who need access to clinical and other data.

[41] http://www.drfosterhealth.co.uk/

At a more commercial remove, PatientsLikeMe is a US-based for-profit organisation that provides resources for individuals to share information about their conditions, treatments etc. and promotes research and innovation in medical treatment[42]. Studies are also carried out, sometimes initiated by groups of patients (Brownstein et al., 2010).

Researchers based in academic organisations may have reservations about research and access-to-data agendas driven by commercial interests and advocacy groups. However, it is clear that such groups are now part of the research information landscape, and in many respects they are welcomed, especially by independent researchers outside of the research mainstream, who can find the traditional academic closed shop frustrating.

Moreover, there is a legitimate public interest argument for data transparency, especially in respect of research conducted at taxpayers' expense and bearing on complex and controversial political issues, or where there is a clear public interest, as in the case of medical research into new treatments and therapies. Data transparency is also a safeguard against academic abuses, as the data behind research claims is made more easily available for scrutiny.

MANAGING RESEARCH DATA IN THE PUBLIC SPHERE

Several examples were given of positive responses to the demand for open research data, on the part of research organisations, policy organisations and funders:

- The IPCC Data Distribution Centre[43], founded to enable non climate-researchers to work with climate data.
- In Germany the Federal Ministry of Research and Education (BMBF) founded the Climate Service Center[44], dedicated to refining the knowledge derived from climate research in a practice-orientated way and conveying the findings to decision-makers in politics, administration, the economy and for the broad public.
- The Malaria Atlas Project[45], funded by the Wellcome Trust is a public information project, and all data is made available under CC licences. Some data is derived from public sources, some from commercial sources; in the latter case the data may be licensed with a different permission mix, meeting the requirements of data providers in a way that still preserves the data for public use, albeit with restrictions.

CONCLUSION

Many researchers and research organisations can be uncomfortable with the claims made on data produced in the course of research by individuals and organisations outside the research community, often pursuing agendas that conflict with the researchers' own or with the perceived wider public interest. But in many respects the walls are coming down that once divided researchers based in research organisations and those outside the salaried academic community. The products of publicly-funded

[42] http://www.patientslikeme.com/
[43] http://www.ipcc-data.org/
[44] http://www.climate-service-center.de/index.html.en
[45] http://www.map.ox.ac.uk/

research are a public good, and data should be managed on the assumption that it can be accessed and used by the wider public, for purposes not intended by the data producers. Data transparency needs to be managed through national and organisational policies, and especially in politically sensitive areas monitoring, audit and review procedures need to be in place to guarantee data integrity and protect against abuses.

4.6.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Societal benefits;
- Academic benefits;
- Research benefits;
- Organisational incentives
    - to be seen to manage data responsibly, especially where it may be sensitive, as in the case of climate data.

BARRIERS

- Trustworthiness of the data, data usabilitity, pre-archive activities;
- Academic defensiveness.

ENABLERS

Stakeholders	Action points
Researchers	Approach data management data on the assumption that data will be made publicly available.
Research and education organisations	Establish data transparency policies and procedures, with clear audit trails for modification and use of sensitive data.
Funders	Fund public understanding of science initiatives and other public engagement and communication activities.
Policy-makers (national and regional)	Establish clear research data transparency and audit policies.
Service providers (infrastructure and data management)	
Publishers	

4.7 DATA SHARING CULTURE

4.7.1 SUMMARY

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.

4.7.2 DISCUSSION

The social dimensions of research and working relationships affect data practice, and this was reflected in the prevalence of comments by respondents on the existence or otherwise of a data sharing culture within given research communities. 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 data sets 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 data sets 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 too 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 are

a commodity, with actual or potential commercial value, there is necessarily a presumption against sharing.

Technological development is in itself a driver for cultural change, and this is reflected in the observation that in some areas 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. Where cultures and presumptions of open data use encounter closed data cultures there are bound to arise challenges to customary attitudes and practices that may lead to change. One symptom of such tensions may be the high-profile uses of Freedom of Information legislation to force access to sensitive data, such as the climate data held at the University of East Anglia, or the smoking data held at the University of Stirling (see Section 4.6.2 *supra*). Although such controversies may indicate that allowing access to data is not an unmixed blessing, such instances are in fact forcing researchers and research organisations to manage their data on the presumption that it will be publicly available. This is turn reinforces 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 users of data and the uses to which data are put transparent and accountable.

CONCLUSION

On a general level, it may take a long term for cultural 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 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.

4.7.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Research benefits
    - A stronger data sharing culture enhances and accelerates research impact.

BARRIERS

- Trustworthiness of the data, data usability, pre-archive activities.

