Project acronym:	RECODE
Project title:	Policy RECommendations for Open access to research Data in Europe
Grant number:	321463
Programme:	Seventh Framework Programme for Science in Society
Objective:	SiS-2012.1.3.3-1: Scientific data: open access, dissemination, preservation and use
Contract type:	Co-ordination and Support Action
Start date of project:	01 February 2013
Duration:	24 months
Website:	www.recodeproject.eu

Draft Deliverable D4.1:
Institutional barriers and good practice solutions

Author(s):	Merel Noorman (eHumanities group, Royal Netherlands Academy of Arts and Sciences), Vasso Kalaitzi, Marina Angelaki, Victoria Tsoukala (National Documentation Centre, Greece), Peter Linde (Blekinge Institute of Technology), Thordis Sveinsdottir, Lada Price, and Bridgette Wessels (University of Sheffield).
Dissemination level:	Public
Deliverable type:	Final
Version:	1
Submission date:	2 September 2014 (Due date 31 August 2014)

Table of Contents

LIST OF ACRONYMS

ACRL - Association of College & Research Libraries
AHRC - Arts and Humanities Research Council
ANDS – Australian National Data Service
APARSEN – Alliance for Permanent Access
ARL - Association of Research Libraries
ATLAS - particle physics experiment at the Large Hadron Collider at CERN
AUP - Amsterdam University Press
BBSRC - Biotechnology and Biological Sciences Research Council
CDL – California Digital Library
CERN - European Organization for Nuclear Research
CESSDA - Council of European Social Science Data Archives
COAR – Confederation of Open Access Repositories
COPD - Chronic Obstructive Pulmonary Disease
DANS - The Data Archiving and Networking Services institute in the Netherlands
DARIAH - DigitAl Research Infrastructure for the Arts and Humanities
DCC – Digital Curation Center
DNFR - The Danish National Research Foundation
DOAJ – Directory of Open Access Journals
DOI - Digital Object Identifiers
DPC – Digital Preservation Coalition
DRIVER – Digital Repository Infrastructure Vision for European Research
DRIVER-II - Digital Repository Infrastructure Vision for European Research II
DSA – Data Seal of Approval
EC - European Commission
EDINA - is the JISC-designated national data centre at the University of *Edinburgh*
EIFL – Electronic Information for Libraries
ESF - European Science Foundation
ESRC - Economic and Social Research Council, UK
EU - European Union
EU JRC or JRC – European Union Joint Research Centre
EUDAT – European Data Infrastructure
EuroCRIS – The European Organisation for International Research Information
EvA - Emphysema versus Airways disease
FNRS - Fonds de la Recherche Scientifique, Belgium
FOIA - Freedom of Information Act
FOSTER - Facilitate Open Science Training for European Research
FWF - Austrian Science Fund
GEOSS - Global Earth Observation System of Systems
GIS – Geographic Information Systems
HE – Higher Education
HEI - Higher Education Institutes
ICORDI – International Collaboration on Research Data Infrastructure
ICES- International Council for the Exploration of the Sea
ICPSR - Inter-university Consortium for Political and Social Research
ICT - Information and Communication Technology
IDMB - Institutional Data Management Blueprint Project
IFLA - International Federation of Library Associations
IGO – Intergovernmental Organisations
INSPIRE - Infrastructure for Spatial Information in the European Community

IOC/IODE- Intergovernmental Oceanographic Commission/ International Oceanographic Data and Information Exchange
IPR - Intellectual Property Rights
IT - Information Technology
JISC – Joint Information Systems Committee
KE – Knowledge Exchange
LERU - The League of European Research Universities
LHC - Large Hadron Collider
LIBER - Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries
LIS – Library and Information Sciences
NEH - National Endowment for the Humanities
NERC - Natural and Environmental Research Council
NODCs - National Oceanographic Data Centres
OA – Open Access
OAPEN - Open Access Publishing in European Networks
OAR - Open Access Repositories
ODE – Opportunities for Data Exchange
OECD - Organisation for Economic Co-operation and Development
OpenAIREplus - 2nd Generation of Open Access Infrastructure for Research in Europe
OpenDOAR - The Directory of Open Access Repositories
PARSE – Permanent Access to the Records of Science in Europe
PPPA - Particle Physics and Particle Astrophysics
RDA – Research Data Alliance
RDM – Research Data Management
RECODE - Policy RECommendations for Open access to research Data in Europe
RCUK - Research Councils United Kingdom
SHERPA - Securing a Hybrid Environment for Research Preservation and Access
SHERPA/JULIET - Securing a Hybrid Environment for Research Preservation and Access (Research funders' open access policies)
SHERPA/ROMEO - Securing a Hybrid Environment for Research Preservation and Access (Publisher copyright policies & self-archiving)
SiS - Science in Society
SPARC - Scholarly Publishing and Academic Resources Coalition
SSH - Socioeconomic Sciences and Humanities
SURF - collaborative organisation for ICT in Dutch higher education and research
UCL - University College London
UK – United Kingdom
UKDA - UK Data Archive
USFD – University of Sheffield
VPH - Virtual Physiological Human

LIST OF FIGURES AND TABLES

EXECUTIVE SUMMARY

Open access to research data provides many benefits to science and society, but as the open access trend grows it becomes increasingly clear that providing unrestricted access to research data is not inherently a "good thing", and it is certainly not easy to achieve. The RECODE project looks at the grand challenges associated with open access and data preservation and dissemination, including technological and infrastructural, legal and ethical, and institutional and policy issues. In particular, it seeks to understand and use the fragmentation between and within disciplines in order to address these challenges. The aim is to produce policy recommendations for open access to research data, based on existing good practice.

In this fourth RECODE deliverable we focus on the challenges faced by institutions, such as archives, libraries, universities, data centres and funding bodies, in making open access to research data possible. Policy makers and the scientific community expect these institutions to play an important role in creating and funding data sharing infrastructures and stimulating and assisting researchers to make their research material public. They look towards these institutions to curate and preserve information, and provide guidance to researchers in managing their data.

Based on an initial review of the literature we identified four key challenges:

- financial support;
- evaluating and maintaining the quality, value and trustworthiness of research data;
- training researchers and other relevant stakeholders;
- creating awareness about the opportunities and limitations of open research data.

In this report we analyse these four challenges and how they can be addressed. More specifically we explore current strategies, the remaining barriers and possible solutions for overcoming these barriers. This analysis serves as the basis to formulate policy recommendations that support institutions in addressing the challenges of making more research data openly accessible.

To guide the analysis of the challenges, we have used the following case studies, also featured in previous RECODE deliverables, from five different scientific disciplines:

- Particle physics and high energy computing
- Health sciences
- Bioengineering
- Earth Sciences;
- Archaeology

Each case illustrates some of the grand challenges identified by the RECODE project and provides an access point into a network of institutions that collaborate to make certain kinds of research data accessible to various stakeholders. The cases studies also offer an interdisciplinary grounding that helps us to address disciplinary fragmentation in the area of open access to research data as well as maintain awareness of discipline-specific issues and practices. Moreover, we have used them to identify policy gaps and evaluate good practice solutions that will contribute to an inclusive and participatory development of policy recommendations.

The research approach has been two-fold. We first conducted a review of policy documents, reports, scholarly literature, other relevant documents and websites to provide an overview of current institutional approaches to making open research data possible and the gaps between these approaches and practice. We supplemented this review with 15 interviews with key individuals from each of the case studies, including data centre managers, project coordinators and division managers. To validate and fine-tune the results of this two-step analysis, we organized a one-day workshop with representatives from different stakeholder groups.

Chapter Two discusses the challenge of organizing funding for open access to research data. It shows that various funding bodies have taken up this challenges and are making increasingly more resources available to both researchers as well as institutions that are developing the required infrastructure. Research institutes, data centres and libraries are also directing funds towards building data infrastructures and organizing support for researchers and other members of staff. Moreover, institutions are developing new business models to recover the cost of publishing and maintaining open data based, for instance, on memberships fees or collaborations with other institutions. Nevertheless, the analysis in this Chapter also demonstrates that there are still barriers to overcome in order to secure sustainable open access to a wide range of research data. Not all funders and research producing institutions make funds available or aware of the costs involved in data curation. Another barrier is that current funding models are generally project-based and there are few institutions that provide preservation and curation services for the long-term. In addition, few institutions have sustainability strategies or policies that take into account growing volumes of data and the additional costs they will generate to keep them findable, accessible, and reusable. Moreover, the distribution of responsibilities is yet to be clearly established.

In Chapter Three we turn to the challenge of maintaining and evaluating the quality and integrity of open research data. Ensuring the quality of research data is a prerequisite to achieving the promises of open access to research data. In many disciplines, formal and informal mechanisms are already in place to check the quality of digital research data produced. Research communities may perform several review processes, manually and automatically validating data at various stages in the data life cycle. Several institutional stakeholders play an active role in these processes, including data repositories and centres, consortia, and publishers. The analysis in this Chapter shows that institutions have focused primarily on developing strategies to ensure the technical quality of data deposited (e.g. are the correct formats used, is the metadata complete, etc.). Less effort has gone into establishing review practice that focus on the scientific value of data, partly because it is a time consuming and difficult task. An important barrier that has to be overcome in order to move forward is the lack of incentives for researchers to engage in data review processes. The Chapter also showed that the long-term perspective has yet to be further developed. Issues such as how to deal with increasing volumes of heterogeneous data or how to deal with data selection and retention have been less of a priority for many institutions. The distribution of responsibility between various stakeholders is also an area that requires further attention. In this Chapter we highlight some recent developments that offer promising solutions, including data journals and new mechanisms to assist researchers in evaluating openly accessible data.

We look at the challenge of educating and training researchers, data scientists and other professional staff in Chapter Four. Producing, curating, evaluating and using open research data require considerable skills and expertise that have to be acquired and maintained. As the

Chapter showed, several institutions have developed strategies for educating and training researchers, librarians, information and data scientists and other professionals, building on existing and emerging digital data management practices. Libraries, data repositories, data centres and dedicated organizations, in particular, play an important part in offering workshop, training materials and other kinds of support. Nevertheless, several barriers have yet to be overcome, including distributing responsibility between stakeholders, engaging researchers and bridging the skill gap in libraries and data centres. Moreover, data management and curation skills have to be better and more commonly embedded in post-graduate education and new curricula and professional qualifications must be developed. All the different stakeholders with their organizations will need to cooperate to overcome these barriers, as they are multiple and complex.

Chapter Five looks at the challenge of creating awareness about the opportunities and limitations of open research data. As we have seen in previous RECODE reports, some of the concerns researchers as well as institutions have about sharing data are based on a partial understanding of what open access entails and what the possibilities and risks are, because technical skills and knowledge are lacking or because there are few good examples available. This Chapter discusses some of the initiatives that institutions have taken to promote and advocate open access within their own organizations and among research communities. Libraries, in particular, often consider it their responsibility to encourage researchers and university departments to make their data openly accessible. In addition, there are a number of professional organisations that play an active role in creating more awareness about open access. Yet, our review of the literature and the feedback during the interviews also indicate that creating awareness is not a priority in many institutions. Moreover, top-down approaches and advocacy may have adverse effects and it can be difficult to reach particular stakeholders. Different stakeholders have different needs and interests, which require tailored approaches to creating awareness. Another key barrier is the lack of incentives for researchers and institutions to take an interest in open research data. Data management mandates as well as awards and professional recognition for open research data contribute to creating incentives, but institutions will have to work together to create an environment in which the various stakeholders can discuss what open access to research data should look like for their particular area of interest and what that would entail for them.

