Grant agreement no. 261530

Opportunities for Data Exchange

ODE - Opportunities for Data Exchange

Theme: Research Infrastructures

Topic: INFRA-2010-3.3 Coordination actions, conferences and studies supporting policy development, including international cooperation, for e-Infrastructures

D6.1 SUMMARY OF THE STUDIES, THEMATIC PUBLICATIONS AND RECOMMENDATIONS

SEVENTH FRAMEWORK PROGRAMME

Document identifier:	**ODE-WP6-DEL-0001-1_0**
Date:	**26 Oct 2012**
Work package:	**WP6**
Partners:	**APA, CERN, CSC, HA, STFC**
WP Lead Partner:	**CSC**
Deliverable:	**D6.1**
Document status:	**Final Version**

Document Status Sheet

Issue	Date	Comment	Author
1_0	26 Oct 2012	Final version	Sunje Dallmeier-Tiessen Robert Darby Kathrin Gitmans Patricia Herterich Simon Lambert Salvatore Mele Josefine Nordling Hans Pfeiffenberger Sergio Ruiz Eefke Smit

Document Change Record

Issue	Item	Reason for Change

Project information

Project acronym:	**ODE**
Project full title:	**Opportunities for Data Exchange**
Proposal/Contract no.:	**261530**

Project Officer: Carlos Morais-Pires

Address:	Information Society and Media European Commission (BU25 - 4/85) Brussels, Belgium
Phone:	+32 2 29 63 401
Fax:	+32 2 29 93 127
E-mail:	Carlos.Morais-Pires@ec.europa.eu

Project Co-ordinator: Dr. Salvatore Mele

Address:	CERN Open Access Section C27900, CERN CH1211 Geneva 23
Phone:	+41 22 767 8603
Fax:	+41 22 766 8700
E-mail:	Salvatore.Mele@cern.ch

1. INTRODUCTION

1.1 Data and its reuse in science

Data are central to all fields of science, whether measurements taken on highly specialised facilities, observations of the cosmos or the earth, results of surveys, records of communications or transactions, quantitative analyses of texts, or infinite other varieties. They are both the raw material and the end-product of science: the raw material, because of their centrality in testing hypotheses, validating observations and inspiring theories; the end-product, because the accumulation of reliable data on aspects of the observable world is part of the heritage of scientific endeavour.

The term 'data deluge' or even 'data tsunami' has been coined to encapsulate in vivid terms the vast quantities of data generated from the latest generations of experimental or observational facilities. Indeed, the recent observation of the Higgs-like particle at the Large Hadron Collider at CERN is underpinned by 34 petabytes of data (and simulation) written to tape in one year; in the foreseeable future, the Square Kilometre Array telescope will generate 1 exabyte of data every day. In addition to the data-intensive side of the spectrum, another end exists: a long tail of small-scale data typically collected by manual processes, which should not be eclipsed. This data may be just as reusable and potentially useful as large-scale data; testimony to a comparable vast effort collectively invested by many communities into its collection and validation, as exemplified by a recent effort[1] to collate global data on biomass of plankton, which accounts for half of all carbon in the biosphere..

The emergence of e-Infrastructures for data recording and processing has had enormous implications for the transition from science to data-intensive science. Data can be copied, transformed, refined, combined and analysed in ways that were unimaginable until recently. All these potentialities lend themselves to data sharing, or data exchange: the productive opening up of data for reuse beyond the immediate purpose for which it was collected. This is important for several reasons:

The scientific method: reproducibility. Allowing results to be reproduced and anomalies to be further (and independently) investigated.

Avoiding unnecessary replication of work. If a particular experiment under particular conditions has already been conducted, future researchers can reuse the detailed results without needing to repeat it.

[1] MAREDAT – Towards a world atlas of marine plankton functional types, http://www.earth-syst-sci-data-discuss.net/special_issue9.html

Long-term reuse. The accumulation of data relating to particular points in time which *ipso facto* cannot be reproduced—the obvious example being in climatology.

Bringing data together in new ways. Aggregating and combining datasets that have not hitherto been connected to open up new perspectives, in particular interdisciplinary ones.

However it should not be imagined that data sharing can be achieved without effort, or that no difficulties arise even at the stages of preparing data for sharing. The data may require processing in order to be acceptable for wider release: for example, anonymisation of data relating to individual persons. It may also require some degree of enriching with supplementary information to make it possible to reuse, for example, explicit statements about calibration and uncertainties, or the conditions where the data were collected, which might be well known to the original project team but certainly not to potential reusers of the data.

This centrality of data and its various roles has led to the idea that we are entering a fourth paradigm of science, an idea originally described by Jim Gray[2]. He recognised that science has passed through three stages: the empirical, theoretical and computational, and is now entering the stage of data exploration, which may also be labelled 'e-science'. The reasons for the passage from one stage to the next are distinct, but they all reflect some limitation in the methodology of the previous stage. Thus 'the theoretical models grew too complicated to solve analytically, and people had to start simulating.' Now the limitation is the difficulty in managing vast amounts of data, which might be either observational or arising from simulations. Data, as Gray points out, has become a vital resource in its own right: 'People now do not actually look through telescopes. Instead, they are "looking" through large-scale, complex instruments which relay data to datacentres, and only then do they look at the information on their computers.' This leads to data-intensive science, which in fact enhances the methods of the previous stages of experiment, theory and simulation at a scale which permits the classification as a new paradigm.

Once the data is collected and managed in such an environment, the possibility of exchange and reuse inevitably arises. Data may be reprocessed, transformed and synthesized to make it suitable for different purposes. Interdisciplinarity thrives in such an environment.

Different scientific disciplines have different degrees of readiness for data exchange. The potential exists, but the benefits cannot be obtained without some

[2] The Fourth Paradigm: Data-Intensive Scientific Discovery, http://research.microsoft.com/en-us/collaboration/fourthparadigm/

effort and costs. Data that is neglected will not be reusable, or not for very long, or not beyond the restricted circle of the researchers who first collected it. Curation and preservation of data must keep up with the trend to openness, or the masses of data on all scales might become nothing more than vast mausoleums of scientific endeavour.

1.2 Stakeholders and infrastructures

Data sharing requires conscious decisions and actions from all those stakeholders who enable, influence or benefit from exchange of data. These actions all relate to different aspects of the same scientific infrastructures: that is, 'technical tools and instruments and socio-economic systems for organising and sharing knowledge'[3]. These stakeholders include, centrally, the researchers who generate and use data in their studies, but also the actors who manage data for them and those who finance storage and curation or set policies on its accessibility or long-term retention; and the long-term actors of scholarly communication: academic publishers whose journals report the findings and libraries who collect and organise this knowledge. All these have an interest in what is done with data, and also an ability to influence what is done, either through direct requirements expressed through policies or through their day-to-day actions and behaviours.

The PARSE.Insight roadmap for a science data infrastructure focussed on long-term preservation[4] and conceived the technical components of the infrastructure in terms of what threats they were designed to counter: for example, a component to provide evidence of authenticity of a digital object counters a threat that the chain of evidence may be lost and there may be lack of certainty of provenance or authenticity. But the infrastructure also encompassed other aspects: financial, organisational/social and policy, even while accepting that some of these are difficult to make concrete.

The *Riding the Wave* report of the High-Level Expert Group on Scientific Data, submitted to the European Commission in October 2010, urges an international framework for a collaborative data infrastructure, which emphasizes the different roles of stakeholders in the creation and sustaining of the technical infrastructure, which must be 'flexible but reliable, secure yet open, local and global, affordable yet high-performance'.

It is from this starting point that the ODE project was conceived. The motivation came from the realisation that the full potential of science can now only be achieved through scientific data infrastructures which present a layer where data can be shared, and that the driver and barriers to such sharing have to be understood, to allow this vision to be realised.

[3] EC Commissioner Neelie Kroes quoted in *Riding the Wave: Final Report of the Export Group on Scientific Data*, October 2010

[4] http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

These obstacles, once identified, will not vanish without positive action, and these actions must be based on information that is reliable, timely and focussed. On this basis decisions can be taken to shape policy and investment.

1.3 The work of the ODE project

The ODE project has characterised and understood the key factors in data sharing relating to stakeholders and infrastructures. Its objective, as described by the results presented in this report, was to provide timely and relevant information to allow informed decisions to be made. Five themes were recognised as significant for data sharing through a European ecosystem of interoperable e-Infrastructures and data repositories:

- the sociological theme
- the outreach theme
- the scholarly communication theme
- the critical mass theme
- the theme of "why should we bother?"

Each of these themes gave rise to a number of questions for which ODE provided the ability to explore and respond to these questions, accurately, efficiently and promptly.

The relevant factors were conceptualised in terms of what forces would encourage, discourage or prevent the practice of data sharing. A conceptual model was developed as a framework in which to explore these forces. The forces themselves were categorised as *drivers, barriers* and *enablers*, and were fitted within a *process model* describing the functional logic of data sharing in terms of agents, actions and objects, and a *context model* mapping 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 was developed from examination of a wide range of previous studies.

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.

The conceptual model (described in deliverable D5.1 Compilation of results on drivers and barriers and new opportunities[5]) was validated, refined and elaborated through a process of consultation and review with expert and interested members of the key stakeholder groups. This validation process was conducted in two stages: a workshop on data sharing was held in conjunction with the conference of the Alliance for Permanent Access 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.