ENABLERS

Stakeholders	Action points
Researchers	Engage in community discussions about practices and policies, and lead by example.
Research and education organisations	Put in place policies to encourage and mandate good data practice. Provide data skills training for researchers.
Funders	Include data management requirements in research grant conditions.
Policy-makers (national and regional)	Use policy instruments to incentivise good data practice, e.g. through recognition of data publication in national research assessment.
Service providers (infrastructure and data management)	Support practices in the communities – take the burden off the researchers
Publishers	Develop the publications and services that enable researchers to publish, discover and use data.

4.8 NATIONAL AND REGIONAL POLICY AND LEGAL FRAMEWORKS

4.8.1 SUMMARY

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.

4.8.2 DISCUSSION

Data can function both as a commodity, to be exploited for personal or corporate advantage, and as a public good, given up for community benefit. How data functions in scientific discourse and exchange will depend on the stakeholders that assert ownership of the data, the context of use, and the stakeholder majority. Stakeholder interests may align or compete, and research policies or laws framed at the national and supra-national level must provide models and procedures for mapping and negotiating between competing interests. Such frameworks are especially important where personal and commercial interests protected in law are involved.

Two clear points emerge:

- that the trend towards greater data openness is being reflected to a greater or lesser degree in national research and education policies encouraging or mandating data management and sharing practices;
- that the variation in national policies and approaches to data sharing can be a cause of friction given the supranational nature of research and data use, and that greater harmonisation of national policies and legal structures would facilitate global data traffic and the growth of data sharing cultures.

There emerged from the evidence the sense of a clear need for high-level policy and legislative approaches to enable or even mandate different aspects of data sharing practice. Although there are positive examples of national initiatives,[46] many respondents stressed a lack of national leadership and investment, and the fact that different countries were at different stages of policy and legislative evolution in respect of data sharing. Austria and Denmark were held to suffer from an absence of data sharing regulation; the UK picture was viewed as mixed, with each of the seven Research Councils having different data management and data sharing policies; the Netherlands Organisation for Scientific Research (NWO) promulgates a strong national policy; whereas the national policy formulated by the German Research Foundation (DFG) is softer, and only encourages researchers to state their research data management plan within the proposal.

[46] Among those mentioned by respondents were the NIH data sharing mandate, and a task force in Spain that is working on national policy and mandates.

The other complicating factors at national framework level are differences in legislation, particularly in respect of intellectual property in the data, protection of individual confidentiality, and national security. Thus, for example, pharmaceutical or engineering data may be subject to intellectual property protection; personal data confidentiality issues particularly apply in medical research using patient data and in social sciences; and national security concerns may apply in respect of satellite data.

All sorts of data may legitimately lay claim to legal protection, but such protections as are required can be applied without necessarily thereby rendering data wholly unusable. In practice different countries strike the balance between protection and openness at different points and this can lead to bewildering and often frustrating experiences for the producers and prospective users of data. One can easily imagine the potential difficulties involved in a scenario where, for example, data produced by researchers in country A working in country B may be held in a repository in country C that is funded by an organisation in country D.

Several researchers cited instances of researchers being unable to access data either because of legislation perceived to be obstructive, or because there was no legal requirement placed on the data producer or publisher to make their data available or accessible for practical use, such as data mining. The general view appears to be that there should be in law an open data presumption, with reasonable qualifications for the protection of IP, confidentiality and national security, and that data openness should be actively enabled and even in cases enforced through the medium of political and legal instruments.

CONCLUSION

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 the EU is a supranational organisation with a political and legal mandate to establish 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.

4.8.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Societal benefits;
- Legislation/regulation
    - It can have a positive function in e.g. obliging researchers to share data, or requiring publishers to open data for mining.

BARRIERS

- Legislation/regulation.

ENABLERS

Stakeholders	Action points
Researchers	
Research and education organisations	
Funders	Make the case to European Governments for a coherent co-ordinated approach to developing regional frameworks.
Policy-makers (national and regional)	Co-ordinate national legal and policy frameworks to reduce friction and blockages in international data traffic.
Service providers (infrastructure and data management)	
Publishers	

4.9 INCENTIVES IN THE ACADEMIC REWARD SYSTEM FOR GOOD DATA PRACTICE

4.9.1 SUMMARY

Recognition and reward for data sharing is very much a personal driver, bound up with the self-image and status in the community of the individual researcher and his/her immediate collaborators, and the barriers are personal too. There is a strong connection with data citation, since citation of publications is a key to recognition, particularly when it comes to formal research evaluations and the funding decisions based on them. However less formal types of recognition have their value too. The other side of the coin is academic defensiveness: fear of others benefitting at one's own expense.

If it is wished to elevate data to the same status as publications, then a similar degree of formality and standardisation is required as for publications, where the system of citations and the understanding of their value are very well developed.

In general, the effort in preparing data for sharing must be balanced by the rewards. At present the rewards do not always appear sufficiently concrete.

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.