In the final Chapter we draw together the analyses in the preceding Chapters and provide a set of recommendations. Our main observation from the review, the interviews and the workshop is that open access to research data is still in the earlier stages of development and solving the harder problems, such as funding long-term preservation of data, evaluating the scientific quality of data or getting the more reluctant researchers to experiment with open access is put on hold. The early stage development also means that various institutions acquire and are still adjusting to new roles and responsibilities, as they begin to offer data services, establish infrastructures and issue policies. In developing data management policies and services, institutions will need to consider how to give shape to these responsibilities. The analyses of the four challenges also suggest that in terms of financial resources, as well as knowledge and expertise, institutions will have a hard time addressing some of the challenges on their own. They will have to engage in collaborative efforts to develop data repositories, data management services, training programmes, etc. Another observation pertains to the value and role of data management in scientific practices. The awareness about open research data and the incentives to make data open will increase when research communities start valuing data produced as much as they value publications, and when research institutions, universities, funding bodies and scholarly societies start evaluating and rewarding researchers and research groups based on their data management efforts. The final observation is that a primarily top-down and centralized move towards open research data will only be effective to a certain extent. In order to have a vibrant open research ecosystem, institutions will have to acknowledge disciplinary heterogeneity and autonomy.

Taking these observations into account and based on the feedback we have received during the fourth RECODE workshop, we make the following recommendations:

- *Where possible, institute data management mandates and policies with open research data as the default and clear lines of responsibility, while ensuring that the required resources are available.*

- *Stimulate and ensure compliance with mandates and policies and make data practices part of the evaluation and reward systems*

- *Create incentives for researchers to publish their data and make use of available open research data.*

- *Create room for innovative ideas and bottom-up initiatives to further develop data management services and sustainable business models.*

- *Start planning for long-term preservation and curation of open research data.*

- *Pursue collaborations between and within institutions.*

- *Develop strategies that support the evaluation of the quality of data and data repositories both in terms of technical quality as well as scientific value.*

- *Create environments that stimulate open access and provide support and training for researchers and other relevant staff in their specific practices.*

1 INTRODUCTION

The emphasis on free and unrestricted online sharing of research data in the European Commission's plans for European research reflects a more general consensus about the value of open research data. Open access (OA) to research data provides many benefits to science and society, as it offers scientists a wider range of data to use for re-analysis, comparison, integration and testing. It can contribute to the quality and integrity of scientific practices, as it increases transparency and accountability. It can also improve the way science and scientific data can be used to achieve social goals, and thus increase the value of the contribution that science makes to society. Moreover, there is a strong belief that open access to research data will be beneficial to innovation and economic growth.[1]

However, as the OA trend grows it becomes increasingly clear that providing unrestricted access to research data is not inherently a "good thing", and it is certainly not easy to achieve. The benefits of open access are contingent on certain social, cultural, economic and infrastructural conditions. Thus, a culture of freely sharing data may have evolved, seemingly organically and from the bottom up for certain disciplines, but in other disciplines open access presents challenges that are not easily overcome. The sensitive nature of the research or the rights and interests of human research subjects, for instance, may form a barrier for researchers to share their data. Researchers and research institutes may lack the resources to prepare data sets and curate them for longer periods of time. The research landscape is made up of a wide variety of academic disciplines and sub-disciplines that all have their own practices and standards, shaped by the specific characteristics of their research topics. What works well in one field might not be easily applicable in other.

The RECODE project looks at the grand challenges associated with open access and data preservation and dissemination, including technological and infrastructural, legal and ethical, and institutional and policy issues. In particular, it seeks to understand and use the fragmentation between and within disciplines in order to address these challenges. The aim is to produce policy recommendations for open access to research data, based on existing good practice. In previous RECODE reports we surveyed the motivations, drivers and barriers of the various stakeholders and we addressed the technical and infrastructural challenges as well as the ethical and legal issues.[2]

In this fourth RECODE report we focus on the challenges faced by institutions in making open access to research data possible, in particular regarding funding, the quality and integrity of research data, training and creating awareness. We examine each of these issues in order to arrive at policy recommendations that provide support for institutions.

[1] Sveinsdottir, Thordis, Bridgette Wessels, Rod Smallwood, Peter Linde, Vasso Kalaitzi and Victoria Tsoukala, *Stakeholders Values and Ecosystems*, D1.1 RECODE Project, 30 September 2013

[2] Sveinsdottir et al. 2013; Bigagli, Lorenzo, Thordis Sveinsdottir, Bridgette Wessels, Rod Smallwood, Peter

[2] Sveinsdottir et al. 2013; Bigagli, Lorenzo, Thordis Sveinsdottir, Bridgette Wessels, Rod Smallwood, Peter Linde and Jeroen Sondervan, *Infrastructure and technology challenges*, D2.1 RECODE project, 31 March 2014; Finn, Rachel, Kush Wadhwa, Mark Taylor, Thordis Sveinsdottir, Merel Noorman and Jeroen Sondervan, *Legal and ethical issues in open access and data dissemination and preservation*, D3.1 RECODE project, 30 April 2014.

1.1 Institutions and open access

In this report we look at how institutions, such as archives, libraries, universities, data centres and funding bodies, can enable and support the further development of open access to research data in Europe. Policy makers and the scientific community expect these institutions to play an important role in creating and funding data sharing infrastructures and stimulating and assisting researchers to make their research material public[3]. They look towards these institutions to curate and preserve information, and provide guidance to researchers in managing their data. The first RECODE report showed that institutions are increasingly aware of the practical issues and are beginning to create an overall framework of data management and curation. They recognize the differences between disciplines and within disciplines when it comes to the generation, preservation and use of data. They understand the value of open research data, but also see the costs involved in making data public. However, for many institutions this is a relatively new and unchartered area and most institutions have only taken initial steps in exploring and giving shape to their new roles and responsibilities.

1.1.1 Open access to research data

Institutions first need to consider what open access to research data entails. Various definitions have been proposed for open access to research data, each emphasizing different aspects. Some definitions refer to the absence of barriers that restrict access to data, such as pay walls and copyright. The European Commission, for instance, defines "open access" as "free internet access to and use of publicly funded scientific publications and data".[4] Others, such as the Royal Society in the UK, stress that open research data should not only be freely accessible, but should also be easy to access and use. It proposes a more specific definition for "open data" as referring to data that is accessible, usable, assessable, and able to be evaluated.[5]

Similarly, there are many definitions for research data and it is hard to succinctly circumscribe which data can and should be made publicly available. Within and between the different disciplines there is a huge variety in what researchers regard as data. They may make distinctions between raw and processed data, or between field notes and observational data. Moreover, data are meaningless without knowledge of the context. To properly interpret the data, knowledge of the conditions under which the data are generated is required. Thus metadata are essential in providing open access to data that are re-usable and interpretable. Even then, some researchers would argue that data without models or software code are useless. Other researchers, for instance in the social sciences, argue that the contexts in which they generated their data are almost impossible to capture in metadata because they involve considerable tacit knowledge.

[3] European Commission. *Online survey on scientific information in the digital age*, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

[4] European Commission, *Commission Recommendation on access to and preservation of scientific information,* C(2012) 4890 final, Brussels, 17 July 2012, p.13. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[5] The Royal Society, *Science as an open Enterprise*, London, 2012.
http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

In the RECODE project we started from the definition provided in the Berlin Declaration on Open Knowledge in the Sciences and Humanities. It defines open access as "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community".[6] Sources of knowledge encompass: original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material. The Declaration also presents two key criteria for a contribution to be open access: 1) the authors and rights holders must grant users free access to the materials including a license to copy, use, distribute, and display material subject to proper attribution of authorship and responsible use; 2) a version of the work should be in an appropriate standard format and submitted in an online repository with suitable technical standards that seeks to enable open access, unrestricted distribution, interoperability and long-term archiving.[7] This definition, similar to the Royal Society's understanding of open data, makes it clear that open access is about more than just placing a table with numbers on a website. Others should be able to find, interpret and re-use data.

The ideal of open access as presented in some of the definitions above is hard, if not impossible, to realize for all the different disciplines, as many disciplines have to deal with constraints on the publication of data, such as protecting the privacy of research subjects or the insurmountable costs and effort involved in making huge volumes of data accessible for longer periods of time. Most policy documents therefore are careful to note that not all data can or should be made open. Throughout the project we have therefore also looked more generally at efforts to limit the restrictions on and difficulties of data sharing as much as possible.

The ambition to make more research data openly available is not straightforward. It requires considerable work to make data easy to access, use and evaluate. In order for data to be accessible, it has to go through several phases, including ingestion, storing and providing access. The Digital Curation Centre has provided a useful illustration of these different phases (see Figure 1.1). Each phase involves a range of different activities. They have to be digitally generated or converted into standardized and machine-readable formats and metadata have to be added as part of the ingestion and selection phase; the data and the metadata have to be reviewed and checked for inconsistencies, noise or errors and if possible linked to other data sets. To store data, and make them accessible, infrastructures must be funded, built and maintained. Tools need to be developed to make the data searchable and re-usable. To preserve the quality of data, multiple versions of data sets have to be managed and occasionally migrated to other and new technological platforms. Additionally, making research data available for unrestricted use requires that researchers and data managers require the motivation, skills and support to publish their data. Copyright issues and informed consent must also be clarified and managed. Strategies have to be developed to evaluate the quality of data sets, models or code and to measure their impact. Once the data is online different levels of access may have to be managed and the security of the data maintained.

The various activities involved in providing open access to research data take place within networks of institutions. In these networks, institutions will perform varying and multiple roles and functions, often in collaboration with other institutions.

[6] Max Planck Society, *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003. http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf

[7] Ibid.

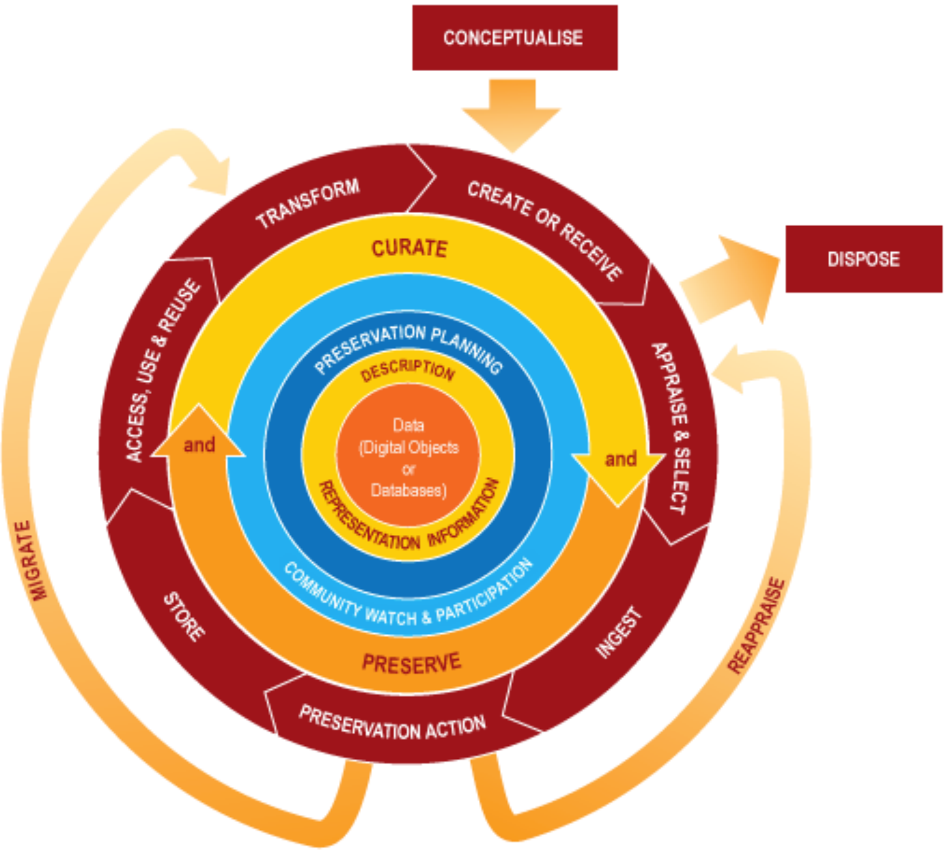

Figure 1-1: Data Curation Life Cycle Model[8]

1.1.2 Institutions

In this report, the term institution refers broadly to formal organisations involved in scientific and scholarly data practices. This includes, but is not limited to, (digital) libraries, universities, research institutes, national and international data centres or archives, (data) publishers, funding bodies, research councils, scholarly societies and consortia. These can be organizations with a long history, such as university libraries or archives, or newly established data centres, data publishers or collaborations between institutes. Some institutions may be part of other larger institutions and not able to act entirely independently from their host institutions, such as university libraries or subject data centres. Other institutions, including data centres and research institutes, often have multiple sources of funding and are managed and maintained by international consortia of universities, research councils and other institutes. Institutions that make public sector data available, such National libraries, National Weather Services and National Offices for Statistics, are also relevant for the scientific community and increasingly more so as they are making their data openly accessible in accordance with recent Open Data policies. Each of the mentioned institutions plays different roles, in accordance with its background and specific expertise, interests and goals.