Within this conceptual framework, the work of the ODE project followed a logical arc. The first step was to develop a broad, shared understanding of the issues to be addressed by the project and to produce a concise analysis of the questions arising. A public report (Baseline report on Drivers and Barriers on Data Sharing[6])was produced that provides an inventory of material collected and generated and people contacted during the work, with conclusions regarding the baseline situation today. It takes the form of a collection of 21 stories of success, near misses and honourable failures in data sharing.

A selection of these stories was edited to illustrate the most important points across a variety of disciplines and is available online and in hard copy (Ten Tales of Drivers and Barriers in Data Sharing[7]).

A parallel line of work considered the impact that data sharing, re-use and preservation is having on scholarly communication, with a view to identifying incentives for researchers and other stakeholders that will help to optimise the take-up of future e-infrastructure. This was summarised to two reports:

- Report on integration of data and publications[8]
- Report on Best Practices for Citability of Data and on Evolving Roles in Scholarly Communication[9]

The central activity on 'questions and answers' sought to reach consensus among experts in an understanding of the drivers and barriers for data sharing, re-use

[5] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Compilation+of+Results+on+Drivers+and+Barriers+and+New+Opportunities
[6] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Baseline+report+on+Drivers+and+Barriers+on+Data+Sharing
[7] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Ten+Tales+of+Drivers+%26+Barriers+in+Data+Sharing
[8] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+Report+on+Integration+of+Data+and+Publications
[9] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Report+on+Best+Practices+for+Citability+of+Data+and+on+Evolving+Roles+in+Scholarly+Communication

and preservation, starting from the information on views gathered previously, available in D5.1 Compilation of results on drivers and barriers and new opportunities[10].

The final step was to demonstrate the value of information gathered and to distil the results in order to ensure that each of the project's target audiences can make informed decisions about the future of e-Infrastructures for data sharing and preservation. This was done through five different briefing sheets aimed at each stakeholder group:

- Researchers[11]
- Data Centres[12]
- Libraries[13]
- Funders[14]
- Publishers[15]

After a reminder of the approach of the project, this report summarises this work in five chapters, one for each of the stakeholder groups. Each chapter discusses the current situation and approach to data sharing within each stakeholder group, giving some examples of best practices, examples, success stories and lessons learned on data sharing. Suggestions on how to implement and scale these successful approaches are given. A final conclusion outlines the main finding of the project.

[10] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Compilation+of+Results+on+Drivers+and+Barriers+and+New+Opportunities
[11] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+briefing+sheet+for+Researchers
[12] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+briefing+sheet+for+Data+Centres
[13] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+briefing+sheet+for+Libraries
[14] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+briefing+sheet+for+Funders
[15] http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+briefing+sheet+for+Publishers

2. APPROACH

The primary challenge for the ODE project was to ensure that credible evidence about attitudes towards and experience of data sharing and (re)use would be gathered, interpreted and summarised in meaningful and relevant ways for each of the ODE target groups, discussed in the introduction, and communicate it back to them.

The project partners' researchers could build on and, during the course of the project, expand their network of relationships with members of these target communities.

The approach had to ensure that the information flows between the project and those communities were managed properly, added value to and disseminated with an impact beyond those persons and organizations engaged with in the course of the project.

To this end, there was heavy emphasis on the two parallel and iterative threads of the project, external communication of (intermediate) results – mainly organized and performed by Work Package 2 (WP2) – and the research generating the results. Research, then, was performed in two phases: generating baseline information in WP3 and WP4 and conceptualising, and generating in-depth qualified results in WP5. At the end of both phases, extensive intra-project and external discussion were organized to support conceptualization, perform interpretation and summarization.

In a third phase engaging all project partners, the general and target-group specific final products were derived in WP6 and tailored dissemination took place, again organized and documented together with WP2. It is expected that the material generated and further publications and presentations of project participants will create notable impact well beyond the official end of the project.

2.1 Phase 1

To develop a broad, shared understanding of the issues to be addressed by the project, WP3 strove to gather as much as possible fresh, unforeseen, hitherto undiscussed insights and examples of sharing or not sharing of data. In order to "break out of boxes" the experienced partners might be caught in, WP3 conducted 21 interviews, which were structured to be as open as possible with influential persons from all groups, with an emphasis on researchers and funders. It is to be noticed that senior members of both groups tend to be members of the so called group of "policy makers". A number of "service providers" and a researcher from a project mobilizing amateur scientists completed the field of view.

In parallel, WP4 performed a deep, partially quantitative, study of attitudes and findings in constituencies traditionally in the business of managing research information: libraries and publishers. At first, it re-used data from the PARSE.Insight project on the status and outlook of preservation and re-use of data, and then built on this expertise by conducting a survey and interviews. This research particularly focused on members of the LIBER group of research libraries, STM publishers and journal editors.

Both WPs did not just package their results but before they did this, discussed and analysed them - with all partners, intensively, remotely and finally at an all-hands meeting in early July 2011 - to derive questions about and categories of "drivers and barriers".

2.2 Phase 2

At the same meeting, WP5 presented and the consortium discussed a first version of its conceptual model of data sharing, based on which the questions for its later series of interviews were to be structured and formulated. Preliminary alignment of WP3 and WP4 results and conceptual model took place at this juncture.

From May 2011 up to the first conference, under the auspices of APA, London, in November 2011, preliminary and partial results from WP3 and WP4 were presented, discussed and validated at various workshops and conferences of target groups as diverse as libraries, high energy physics researchers and the European Parliament. The feedback was gathered and used to improve the conceptual model.

It was found that half of the interviews from WP3 provided appealing and revealing stories or "Tales". The glossy brochure "Ten Tales" was developed and immediately found an interest, in particular in the political and other decision-making domains – up to a group of officials preparing a G8 summit.

At the first conference, a workshop composed of ODE members and external experts discussed the second version of the WP5 conceptual model, in depth, and finals results of WPs 3 and 4 were presented as planned.

In the first quarter of 2012, more than 50 semi-structured interviews were conducted by project members as part of WP5. The interviews were analysed subsequently and qualified themes in the conceptual model. Furthermore, they aimed at detecting salient themes of particular importance.

2.3 Phase 3

The third phase (WP6) was launched with an all-hands meeting (AHM) in April 2012 to summarize and categorize findings and to find the most effective formats

besides this report and venues besides the second conference for the dissemination of ODE results.

It was decided soon afterwards to prepare tailored high-quality hand-outs. They comprise a folder with 2-page summaries of most important findings, directed individually at each of 5 stakeholder groups defined at the AHM and a number of subsequent teleconferences. It was here that a third category of findings, "enablers", besides drivers and barriers, was defined and made use of in "what you can do"–sections of the hand-outs.

Again, between April and November 2012 preliminary versions of findings and the hand-outs were presented and discussed at a range of conferences and meetings of national and international groups, other projects and political meetings, e.g. in Brussels, prior to the second conference. At the Frascati APA conference in November 2012 the full and final results and deliverables were presented publicly and to European Commission representatives and subjected to official review. The core and conclusions of ODE findings are discussed in the remainder of this document.

3. STAKEHOLDERS

3.1 Data centres

3.1.1 Introduction

Data centres are essential components of a data sharing system. Data for long-term preservation and sharing, and data management tailored to specific discipline and data requirements can best be provided by specialist data service providers. This is axiomatic, as is demonstrated by the existence of large numbers of data centres serving a wide variety of data management needs across a broad range of communities.

Data centres come in many shapes and sizes, from nodes in the vast distributed networks of big science, to modest institutional repositories housing data files generated by local researchers. At their most basic data centres store research datasets for a defined community and make these datasets accessible for others to discover and use. But data centres are also active components of research communities, defining, preserving and communicating the intellectual capital that researchers draw on to advance the frontiers of research. As such data centres are also defined – to a greater or lesser degree – by some or all of the following features:

- they train and support researchers to ensure the data they ingest are of required quality and meet specified standards, and comply with any legal or other requirements;
- they assign DOIs to datasets and create and enhance the metadata that give data meaning and enable other researchers and services to discover and process the data;
- they curate data, managing the integrity of datasets, controlling access to them and how they are used, and ensuring that datasets remain accessible and usable in the long-term, and can be persistently referenced from other sources;
- through participation in their communities they shape and promulgate data and metadata standards; and
- they forge the links with allied data centres, publishers and other service providers that allow data to function within a global scholarly communication network.

The effectiveness of data centres as agents of data sharing is reliant on the knowledge and expertise of the people who operate them, and the strength of their working relations with the communities they serve. The ODE project has found that data centres that successfully promote data sharing:

- know their communities intimately, are trusted by them, and tailor their services to the needs of those communities;
- strive to maintain the highest standards of data management and work with their research communities to promote those standards;
- are funded and managed sustainably; and
- actively seek to work within the global research network by forging links with other data centres, with research institutions, with publishers, and with complementary service providers.

3.1.2 Current situation among data centres

Data centres containing datasets that can be accessed online are now widely distributed among research communities, and there are examples of disciplines where data centres play a central role in an active data sharing community. But there is no cause for complacency. The 2009 PARSE.Insight survey of researchers in all disciplines worldwide found that only 20% of respondents stored their research data in a digital archive, with 14% using an organisation archive, and just 6% a discipline archive[16]. Although data preservation and sharing activity might be expected to have increased somewhat since then, this baseline figure should be borne in mind when considering the current role of data centres in data sharing.