4.9.2 DISCUSSION

Many respondents were emphatic that academic recognition and reward is or could be a powerful incentive for data sharing. The thinking always seemed to be in terms of benefit to the individual—the possible driver of organisational benefits such as 'Publication of high quality data enhances organisational profile' was little mentioned.

However it is also notable that many respondents did not select individual contributor incentives as a driver. Some in fact declared that there is no personal benefit – though this is probably because currently the effort involved in sharing data is high and the rewards uncertain. Data sharing and provision are competing with paper writing on the priority list of researchers.

Although elevating data to the same status as publications in terms of capability for evaluation through citations etc. is one goal of aspiration, it was clear that there are in fact other drivers that could motivate data sharing. It might offer an alternative route to academic prestige for those scientists who for whatever reason do not climb high on the publications ladder. Moreover, beyond direct recognition equivalent to citations there are wider if more diffuse possibilities such as general visibility with one's peers, and 'marketing' for a research project or programme by putting data out. In some fields at least, older researchers may wish to leave a personal legacy, and releasing their accumulated datasets is one of way of achieving this.

It was pointed out that not only data but software could also offer similar incentives.

CITATIONS AND RESEARCH EVALUATION

If the goal is to elevate data on to a par with publications, then there is a need for a well-developed and standardised method of citations, just as for papers. Here there is a link with the theme 'Data citation and description for discovery and reuse'. Indeed the lack of proper citability is seen as a barrier. The logical conclusion of this view is that eventually data should be considered in the same way as journal papers in the evaluation of research by governments and funders, influencing quality rankings and future grants of funding. There is however also a more *laissez-faire* opinion that different forms of data citation will emerge to correspond to varieties of research data.

The benefits to individuals would be made apparent by services to track the impact of publications and datasets. These emerging forms of impact assessment are broadly grouped under the rubric of 'altmetrics'[47]. An example of an altmetrics service is Total Impact[48], which aggregates a variety of impact measures across the spectrum of formal and informal communication, including articles, data sets, blog posts and other publications.

ACADEMIC DEFENSIVENESS

The other side of the coin to recognition and status is the fear of seeing one's reputation lowered, or of losing the advantage of holding one's data. Indeed academic defensiveness and protection of reputation was identified as a barrier by some. It was observed that researchers are on the whole reluctant to share data if they feel others might benefit to their detriment, or if they feel they have not fully exhausted its use value for their own research.

As well as this general concern, some particular notes of caution were sounded. The rights of the data originator might not be the same as those of the author of a paper. And peer visibility and status might be illusory if the data is not curated properly—in this case the only real reward might be self-re-discovery and re-use of the data in the future.

CONCLUSION

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 formalised 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 formalisation of citation and of evaluation is required.

4.9.3 CONCEPTUAL MODEL ANALYSIS

[47] http://altmetrics.org
[48] http://total-impact.org/

DRIVERS

- Individual contributor incentives
    - It is possible to achieve visibility and respect without formal publication or citation of data, by being seen to be behaving generously and by having keeping a high profile through releasing one's data.

BARRIERS

- Individual contributor incentives;
- Availability of a sustainable preservation infrastructure
    - A suitable infrastructure is required to support citability of dataset;
- Academic defensiveness.

ENABLERS

Stakeholders	**Action points**
Researchers	Be aware of the benefits, formal and informal, of sharing/publishing data.
Research and education organisations	Reward researchers for sharing data.
Funders	Take data into account as well as publications when planning research evaluations.
Policy-makers (national and regional)	Declare the importance of data as a measure of academic prestige.
Service providers (infrastructure and data management)	Make sure that datasets are citable. Put in place good curation of datasets to preserve their long-term value. Clarify the rights of the original gatherers of the data.
Publishers	Encourage proper data citation.

4.10 QUALITY ASSURANCE OF DATA

4.10.1 SUMMARY

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 reusers of data will want to know about both. The need is particularly acute for cross-disciplinary reuse, when the potential reuser 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. This is dealt with by standards such as the Data Seal of Approval[49], the Deutsche Initiative fűr Netwerkinformation (DINI) Certificate[50], *Trustworthy Repositories Audit and Certification* (TRAC)[51] and ISO 16363:2012 (*Space Data and Information Transfer Systems: Audit and Certification of Trustworthy Digital Repositories*)[52].

4.10.2 DISCUSSION

There are two key questions relating to data quality that arise when considering whether to reuse data that originated from outside the reuser's own circle:

- Whether it is good enough for reuse;
- Whether it is trustworthy.

These are particularly important barriers to reuse of data outside the discipline of its origin. Combining datasets from different sources offers new opportunities but also special problems. Often it is assumed that specialist expertise in the discipline of origin of the data is a prerequisite for being able to reuse it.

Part of the problem in preparing data for reuse is not knowing what another researcher, possibly from a different discipline, might want to do with it in future. A more sanguine view is that such situations are inevitable in the age of being able to find anything on the internet.