[8] Digital Curation Centre, *DCC Curation Lifecycle Model*, 2004-2012. http://www.dcc.ac.uk/resources/curation-lifecycle-model

It is hard to draw clear lines around the various institutions in terms of what they are and do and with regard to open access to research data. A data centre's primary activity, for instance, may be the curation of large data sets, but it may also have a secondary role as a publisher (see for example the UK Data Centre). Some institutions focus primarily on archiving and preserving data for the long term, whereas others are concerned with providing more immediate access to data and facilitating their (re-)use. They may have larger or smaller repositories or digital libraries and provide a variety services for depositing, curating and sharing data. Some institutions do not store data themselves but play an important role in advocating open access, funding resources or setting guidelines or mandates for data management.

To bring into view the various different stakeholders and analyse their roles or functions with regard to open research data, we introduced a functional taxonomy in the first RECODE report.[9] The original taxonomy consists of five functions that stakeholders can fulfil. As depicted in Figure 1.2, the categories are not mutually exclusive and at any given time stakeholders may operate and interact from within different functional categories. In this taxonomy stakeholders have one primary function (PF) and can have several secondary functions (SF). The five basic functions in the Open Access ecosystem are: 1) Funders & Initiators (F&I), 2) Creators (Cr), 3) Disseminators (D), 4) Curators (Cu) and 5) Users (U). Data creators can, thus, also act as users, disseminators and/or curators within the open data ecosystem. For this report we include an additional sixth function: Enablers (E). This function encompasses outreach and advocacy activities that are performed to enable researchers and institutions to share and publish data, including advocacy, information and guidelines provision, and training.

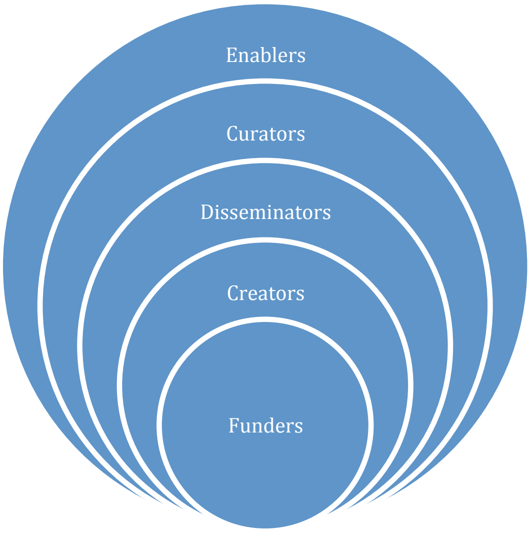

Figure 1-2: The RECODE stakeholder functions[10]

The taxonomy can be used to give some insight into the various existing roles of institutions. Table 1.1 provides an initial overview of some of the key types of institutions. Each institution is listed as having either a primary function (PF) or a secondary function (SF).

[9] Sveinsdottir, et al., op. cit., 2013.
[10] Adapted from Sveinsdottir, et al., op. cit., 2013.

This overview is primarily intended to serve as heuristic to identify relevant institutes in the case studies, rather than to provide a precise and all-encompassing taxonomy.

Institution	**Primary function**	**Secondary functions**
Research councils	F&I	E, D
Foundations	F&I	U, E, D
Funding bodies	F&I	E
University/Academy	Cr	U, F, E, D
Research Institute	Cr	U, F, E, D
Scholarly Society	Cr	E, D
Research consortia	Cr	Cu, D, E, U
Data Centres	Cu	Cr, E
University Libraries	Cu	D, E, F, U
Publishers	D	E, Cu, U
Professional associations	E	D

Table 1-1: Types of Institutions

1.1.3 New roles and responsibilities

As various reports, roadmaps and policy documents have pointed out, enabling open access to research data and successfully exploiting the various approaches, requires changes in research cultures, infrastructures and funding models.[11] In the discussions, key roles and responsibilities are assigned to various institutions in making these changes happen.

Universities and research institutions are assigned a multitude of different responsibilities when it comes to operationalizing open access to research data. Institutions should set internal policies for data management and these should "encourage a data sharing culture, facilitate re-use of data and enable researchers to share their data".[12] The Royal Society, for instance, recommends that:

> *Universities and research institutes should play a major role in supporting an open data culture [...] Learned societies, academies, and professional bodies should promote the priorities of open science amongst their members and seek to secure financially sustainable open access to journal articles. They should explore how enhanced data management could benefit their constituency and how habit might need to change to achieve this.*[13]

The Royal Society sees a role for universities and research institutes in adopting open data as a default position; evaluating researchers also on their data sharing; developing strategies and

[11] The Royal Society, op. cit., 2012; High level Expert Group on Scientific Data, op. cit., 2010; Organization for Economic Cooperation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD, Paris, 2007. http:/www.oecd.org/dataoecd/9/61/38500813.pdf

[12] McAllister, D., Implementing an Effective Data Sharing Policy: A UK Public Funder's Perspective, BBSRC, no date. http://eidcsr.oucs.ox.ac.uk/docs/EIDCSRWorkshop29-03-10%20-%20DavidMcAllister.pdf

[13] The Royal Society, op. cit. 2012, p. 10

policies for taking care of their own knowledge resources and offering services to support researchers in the management of their data.[14] They should also provide funds to encourage the culture change and for the sustainability of any resources and infrastructure developed.

Libraries have also become increasingly more important in research data management.[15] It is widely acknowledged that the role of research libraries is changing and data management is one area in which they can expand their activities.[16] LIBER (Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries) considers libraries to be institutions that are perfectly positioned for data curation and preservation, while also providing training and assistance to researchers when it comes to managing their data and searching for data for re-use. Moreover, libraries are often seen as the executors of institutional strategies and policies and many of the responsibilities assigned to research institution or universities are realised through the university library or data repository. Based on a review of the literature, Cox and Pinfield find that the role of libraries in Research Data Management (irrespective of whether the data will then be open access) may involve activities and services such as offering advice on funding sources; advocacy for open access; data analysis advice; research data citation, advice on copyright and licensing issues, as well as technical advice on data formats and metadata.[17] In performing these activities they will have to collaborate increasing more intensively with institutions' IT services departments, which have also become key players in the data ecosystem.

Funding bodies are not only seen as sources for funding open access to research data, but also as enablers and are assigned a range of responsibilities. Various reports and roadmaps assign funders the responsibility of providing funds for data preservation and infrastructure development, issuing mandates, setting guidelines for research data management, and promoting open research data. Lyons et al. state that the responsibilities of funders also lie within wider public policy making, especially when it comes to assessing and meeting stakeholder needs.[18] Funders should partake in strategy and policy co-ordination along with stakeholders, and should act as advocates for data curation and fund expert advisory services. Funders should also support the development of the data curation workforce.[19] In 2005 The US National Science Board assigned a range of responsibilities to the US National Science Foundation, including creating a culture in which the publishing and preservation of digital data are of primary importance and ensuring that both the physical resources and the necessary training are broadly available.[20] This demonstrates how broad the funders' role is seen as being when it comes to implementing open access to research data, ranging from support to advocacy, driving the creation of data standards as well as making sure that data publications receive adequate credit.

[14] The Royal Society, op. cit. 2012

[15] See e.g. LIBER Europe, "Reshaping the research library", no date. http://www.libereurope.eu/committee/reshaping

[16] Ibid.

[17] Cox, Andrew M. and Stephen Pinfield, "Research data management and libraries: Current activities and future priorities". *Journal of Librarianship and Information Science*, Published online before print, 28 July 2013.

[18] Lyon, Liz. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, UKOLN Consultancy Report, 2007. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

[19] Ibid.

[20] National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, 2005. http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

There are also key intermediary roles emerging for other institutions, including publishers, data centres and newly established special purpose organisation that aim to enable data sharing and open access. Through their traditional role of coordinating peer-review, publishers are assuming a significant role in providing access to open research data and ensuring their quality, through requiring the deposit of research data in certified open repositories, and/or with dedicated publications that focus on research data. Apart from the more traditional type of publications undertaken by professional publishers, disciplinary repositories and data centres are increasingly important in managing, evaluating of and providing access to research data as well as in the education and training of researchers and other professionals working with data. In fact, as recent work on oceanographic research data shows, the increasing citations of research data directly from no longer the only or best way to publish research data.[21] Special purpose organisations, such as the Digital Curation Centre in the UK and the international Research Data Alliance, provide support for researchers and institutions in their efforts to make data openly accessible, but they also conduct research, produce reports, guidelines and standards, and organize meetings and discussions to bring open access to research data to the next level.[22]

For many institutions the expectations expressed in policy guidelines and roadmaps mean they have to take on a new roles and explore and develop new activities. Jones et al. state: "research data management represents new demands for HEIs [Higher Education Institutions] in terms of technical and organisational infrastructure, the provision of specialist data curation skills and long term planning for sustainable services".[23] The Royal Society notes that "the traditional role of the library has been as a repository of data, information and knowledge and source of expertise in helping scholars access them. That role remains, but in a digital age, the processes and the skills that are required to fulfil the same function are fundamentally different".[24]

In practice, changing roles and practices result in overlapping and shifting responsibilities, which are not always clear-cut.[25] Many issues cross stakeholder groups due to the complexity of the data journey (from collection through to making the data open access for re-use). For example, training and skill development are seen to be the responsibility of the various stakeholders: governments should adapt new policies for data management skills to be taught at university and secondary school level,[26] funders should educate their grantees on data management and institutions, with the help of libraries and IT departments, should provide training and educate their researchers and other staff on data management[27.] Researchers should also serve as mentors to early investigators and students who are interested in

[21] Belter Christopher W., "Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets", *PLoS ONE*, Vol. 9, Nr. 3, 2014. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0092590

[22] Digital Curation Centre, "About the DCC", no date. http://www.dcc.ac.uk/about-us; Research Data Alliance. "About", no date, https://rd-alliance.org/about.html

[23] Jones, Sarah, Graham Pryor and Angus Whyte, "Developing Research Data Management Capability: the View from a National Support Service*", iPres Conference*, 2012, p. 142. http://www.dcc.ac.uk/sites/default/files/documents/institutional-engagements/Institutional-engagements-iPres.pdf

[24] The Royal Society, op. cit., 2012, p. 63.

[25] Pearlman, Jay, Albert Williams III and Pauline Simpson (eds.), *Report of the Research Coordination Network RCN OceanObsNetwork: Facilitating Open Exchange of Data and Information*, 2013. https://darchive.mblwhoilibrary.org/bitstream/handle/1912/5937/RCN_Open_data_report_final.pdf?sequence=1

[26] High level Expert Group on Scientific Data, op. cit, 2010.