The distribution of data centres varies by discipline and community. Well-established data centres are thriving in areas where there is a mature culture of data sharing – such as molecular biology, earth observation and some social science disciplines; whereas in areas where the willingness or imperative to share data is weak, data centres are correspondingly less well established, less engaged with their communities, and less successful.

The nature of data, their significance within a given research activity, and the data requirements of researchers vary widely, and determine the nature and function of the data centre. Different kinds of data require different kinds of management. Data derived from scientific instruments need quite different kinds of handling to clinical trial data or social science survey data.

For example, PANGAEA[17] collects earth observation data from remote sensors with complex calibrations, and the final registered dataset is the outcome of an intensive process of collaborative normalization, in which in-house data editors and experts embedded in major projects work alongside research teams to generate datasets. This requires expert specialist knowledge applied in parallel to actual research processes, both to form the data and to document the

[16] Kuipers, T.van der Hoeven, J. (2009), *Insight into Digital Preservation of Research Output in Europe. Survey Report*, p.32. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf .
[17] http://www.pangaea.de/ .

instruments and transformations used in such a way that data can be clearly understood by other researchers. Long-term curation of instrumental data can also be very effort-intensive, and may require highly engineered monitoring and preservation of bespoke software and hardware environments.

Other kinds of data, such as social science data or clinical trial data, may be relatively simple to record and understand, and present few technical demands for long-term preservation. But if such datasets include personal data, then ethical and legal guidance may be required at the outset of data collection, and intensive processing may have to be undertaken to aggregate and anonymise data, and to ensure that appropriate consents for data sharing are obtained. This again is very labour-intensive and requires a great deal of engagement with researchers.

A data centre as commonly understood is an organisation serving a community defined by discipline or subject area. Here are just a few examples encountered by the ODE Project:

- the Dryad repository of data referenced in published papers in the biological sciences[18];
- the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL)[19];
- the data centre of the International Council for the Exploration of the Sea (ICES)[20]; and
- the Dutch Data Archiving and Networked Services (DANS) data archive[21], which collects datasets in the humanities, archaeology, geospatial sciences and behavioural and social sciences from various research organisations with which it has contracts. Subject collections are separately searchable.

Dryad is funded by project grants; EBI and ICES are both funded by contributions from international consortia of partner countries; the DANS data archive is funded by the Dutch government. The data centre may collect a narrow or wide variety of data types from a diverse range of sources: Dryad collects data referenced in published papers across a broad disciplinary range, whereas EMBL-EBI collects only nucleotide sequence data. The sources of data may be national or international; but the orientation is to the subject community (the DANS data archive may be thought of as a network of subject-defined data centres).

[18] http://datadryad.org/ .
[19] See ODE Project (2011), *Ten Tales of Drivers and Barriers in Data Sharing*, p. 6-7; Schäfer, A., Pampel, H., Pfeiffenberger, H. et al. (2011), *Baseline Report on Drivers and Barriers in Data Sharing*, p. 17-19.
[20] See *Ten Tales*, p. 12-13; *Baseline Report*, p. 32-24.
[21] http://www.dans.knaw.nl/en/content/data-archive.

Although data centres often emerge from a local community and may be locally or nationally funded, their success is built on openness to the global community and commitment to working with other stakeholders in the community. This may take the form of agreeing and promoting common metadata and citation standards, or may extend to more collaborative relationships with peer data centres in order to create sustainable service networks and foster the adoption of common standards. For example, the International Nucleotide Sequence Collaboration (INSDC) between the DDBJ in Japan, GenBank in the US and EMBL-EBI in the UK, allows the three databases to exchange and synchronize their data daily, so that researchers can access up-to-date data and information from around the world.

Some data centres are also facing the challenge of enabling interdisciplinary use of data. The ICES data centre operates at a disciplinary intersection between physical oceanography and the biological disciplines of fisheries and marine ecology. On the one hand this has proven challenging, with each discipline having very distinct cultures of data sharing; but on the other hand it has forced the disciplines to find ways of working together, and to adopt common approaches to issues of data interdisciplinarity and standardisation.

Data centres such as ICES maintained by international partnerships are evidently well-established for the long term; but for many data centres funding is often provided at a local or national level from a single source or a small number of sources, and may be at risk from changes in economic and political priorities. The Arts and Humanities Data Service (AHDS) in the UK is one example of a data centre that was obliged to close in recent years because its funding was withdrawn. One of the great challenges for data centres is to find funding models that are sustainable over the long term.

Critical to the success of data centres has been their engagement with publishers to facilitate reciprocal linking between datasets and publications. Dryad and PANGAEA are two examples of data centres that have successfully implemented reciprocal relationships with publishers[22]. But even though there are evident synergies between publishers and data centres, ODE found evidence of very few successful relationships. There were indications in some instances of a lack of trust in commercial publishers on the part of data centres and a reluctance to work with them; while publishers that were willing to engage with data centres reported difficulties building viable partnerships, because they were unable to find sustainable well-used data centres that employed best practice in metadata standards and persistent identification.

[22] See Dryad's Joint Data Archiving Policy: http://datadryad.org/jdap; and for the PANGAEA-Elsevier partnership:
http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01434.

It should be remembered that, as the PARSE.Insight survey indicates, the 'big science' and disciplinary data centres may in fact account for only a minority share in the total aggregate of stored data, and this in turn may partly explain why publisher-data centre partnerships are currently at a very modest stage of evolution, and data sharing overall has been slow to evolve. It may be that far more data is actually stored in institutional data repositories which service the needs of local researchers, and act as storehouses especially for the long tail of smaller data sets that would not otherwise have a home. These data stores may not currently interact as well as they might with the discovery services that facilitate data sharing and reuse.

The institutional data repository can come in different forms: many institutions nowadays simply collect datasets alongside articles and other research outputs in a single institutional repository; but there is also evidence that some universities are beginning to adopt a total lifecycle approach to data management, and implementing systems that help researchers to manage data within research processes as well as to collect datasets in dedicated data archives. So-called Research Data Management (RDM) platforms embrace the entire data lifecycle in what might be thought of as a content management system for data, providing tools and workflows to automate data capture, transformation, curation and preservation processes, facilitate licensing, publication and citation, and enable discovery and access for reuse. It has been argued that research institutions are better placed to support researchers throughout the data lifecyle from the very start of a research project, in contrast to data centres that have traditionally been focused on curating and disseminating the end products of research.[23]

Research institutions may be motivated to collect the research data produced by their researchers for the same reasons they are motivated to collect primary research outputs: to preserve valuable assets, to enhance the prestige of the institution, to increase research impact, and to support local research management needs. Such data repositories can be relatively easily accommodated within existing infrastructure and service funding models, and can interact richly with local systems, affording institutions a low-risk, high-benefit data management solution.

But while local data repositories present an attractive solution for the capture of data outputs, they may be less effective as vehicles of effective data sharing. They are likely to contain a disparate collection of datasets in a variety of disciplines, which makes it hard to apply data management and curation procedures tailored to specific disciplinary needs; nor can institutions necessarily supply the subject expertise to provide rigorous quality assurance of data and appropriate

[23] See example Wilson, J.A. J., Fraser, M. A., Martinez-Uribe, L. et al. (2010), 'Developing infrastructure for research data management at the University of Oxford', in *Ariadne* 65. http://www.ariadne.ac.uk/issue65/wilson-et-al.

metadata. Disciplinary resources, on the other hand, may be better placed to leverage buy-in and loyalty from a much larger community and to develop the community-oriented services that enable their users to engage in data sharing. There is evidently a large amount of data stored in local repositories, and the challenge for both research institutions and disciplinary data centres will be to find ways of working together to maximise sharing and discovery of data wherever it is stored.

3.1.3 The way to successful data sharing

Data centres must know their communities and provide the resources and services to meet their needs. The services will follow from the needs of the community. Data specialists should have an intimate understanding of how researchers work and communicate within their community, and should be in constant dialogue with the community, so that their services can evolve as the requirements of users change.

Date centres should in addition be founded on clear mission statements and service definitions, which describe the 'designated community'[24], the services that are provided, and the standards that are applied. In some cases requirements may be minimal, largely focused on data storage at the end of a project; other kinds of research may require participation in data planning from the outset of the project, with expert assessment of data quality, even active involvement in defining and creating the dataset. Some data centres may provide a range of support and skills training, assisting users to develop data management plans, and providing consultancy services and training, or expert advice on ethical and legal issues.

Whatever the service the data centre provides, it must be trusted. Researchers must be confident that data is authoritative, correctly attributed, well-managed and secure. Data centres that have the trust of their communities adopt high standards of governance and provide a high quality of service. They adopt and promulgate standards that are accepted and used by the community for the preservation, description, publication and citation of data. Where relevant they seek and obtain accreditation to established standards, such as the Data Seal of Approval[25], the Deutsche Initiative fűr Netwerkinformation (DINI) Certificate[26],

[24] The 'Designated Community' is defined in the OAIS Reference Model as: 'An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time' (Consultative Committee for Space Data Systems (2012), *Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2*. http://public.ccsds.org/publications/archive/650x0b1.pdf).

[25] http://datasealofapproval.org.

[26] http://www.dini.de/dini-zertifikat/english/.

Trustworthy Repositories Audit and Certification (TRAC)[27] and ISO 16363:2012 (Space Data and Information Transfer Systems: Audit and Certification of Trustworthy Digital Repositories)[28].