One particular aspect of data quality is that of provenance: maintaining a record of the origins and successive steps taken in curating data, but also in the case of processed data

[49] http://datasealofapproval.org
[50] http://www.dini.de/dini-zertifikat/english/
[51] http://www.crl.edu/PDF/trac.pdf
[52] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

tracking the methods, software and calibrations used. This is perceived as a difficult problem, but is particularly necessary for data with potentially a very long lifetime of usefulness, such as earth observation data, or where complex processing has been carried out, when it might be necessary to distinguish different versions of software.

Apart from potential reuse, prevention and detection of scientific fraud is important. The reasons why fraud is possible include difficulty in going behind the publication to the raw data on which it is based; and a lack of skills in producing and interpreting data when it is accessible.

It is worth noting that some interviewees did not single out the barrier 'Trustworthiness of the data, Data Usability and Pre-archive activities', at least within a single discipline. This might be because they were thinking from the perspective of data archives with a working assumption that the incoming data is of reasonable quality, it not being the archive's job to check that.

The need for data quality can be linked to the recognition and reward due to researchers for sharing their data (see Section 4.9). Although such behaviour is laudable, there is no quality standard expected.

APPROACHES TO SOLUTIONS

There is a general feeling that 'professional' handling of data by third parties such as data centres and libraries could be part of the solution in a good repository infrastructure. Such organisations could for example ensure that adequate metadata is captured. However it might be impossible for repository staff to have detailed technical knowledge.

Complementary to this is the need for the researcher to capture relevant information at all stages of the data lifecycle, rather than leaving everything to the end when they wish to release the data. However preparing data for sharing, including metadata needed as a basis for quality assurance, is time- consuming.

Standards obviously have a role to play: as one interviewee remarked, trustworthiness is linked to data standards and universal agreements: 'A CIF file cannot be misinterpreted or misused.' But different disciplines have different degrees of standardisation.

Publishers recognise that they might have a role in establishing standards of data referencing and description and quality assurance of published data.

Peer review of data was not much discussed, though one publisher commented on the lack of clarity about what is expected and at what stage in the publishing process. Data journals were not much mentioned, perhaps because currently they exist only in a few domains.[53]

CONCLUSION

[53] This is a subject that has been studied in the APARSEN Project (http://www.alliancepermanentaccess.org/index.php/aparsen/). See Pampel et al. (2012).

If data reuse is to be encouraged then it is vital that the potential reusers 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.

4.10.3 CONCEPTUAL MODEL ANALYSIS

DRIVERS

- Research benefits
    - Researchers will be more willing to use and rely on data that are certified as trustworthy or as meeting defined published standards, and this will facilitate the growth of a data sharing culture;
- Organisational incentives
    - Accreditation of data centres as meeting defined standards can be a valuable component of their sustainability, and will encourage data producers and consumers and other service providers (such as publishers) to establish relationships with them.

BARRIERS

- Availability of a sustainable preservation infrastructure
    - The variety of types of data requiring different models of validation and quality assurance and highly localized or specialized knowledge present significant challenges to development of scalable preservation infrastructure or to data centres and publishers seeking to implement standard, rigorous quality assurance processes.
- Trustworthiness of the data, data usability and pre-archive activities
    - This barrier really exists in two aspects: fitness for purpose and trustworthiness.

ENABLERS

Stakeholders	Action points
Researchers	Adopt the practice of systematically recording information about the origins and processing of datasets throughout their lifecycle, as a basis for quality assurance in future.
Research and education organisations	Provide incentives and reward systems for recognition of data quality assurance activities.
Funders	Fund research in recording and practical use of provenance information for reuse of datasets. Provide incentives and reward systems for recognition of data quality assurance activities.
Policy-makers (national and regional)	
Service providers (infrastructure and data management)	Seek ISO 16363 certification and engage with communities and accreditation bodies in the establishment and development of standards. Co-operate with publishers in development of data publications. Preserve provenance of datasets. Obtain good quality metadata on ingest. Experiment with new approaches such as crowd-sourcing for metadata and quality assurance.
Publishers	Encourage more systematic handling of datasets, when accepted, with a view to establishing their quality. Develop data journals and other specialist data publication services.

5 CONCLUSIONS

The ODE Conceptual Model presents different analytic tools for conceptualising the sharing of research data. The process and context models describe data sharing in terms of activities and agents that are generic to data sharing across all domains. The model of drivers, barriers and enablers presents a comprehensive enumeration of the factors that can motivate or inhibit data sharing and those enablers by which barriers might be surmounted or reduced.

The Conceptual Model 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 to develop 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.

A further outcome of the research undertaken in this phase of the ODE Project, in addition to the Conceptual Model, has been the formulation of a simple tool for evaluation of a data sharing domain, which is presented in Annex 2. The data sharing domain evaluation tool affords a means to assess the maturity of a data sharing domain by the presence and strength of certain indicators. This is proposed as a high-level domain analysis tool that may be useful in identifying areas that need to be addressed in policy. It should be seen as complementary to the main findings presented in the body of this report.