[27] Jones, et al., op. cit., 2012.

pursuing data sciences.[28] Such diffuse and sometimes conflicting roles and responsibilities can complicate the process of making research data freely accessible. Moreover, for each research project, different answers can be given to questions about responsibility, for example about who has responsibility for maintaining the software and for correcting errors.

1.1.4 Challenges facing institutions

Some institutions have already made considerable progress in sharing data and providing open access to data. Several large international consortia have, for example, created infrastructure to manage the huge amounts of research data, such as the CERN data centre and the European BioInformatics Institute. Other institutions have embraced the idea of open access and have begun to take steps towards transforming data-sharing practices: libraries are developing data management plans and creating repositories; funding bodies and research associations have adopted guidelines and mandates to provide open access. Yet, developments are still in a very early stage and making research data publicly available has proven to be a considerable challenge in most disciplines.

Several reports and roadmaps have identified a range of challenges to overcome. In its report to the EC, the High Level Expert Group on Scientific Data identified issues ranging from the preservation of data, protecting their integrity and conveying their context and provenance to funding open research data to dealing with the legal and ethical issues.[29] Knowledge Exchange, a collaboration between partners in four different countries, identified making sharing datasets an accepted and integrated part of the academic culture a "major challenge".[30] The report, which is a response to the vision of High Level Expert Group, also notes other challenges, including gaps and connectivity issues in the research data infrastructure and figuring out who will fund what infrastructures. The League of European Research Institutes (LERU) also notes a number of challenges facing institutions with regard to data management, such as the growing volumes of data, creating awareness and gaining a better understanding of the costs and benefits.

Some of the challenges have been addressed in previous RECODE reports, namely the technical, infrastructural, legal and ethical issues. In this report, we focus on some of the organizational challenges that confront institutions.

1.2 Methodology

The approach to the research that this report is based on has been two-fold. We first conducted a review of policy documents, reports, scholarly literature and other relevant documents and websites to provide an overview of current institutional approaches to making open research data possible and the gaps between these approaches and practice. We supplemented this review with interviews with representatives of the various stakeholders. To validate and fine-tune the results of this two-step analysis, we organized a one-day workshop with representatives from different stakeholder groups.

[28] National Science Board, op. cit. 2005.

[29] High level Expert Group on Scientific Data, op. cit. 2010.

[30] De Graaf, Maurits and Leo Waaijers, *A Surfboard for Riding the Wave: Towards a four country action programme on research data,* A Knowledge Exchange Report, November 2012. www.knowledge-exchange.info/surfboard

Various strategies and policies have been developed in support of open access in the European scientific landscape. As mentioned, the European Commission, individual members states as well as scientific institutions have taken steps to clarify roles and responsibilities, to outline workflows and organize funding. In the review we examine these (formal) strategies. We also examine how these strategies and policies take shape in practice and what barriers and obstacles institutions encounter when implementing them. We identify the barriers and good practices by looking at the discrepancies between formal strategies and policies on the one hand and current practices within institutions on the other hand.

To support and further deepen our analysis we drew on the five case studies that have also been used in the previous RECODE work packages. The case studies provide a source of illustrations of some of the issues. As each case study brings together a wide variety of institutions in various configurations, we have not exhaustively analyzed how they are interrelated and how each institution deals with (demands for) open access to research data. Such an analysis would be beyond the scope of this research, given the time constraints on this work package. Rather, we have conducted 15 semi-structured interviews with representative stakeholders to guide and crosscheck the analysis of the literature. More specifically, we spoke with individuals that are actively involved in the coordination and management of research data, including data centre managers, project coordinators and librarians. The questions for these interviews were based on an initial literature review. The interviews served to examine the institutional issues and solutions identified in the review in more depth. Moreover, they helped to understand how practices differ from stated policies.

The final workshop took place in Riga, Latvia on July 1, 2014 prior to the annual LIBER conference. More than 40 stakeholder representatives from various countries throughout Europe, as well as a few participants from outside of Europe, attended the workshop. As the workshop was tied to the LIBER conference, university and national libraries were well represented, but there were also participants from data centres, universities, research institutes and commercial companies. During the first part of the workshop, five invited speakers presented their view on the institutional challenges in open access to research data and participated in a panel discussion. Two weeks before the workshop, we sent the participants a draft version of this report and they were invited to provide feedback and suggestions during two breakout sessions in the second part of the workshop. They were also encouraged to reflect on the four challenges described in this report and potential strategies to overcome these solutions. We have used the feedback from this workshop to fine-tune our analysis and recommendations.

1.2.1 Four challenges

Based on an initial review of policy documents, journal papers and reports as well as the results of the earlier RECODE work packages we identified four key challenges that we focus on in this report:

- financial support;
- evaluating and maintaining the quality, value and trustworthiness of research data;
- educating and training researchers and other relevant stakeholders;
- creating awareness of the opportunities and limitations of open research data.

Financial support

Funding is a key issue in making and keeping research data openly available. While the creation of open access and data preservation repositories is clearly advantageous for institutions and those they serve, a significant financial outlay is needed, as setting up and maintaining open access repositories can cost millions of Euros per year. Although governments are investing in the preservation and dissemination of scientific information, the Directorate General for Research and Innovation recognizes that "research libraries often have to find creative solutions with a limited budget, and despite their increasing responsibilities in access and dissemination".[31] Data-sharing agreements between institutions with an associated sharing of the costs or the re-use existing ICT infrastructures may be more cost effective than creating new systems from scratch. However, such existing infrastructures may have problems of their own, such as technological obsolescence, or may also require additional staff training. Moreover, the long-term preservation of growing volumes of data introduces a whole new set of issues for which there is currently no clear funding strategy.[32]

Evaluating and maintaining the quality, value and trustworthiness of research data

In order for researchers to use data effectively they need to have some level of confidence in the accuracy and soundness of open research data. The second RECODE report focused on the technical aspects of evaluating and maintaining the quality and integrity of data. In this report we look at the role that institutions play in data quality management.

Institutions have developed various measures and strategies for evaluating and maintaining the quality and integrity of data as well as determining their value and impact. These include adherence to established best practice, peer review procedures, citation records, clear origins of data, transparent review and publishing practices, standard metadata, etc. Issues still remain, though. For example, some institutions have established practices for providing peer review of data, but the standards of such evaluations have not yet been agreed by stakeholders and vary considerably.[33] Also what constitutes high-quality data is a matter of debate. Most existing data management policies focus in particular on the technical requirements for reusability, but say little about scientific quality. However, for researchers to want to re-use the data, they will have to have some level of confidence in the scientific value of the data.

Educating and training researchers and other relevant stakeholders in the knowledge/skills required for open research data

Data sharing is not yet a common practice in most disciplines. This is partly because researchers often lack the skills and knowledge to share their data.[34] It requires considerable technical skill to generate digital data or convert data into machine-readable formats and to use the software tools to access and analyze the data. Researchers who wish to make their data publicly and digitally available and re-usable have to become acquainted with software tools and data formats that might not easily fit their existing research practices. Re-using data, in turn, requires researchers to learn how to search for and use data through web-based tools.
Researchers may also lack knowledge on data management policies and the legal and ethical aspects of dealing with their particular kind of research data. It can take considerable effort

[31] Directorate General for Research and Innovation, *National open access and preservation policies in Europe: Analysis of a questionnaire to the European Research Area Committee*, 2011.

[32] Bigagli, et al., op. cit., 2014.

[33] High level Expert Group on Scientific Data, op. cit., 2010.

[34] Borgman, Christine L., "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 6, 2012. http://dx.doi.org/10.2139/ssrn.1869155

and time to figure out how to adhere to policies and laws on data protection and intellectual property rights.

Institutions play an important role in providing training for open access. However, they encounter a range of issues in developing new practices in these areas, such as the diverse needs and knowledge levels between and within different disciplines, the rigidity of existing research cultures and the fast pace of technological developments. Moreover, librarians have had to develop a new skill set in order to support researchers in their data management activities.

Creating awareness about the opportunities and limitations of open research data
Besides a lack of skill, there are many other reasons why data sharing and open access are still not the norm in most disciplines, as previous RECODE reports showed. Researchers are reluctant to make their data publicly available because of concerns ranging from their work being scooped or misused, to not having enough time or funding to make their data accessible, to maintaining the privacy and confidentiality of their research participants.[35] However, in many cases a lack of experience or good examples cloud the discussion. Researchers as well as institutions are often not aware of the possibilities for publishing and sharing research data. In order to create open research ecosystems, institutions and research communities have to become aware of what open access to research data entails, how it can best be achieved and what the limitations are.

Institutions can play a role in stimulating change in existing practices, by advocating the benefits of sharing and providing open access. However, top-down approaches may not work well to inspire researchers. It may be hard to reach out to those researchers who are not yet accustomed to making their data publicly available. Moreover, researchers and research groups in many disciplines currently have few incentives and rewards for publishing data and maintaining their quality. Creating awareness requires collaboration with multiple stakeholders.

In this report we analyse the four challenges and how they are being and can be addressed. More specifically we explore current strategies, the remaining barriers and possible solutions for overcoming these barriers. This analysis serves as the basis to formulate policy recommendations that support institutions in addressing the challenges of making more research data open access.

1.3 Case studies

The fragmentation within and between disciplines when it comes to developing open access can be a barrier to realising the benefits of curating and freely sharing research data. The different disciplines and their various fields each have their own particular methodological and data practices. They thus have their own particular requirements for and expectations about data management and curation. In some fields this has led to a flourishing open data culture in which data are commonly shared using standardized formats and ontologies; in other fields open access is the exception and researchers are more inclined to use their own notations and only share their data upon request, if at all. Institutions, thus, engage with a wide range of stakeholders with different and at times conflicting values, drivers and

[35] Nelson, Bryn, "Empty Archives", *Nature*, Vol. 461, September 2009, pp. 160-163; Kuipers, Tom and Jeffrey van der Hoeven, *Survey results,* D3.4 PARSE.Insight, December 2009.

interests, which complicates the process of giving shape to their new roles and responsibilities. Nevertheless, institutions taking their first steps towards open access can learn from those stakeholders that have already developed successful open access platforms and established good practices.

One of the aims of the objectives of the RECODE project is to reduce stakeholder fragmentation. In order to do so, we have selected five case studies from across scientific disciplines and each case illustrates some of the grand challenges identified by the RECODE project. Each case study provides an access point into a network of institutions that collaborate to make certain kinds of research data accessible to various stakeholders.

1. *Particle physics and high energy computing*
Particle physics produces extremely large volumes of data (e.g. the Large Hadron Collider (LHC) at CERN produces about 15 petabytes of data per year). The LHC Computing Grid is the world's largest computing grid, and the Particle Physics and Particle Astrophysics (PPPA) Group at the University of Sheffield is a member of one of four regional Computing Grid Groups in the UK.[36] In this report, we use this case study to gain further insight in the institutional challenges involved in collecting, disseminating, storing and processing large quantities of numerical data from experiments related to particle physics where the expertise and resources necessary for storing and processing the data are only available to established experts in the field and/or very large consortia. The PPPA group is a part of a large network of institutions connected through the LHC. The networks include universities, companies, funding bodies, and others. These stakeholders collaborate in various big experiments such as the ATLAS experiment at CERN. To explore the challenges that various institutions encounter in this case study, we have also looked at institutions other than the University of Sheffield associated with the ATLAS experiment.

2. *Health sciences*
The collection and validation of personal data in clinical, health and biological contexts and their use in research poses various problems, including data protection, security, conflicting interests and data quality control. This case study focuses on the FP7 project EvA (project number 200605), which brings together a range of disciplines to look for genetic markers specific for emphysema and for airway disease in Chronic Obstructive Pulmonary Disease (COPD) patients. The consortium includes medical research institutes and university departments throughout Europe. It also collaborates with and is part of other consortia in data sharing, such as AirProm (Airway Disease Predicting Outcomes through Patient Specific Computational Modelling). This latter project brings together experts and current research to build a multi-scale computational model of the lung to characterise asthma and COPD.[37] The EvA projects illustrates how an interdisciplinary consortium consisting of a variety of institutions with often conflicting interests, manages to share and re-use data within and outside of the consortium and what open access would require from these institutions in this context.