Implicit in this is that the data centre has a sustainable business model, so that researchers can be confident a service will exist for the long term. Data centres that are over-reliant on project funding or that have a localized community base may be particularly at risk. They may have to adopt business models that exploit a number of strategies for covering costs, for example through partnerships with other service providers in their communities, through the provision of tailored content and consultancy services to client organisations, or through charging journals for data management.

Successful data centres are well-integrated within the larger networks of their research communities and adopt common metadata and citation standards that facilitate free circulation and interdisciplinary use of data. Data centres have successfully worked with publishers to implement standard citation practices, which enable cross-linking between research papers and datasests. Publishers and other service providers are also beginning to develop the service layer that interacts with data centres and makes the data they contain easier to discover, access and use: examples of recent and current initiatives include the prospective Registry of Research Data Repositories[29], the DataCite Metadata Store[30] which provides a data discovery interface, and Thomson Reuters' recently-launched Data Citation Index[31]. One of the key drivers for researchers to share data will be the ability to make datasets citable and to create services that track and measure data usage and citation. Data centres should maximise their potential to interact with discovery and other services.

Data centres can also develop their systems and processes to integrate with researcher workflows, for example by building tools that allow researchers to manage the entire data lifecycle, from the creation of a data management plan through data capture, transformation, quality assurance and documentation, to final publication and deposit of a dataset with assigned metadata and DOI. They should have user-friendly, low-threshold data publication or data deposit services, with clear guidelines and recommendations for best practice, e.g. in metadata and citation formats.

Research communities also need to find ways of accessing the large reserves of data that are held in institutional data repositories. If it becomes more common for institutions to offer their researchers local data management systems that

[27] http://www.crl.edu/PDF/trac.pdf.
[28] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510.
[29] http://www.re3data.org/.
[30] https://mds.datacite.org/.
[31] http://wokinfo.com/products_tools/multidisciplinary/dci/.

allow them to manage data throughout the research process, then it is important that the locally-held datasets can interact with data discovery and sharing services – for example, by using metadata standards common in the discipline, and by assigning DOIs to make datasets citable. It should be possible to automate interactions between data centres and discovery services and institutional data repositories that will maximise data sharing.

Data centres are repositories of valuable subject knowledge and expertise, and they need to find ways to transmit that intellectual capital to their communities. Data specialists can deliver training or support directly to researchers, and can advise institutions on how to teach the basics of good data management to postgraduate students. Data centres also need to be the vectors through which their knowledge and professional skills are transmitted to new generations of data specialists. They should support the development of data specialist qualifications and training, and professional career paths for data specialists and data librarians.

3.1.4 How to implement this?

In order to participate as effective components of a data sharing system, data centres can do a number of things:

- Data centres should know their designated community intimately and its data management needs, and define their service model accordingly. The nature of the offering may depend on the requirements of the community and how they are being met by other service providers. A data centre may offer a range of services, from basic preservation to support in the creation and management of data, expert advice on legal and ethical issues, training and consultancy. The designated community and services offered should be clearly defined in mission statements and policies.
- Data centres must be trusted authoritative members of their community. They should seek accreditation to appropriate standards such as the Data Seal of Approval or ISO 16363, and they should participate in the life of the research community to the fullest possible extent. Data centre specialists should have intimate knowledge of the research community they support; data centres should have established channels of communication with community representatives, attend major community events, and actively participate in the development of community standards.
- Data centres must have sustainable business models if they are to become trusted by the research community, by other data centres with which it may be advantageous to establish relationships, and by publishers with which they can establish linking relationships. Data centres may need to develop imaginative business models, and look to spread costs among the

community or to market tailored services and consultancy activities that become sources of additional revenue.

- Data centres should promote the adoption of common citation and metadata standards, and should follow best practise in assigning DOIs to datasets. They should publish clear data citation guidelines and recommendations and provide support to their stakeholders. Data centres should work together and with the research community to embed common standards for the description of datasets.
- Data centres should build relationships of trust with publishers. They should promote approved data citation practice using DOIs and develop relationships for reciprocal linking of publications and datasets. They should adopt open standards that facilitate the development of data discovery and interrogation tools by publishers and other service providers.
- Data centres should adopt a total data lifecycle approach to data management. They should build their systems and services in ways that integrate with researcher workflows and make it easy for researchers to manage their data throughout the lifecycle, from the early planning stages of a project to its completion and beyond. This requires intimate understanding of working practices and processes. Research Data Management platforms that apply content management systems approaches to data management afford useful examples of how data centres can provide a more integrated environment for researchers to manage their data.
- Data centres and other discovery services should work together with institutional data repositories to integrate them into the data sharing infrastructure. Local repositories might interact with disciplinary data centres to share and validate data and metadata to common standards, register dataset DOIs, and leverage the capacity of disciplinary services to make local datasets easier to discover and use.
- Data centres should work with the institutions where students and researchers are based to deliver their services. Data specialists can work with institutions to advise on training in data management skills for postgraduate students and early-career researchers; they can deliver data skills training, and provide dedicated support to students and researchers in their data management needs. Data centres may derive benefit from working with institutional libraries, which have a detailed understanding of their researchers' needs, and are well placed to mediate access to more specialised support services.
- Data centres should support the training and professional development of data specialists. Professional qualification might be achieved at a basic level within a Library and Information Studies course, or even at a more advanced level as a postgraduate qualification in its own right.

Professional development training courses could be provided; data centres could contribute to a broad community of data specialists with their own professional fora and community events.

3.2 Funders and Policy Makers

3.2.1 Introduction

Policy makers and funders exercise great influence in shaping the research agenda: not only by determining the fields of research that are explored, but also by prescribing the methods and practices employed by researchers, and the kinds of output that are produced. Funding agencies are instrumental in translating broad policy objectives into research programmes with defined areas of research and specific criteria for outputs and outcomes that must be met by researchers in order to fulfil their contractual requirements.

Policy makers and funders may be governmental or non-governmental. Public research policy in its broadest sense is formulated both nationally and internationally: by national governments and by supranational bodies such as the European Commission. Public policy makers are the democratically-accountable disbursers of public funds, and they are required to invest the funds they hold in trust for maximum return to society. In respect of research and research data this requires strategic and sustainable investment in:

- the productive activities themselves: data-productive research carried out by researchers and financed by research funding agencies;
- the infrastructure that allows the value of the data products to be realised: the internet-based research networks governed by common standards and interoperability protocols; the data repositories and other services that allow data to be stored, discovered and re-used; and the skills base that enables researchers to create and manage data effectively, and which provides the specialists who curate the data for the long-term and transmit data skills to future generations.

In the independent sector, large non-governmental agencies and charities, such as the Wellcome Trust, can function as both policy makers and funders of data-intensive research, and through their activities contribute to the production and sharing of research data. These activities are rarely formally aligned with governmental policies and practices, and while this undoubtedly contributes to the overall diversity and wealth of outcomes, there are risks of redundancy and strategic incoherence where a proliferation of research agendas has little reference to a common framework.

How all policy makers and funders can target their limited resources at so many points of the data sharing ecosystem for maximum social and economic benefit is an enormous question to which there are no simple answers. But two things are

clear: that investment at all these points is necessary to create a fully realised data sharing system; and that gaps and redundancies in investment can best be avoided by a co-ordinated approach on the part of all agencies – governmental and non-governmental – that make research policy and fund research activities.

3.2.2 Current situation among funders and policy makers

The political priority accorded to research data is reflected in national programmes and international partnerships that aim to develop education and research infrastructure and services. For example, in the UK JISC, which is funded by the UK higher education funding bodies and research councils, is addressing data policy and practice through its Managing Research Data programme 2011-2013.[32] In a broader international perspective, the Alliance for Permanent Access[33] has a membership which aims to develop a shared vision and framework for a sustainable organisational infrastructure for permanent access to scientific information, and is engaged in a variety of projects (including ODE) guided by a strategic plan; and the Knowledge Exchange, a European partnership [34] dedicated to making scholarly and scientific content openly available on the Internet, is engaged in a whole raft of activities whose objective is 'to make substantially more research data available and re-usable in the public domain'.[35]

These activities are exemplary and representative of numerous data-related policy programmes and initiatives ongoing throughout the European Research Area. Both the diversity of these activities and the synergies between them indicate how high on the agenda the question of data is for those who have a stake in its production and use, and how in many respects there is a convergence towards a common ground of policy and practice.

Many funders of research are also addressing data sharing challenges directly in their own data policies and in the conditions of research grants, which often require researchers to submit data management plans, and to undertake specified data preservation activities. A smaller number of funders make dedicated financial provision for data management and publication activities; and an even smaller number of these recognise data preservation and publication as valid research outputs in their own right.

It is evident from these examples and numerous others that have been encountered in the course of research undertaken by the ODE Project that the political will to realise the potential of data through sharing and re-use is being

[32] http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.asp
[33] http://www.alliancepermanentaccess.org/
[34] Knowledge Exchange partners are: Denmark's Electronic Research Library (DEFF), the German Research Foundation (DFG), JISC, and SURF in the Netherlands.
[35] http://www.knowledge-exchange.info/Default.aspx?ID=284

reflected in national and international research and education policies and in the policies of both state and independent research funders. The translation of this political will into concrete policies with specific data management requirements is taking place, although there is still a long way to go if data management activities are to become embedded in research workflows and the benefits of data sharing are to be realised on a wide scale.