The process of validating and qualifying the Conceptual Model in discussion with members of data sharing stakeholder groups has also served to identify a number of salient themes in respect of data sharing today and in the future. The thematic presentations in Section 4 synthesise this material into a number of coherent views on aspects of data sharing that the interviews threw into bold relief. Under each thematic treatment a Conceptual Model analysis demonstrates how the Model can help to structure thinking about the operative drivers and barriers, and indicates the key action points required of stakeholders to enable progress in data sharing.

It is not in the scope of this work to produce recommendations. The action points are those that 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. It is short step for stakeholders to take from these action points to arrive at recommendations for their communities. In this light it is hoped that the work of the ODE Project might represent a foundation on which others can build.

That there is much building to be done is indisputable. Validation of the Conceptual Model underlined the highly fragmented nature of the data sharing field. The level of interdisciplinary data sharing is a telling indicator in this respect. It certainly appears to be the case that interdisciplinary data use is not currently taking place systematically, although there are examples in areas such as Earth and environmental sciences,

Economics and Archaeology, where there is widespread use of existing data sets from a variety of disciplinary and public data sources. In perhaps the most high-profile example researchers at the Intergovernmental Panel on Climate Change (IPCC) have constructed vast multidimensional data sets by combining atmospheric, solar, agricultural, biological and socio-economic data[54]. Such interdisciplinary activities can play a progressive role in the evolution of data sharing more generally, precisely because they promote the development of common infrastructure and technologies and the acceptance of common standards and languages for data.

The aim of the ODE Project is to support strategic policy-making and investment in the development of a European e-infrastructure for data sharing. This directly relates to the European Commission's 'Horizon 2020'[55] initiative for a global data infrastructure and a digital research area for Europe. This is currently being widely consulted, and one of the stimulants to discussion is the proposed vision for global data e-infrastructure in 2030, expressed as a series of desired goals to be achieved by that date (High Level Expert Group on Scientific Data, 2010).

It is clear that ODE relates to many of goals of Horizon 2020, and the relationship may arise in different ways: the recognition of drivers as being key for the achievement of the Horizon 2020 goals, the overcoming of barriers, or the strengthening of enablers. For example, one of the stated goals of Horizon 2020 addresses the issues of data discovery, data usability and trustworthiness of data:

> *Researchers and practitioners from any discipline are able to find, access and process the data they need in a timely manner. They are confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.*

Implicit in this vision is an overcoming of those barriers that currently prevent researchers from discovering, accessing and using the data they need, and the analytical models provided by ODE offer a number of tools through which policy-makers and funders can identify those barriers and put in place the policies and practices that will enable stakeholders to overcome them.

It is hoped that the ODE Conceptual Model can be of value in elucidating the relationships between the Horizon 2020 goals and the conditions needed to bring them about, thus providing further input to the realization of the Horizon 2020 vision.

[54] http://www.ipcc-data.org/
[55] http://ec.europa.eu/research/horizon2020/index_en.cfm

6 BIBLIOGRAPHY

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H. et al. (2011), 'Public availability of published research data in high-impact journals', *PLoS ONE* 6(9): e24357. doi:10.1371/journal.pone.0024357

Arzberger, P., Schroeder, P., Beaulieu, A. et al. (2004), 'Promoting access to public research data for scientific, economic, and social development', *Data Science Journal* 3: 135-152. doi:10.2481/dsj.3.135

Attwood, T. K., Kell, D. B., McDermott, P. et al. (2009), 'Calling International Rescue: knowledge lost in literature and data landslide!', *Biochemical Journal* 424: 317–333. doi:10.1042/BJ20091474

Beagrie, N., Chruszcz, J., and Lavoie, B. (2008), *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. Final Report. Available at: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Lavoie, B., and Woollard, M. (2010), *Keeping Research Data Safe 2.* Final Report. Available at: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Berman, F., Lavoie, E., Ayris, P. et al. (2010), *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. Available at: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Birnholtz, J. P. and Bietz, M. J. (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*. New York, NY: ACM. doi: 10.1145/958160.958215

Borgman, C. L. (2010), 'Research data: who will share what, with whom, when, and why?', in *China-North America Library Conference*. Beijing. Available at: http://works.bepress.com/borgman/238

Brownstein, C. A., Brownstein, J. S., Williams, D. S. et al. (2009), 'The power of social networking in medicine', *Nature Biotechnology* 27(10): 888-890. doi:10.1038/nbt1009-888

Consultative Committee for Space Data Systems (2009), *Audit and Certification of Trustworthy Digital Repositories: Draft Recommended Practice. CCSDS 652.0-R-1. Red Book*. Available at: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS 6520R1/Attachments/652x0r1.pdf