3. *Bioengineering*
Biological engineering or bioengineering uses engineering approaches to study biological functions. Using computational models of biological processes researchers aim to further

[36] Particle Physics and Particle Astrophysics Research, "Research in particle physics and particle astrophysics", University of Sheffield, no date. http://www.hep.shef.ac.uk/research/

[37] The AirPROM project, "Home|Airprom", no date. http://www.europeanlung.org/en/projects-and-research/projects/airprom/home

develop scientific understanding of these processes. There is a perception that the data used for developing computational models of human physiology are, in a sense, fragile, and that the outputs of computational models of extremely complex systems may not be repeatable in the manner that is expected for acceptance in the current scientific paradigm. Besides various model repositories, several data repositories have been developed and there are a number of tools and standards to share data. Institutional issues in this case study include funding the infrastructure, training researchers and maintaining the quality of data. We will explore these issues with the Bioengineering Institute at the University of Auckland[38], associated institutions, and colleagues in the Virtual Physiological Human (VPH) community. This institute has a relatively long tradition of publishing its models in open source. It has collaborated with various institutes throughout the world to set up its data-sharing infrastructure. In New Zealand, for example, it works with the NeSi and the eResearch Centre and in Europe it has multiple partners in the VPH project. In this work package we also look at these associated institutions.

4. *Earth Sciences*

GEOSS[39] is an initiative that seeks to make existing systems and applications for geographic observation, including observations around drought, forestry and biodiversity, interoperable. In addition to providing interoperable access to data, GEOSS also seeks to develop an advanced operating capacity that provides access to analytical models that scientists from different disciples have used to make the data more understandable. In order to do so, GEOSS uses advanced modelling from a range of heterogeneous data sources to make data models usable by other communities, including through the use of natural language interfaces. It focuses particularly on the problem of data discovery and access, analysing search tools and techniques involving use of metadata, relevance indicators, keyword searches, to enable researchers and the general public to find their data of interest through the mass of available scientific data and information, and to access disparate content (e.g. heterogeneous encoding formats) through the same platform. The initiative also considers specifically the problems of technological sustainability and obsolescence, in relation to ensuring continued, coordinated and sustained access to research data as it ages.

GEOSS brings together around 80 nations and other international organisations. One such international organisation and active contributor is the Joint Research Centre. This Centre covers the role of data disseminator as the technical coordinator of the scientific and technical development of the European Directive INSPIRE. This Directive establishes an infrastructure for spatial information in Europe, to support EU environmental policies and activities that may have an impact on the environment. It aims to deliver integrated spatial information services to its target users, which include policy-makers at European, national and local level and citizens. INSPIRE contributes to GEOSS by making data a services operated by European Members States interoperable and accessible; developing standards and specification relevant to GEOSS and providing a portal and catalogue connected to GEOSS.
This case study provides insights in to the challenges that institutions, like the JRC, face in making data from heterogeneous research fields and research groups openly accessible and interoperable.

[38] The University of Auckland, “Auckland Bioengineering Institute”, no date. http://www.abi.auckland.ac.nz
[39] Group on Earth Observations, “Group on Earth Observations”, 2014. www.earthobservations.org

5. *Archaeology*

Open Context (OC) is a free, open access online platform for researchers in archaeology and related disciplines, to electronically publish primary field data and documentation. It offers researchers various services to help them prepare and publish their data, such as web services and editorial review. The platform also serves as a portal for easy browsing and searching. The aim of OC is to make archaeological field data freely and easily accessible on the Web. Additionally, it wants to encourage data sharing and (re)use. It therefore strives to publish archaeological datasets as Linked Open Data, such that the data sets provided on the website can be easily referenced by unambiguous identifiers and they include links to other resources on the Web. OC is the result of a project, funded by the National Endowment for Humanities and the Institute of Museum and Library Services. Currently, OC is maintained and administered by the Alexandria Archive Institute[40], a not-for-profit organisation[41], based in Berkeley, California, while IT development is carried out in collaboration with the Berkeley School of Information. OC furnishes useful information regarding attitudes, practices and policies within the ecosystem of archaeology, as well as significant information regarding the technical approach adopted for the deposition of, accessibility to and preservation of the data it contains. OC relies on other repositories, like the California Digital Library (CDL) at the University of California, for the preservation of data and maintaining its quality. CDL, established in 1997, provides data archiving and curation services. Such services include persistent identifier services, data storage and guidance on data management planning.

We use these case studies to guide our analysis of the challenges facing institutions. They provide an interdisciplinary grounding that helps us to address disciplinary fragmentation in the area of open access and data preservation and dissemination as well as maintain awareness of discipline-specific issues and practices. Each case study brings together a wide variety of institutions from different countries and different disciplines. The case studies help identify policy gaps and evaluate good practice solutions that will contribute to an inclusive and participatory development of policy recommendations.

In the following chapters we will look at each of the four challenges in more depth and discuss potential ways of addressing them. Each Chapter will give an overview of policies and strategies that address the particular challenge, present some of the remaining barriers and highlight some current developments to may help to overcome these barriers. The final Chapter will draw together the analyses of the challenges and provide a set of recommendations.

[40] The Alexandria Archive Institute, “The Alexandria Archive Institute”, no date. http://www.alexandriaarchive.org/

[41] Open Context is financially supported by The William and Flora Hewlett Foundation, The National Endowment for the Humanities and The Institute of Museum and Library Services.

2 FINANCIAL SUPPORT

One of the main drivers for open access to research data is the possibility of efficiency gains in publicly funded scientific research and business development. Efficiency gains can be realized through the re-use of data and avoiding duplication of data collecting and producing efforts.[42] In its vision for 2030, the High Level Expert Group on Scientific Data projects a future in which "Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data".[43] Initiatives such as the Human Genome project and Elixir provide inspiring illustrations of such visions. The $3.8 billion investment in the Human Genome Project is claimed to have had a $796 billion economic output in the U.S.[44] The inter-governmental organisation behind Elixir aims to "orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments".[45] The organisation argues that the cost of storing and managing experimental data in one open infrastructure is much less than funding re-runs of experiments and reproducing lost data.

Despite the potential cost savings that open access to research may bring in the future, one of the key challenges for institutions is securing funds for open research data. Preparing and archiving data as well as making them freely accessible can be expensive, depending on the characteristics of the data to be stored, searched and used. In particular in data-intensive research fields, the costs can be high, because of the extremely large data sets and the high-tech computing equipment required to processes and interpret the data. In particle physics, for instance, making and keeping data available to a wider public is an expensive process, because of the amount of effort and resources required for the production and storage of the data. Moreover, users need extensive computational resources and specialised knowledge to access and interpret the data. However, even for smaller individual projects the costs of making data freely available and easily accessible may press relatively heavily on the available research budgets.[46] Preparing data such that a wider range of users can access and properly interpret them may require considerable resources both in time and money.

Funds must be secured for various phases in the data life cycle in order to make open access possible, including the preparation, ingestion, sharing and archiving of the data. Researchers need to spend time on formatting data, adding metadata and making them accessible. Archives, data centres and repositories incur significant expenses for acquisition, ingestion and access, including personnel wages, training costs for researchers and (data) librarians and outreach programmes. In a cost-benefit study of digital preservation for research data, JISC concluded that the actual costs of archiving activities are small compared to the cost involved in acquisition, ingest and providing activities.[47] Staff costs are comparatively high. For some

[42] OECD, *OECD principles and guidelines for access to research data from public funding*, 2007. http://www.oecd.org/science/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

[43] High level Expert Group on Scientific Data, op. cit., 2010, p. 5.

[44] Battelle Technology Partnership Practice, *Economic impact of the Humane Genome project*, Batelle Memorial Institute, May 2011. http://web.ornl.gov/sci/techresources/Human_Genome/publicat/BattelleReport2011.pdf. More conservative estimates have also been made. See for instance Wadman, Meredith, "The Economic Impact of the Human Genome grows", *Nature,* 12 June 2013. http://www.nature.com/news/economic-return-from-human-genome-project-grows-1.13187

[45] Elixir, "Rationale: The Importance of Biological Data", no date. http://www.elixir-europe.org/about/rationale

[46] Sveinsdottir, et al., op. cit., 2013.

[47] Beagrie, Neil, Brian Lavoie and Matthew Woollard, *Keeping Data Safe (Phase 2),* JISC, 30 April 2010.

archives the latter can be as much as 50% of the overall costs.[48] JISC argues, though, that preservation costs of a particular archive are likely to significantly decline over the years. The heterogeneity of data sets and associated requirements that are produced in the wide variety of disciplines add to the challenge of financing the various phases of open research data: how can institutions be flexible in providing various services for different kinds of research on a limited budget?

Archives, data centres, and repositories also incur expenses for the development and use of the required technical infrastructure, including the hardware needed to store the data as well as the software tools to use them. In its roadmap, PARSE.Insight distinguishes between three funding stages of data management infrastructure development: prototypes, emerging infrastructures used by early adopters and long-term funding.[49] It notes that funding for sustainable infrastructure is difficult. Most data sets are created as part of a project that only last a few years, but the costs for curating the data will continue to be made after project funding ends. The data needs to be maintained, potentially updated, transferred to newer technologies, etc. It is not immediately clear which institution will bear the responsibility for this.

Moreover, increasing volumes of open research data may change existing practices and introduce new ones that will require funding. For instance, monitoring access to data or maintaining the integrity of data may generate new costs. It may require the establishment of an ethics board or the development and implementation of additional administrative and editorial procedures.[50] Bringing all associated costs, as well as the benefits, of open access to research data into view is part of the challenge.

As the above shows, the challenge of funding open access has multiple aspects. In the following section, we discuss several policies and strategies that institutions have developed to address some of these aspects We then turn to the remaining barriers and subsequently highlight a few recent developments that may help to overcome these barriers.

2.1 POLICIES AND STRATEGIES

Organizing funding for open access is a joint responsibility between researchers and institutions. Many institutions consider researchers to be responsible for organizing the financial resources for the publication and curation of their data. As data producers, scientists are viewed as the starting point of the data journey and they are deemed responsible for ensuring data quality, ethical data collection and clear communication of data, e.g., writing of metadata and context. Nevertheless, the top-down push for open access has also created responsibilities in terms of funding the required activities for the various institutions. Research communities look towards national and transnational funding bodies as well as research institutes to provide the resources to implement the various mandates and policies. Funding bodies in particular are able to set the agenda and provide a starting point for a culture change by mandating data management plans and the publication of data from funded research, but also by making the funds available to make it possible. They are also seen to be

[48] See also C. Rizzuto, *Research Infrastructures and the Europe 2020 strategy*, ESFRI, 2010. http://ec.europa.eu/research/infrastructures/pdf/esfri/publications/esfri_inspiring_excellence.pdf.

[49] PARSE.Insight consortium, *Science Data Infrastructure Roadmap,* 5 June 2010. http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

[50] Finn, et al., op. cit., 2014.

responsible for providing or at least investing in infrastructure in the form of data repositories for data from their funded research.[51]

As the open access movement grows, more resources are indeed becoming available to fund the required infrastructure and the curation of data. Governments, funding agencies, universities and libraries have allocated funds and developed policies to stimulate open data sharing. They commonly perceive funding interventions as an effective tool to encourage researchers and institutions to make their data openly available.[52]

The various institutions use roughly two kinds of funding strategies: 1) funding researchers and their projects and 2) funding the development of data infrastructures. These two strategies are also part of the European Commission (EC) initial effort to encourage open access to research data in Horizon 2020. In the new funding programme framework, a Pilot is included in which open access is required unless there are valid reasons for opting out.[53] For 2014-2015, topic areas participating in the Open Research Data Pilot will receive funding of around €3 billion.[54] Other instruments to encourage open access in the framework programme include funding for proposals that aim to develop e-Infrastructures for Open Access and coordination and support proposals that aims to strengthen and consolidate Europe’s contribution to the international Research Data Alliance.