There are also considerable variations between countries and between funders in both policy and practice. Policy makers and funders may encourage or they may mandate different activities, and the nature and force of legal, ethical and commercial restrictions on data use and publication varies widely between countries. These variations can be a cause of friction given the supranational nature of research and data use: many of those who expressed views about the state of data sharing today or recounted their experiences of data sharing highlighted the fragmented state of policy and funders' practices, and a widespread lack of clarity and consistency about what could and should be done with research data.

Although ODE encountered many positive examples of national policies and initiatives, a number of individual researchers and other stakeholders voiced a perception that there was a lack of national leadership and international co-ordination in this area. This may reflect the fact that different countries are at different stages of policy and legislative evolution in respect of data sharing. Differences in national legal frameworks, especially as they defined and protected national, commercial and personal interests, were often seen to impede the free traffic of research data; whereas greater harmonisation of national policies and legal structures was regarded as a means to accelerate the growth of data sharing. There emerged from the ODE evidence the sense of a need for high-level policy and legislative approaches to enable or even mandate different aspects of data sharing practice.

The other principal area where a more strategic approach was felt by many stakeholders to be lacking was in investment and funding policy. The direct funders of research were often felt to be constrained by short-term and project-focused business needs in conflict with the long-term perspective required for sustainable management of research data. This limitation was reflected in the relative under-investment in sustainable infrastructure and services: evidence indicated a number of research areas and disciplines lacking in well-developed data infrastructure, sustainably-funded high-quality data centres, and the skills and resources to support researchers in their data management activities.

It also often appeared to be the case that where funders did assert explicit data management requirements as conditions of their grants, the procedures to monitor and enforce compliance were lacking or insufficiently robust.

3.2.3 The way to successful data sharing

At the highest level, there is a need for national and international research data policies to work together in order to frame data preservation and sharing requirements on a common basis. It is on this common ground that an optimal balance can be achieved between the requirements of researchers for unrestricted access to data, and the needs of other stakeholders to control access to data in order to protect national, commercial, personal and other legally safeguarded interests.

But it is perhaps even more important for governmental funding agencies of countries and international organisations to create the conditions that encourage researchers to share and make use of data. Given the incentives and the opportunities to share data, researchers will do so, and where they encounter barriers, they will find solutions within their communities.

Funders can foster this culture of data sharing by applying consistent research grant policies, which provide standard definitions of data management activities and requirements, clarity as to what is encouraged and what is mandated, incentives to reward good data practice, and appropriate procedures for monitoring and policing compliance.

Governmental funders of research should be directed by national policy in strategic investment in infrastructure, services and skills; they should apply coherent, consistent and accountable data management policies through their research grants; and they should develop the understanding and business processes to support long-term preservation and sharing of data.

Governmental policy makers and funders should also reach out to their non-governmental counterparts: there are opportunities here for the state and independent sectors to agree a common framework within which they can develop complementary policies and find other ways to work together, for example, through co-funding infrastructure and services.

Policy makers and funders must adopt coherent strategic approaches to investment in sustainable data preservation and sharing infrastructure and services. Investing at this level requires an international perspective embracing both state and independent sectors. Data sharing is premised on the existence of a preservation infrastructure that is international and sustainable over the long term, and the investment strategy this involves, which requires multiple international partners to make an on-going commitment, is too often in conflict with the narrower funding horizons and results-oriented policies of national governments and funders.

3.2.4 How to implement this?

For policy makers and funders to create the conditions for a rich data sharing culture, activities need to be undertaken on a number of fronts:

- National and international public policy makers should adopt and promulgate clear and consistent messages about the value of research data and should place requirements on funders to put into effect funding policies and programmes that recognise data as primary research outputs and reward good data practice on the part of researchers.
- Structures of high-level public policy co-ordination should be provided, whether policy is cascaded within a regional political structure, from supranational to national levels, as in some instances through the European Union, or whether policy is agreed among partners through participatory mechanisms. This is an area in which the European Commission can take a lead.
- As different countries have differing legal approaches to the protection of national, commercial and personal interests in data, there should be some forum in which countries can negotiate a consensus about what data falls within the sharing domain, and what remains outside. There is a need for shared definitions, and commonly-accepted borders, and this is an area in which the European Commission can take a lead.
- Public policy makers and funders should direct their funding in a co-ordinated and complementary fashion to create data sharing infrastructure and services. This involves ensuring the provision of sustainable funding for data centres and other parts of the data sharing infrastructure, which may best be achieved through international partnerships.
- Policy must also be directed towards developing the skills base to support data management. This might be achieved, for example, through working to define professional training standards and accreditation for data specialists. Data management training might be incorporated into library and information studies; and professional data specialist qualifications could be taught and examined at universities.
- Public policy makers and funders should use their political and funding instruments to foster a data sharing culture. This can be done by a variety of means: placing requirements on research organisations to provide data skills training in postgraduate education; funding data management skills training for students and researchers in universities, data centres and libraries; and providing incentives in the academic reward system for good data practice, e.g. through certification and audit of research data against recognised standards of preservation, quality and transparency, and recognition of good data publication in national research assessments.

- All funders' research grant policies should be explicit about the value of research data as primary research outputs: this involves placing clear data management requirements on researchers as part of research grants, at a minimum requiring all researchers to submit a data management plan for approval by the funder, and preferably mandating the deposit of citable data sets in accredited data repositories; putting in place procedures to monitor compliance with data requirements and to assess the quality of data management; providing incentives for good data management; and evaluating data quality as part of the overall research assessment.
- Because funders are in constant conversation with researchers, they are well placed to act as a bridge between researchers and policy makers. They can report to policy makers on researchers' data needs that are not being met by existing infrastructure and services, and so inform ongoing development of high-level policy; and they can convey the messages of high-level policy in terms that speak directly to the interests of researchers.

3.3 Libraries

3.3.1 Introduction

Librarians have transferred their skills in collecting, organising, preserving and making available printed material to new types of collections, e.g. digital material. They have not only adapting their skills to the digital environment, they are also broadening their horizons by acting as publishing agents in helping researchers make their research outputs widely available. Within the context of their historical role as facilitators of knowledge sharing and creation, the emergence of data sharing represents both a continuation and a shift in the role of libraries.

This chapter focuses on libraries and librarians and their role in data management and data sharing. After a short summary of the current situation, we present results of the ODE project regarding successful data sharing cases and propose a number of ways in which libraries can contribute to the development of a rich data sharing culture.

3.3.2 Current situation

In recent years, libraries have begun to rethink their role and how to reposition themselves within the increasingly digital environment which is influencing the research process. Several studies[36] have interviewed librarians to explore their

[36] Such as RIN 2007: Researchers' Use of Academic Libraries and their Services. A report commissioned by the Research Information Network and the Consortium of Research Libraries (2007).
http://www.rin.ac.uk/system/files/attachments/Researchers-libraries-services-report.pdf or

role in the changing information landscape, and to ask whether data management is part of their role.

The results[37] of the ODE survey, which were validated through workshops and interviews, show that European librarians are being asked by their users to provide data management support services. Many of these libraries still have a gap to fill when it comes to meeting this demand, especially when compared to those libraries in other parts of the world that are leading the way in this field. Aware of this gap between demand and supply of data management support services, libraries are taking steps forward to develop their role in providing support for data sharing and reuse.

Libraries are willing to play an important role in data management, sharing and preservation. However, in many cases a discrepancy between this claim and current practice is usually found. Datasets are rarely part of the digital material currently stored by libraries (although this is changing) and many institutions do not have a digital preservation strategy put in place.

Support for finding data sets, which comprises traditional library skills such as metadata and cataloguing is already seen as part of the library's core role. However, librarians are aware that they need to develop skills that go beyond the scope of the competencies they have applied so far and are making efforts to develop these skills. Subject expertise as well as data curation and archiving skills, are required for successful data management.

One of the recurrent barriers that keeps libraries from offering more support for data management is the lack of funding. The PARSE.Insight survey[38] found that libraries in particular consider themselves responsible for data preservation. They would like to provide more services but the ODE findings show that they hesitate to enter this new field, because they have not had the resources to do so.

3.3.3 The way to successful data sharing

Through engagement in dialogue and the above mentioned survey, ODE has learned from librarians that they feel insufficiently prepared to support their researchers in managing their data. Although there is much expertise to build on, librarians perceive a need to develop skills in certain areas to address their new role in data management. Priority areas for skills development are data curation and archiving. Experience shows that subject expertise is highly

CLIR 2008: No Brief Candle: Reconceiving Research Libraries for the 21st Century. Council on Library and Information Resources (2008). http://www.clir.org/pubs/abstract/pub142abst.html

[37]Kotarski R, Reilly S, Schrimpf S, Smit E, Walshe K (2012). Report on best practices for citability of data and on evolving roles in scholarly communication. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-ReportBestPracticesCitabilityDataEvolvingRolesScholarlyCommunication.pdf

[38]http://www.parse-insight.eu/downloads/PARSE-insight_survey_questions_datamanagement.pdf

relevant too. ODE has learnt that many libraries are investing in developing these skills through professional training or are planning to do so.

Libraries are well positioned to predict future trends in scholarly communications. The vast majority of participants in the survey run by ODE believe that datasets will become separately citable items and will enrich scientific publications. Some best practice examples illustrating how to connect publications and the underlying datasets exist already as the data repositories PANGAEA[39] and Dryad[40] show. There are opportunities for libraries to support the linking of data and publications and to provide support for the citation of data sets.