Deutsche Forschungsgemeinschaft (2012), *Proposal Preparation Instructions: Project Proposals*. Available at: http://www.dfg.de/formulare/54_01/index.jsp

Gardner, D., Toga A. W., Ascoli, G.A. et al. (2003), 'Towards effective and rewarding data sharing', *Neuroinformatics* 1(3): 289-95. Available at: http://hinxtongroup.files.wordpress.com/2010/10/gardner_neuroinformatics_20031.pdf

Goldstein, S. J. and Ratliff, M. (2010), 'DataSpace: a funding and operational model for long-term preservation and sharing of research data'. Available at http://arks.princeton.edu/ark:/88435/dsp01w6634361k

Hartig, O. (2008), 'Trustworthiness of data on the Web', in *Proceedings of the STI Berlin & CSW PhD Workshop, Berlin, Germany*. Available at: http://www2.informatik.hu-berlin.de/~hartig/publications.html

High Level Expert Group on Scientific Data (2010), *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission. Available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Hodson, S. (2009), 'Data-sharing culture has changed', *Research Information* December 2009/January 2010. Available at: http://www.researchinformation.info/features/feature.php?feature_id=243

International Organization for Standardisation, *Space Data and Information Transfer Systems – Open Archival Information System – Reference Model*. ISO 14721:2003. Available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

Kaye, J., Heeney, C., Hawkins, N. et al. (2009), 'Data sharing in genomics – re-shaping scientific practice', *Nature Reviews Genetics* 10: 331–335. doi: 10.1038/nrg2573

Key Perspectives Ltd (2010), *Data dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability.* SCARP project synthesis report. Edinburgh: Digital Curation Centre. Available at: http://www.era.lib.ed.ac.uk/bitstream/1842/3364/1/SCARP%20SYNTHESIS.pdf

Kuipers, T, and van der Hoeven, J. (2009). *Survey Report*. PARSE.Insight: Insight into issues of Permanent Access to the Records of Science in Europe. Available at: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

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

National Science Foundation (2011), *Data Sharing Policy*. Available at: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Newton, M. P., Mooney, H. and Wittz, M. (2010), *A Description of Data Citation Instructions in Style Guides*, poster presented at the *International Digital Curation Conference (IDCC), December 7-8, 2010, Chicago, Illinois*. Available at: http://docs.lib.purdue.edu/lib_research/121/

Organisation for Economic Co-operation and Development (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Available at: http://www.oecd.org/dataoecd/9/61/38500813.pdf

Pampel, H., Pfeiffenberger, H., Schäfer, A. et al. (2012), *Report on Peer Review of Research Data in Scholarly Communication*. Available at: http://epic.awi.de/30353/

Parr, C., Cummings, M. (2005), 'Data sharing in ecology and evolution: why not?'. Technical Report HCIL-2005-06, CS-TR-4708, UMIACS-TR-2005-16. Available at: http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2005-06

Phippena, A., Razaa, A., Butela, L. et al. (2011), 'Impacting methodological innovation in a local government context – data sharing rewards and barriers, *Methodological Innovations Online* 6(1): 58-72. Available at: http://www.pbs.plym.ac.uk/mi/pdf/09-06-11/8.%20Article%20-%20Phippen%20et%20al%2058-72%20(proofed).pdf

Piwowar, H. A., Day R. S. and Fridsma, D. B. (2007), 'Sharing detailed research data is associated with increased citation rate, *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308

Piwowar, H. A. and Chapman, W. W. (2008), 'A review of journal policies for sharing research data', in *ELPUB 2008: International Conference on Electronic Publishing*. hdl:10101/npre.2008.1700.1. Available at: http://precedings.nature.com/documents/1700/version/1/files/npre20081700-1.pdf

Polaiologk, A., Tjalsma, H. and Sesink, L. (2010), 'Costs of digital archiving: the case of DANS', in *Proceedings IASSIST 2010: Social Data and Social Networking: Connecting Social Science Communities across the Globe*. Available at: *http://*www.slidefinder.net/I/IASSIST2010ABCmodel/23825444

Pundt, H. and Bishr, Y. (2002), 'Domain ontologies for data sharing – an example from environmental monitoring using field GIS', *Computers & Geosciences* 28: 95–102. doi:10.1016/S0098-3004(01)00018-8

Reilly, S., Schallier, W., Schrimpf, S. et al. (2011), *Report on Integration of Data and Publications*. Available at http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Schäfer, A., Pampel, H., Pfeiffenberger, H. et al. (2011), *Baseline Report on Drivers and Barriers in Data Sharing*. Available at http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf

Schofield, P. N., Bubela, T., Weaver, T. et al. (2009), 'Post-publication sharing of data and tools', *Nature* 461: 171-173. doi:10.1038/461171a