The strategy of providing funding for researcher to make their data openly accessible is slowly gaining ground. The DCC has provided an overview of the data policies of research funders in the UK.[55] The overview shows that although most funders explicitly acknowledge that data sharing involves additional costs and allow or require these costs to be part of the proposal, some funders have no explicit statements on this.[56] The Wellcome Trust, for instance, is one of the funding bodies that has an explicit policy about including data management costs in project proposals. The Trust “considers that timely and appropriate data management and sharing should represent an integral component of the research process. Applicants may therefore include any costs associated with their proposed approach as part of their proposal”.[57] This approach fits within the mission of the Trust, as it supports “unrestricted access to the published output of research as a fundamental part of its charitable mission and public benefit”. Cancer Research UK, in contrast, considers timely and appropriate data management and sharing an integral component of the research process so will not provide additional funds for these activities. The Arts and Humanities Research Council's makes no explicit statement about supporting data management and sharing costs. However, the provided Technical Appendix for project applications asks how required

[51] Open Knowledge Foundation Blog, “EC Consultation on Open Research Data”, July 16, 2013. http://blog.okfn.org/2013/07/16/ec-consultation-on-open-research-data/; Lyon, op cit. 2007.

[52] Mossink, Wilma, Bijsterbosch, Magchiel, and Nortier, Joeri, *European Landscape Study of Research Data Management,* SIM4RDM, May 2013.

[53] European Commission, *Guidelines on Data Management in Horizon 2020*, version 1.0, 11 December 2011. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[54] European Commission, “Commission launches pilot to open up publicly funded research data”, European Commission Press Release IP/13/1257, 16 December 2013. http://europa.eu/rapid/press-release_IP-13-1257_en.htm

[55] Digital Curation Centre, “Curation policies and support services of the main UK research funders”, 2012. http://www.dcc.ac.uk/sites/default/files/documents/RC%20policy%20overview%20v2.2.pdf;

[56] The BBSRC, EPSRC, ESRC, NERC and the Wellcome Trust state that these activities can be included in grant proposals.

[57] See more at: Digital Curation Centre, “Wellcome Trust”, no date. http://www.dcc.ac.uk/resources/policy-and-legal/research-funding-policies/wellcome-trust#sthash.Jq9rfezX.dpuf

hardware, software and relevant technical expertise, support and training will be acquired, suggesting that appropriate costs may be budgeted in to grant applications.[58]

The UK is ahead of the rest of Europe when it comes to providing financial resources to researcher for open access to research data. Many funding bodies in Europe have yet to develop a policy on open access or have no immediate intention of doing so. Sim4RDM conducted as survey among funders in Europe and of their 18 respondents nearly half of them had a funding policy that covered data management.[59] Several funders were not planning to develop a policy. One funder did not have a policy because the area they cover is too heterogeneous for detailed recommendations. Others argued that formal requirements were not needed.

In the US several funders have also made provisions to help researchers cover the costs resulting from the mandates and recommendations regarding open research data. The National Institute of Health, for instance, states that it "recognizes that it takes time and money to prepare data for sharing. Thus, applicants can request funds for data sharing and archiving in their grant application".[60] In a review of US funders' data management policies, Dietrich et al. found that four of the ten funders did not have a policy on the matter at all. Of the other 22 policies they found, only eight mentioned whether funding for data management could be written in a proposal explicitly.[61]

Funding of researchers and their projects contributes to development of data infrastructures. Subject data repositories developed as part of research projects tend to get their funds through such project financing. The data collected as part of the AirPROM project in the Health case study provides an example. Researchers within the project share data between multiple projects to generate models of the lung. One work package is focused on establishing and maintaining a secure and sustainable data storage exchange and processing facility.[62]

As a second kind of strategy to stimulate open access, funding bodies have also directly invested in developing and maintaining data centres and repositories that offer data services to researchers, often at no cost to researchers or research groups. Science and medical funders frequently contribute to joint initiatives, for example at the European Bioinformatics Institute. In the Netherlands, the Data Archiving and Networked Services institute, jointly funded by the Royal Netherlands Academy for the Arts and Sciences and the Netherlands Organisation for Scientific Research (NWO) provides free data storage and preservation for datasets in the humanities, the social sciences and other disciplines.[63] However, the DCC overview shows that while most UK research funders provide a publications repository for their funded researchers, the provision of data centres is lacking as very few funding bodies have a full

[58] Digital Curation Centre, "AHRC - Arts and Humanities Research Council", no date. http://www.dcc.ac.uk/resources/policy-and-legal/research-funding-policies/ahrc#sthash.rcApQUDJ.dpuf

[59] Mossink, et al., op. cit., May 2013.

[60] National Institute of Health, "NIH Data Sharing Policy and Implementation Guidance", 5 March 2003. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

[61] Dietric, Diana, Trisha Adamus, Alison Miner and Gail Steinhart, "De-Mystifying the Data Management Requirements of Research Funders", *Issues in Science and Technology Librarianship,* 2012. http://www.istl.org/12-summer/refereed1.html

[62] AirPROM, "Work packages", no date. http://www.europeanlung.org/en/projects-and-research/projects/airprom/who-is-involved/work-packages

[63] Data Archiving and Networking Service, "Data Archiving and Networking Service|Homepage", no date. http://www.dans.knaw.nl

data service in place to support their researchers.[64] The exceptions are the ESRC and NERC, which both provide comprehensive preservation and support services through their data centres. The Arts and Humanities Research Council provides a data service for researchers in the area of archaeology through Archaeology Data Service and there are several services supported by STFC, such as the UK Solar System Data Centre and Atlas data store.

Funder agencies are not the only institutions with a responsibility in making open research data possible and they have their own expectations about the role of other institutions. JISC points out that funders of digital resources often have the expectation that "the project leader's host university, library or institution will bear the ongoing post-grant costs of the project".[65] Moreover, the responsibility of offering data services for the long- and short-term is also likely to fall upon institutions, because there is only patchy coverage of subject-specific data repositories and other data services.[66] For research that falls outside subject data centres' remits, the institutions in which funded researchers are based are often expected to maintain outputs in the long-term. Cox and Pinfield note in this respect: "As a result, higher education institutions (HEIs) in many countries are beginning to develop infrastructures to support researchers to manage their data more effectively, with services ranging from advice to storage repositories".[67] Several universities have established data repositories. In the Netherlands, a number of universities have started using the Dutch Dataverse Network, currently run by DANS.[68] The Dataverse Network is "an open source application to publish, share, reference, extract and analyze research data", first developed at Harvard University.[69] More and more universities push for researchers to deposit their data in institutional repositories, although it is not yet a common practice.

A number of databases or disciplinary repositories are the result of collaboration between multiple universities, research institutes, funding agencies and governments. Some of these consortia have been successful in organizing funding for data preservation for longer periods of time. For instance, the International Nucleotide Sequence Database Collaboration has developed and maintained three databases (DNA Data Bank of Japan, European Nucleotide Archive, and GenBank) for over 18 years. These databases receive funding from member institutions, project grants, funding bodies and governments. For large scale collaborations, like the ATLAS project, the picture becomes even more complex. ATLAS is one of the experiments run at CERN and part of the data preservation takes place at CERN's data centre. CERN is funded by 21 member states and through external funding, like project grants. Other higher-level data is stored at various institutions that provide the resources to run experiments.

Currently, many data centres and repositories receive their funding for open research data primarily from public and private funding bodies, either directly or through grant funding. Few other business models have been developed, although some institutions have started taking first steps to recover costs through other means. Some data centres have started charging for their services for larger data sets. For example, Open Context, one of our case

[64] The Digital Curation Centre *Curation policies and support services of the main UK research funders*, 2012. http://www.dcc.ac.uk/sites/default/files/documents/RC%20policy%20overview%20v2.2.pdf

[65] Maron, Nancy and Matthew Loy, *Funding for sustainability*: *How Funders' practices influence the future of digital resources*, JISC, June 2011, p. 24

[66] Cox and Pinfield, op. cit., 2013, p. 2.

[67] Ibid.

[68] Dutch Dataverse Network. http://Dataverse.nl/dvn

[69] The Dataverse Network project, "History", no date. http://thedata.org/book/history

studies, has mostly received funding through project grants. However, it has also developed a contributor-pays model to supports its open access publishing and archiving. Researchers wishing to publish their material pay between $250 and $6000 "depending on the complexity and size of the contributed database and related content".[70] Moreover, it offers additional fee-based service for "implementations based on Open Context's Web-services (API) or other customizations".[71] The Dryad Digital Repository, a repository that provides open access to research data underlying scientific publications, has developed a business and sustainability plan based on a combination of membership fees, Data Publishing Charges (DPC) and project grants.[72] A diverse range of stakeholders, including journals, research institutions, publishers and scientific societies can become a member and pay a fee, in exchange for a say in the governance of the organization and discounts on submission fees. The non-profit organization has various payment plans for data deposits ranging from a voucher plan to pay-on-submission fees. Dryad notes about these DPCs:

> *The Data Publishing Charge (DPC) is a modest fee that recovers the basic costs of curation and preservation, and allows Dryad to make its contents freely available for researchers and educators at any institution anywhere in the world. DPCs provide a broad and fair revenue stream that scales with the costs of maintaining the repository, and helps ensure that Dryad can keep its commitment to long-term accessibility.*[73]

In sum, more funding is becoming available from the traditional channels of research funding both for researchers as well as data infrastructure. At the same time, various institutions have also started experimenting with alternative ways to cover the cost of open research data. Nevertheless, open access to research data is still relatively new and many institutions have only taken initial steps if any. Funding open access to research data is not yet widespread. Moreover, the strategies and policies discussed above do not address some remaining barriers, which we will discuss in the next section.

2.2 REMAINING BARRIERS AND CHALLENGES

Despite increasing funding for open research data, open access to research data presents some problems that have received relatively little attention, including long-term curation, changing institutional cultures, weighing and modelling costs, and data selection.

Longer-term curation of data remains a key problem. Most open access initiatives pursue the short-term goal of getting more researchers to share their data. Data centres and institutional repositories are often happy to offer free data services, because they support open access and may get recognition for hosting particular data sets. Questions about ongoing technological development, increasing volumes of heterogeneous data sets and long-term preservation have been put on hold. For example, in many cases there are no plans or protocols in place for dealing with updating technological infrastructures and translating data sets to new data formats.

[70] Open Context, "Data Publication Guidelines for Contributors", no date. http://opencontext.org/about/publishing

[71] Ibid.

[72] Dryad, "Business Plan and Sustainability", 5 September 2013. http://wiki.datadryad.org/Business_Plan_and_Sustainability

[73] Wendell, Laura, "The who, what, when and why of Data Publishing Charges", Dryad News and Views, 30 August 2013. http://blog.datadryad.org/2013/08/30/the-who-what-when-and-why-of-data-publishing-charges/

Various large-scale consortia have given some attention to the problem of data preservation. They tend to rely on the continued support of member states and organizations. Examples include GenBank and European BioInformatics Institute who have a more long-term outlook as they have been set up as data centres that curate data for longer period of time. The member institutions and countries are committed to contributing the resources to enable this.