Some university libraries already manage small data sets in their institutional repositories[41] or are involved in developing tools and infrastructure for data management within their institutions.[42] They can also work with institutional management to develop and help to implement appropriate policies for data sharing. As libraries already collaborate with researchers in order to make their publications available, they should also inform them about the benefits of data sharing.

3.3.4 How to implement this?

Drawing on the key findings of ODE on the current and future role of libraries in data management and sharing, the following activities should be undertaken by libraries to meet the demand for data sharing and managing services:

- A first step towards growth in the adoption of data sharing and data management practice is to improve communication with the other stakeholder groups. Talking to data centres, publishers and researches helps to make sure that standards, guidelines, support and training related to data and data management meet the community's needs.
- Collaboration to create an incentive system that gives credit for sharing data should be prioritised. Recommendations for best practice in citation need to be followed by the development of tools and services, such as citation metrics and bibliographic management tools. Librarians already utilise and provide training in bibliometrics and citation and work with other service providers to bring their experience to bear in the development of such tools.

[39] http://www.pangaea.de/
[40] http://datadryad.org/
[41] 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 http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf
[42] See e.g. the Data Management Rollout at Oxford (DaMaRO) Project, http://damaro.oucs.ox.ac.uk/

- The use of persistent identifiers must continue to be promoted, supported and implemented by libraries; both for enabling referencing of data sets as well as for finding data. Librarians could offer training for this and, if demanded by their users, even co-operate with data centres to bring their expertise on data management to the libraries.
- Librarians must ensure that they have the knowledge and ability to inform users about licensing and the conditions under which datasets can be reused. Knowledge about intellectual property rights attained by librarians in supporting researchers to make their publications available can be further extended to accomplish this task.
- In order to address the gap in skills, communication should be fostered within the community, together with researchers, data centres and publishers to agree about the type of skills that need to be developed to support data sharing. Based on this, professional training courses in data librarianship should enable librarians to acquire these skills.
- Libraries should engage in the development of an infrastructure for the long term preservation of data that ensures retrievability of these data via persistent identifiers as well as their understandability and usability.
- Policies are important, but funding is key if libraries are to develop the skills and have the resources in place. Activities such as the writing of data management plans is a good starting point for getting active in managing and sharing research data. Libraries should communicate the importance of data management to their funders and encourage them to include funding for data management in research funding. ODE has learned that the re-creation of data is more expensive than preserving existing data for the long term.
- Those libraries that were selected as a benchmark for excellence in our survey could serve as best practice examples for orientation. Libraries beginning to get involved in data management should share their experiences, both failures and success stories, for example when it comes to prioritising skills and developing appropriate training courses. Therefore, more international communication in this field is needed.

Libraries are aware of the fact that data sharing and management are becoming increasingly important. It has the potential to embed libraries in the research process. Libraries are ready to accept this challenge and to acquire new skills to help researchers in data management. ODE has identified the barriers that are preventing them in engaging in data sharing, but also elucidated drivers and enablers that should act as a potential starting point for libraries to become active within the field of research data management. By taking a multistakeholder approach, ODE has been able to explore the role of libraries in context, to explore synergies, and to identify areas of opportunity where libraries

can work together with other stakeholders to take the next steps in developing an infrastructure to ensure the success of data sharing.

Libraries should be working as a community to identify best practice examples of support service for data management and sharing. They should also be working collectively to engage with other stakeholders to define their roles and to contribute to the dialogue and developments surrounding data sharing.

The expectation is that although there is a gap between the demand and supply of data management support, libraries will work towards filling this gap and mobilise to define and develop the skills needed to do so. This means investment in continuing professional development, but also a new approach to the recruitment of librarians with expertise in research within specific domains.

3.4 Publishers

3.4.1 Introduction

Journals play a pivotal role in scholarly communication and many regard the research articles in these journals as the core kernels of the ever growing body of the official Record of Science. Increasingly, underlying research data is added to or made part of research articles reflecting the overall growth of data available in digital form in science and research, while voices emerge that data deserve to be marked as primary research output similar to official publications. As a result, editors and publishers of journals are experiencing a vastly growing demand from authors to include the underlying data to their papers.

For many publishers this has created a new challenge. Some have enthusiastically embraced the new opportunity and have found useful ways to add data to papers and to launch new data publications. Several have entered into successful collaborations with data repositories, mostly in the molecular biology areas, and are developing novel ways to present data within the context of the article. Others struggle with an overflow of over-sized supplementary files with data and seek the best ways to handle them: how to keep all these multi-format files stored, available and preserved for the future. Many editors and publishers are looking for new conventions in how to treat data in the context of articles..

A great number of initiatives have been started to ensure better integration of publications and research data. Their main aim is to enable research data to become part of the growing body of the official Record of Science and thus be treated as a first class research object. In the recent time span, consortia like DataCite and the World Data System have established themselves with the aim to ensure good data sharing in the partner-repositories and consistent adding of persistent identifiers for the data registries, while working committees like those of NISO/NFAIS and CoData work on best practice recommendations for matters

such as Supplementary Journal Information and Data Citation. The publishing community has been actively supporting and participating in these initiatives, aware of the importance of the topic.

3.4.2 The current situation in the publishing landscape

The amount of data connected to publications is growing tremendously. But even if this is the case, data are not an entirely new element in research publications. In fact it is difficult to find research publications, even the very old ones, where data do *not* play a role. The traditional scholarly article would normally refer to data that underpin the scientific claims made in the article, be it usually in a very aggregated form. Only in recent years, mostly so in the last decade, have authors started to add underlying data to their articles, in some case even the original, raw data.

For project ODE an inventory was made of the way in which data appears in relation to publications. This is best depicted in the following Data Publication Pyramid[43]:

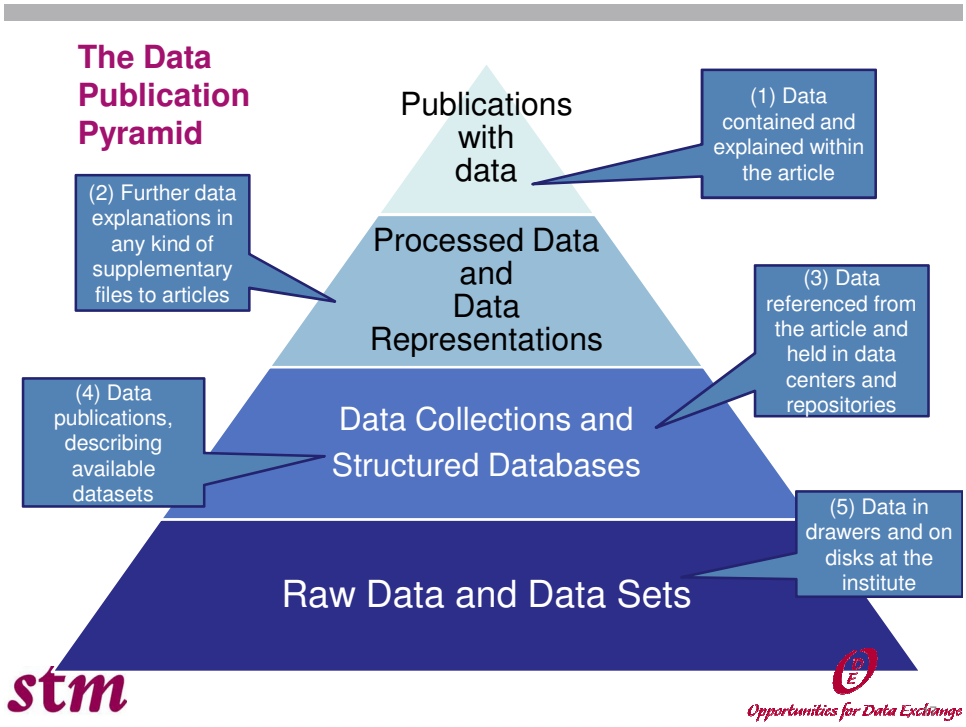

Illustration 1: Data Publication Pyramid

In essence, this illustration shows how the manifestation of data can broadly take 5 different forms. At the bottom of the pyramid, see (5) there is the category of raw data from a research project, usually not (yet) published but shared with colleagues of the project or research group awaiting further cleaning and

[43] ODE report: Integration of Data and Publications, http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+Report+on+Integration+of+Data+and+Publications

processing to allow for analysis and interpretation. On the top of the pyramid, see (1), we find the traditional research paper containing the most important data from the project in usually a highly aggregated form: a graph, a table or other summary of only the most important data points usually that underpin the claims and conclusions of the article.

Two intermediate categories have occurred strongly in the past decade: since the turn of the century an enormous increase has occurred in supplementary material for journal articles in digital form, very often containing data, see (2). More and more disciplines are entering some form of computational science and digital environments make the exchange of data really easy. Also: data repositories and structured databases have been established for certain disciplines where data is held to which publications refer and link, see (3). Good examples of these are Genbank and WorldProteinData bank in the life sciences and the Pangaea database in earth sciences. These data repositories have set up bilateral linking mechanisms widely used by the journals in these areas. New on the horizon here are the special data-publication journals, see (4), that have been launched recently which aim to describe datasets available in these repositories. Examples are ESSD[44](Earth Systems Science Data journal) and the Gigascience journal[45]. More data journals are expected and have been announced.