Tenopir, C., Allard, S., Douglass, K. et al. (2011), 'Data sharing by scientists: practices and perceptions, *PLoS ONE* 6(6): e21101. doi:10.1371/journal.pone.0021101

Van den Eynden, V., Corti, L., Woollard, M. et al. (2011), *Managing and Sharing Data: Best Practice for Researchers*. 3rd edition. Colchester: UK Data Archive. Available at: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Van House, N. A. (2002), 'Trust and epistemic communities in biodiversity data sharing', in *JCDL '02 Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY: ACM. doi:10.1145/544220.544270

Walport, M. and Brest, P. (2011), 'Sharing research data to improve public health', *The Lancet* 377(9765): 537-539. doi:10.1016/S0140-6736(10)62234-9

Wellcome Trust (2010), *Policy on Data Management and Sharing*. Available at: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Wheatley, P. and Hole, B. (2009), 'LIFE3: predicting long term digital preservation costs', in *iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*. Available at: http://www.escholarship.org/uc/item/23b3225n

Wilson, M.D. (2008), 'A case study in the rewards of long term data sharing - 26 years of the MRC Psycholinguistic Database', in *UK e-Science All Hands Meeting '08 (AHM '08), Edinburgh, 08-11 Sep 2008*. Available at: http://epubs.stfc.ac.uk/work-details?w=49008

Wright, M. (2011), *Research Data Sharing in the UK*, presentation to the heads of the *G8 Research Councils Workshop on Access to Research Data: New Developments and Old Problems, 10-12 July 2011, Scarman House, University of Warwick, UK*. Available at: http://securedata.data-archive.ac.uk/media/21925/horcg8mwr.pdf

ANNEX 1: INTERVIEW PRO FORMA

The interviews conducted in WP5 used a standard script as a basis, derived from the Conceptual Model of drivers and barriers as it was at the time of the interviews. It should be noted that this differs from the final version presented in this report by not separating out enablers from drivers and barriers. Interviewers used a pro forma for recording the results of the interviews against the questions of the script. There are separate sections for researchers and non-researchers.

(a) Interview with an academic researcher

About the interview itself

Name of interviewer	
Date of interview	
Method of interview (telephone, Skype, …)	
Approximate duration of interview (minutes)	

Questions to be filled in prior to interview from reply to invitation

1. Name	
2. Organisation	
3. Country	
4. Age group	
5. Role	
6. Academic discipline	

Multiple choice questions

7. Which of the following applies to the digital research data of your research? (multiple answers possible)	a) My data is openly available for everyone. b) My data is available for a fee. c) My data is openly available for my research discipline. d) My data is openly available for my research

	group / colleagues in research collaboration. e) Access to my data is temporarily restricted Which restriction?

8. Which of the ODE drivers motivate you in sharing your data with others? (multiple answers possible from list—record code number(s))	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
9. Which of the ODE drivers motivate you in using other people's data? (multiple answers possible from list—record code number(s))	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
10. Do you presently/or in the past make use of research data gathered by other researchers WITHIN your discipline?	Yes / No
11. Which of the ODE barriers have you encountered in sharing data within your discipline? (multiple answers possible from list—record code number(s))	f) Individual Contributor barriers 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
12. Do you presently/or in the past make use of research data gathered by other researchers in OTHER disciplines?	Yes / No

13. Which of the ODE barriers have you encountered in sharing data outside your discipline? (multiple answers possible from lis—record code number(s))	f) Individual Contributor barriers 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

Questions for free discussion

14. Can you give any examples of barriers preventing sharing, or of overcoming those barriers, in your discipline or across disciplines please? (one or more story answer(s))

If the interviewee needs a prompt, then: what was the data, who was the data provider & data consumer, when did this happen, what was the research topic, why was the sharing important, what were the barriers, how were they overcome?

15. From your perspective, what does good data citation look like, and why?

If the interviewee needs a prompt, then: This could be in terms of HOW (form of citation, identifiers used or technical solutions, WHERE (in reference lists, acknowledgements or in-line) and WHEN data is cited within a paper, or data citing resulting research.

Additional free discussion response here

(b) Interview with a non-researcher

About the interview itself

Name of interviewer	
Date of interview	
Method of interview (telephone, Skype, …)	
Approximate duration of interview (minutes)	

Questions to be filled in prior to interview from reply to invitation

1. Name	
2. Organisation	
3. Country	
4. Age group	
5. Role	

Multiple choice questions

6. Do the researchers that you support or work with share their data?	Yes / No
7. Which of the ODE drivers motivate them in sharing their data with others? (multiple answers possible)	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives

8. Which of the ODE drivers motivate them in using the data of others? (multiple answers possible)	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
9. Do they make use of research data gathered by other researchers WITHIN their discipline?	Yes / No
10. Which of the ODE barriers have they encountered in sharing data within their discipline? (multiple answers possible)	f) Individual Contributor barriers 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
11. Do they make use of research data gathered by other researchers in OTHER disciplines?	Yes / No
12) Which of the ODE barriers have they encountered in sharing data outside their discipline? (multiple answers possible)	f) Individual Contributor barriers 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

Questions for free discussion

13. Can you give any examples of barriers preventing sharing, or of overcoming those barriers, within or across disciplines please? (one or more story answer(s))

If the interviewee needs a prompt, then: what was the data, who was the data provider & data consumer, when did this happen, what was the research topic, why was the sharing important, what were the barriers, how were they overcome?