However, even for such established institutions and for longer running projects and experiment like ATLAS at CERN, long-term curation and preservation pose a problem. One of our interviewees in the Particle Physics case study notes that although several institutions are committed to contributing resources for an experiment like ATLAS, it does not mean that they will also provided the resources for the preservation of data:

> *I mean in the UK, the Research Councils UK had a policy that said, 'Well, the Research Councils fund your research and the institutions then have to preserve the data that comes out of those research projects.' This is not sensible in more than one way for something like the LHC [Large Hadron Collider] data because, the data is [sic] across many countries. The fact that I have a petabyte of it here [...] is because we've stepped up to the plate and we're providing resource to the experiment to help it do its analysis and so on. That does not mean an ongoing commitment from [the University] to preserve a petabyte of data into perpetuity, and nor would that be sensible, nor could we easily make that freely available to everybody.*
> (Interview 3, physicists involved with ATLAS data preservation, particle physics)

He argues that preservation of open access data for after the experiment ends would require an international solution, which is currently lacking:

> *This is an international collaboration and the solution needs to be on an international basis. So I know that there are various institutions who would be interested in being long-term archives. I know there's some interest from the UK. I know there's interest from Italy. I'm sure that there are others that are interested in doing that. So long as that is done in the appropriate international context and in collaboration with the experiments who [sic] have to implement it, that's fine, but I must say at the moment, it's not obvious where those lines are.*
> (Ibid.)

Research consortia often only last as long as their funding does and data preservation plans for after the consortium ends are not usually in place. This is even more of an issue for smaller scale data sets. Individual researchers or research groups may provide the initial funding for the creation of a data set through their research grants. Some researchers reserve some time to maintain a data set after the project ends, but more often then not researchers will move on to other projects and other universities leaving their data sets orphaned.

Universities, national data centres and libraries have stepped in to provide infrastructure and financial resources to preserve such orphaned data sets. However, as the number of data sets grows this may not be a sustainable option for all orphaned data sets, because the curation of open data continues to require resources. Staff or volunteers with particular skills and expertise are needed to keep the data up-to-date and accessible. They will have to make decisions about what data to keep, how to transfer them to new technologies or formats, etc. To stimulate their use, an effort has to be made to bring it to the attention of relevant audiences. The Royal Society therefore points out that universities and libraries will need "larger budgets and highly skilled staff if the roles that are suggested are to be fulfilled by institution, such as universities".[74] Institutions are still searching for sustainable strategies for preserving and curating data for longer periods of time.

JISC points out that funders of digital resource projects generally do not ask grant applicants to outline a clear and plausible vision of the activities involved in curating a digital resource after the grant is finished. The report argues that such resources require sustainability plans that do not only specify strategies for dealing with the storage of data, but also with adding to, keeping the data up-to-date, etc. JISC also points out that funders do not always have a strong incentive to care about the sustainability of digital resources. The more the digital resource is at the heart of their mission, the more they care.[75]

Libraries are under pressure as well. In realigning their role within the more data-intensive scientific and scholarly landscape, some libraries have taken it upon themselves to provide data services and repositories for researchers. These new activities add to the growing workload of libraries and put additional claims on already shrinking budgets. Libraries and universities are under cost pressure from scientific journals that are continually increasing subscription costs for libraries as well as from general austerity measures. Governing boards may be reluctant to reserve large amounts of funds for the preservation of open research data, in particular in areas where the level of data sharing is still low. It can be a major problem to convince university administrations to gather economic resources for developing data curation models. In fact, most of the scarce funding for research data management is coming from libraries themselves.[76] Usually there is no extra seed money available inside the organization and libraries either have to reallocate internal resources or find external funding, e.g. cooperation with outside partners. Therefore the initiation of grants and funding for libraries on national or international levels will be an important factor for getting data curation to gain speed on a broader level at universities.[77]

Another significant barrier to open access to research data is that institutions are often reluctant to give up their valuable data. Arguing for open data, Neelie Kroes, Vice President of the European Commission responsible for the Digital Agenda for Europe, stressed that "data takes on a new importance and value in the digital age" and called it the new gold.[78] Universities and research institutes have also come to recognize research data sets as valuable assets that can be used as alternative sources of funding. They have had to deal with shrinking research budgets and restructuring. Moreover, they come from a period in which they were supposed to generate part of their own income. Selling licenses on data on desirable data sets can be a productive way to do so. Within some universities this can create barriers to opening up research data. An interviewee in the bioengineering case study points out:

[74] The Royal Society, op. cit., 2012, p. 67.

[75] Maron and Loy, op. cit., 2011, p. 22

[76] Walters, Tyler, "Data curation program Development in U.S. Universities: The Georgia Institute of Technology Example", *The International Journal of Digital Curation*, Vol. 4, No. 3, 2009.

[77] Asher, Andrew, Kiyomi Deards, Maria Esteva, Martin Halbert, Lori Jahnke, Chris Jordan, Spencer Keralis, Siva Kulasikaran, William Moen, Shannon Stark, Tomislav Uran, and David Walling, *Research Data Management: Principles, practices, and prospects*, Council on Library and Information Resources, 2013. http://www.clir.org/pubs/reports/pub160

[78] Kroes, Neelie, "Data is the new gold Opening Remarks", Press Conference on Open Data Strategy Brussels, 12 December 2011. http://europa.eu/rapid/press-release_SPEECH-11-872_en.htm

> *Most of the lighter institutions are still suffering from the leftovers politics of the '90's, where every agency had to recover the costs by selling the data. And most of those mandates have gone now. But the thinking is, 'well no this is our resource, we need to sell it, rather than give it away'. It's still very much entrenched. So most of those institutions have really good infrastructure for sharing data that they don't make available to the wider public or the wider research sector out of the country.* (Interview 2, director centre for eResearch, bioengineering)

Another of our case study respondents noted the difficulty he had in convincing his host institution to give up millions of dollars they could earn from selling licenses for the use of the datasets his research groups had produced, in favour of making the data freely available. He managed to convince his university by arguing that the revenues gathered through selling licenses couldn't be matched by the amount of international funding available for open research data initiatives (Interview 1, project coordinator, bioengineering).

Similarly, as a respondent in the archaeology case study noted, it is difficult to change established organizational structures to make cost recovery possible.

> *This is a huge problem. Currently, most of our services are offered to the university community free of charge, but this is slowly changing. We are in the process of moving most of our services over to a full or partial cost recovery basis. For the digital repository, for example, we are going to start charging for the amount of storage capacity that is consumed or allocated. […] Getting the approval from the University administration to move to cost recovery has turned out to be much more difficult than we had anticipated and it's still not done yet.*
> (Interview 3, associate director data centre, archaeology).

As large-scale open access to research data is still in its infancy, few business models have been developed to sustain long-term preservation of data. Funding resources are beginning to be made available, but issues of who will pay for what still have to be sorted out.

Weighing the costs and benefits of open access is also a remaining barrier. It is important to identify the costs of open access in order to speak about effective management. However, few cost models have been developed to calculate the costs and benefits of supporting data management and open access. One notable exception comes from the Australian National Data Service (ANDS), which has take a proactive approach to make research data and public data openly available and sharing its experiences in doing so with others.[79]

Finally, a key challenge is establishing what should be funded. Some disciplines produce petabytes of data that cannot all be stored and some data sets might not be interesting enough to store. As the volume and number of datasets grow, institutions will have to start making decisions about what data to keep and how to store it. They will have to develop strategies for choosing what to invest in. In doing so they may be affected by public demands for outcome and results, but what is valuable data cannot always be predicted or anticipated beforehand. Currently, few explicit strategies have been developed to deal with such questions.

[79] Houghton, John, *Costs and Benefits of Data Provision: Report to the Australian National Data Service*, 2011. http://www.ands.org.au/resource/cost-benefit.html

2.3 Working towards solutions

Several initiatives offer some potential solutions to address some of the remaining barriers. Based on its analysis of the costs in data preservation, JISC recommends institutions to "take advantage of economies of scale, using multi-institutional collaboration and outsourcing as appropriate".[80] In an effort to reduce costs and create economies of scale institutions have started collaborating in offering data services. An example is the collaboration between DANS and several archives and libraries in a federated data infrastructure, based on a Front-Office/Back-Office model.[81] The federated system comprises a network of local data stewards close to scientific practices combined with centralized data services.

Cost modelling has also received some attention. LERU recommends collecting good practices in cost modelling and exchange of information on costs to build a knowledge base to inform their development. Collecting and sharing information on costing for research data is particularly a responsibility for the Chief Information Officer community, according to LERU. The Collaboration to Clarify the Cost of Curation has taken up the challenge of supporting institutions in gaining more insight in their data curation costs. The aims of the project are a) to ensure that where existing work on cost modelling "is relevant, that stakeholders realise and understand how to employ those resources" and b) "to examine more closely how they might be made more fit-for-purpose, relevant and useable by a wide range of organisations operating at different scales in both the public and the private sector".[82]

In order to move forward, funding models will have to be developed that take into account the long-term curation of research data. University College London has attempted to address part of this challenge by offering three different types of services: data storage services for the run-time of the project, data preservation services and access services. By offering storage for the run-time of the experiment UCL aims to encourage researchers to think about what will happen to the data after the project ends. Several funding bodies are also requiring applicants to specify how their data will be preserved for the longer-term.

2.4 Conclusion

Funding is one of the key challenges for making open access to research data possible. Currently considerable attention has been given to creating the required data infrastructures and getting researchers to share their data through funded mandates. Yet, although more resources are becoming available, it is clear that funding for open access research data is still in the early stages of development. Not all funders and research producing institutions recognize or are aware of the costs involved in data curation. Current funding models are generally project-based and there are few institutions that provide preservation and curation services for the long-term. In addition, few institutions have sustainability strategies or policies that take into account growing volumes of data and the additional costs they will generate to keep them findable, accessible, and reusable. Moreover, the distribution of responsibilities is yet to be clearly established.

[80] Beagrie, "Keeping Research Data Safe Factsheet", no date. http://www.beagrie.com/static/resource/KRDS_Factsheet_0711.pdf

[81] Dillo, Ingrid, Rene van Horik and Andrea Scharnhorst, "Training in Data Curation as Service in a Federated Data Infrastructure – the FrontOffice/BackOffice Model", 2013. http://arxiv.org/pdf/1309.2788.pdf

[82] Ibid.

Solutions to some of the remaining barriers lie in gaining a better understanding of the cost involved in the curation of large of volumes of heterogeneous data as well as developing new sustainable business models. Moreover, policies for selecting which data will be funded will have to be developed. Sharing experiences and exploring and establishing collaborations between institutions to reduces costs can also contribute significantly to further developing these solutions.

3 EVALUATING DATA QUALITY, VALUE AND TRUSTWORTHINESS

The increasing volume of openly available research data raises the question of how to ensure their quality, in view of enabling their re-use.[83] In order for open research data to be of value for research communities, researchers need to have some level of confidence in the trustworthiness and integrity of data sets as well as data repositories. Data sets and metadata that contain significant inconsistencies, inaccuracies, flaws or that are incomplete are hard to work with and require additional time and financial resources. The technical aspects of this challenge, as mentioned, have been addressed in the second RECODE report. Here we are interested in what this challenge means to various institutions and what role they can play in addressing it.

While most policy documents stress the importance of data quality, there is no uniform definition of the term. One reason for this is that it is hard, if not impossible, to define what constitutes data, because data can take different forms in different disciplines. The European Commission states that research data "may be numerical/quantitative, descriptive/qualitative or visual, raw or analysed, experimental or observational".[84] It gives examples such as digitised primary data, photographs and images, and films. For most disciplines, however, this definition does not capture the complexity and multifaceted nature of data. They may, for example, distinguish between various levels of data, including raw data, processed or derived data. Various researchers also stress that data are meaningless without the appropriate context in which the data were generated or without the software or models that produced the data. The Research Information Network (RIN) points out that research data are generated for different purposes and through different processes: they can be generated during experiments; they may result from models or simulations, or they come from observing unique events. Moreover, data can be generated and collected for different reasons and at different stages in the research process, with variations in the status that is attached to them. At the same time, some researchers, such historians, would argue that they do not generate data, but make use of primary sources, such as publicly available documents or artefacts.