While all these new initiatives are encouraging and a good reflection of the rise of data in the publishing landscape, several developments become clear that call for better solutions. Recent research, also in the context of ODE and PARSE.Insight[46], makes clear that the amount of research data that remains unpublished could be as much as 75% of all data available. Via survey results from this project, we know that 40% of researchers have real problems to share their data with others. Reasons for this can be organisational, technical, legal (ownership, external sponsors), or even ethical (privacy-related).

At the same time, some of the journals that receive data submitted with articles suffer under the strain of getting far too much. The Editor of the journal Cell, Emilie Marcus, makes mention of journals supplements being used as data dumping grounds: "It had become a limitless bag of stuff."[47]. As an effect, Cell introduced strict limits on what authors can submit in supplementary files as underlying material. Another leading publication in its field, the Journal of NeuroScience announced in 2010 in an editorial[48] that they would stop accepting supplementary files as the burden on the peer review system became too large.

[44] Website ESSD: http://www.earth-system-science-data.net/
[45] The GigaScience journal: http://www.gigasciencejournal.com/
[46] PARSE.Insight: http://www.alliancepermanentaccess.org/index.php/community/current-projects/parse-insight/
[47] The Scientist: Supplemental or detrimental? - The Scientist - Magazine of the Life Sciences http://www.the-scientist.com/news/display/58027/#ixzz1NRVKAV6F)
[48] Link to editorial Jnl Neuroscience: http://www.jneurosci.org/content/30/32/10599.full

The Editor of the journal, John Maunsel, declared that "many (reviewers) feel that it is too much to ask them to also evaluate supplemental material that can be as extensive as the article itself."

A simple solution would be for editorial policies to require from authors that data is deposited in reliable and community endorsed data repositories, alongside clear policies that indicate how the data will be treated in the editorial process, as a growing number of journals have started to do. In fact, the success of databases like Genbank, WPDB and Pangaea is often attributed to their excellent collaboration with the leading journals in their field (Nature, Science, Cell) who have pushed their authors to incorporate accession numbers to these databases within their articles from the early beginning.

However, while open and community endorsed repositories would present an effective solution, many areas and disciplines lack such common data archives and the infrastructure around it.

As a result, the most likely short term reality of the data publications pyramid might look like this:

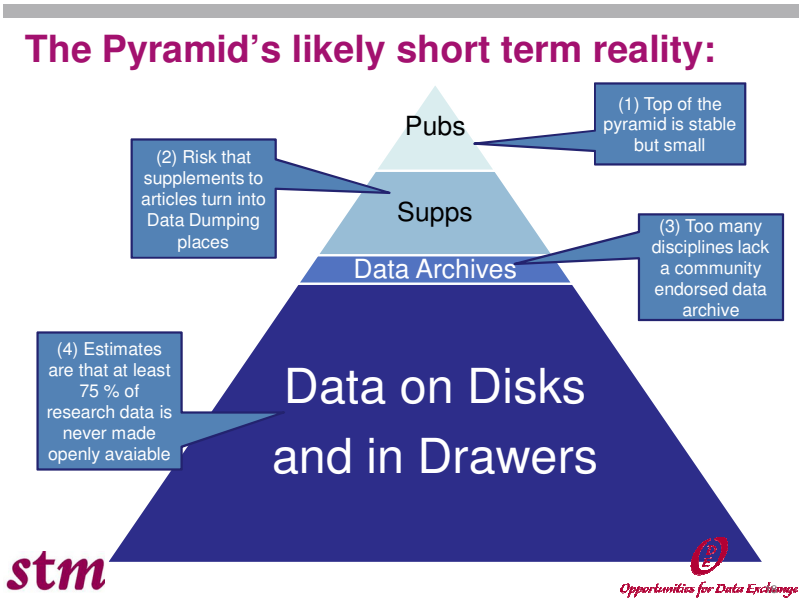

Illustration 2: The likely short term reality for the Data Publication Pyramid

There is a clear gap between this short term reality and the ideal data sharing culture. The majority of data remains hidden on disks and in drawers, very little data sets end up in proper data archives that ensure their future availability and accessibility, and journal supplements carry their own shortcomings in terms of findability or accessibility because of their format variety and access limitations.

3.4.3 The way to successful data sharing

In an ideal world, the data publication would look like this:

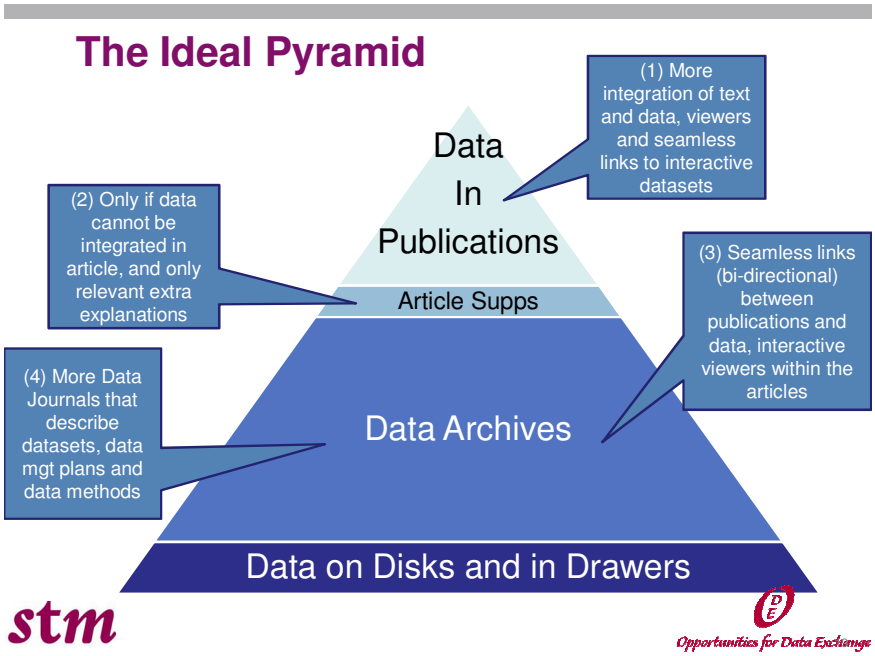

Illustration 3: The ideal Data Publication Pyramid

To create a richer data sharing culture, a number of obstacles need to be removed and a set of incentives need to be put in place. Key-players in this process are journals on the one hand and data archives on the other. Successful interaction between them, with journals encouraging authors to deposit their data in trustworthy Data Archives and Data Archives enabling bidirectional linking from and to publications, will help lift the integration of data and publications to a higher level. As a result, available data sets will be easier to find, to be interpreted and re-used. Similarly, publications will be enriched, offer deeper insight and, as many studies show, will attract more citations[49].

In this ideal situation, data within the article becomes more actionable and is presented in such way that users can easily dig interactively into the data itself, view it, play with it, all within the context of the article. In an ideal world, such integration of data within articles would make the appearance of additional material in articles supplements far less necessary. In that perspective, journal supplements may soon become exceptions rather than the rule and turn out to be something from the past and a clear remnant of the transition era of the first digital publications.

At the same time, Data Archives would be established as a core element in the research infrastructure across all scientific disciplines. Scholarly literature would seamlessly link to and from it, ensuring its integration in the scholarly record of science.

[49] Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

And as to the bottom of the pyramid: this would ideally turn into a fertile ground layer of ever new datasets emerging from research, some of it still on disks and in drawers, not ready yet to be published and awaiting further processing and clean-up. But hopefully, this would only be a small proportion of all available data, and only for a short time until the researchers have been able to validate and process the data properly for sharing and re-use by others.

3.4.4 How to implement this?

Our ODE reports make the following recommendations for a better integration between data and publications:

- Clearer editorial policies on the availability of underlying data: the publication of research results is universally a key-output measure for research projects. To foster a data sharing culture, editorial policies should indicate how they wish the data to be treated alongside the publication. For the success of initiatives like GenBank and WPDB it has been of enormous help that leading journals were supporting and often even requiring the deposit by authors of their data into the archive. Even in areas where commonly endorsed data repositories are absent, it will be useful if editorial policies include clear guidelines on the best way for authors to make their data available, to peer reviewers and to readers of their articles.
- Recommend reliable and trustworthy Data Archives to authors: for the future preservation and availability of data, trustworthy data archives are often a better place to store underlying research data than supplementary files to journal articles. In many cases, a journal website will offer this option as a service to the author, but in only a few cases the curation and preservation of the data is of a similar level as official data archives provide. In the fields where good and trustworthy Data Archives are well known, this is an obvious recommendation. For the areas where no such well-established repositories exist, editorial policies can indicate some of the criteria that authors can use when choosing a repository. Among these would be the registration of persistent identifiers like DOI's, bi-directional linking, and adoption of the emerging certification criteria for repositories.
- Enhance articles for better integration of underlying data: novel ways to integrate data into articles in an interactive way have been appearing and new apps are being supported by publishers to enhance data and text mining possibilities. More collaboration between publishers and the research community will stimulate the development of more workflow oriented research tools that tap into the available data underlying and integrated into publications.
- Ensure persistent identifiers and bi-directional linking: the emergence of DataCite is a very encouraging example for getting a persistent identifier

system in place for deposited data, based on DOIs. For the findability of data it is important that publications link to underlying data. The reverse links, from data to publications, are equally important, most certainly in an environment of more and more subject areas establishing common data archives. Organisations like Crossref[50] have made available applications that enable easy linking between publications and deposited data in a collaboration with DataCite.