14. From your perspective, what does good data citation look like, and why?

If the interviewee needs a prompt, then: This could be in terms of HOW (form of citation, identifiers used or technical solutions, WHERE (in reference lists, acknowledgements or in-line) and WHEN data is cited within a paper, or data citing resulting research.

Additional free discussion response here

ANNEX 2: EVALUATING A DATA SHARING DOMAIN

Data sharing domains may be defined at different levels, according to different variables: typically subject area, stakeholder group, sector (non-commercial research, research, education), and geopolitical context (national/regional policy and legislation, infrastructure, funding). One of the most useful ways to define a domain is by discipline, as this so often corresponds to a community of shared knowledge and practice with specific data requirements.

There are different levels of maturity in data sharing in different disciplinary domains, and the degree of maturity of a domain can to some degree be gauged by the presence or absence of certain indicators. The table below outlines a checklist of stakeholder group indicators which can be used to evaluate a data sharing domain.

Stakeholder group	**Indicator**
Researchers	How evolved is the data sharing culture, i.e. the does a presumption exist that data will be shared? • Is the typical data processing requirement greater or less than the typical data processing capacity of the researcher or research unit (big data/small data)? • Is collaboration between research groups typical? • Are data sets typically subject to sharing constraints (e.g. commercial/legal)? • How well-known and widely-accepted are the data citation and description standards in a given field?
Research and education organisations	Does the organisation provide support to researchers in sharing data? Does the organisation provide training to researchers in data practice?
Funders	Do funders require grantholders to prepare data management plans? Do funders monitor and enforce data management?
Policy-makers (national and regional)	Is there a clear policy commitment to data sharing? Is data publication recognized in national research assessments? Are there international policy agreements or co-ordinated frameworks in place?

Service providers (infrastructure and data management)	Are there well-funded data centres in the field providing high-quality data curation services and specialist support? Are data centres certified as meeting recognised data preservation standards? Are there data centres with established linking relationships with publishers? Are new technologies and tools for managing, sharing, discovering and querying data being developed?
Publishers	Are there established linking relationships between publishers and data centres? Do publishers accommodate supplementary data publication alongside articles? • Do they have clear data publication policies? • Do they have explicit and transparent peer review and quality assurance processes? • Do they support a variety of data file formats, including non-proprietary formats? Do publishers recognise and consistently apply data description and citation standards in the disciplinary journals? Do publishers provide added-value data services, such as discovery interfaces to external data stores?

Some of the indicators in this table, such as general policies, may be present at a high level across a broadly defined domain; others are much more closely tied to very specific disciplines.

As an example, one might evaluate a domain by looking at the indicators against a given stakeholder group, such as publishers. The existence of commercial data services as part of the overall service infrastructure in a given disciplinary domain might be indicative of the general maturity of data sharing practices within that domain. Commercial services exist to meet a demand, and where demand does not exist or is insufficient, there will be no commercial proposition for publishers. This may not be the whole story, of course, as there is an arguable chicken-and-egg quality to the relationship between data sharing services and data sharing practices: in other words, researchers cannot fully realise their data sharing potential in commercially viable ways until the commercial service infrastructure is in place. A demand may exist; a market may be there, but invisible to publishers because they are not providing the means for it to declare itself.

Nevertheless the case holds that where there are services such as established linking relationships between publishers and data centres, as there are in earth sciences and biological sciences, there is a relatively mature and active data sharing culture. This is also reflected in well-established data description and citation standards in the disciplinary journals and the existence of standalone data journals (in the latter case, examples in archaeology and earth sciences were cited by interview respondents). Other indicators may include publishers with explicit and transparent peer review and quality assurance process for supplementary data, and innovation in technologies and services, such as data file formats or discovery interfaces to external data stores.

Conversely, in areas where publishers cannot find viable data repositories with which to establish linking relationships, as in some humanities disciplines, or where publishers do not follow best practice in data citation, or fail to recognise data references in submitted manuscripts, there is likely to be a less mature data sharing domain. Reasons for this may be best elucidated by looking at the researchers themselves: whether there is a presumption that data will be shared; what kinds of data researchers use and produce; the typical size of research units and how collaborative they are with other research units; whether research data are often hedged by legal or commercial constraints; and so on.