- Endorse guidelines for proper citation of data: Initiatives like those of CoData, DataCite and NISO/ NFAIS[51] to define best practice guidelines for the citation of data can work as an enormous stimulus to more data citations and hence a proper acceptance of research data as a first class research object . These guidelines go beyond the recommended use of DOI's as persistent identifiers and also provide instructions where to place data citations within the article in a consistent manner.
- Launch and sponsor Data Journals: Data Journals are becoming more popular and several new titles are being launched and have been announced. We may expect more to come, while a growing number of traditional research journals are opening up to so-called data-publications. Clearly, these will demand an exemplary role in the way they present data and link to data kept elsewhere.
- Partner with reliable Data Archives for further integration of Data and Publications, including interactivity for re-use: findability and accessibility of data and associate representation information[52] serve one real purpose: that of re-use of data. Partnerships between publishers and data archives could focus on many more ways to make stored data truly interactive and re-usable.

3.5 Researchers

3.5.1 Introduction

Research data plays an increasingly important role in scientific and scholarly endeavour. Advances in information technology provide many options for sharing information and data, and can give a powerful boost to further research and additional discovery. This influences new habits in communication and collaboration among scientists. Existing norms of scientific behaviour are challenged by the ways of acquiring, preserving and storing vast data volumes. The diversity of types and quantities of research data from different disciplines does not ease the task of establishing those norms, nor does the recognition that these norms will change over time.

[50] Crossref on www.crossref.org and for Crosscite: http://crosscite.org/cn/
[51] http://www.niso.org/apps/group_public/download.php/8878/RP-15-201x%20Suppl_BWG_final_draft-rev.pdf
[52] Representation Information is defined in OAIS (ISO 14721) as the "information that maps a Data Object into more meaningful concepts."

In recent years a consensus has grown in science, and society generally, that primary research data from publicly funded research are a public good, produced in the public interest, and should consequently be made openly available - or with as few restrictions as possible - in a timely and responsible manner that does not harm the legitimate academic interest of their creators, and a range of other limitations up to national security.

3.5.2 Current situation for researchers

There is an on-going discussion in the scientific community on the challenges of data sharing. Data sharing, re-use and even preservation of data are far from common practice. Many scientists still pursue their research through the measured and predictable steps in which they communicate their thinking within relatively closed groups of colleagues; publish their findings, usually in peer reviewed journals; file their data and then move on. The main barriers to accessing research data that the project identified are:

- insufficient credit given to researchers for making research data available
- unclear trustworthiness of the data
- data usability
- pre-archive tasks and infrastructures
- lack of funding to develop and maintain the necessary infrastructures
- absence of a sustainable preservation infrastructure
- insufficient national/regional strategies/policies on data management
- warranted restrictions on openness related to commercial interests, personal information, safety and national security

Another aspect is a social and cultural dimension to data sharing, and in large part this is determined by the practices that have become established over many years in different communities. Research 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.

Traditional data sharing cultures are also being challenged by greater interdisciplinary communication, facilitated by internet technologies, and the emergence of distinct new and data-heavy disciplines, such as bioinformatics.

3.5.3 The way to successful data sharing

The ODE project has collected views from numerous researchers on the opportunities for data exchange, re-use and preservation. Successful examples, such as sharing data through the Worldwide Protein Data Bank[53], GenBank[54],

[53] http://www.wwpdb.org/

Pangaea[55] and Galaxy Zoo[56], clearly demonstrate that data sharing can provide enormous benefits. It is therefore very important for research data to be preserved and to remain permanently accessible.

ODE also learned about the effort needed for data sharing from its interviews. ODE talked to researchers handling different kinds of data with specific management requirements. Interviewees were able to speak from experience about 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. One recommendation which can be derived from the results is that data should be openly accessible as far as possible. From the researchers' perspective the drivers and benefits of a successful data sharing include the following issues:

- Peer visibility and increased respect achieved through publications and citation
- Research impact increases by citing data in publications
- Status, promotion and pay increase with career advancement
- Status conferring awards and honours
- Stronger data sharing culture enhances and accelerates research impact
- Preserving data for contributors to access later
- Re-usability of data
- Potential increased research funding
- The socio-economic impact of their research
- Changing attitudes to invest effort for possible long-term benefit
- Researchers need to acquire data skills at an early stage in their careers and benefit from on-going training and support. This would enhance their ability to adhere to good scientific practice and raise the impact of their work. Standardisation and interoperability increase effectiveness of data discovery and understanding, facilitate automated processing, and enable interdisciplinary connections and meta-analysis of data sets.
- Co-operation between researchers, funders and service providers, such as data centres, libraries and publishers is needed. More and more funding agencies require proper data management and open access to the data from supported projects. Clear standards, best practices and tools can help researchers to –engage in data sharing. Data centres and libraries need to offer advice and training on finding, preparing and managing data successfully.

[54] http://www.ncbi.nlm.nih.gov/genbank/
[55] http://www.pangaea.de/
[56] http://www.galaxyzoo.org

3.5.4 How to implement this?

The flow of information and ideas amongst researchers is fundamental to the discovery and innovation process. The recommendations of the ODE project outline what funders, libraries, data centres and researchers themselves can do in order to ensure full scientific and economic benefits are being derived through publicly funded research. For researchers to realize in full the benefits of data sharing a successful exploitation of new approaches will come from several activities which need to be undertaken on a number of fronts. According to the results of the ODE project the main changes to establish access to open research data are/can be summarised as follows:

- A shift away from a research culture where data is viewed as a private preserve
- Expanding the criteria used to evaluate research to give credit for useful data communication and novel ways of collaborating
- The development of common and clear standards for communicating data
- More metadata need to be recorded and made openly available to enable other researchers to understand the research and potential for re-use of the data.
- Data scientists/librarians are needed to manage and support the use of digital data in close cooperation with the researchers to create a new career at the service of science
- To ensure that researchers get support and training on data management.
- Improving their skills and understanding in data management. Training should begin in the institutions that train researchers, at the outset of postgraduate study at the latest, possibly even earlier. Researchers should participate in training programs to acquire and develop the skills they need to share and preserve data effectively.
- The development and use of new software tools to automate and simplify the creation and exploitation of datasets.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and preparing data for use by others.

Researchers should identify the best providers of data stewardship and work with them to establish the most useful data formats and descriptions (metadata) that enable their data to be discovered and re-used. Data management including proper data description and formatting must become an inherent research practice. Researchers should take on responsibility of the long term stewardship of their data. If capable providers do not exist or if their capacity is too limited, researchers should communicate with their research funders. Researchers should engage in community discussions about practices and policies and participate in community work to define and establish standards. They should 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. Funding agencies and other stakeholders such as libraries and data centres depend on the views the scientific community in particular about which data should be shared when, or where ethical or other concerns will prevent open data sharing: researchers need to support these bodies representing their communities. They need to help them formulate rules for appropriate data management plans and best practice for sharing that acknowledge the researcher's point of view.

4. CONCLUSIONS

The ODE project provided new insights into data sharing today. The research work done during the project allowed us to detect several drivers and barriers. Special emphasis was given to the role of the individual stakeholders (researchers, libraries, publishers, data centres and funders/policy makers) who also reported on examples and strategies to overcome some of these barriers. This is of particular important as society, policy makers and funders are increasingly demanding open permanent access to research data in Europe and beyond.

The results show that data sharing is fairly advanced in some disciplines, as for example in molecular biology. In many other disciplines data sharing is not yet established though. Researchers hesitate to share data for multifaceted reasons. The barriers are in parts of societal nature, pointing to a missing link to the incentive system in research which prevents researchers from expanding their focus from publications to research data. In addition, technical challenges are apparent, missing standards and interoperability in some particular disciplines but also across disciplines is hindering the advancement of data sharing. Funding for data sharing services and infrastructures, in particular on the long-term, is an important issue spanning all stakeholder groups.

ODE has learned that there are some conditions that can enable data sharing. Service providers like data centres and libraries for example are ready to play a certain role or expand their activities in data management and sharing. Publishers see the added value of data published alongside articles and foresee an adaption of their editorial policies or establish data journals. Researchers and funders view data as a research output in its own right and funding bodies require more and more data management plans specifying data preservation and access. A series of enablers can spread existing best practices, which thrive on collaborations within and across the individual groups. Successful data sharing needs to profit from synergies that arise from such collaborations where every stakeholder group contributes with its expertise, skills and experiences.

A key finding is that incentives for data sharing have to be developed. These have to be linked to the academic incentive system as well as to the research assessment schemes. The technical barriers to share data have to be reduced by simplifying data sharing workflows. Several stakeholders, from publishers or data centres to funders have to be involved in this process to address researchers' hesitation to manage their data and make them available. Researchers have to communicate their needs so new developed services are accessible. Developing a strategy with joined forces creates opportunities to get additional funding, which is especially needed to provide long-term data stewardship. In general, every

stakeholder group shall be aware that more co-operation is needed, and have the will to work together with other stakeholders to profit from their expertise.

In conclusion, the results of ODE provide a comprehensive overview of data sharing today and shed new light on the future challenge of data sharing. This resulted in detailed recommendations for the individual stakeholders involved in data sharing, but also overall topics that need to be addressed in order to develop data sharing in Europe and beyond, and can inform policies and approaches in Horizon2020.