In the absence of a clear definition of data quality, RIN understands quality to pertain to whether data is "fit for purpose". It identifies three key purposes with regard to creating, publishing and sharing data sets:

> *First, datasets must meet the purpose of fulfilling the goals of the data creators' original work; second, the data creators must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers; thirdly, data sets should be discoverable, accessible and re-useable by others*.[85]

[83] European Commission, "Report of the European Commission Public Consultation on Open Research Data", 2 July 2013. http://ec.europa.eu/research/science-society/document_library/pdf_06/report_2013-07-open_research_data-consultation.pdf

[84] European Commission, *Commission Recommendation on access to and preservation of scientific information*, C(2012) 4890 final, Brussels, 17 July 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[85] Swan, Alma and Sheridan Brown, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*, RIN, London, 2008, p. 48. http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs.

According to RIN, "Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated."[86] The H2020 Pilot adopts a similar general definition in its requirement that the "data and associated software produced and/or used in a project is assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data provided in a way that judgments can be made about their reliability and the competence of those who created them)"?[87] Making sure that data is interpretable, assessable and re-usable is essential in making research data open and often requires additional efforts. In order for a wider audience to find, access and use the data, they have to be presented in some standardized format and accompanied by appropriate metadata.

Yet, establishing that open research data are 'fit for purpose' is often a difficult and time-consuming task that requires considerable specialized expertise. It can entail a careful review of the metadata and data content as well as a close inspection of the processes used to create them.[88] The variations in forms of data have different implications for whether and how data can be assessed on their quality. Each discipline is confronted with different challenges. In particle physics, for example, dealing with large quantities of data, only large consortia may have the capacity to provide the necessary quality assurance process, while in bioengineering outputs may not be repeatable in the case of heterogeneous datasets.[89] An additional complicating factor is that ideas about what is a sufficient level of quality will differ, depending on who produces, manages or uses the data. Institutions providing access to public sector data to researchers as well as commercial partners or citizens may have to deal with competing interests in setting their quality standards.

The varying forms of data also raise questions about what and when data should be published. Should raw data be made available as early as possible or should the data first be processed, losing some information but making it easier for others to interpret and re-use the data? A respondent from the particle physics case study noted that in order to make the data accessible and re-usable for other researchers that do not have the tacit knowledge that comes with producing the data, considerable processing of the data is required:

> *Now, speaking personally, I'm aware that if I was somebody coming at this with a background of the Climategate scandal, which I think has driven a lot of the discussion about openness of access to data, you would say, 'But you're hiding that pure raw data and somehow you could be masking all of this from us and you are required to be between us and the data.' I'm aware that that's true. However, I could give one of those people our raw data if they happened to have a 100-petabyte data store that they had free and available […] They'd be able to make almost no interpretation of that raw data. We could give them the software and I doubt they would be able to use it. This is not being obstructive. This is just a simple statement of the complexity of what is going on and so this is why our emphasis is on looking at*

[86] Ibid.

[87] European Commission, *Guidelines on Open Access to Scientific Publications and Research Data in Horizon* 2020, op. cit., 2013 and European Commission, *Guidelines on Data Management in Horizon 2020*, op. cit., 2013; The Royal Society, op. cit., 2012.

[88] Costello, Mark J., William K. Michener, Mark Gahegan, Zhi-Qiang Zhang and Philip Bourne, "Biodiversity data should be published, cited and peer-reviewed", *Trends in Ecology and Evolution*, Vol. 28, No. 4, August 2013, pp. 454-461.

[89] Sveinsdottir, et al., op. cit. 2013.

> *data which has been refined enough to be scientifically valuable for reinterpretation by people who are not completely embedded in all this tacit knowledge*.
> (Interview 3, physicists involved with ATLAS data preservation, particle physics)

Evaluating and maintaining the quality and integrity of data is not straightforward, but it is becoming increasingly more important that institutions take up this challenge in order to fulfil the promises of open access. The following section provides an overview of some of the policies and strategies that have been developed to ensure the quality of research data. We will then explore some of the remaining barriers and discuss several emerging solutions.

3.1 Policies and strategies

Overall, data quality issues entail the involvement of a variety of stakeholders in the data ecosystem such as research funders, universities, data centres, repositories and researchers present in the different stages of the data life cycle. These stakeholders are increasingly giving attention to the quality and integrity of data. Various institutions have developed a range of policies and strategies to make sure that openly accessible data are trustworthy and of sufficient quality. While funders' and institutional research data management policies are not systematically recorded in registries, leading us to assume that such policies are the exceptions rather than the rule, there are, however significant resources in the form of guidelines, training materials, and tools that address the issue of quality and integrity of research data.

In most disciplines it is already common practice that data are assessed at various stages in the data life cycle. They may be checked automatically as part of the automated processes that control scientific instruments. Also, many research communities have formal and informal procedures to assess data. In its report on publication and quality assurance of research data, RIN found that in discipline such astronomy, chemical crystallography and genomics data are assessed automatically during their creation but data creators and curators also manually check the data and metadata.[90] Moreover, the research community also reviews and comments on the data. The report notes that in the classics, which include archaeology and art history among others, there is a tradition of careful editing and proof reading of data, but that "classicists tend to trust the quality of their peers' datasets".[91] Similarly, in a report on responsible data management the Dutch KNAW found that in many disciplines data quality was not perceived as a problem, because scientific practices were considered to have built-in self-correcting mechanisms (e.g., replication of experiments or the use of publicly available sources encourages researchers to produce high-quality data).[92]

Activities at various stages in the data life cycle to ensure or improve the quality of data are usually referred to as quality assurance (QA) or quality control (QC). Such activities can include checking whether values are within the plausible range of the instrument that produced them and the property being measured.[93] The concepts of QA and QC are closely related and often used interchangeably. Waaijers and Van de Graaf, for instance, building on

[90] Swan and Brown, op. cit., 2008.

[91] Ibid., p. 52.

[92] Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), *Responsible Research Data Management and the Prevention of Scientific Misconduct,* 2013. http://www.knaw.nl/en/news/publications/responsible-research-data-management-and-the-prevention-of-scientific-misconduct/@@download/pdf_file/20131009.pdf

[93] DataOne, "Ensure basic quality control", no date. https://www.dataone.org/best-practices/ensure-basic-quality-control

the RIN report on quality assurance, focused in their study on the operational aspects of the concept of, what they call, quality control during the various phases of the data life cycle (i.e., production, management and use/ re-use). The authors identified three types of quality control: quality during the production phase focusing on the documentation accompanying the dataset; data management planning focusing on managing the data to ensure permanent access to data; and quality control of datasets focusing on the scientific/ scholarly quality of research data.[94] Others argue that the key difference between QA and QC is that QA is often process-oriented and focuses on avoiding problems during the creation of data or data sets, while QC is product-oriented and tends to refer to the procedures that check whether created data or data sets meet the quality requirements outlined by the end-users.[95] For the sake of simplicity, we will use the term quality assurance in the following to refer to both the activities involved in checking the quality of data while it is being produced, ingested and made available as well as in checking the data after they have been produced and made available.

Several repositories and data centres have developed a variety of quality assurance measures and offer a range of services to evaluate the technical quality of data sets. These include providing process documentation, completeness/consistency checks, data management and sharing training, file format validation, metadata checks, storage integrity verification and tools for annotating the quality information.[96] PANGAEA data publisher for earth and environmental Sciences, for example, offers quality assurance on metadata (e.g. citation, references, geo-location, and standard parameter vocabularies) in addition to providing permanent identification and access with a Digital Object Identifier (DOI) name for each data supplement.[97] Moreover, PANGAEA has a policy that editors check the completeness and consistency of metadata and data. However, the data producer remains responsible for the scientific quality of the data (e.g. the validity of used methods).[98] In its report on peer review of research data, APARSEN notes that overall repositories ensure quality assurance through two complementary measures: the selection of data during the recording process and the curational measures of data management.[99] Such measures vary, though, according to the form, scope and discipline of data, the report points out.

In particular disciplines that make use of computer-controlled instruments and sensors that produces large amounts of data, such as astronomy, particle physics and genetics, it has become increasingly more common for data repositories to use automated quality assurance processes. Such automated procedures can quickly process data and identify and correct problems in real time without introducing human error. Yet, expert knowledge might still be needed to make appropriate decisions on how to treat data flagged as problematic.[100]

[94] Waaijers, Leo and Maurits van den Graaf "Quality of Research Data. An Operational Approach", *D-Lib Magazine*, Vol. 17, No. 1/2, January/February 2011.
http://www.dlib.org/dlib/january11/waaijers/01waaijers.print.html

[95] Campbell, John, L., Lindsey E. Rustad, John H. Porter, Jeffrey R. Taylor, Ethan W. Dereszynski, James B. Shanley, Corinna Gries, Donald, L. Henshaw, Mary E. Martin, Wade M. Sheldon, and Emery R. Boose, "Quantity is nothing Without Quality: Automated QA/QC for Streaming Environmental Sensor Data", *BioScience*, Vol. 63, No. 7, July 2013, pp. 574-585.

[96] APARSEN, "Report on Peer Review of Research Data in Scholarly Communication", 30 April 2012.

[97] Pangaea Data Publisher for Earth and Environmental Science, no date. http://www.pangaea.de

[98] Ibid.

[99] APARSEN, op. cit., 2012, p. 21.

[100] Campbell et al., op. cit., 2013, p. 581.

Another strategy data repositories employ to enhance and maintain data quality relates to the selection of data for long-term data preservation and retention. As not all data that research project generate are necessarily useful and need to be preserved, the value of data re-use needs to be assessed on the basis of its usefulness and scientific value. Data retention is a topic that is often addressed in institutional data management policies and a timeframe for the review of the research data after the completion of a project is set to evaluate whether these will be maintained in the institutional repository and/or data centre.[101] One project that offers some guidance to institutions, in this regard, is the MaRDI Gross project. This is a JISC funded project with the aim "to support Big Science projects in developing data management and preservation plans to manage the long-term curation of data they generate".[102] To that end it has released a report intended to provide guidance for the development of data management and preservation plans for Big Science data, although (as stated in the report) it can be of use to other planners and data architects who wish to implement good practices in the area.[103] The report notes that there are cases in which an experiment cannot be feasibly redone or it may not be feasible to document that analysis in enough detail so that it can be reanalysed. This means that at least some details of the experimental environment "are not reasonably preservable and that little effort should be made in preserving them if well-documented high level data products are available and intelligible".[104] The report also stresses that this does not mean that it advocates deleting data but rather that we should not overstate their value.[105] It could thus be argued that preserving data is quite easy, but that making it useful requires much more work and effort.

In the UK several institutions have developed basic quality standards for data centres and repositories. The UK Research Councils' (RCUK), for instance, recently adopted common principles on research data.[106] These principles emphasise the significance of open access to publicly funded research data and the modalities necessary to provide access to it. Among them, the existence of sufficient metadata that will allow other researchers to understand the research and re-use potential of the data is at the top of the list. The importance of disciplinary standards for the access to and reuse of data is also emphasized. The NERC (Natural Environmental Research Council) in the UK maintains a data centre and has developed a research data value checklist, which informs decisions of the data centre to maintain or dispose of research data that it holds.[107] The UK Data Service, essentially the main research data repository for the ESRC, developed a collection policy, by which the research data ingested into its databases is selected.[108] According to this policy "only the most relevant and highest impact data collections are selected for ingest and curation, ensuring that limited resources are not expended on ingesting potentially low-use data collections".[109]

[101] E.g. in the case of the University of Northampton Data policy: University of Northampton, "Open research at the University of Northampton", no date. http://www.northampton.ac.uk/research/open-research-at-the-university-of-northampto

[102] The MaRDI-Gross Project, "Initial draft op the project report", 22 March 2012. http://mardigross.jiscinvolve.org/wp/2012/03/22/initial-draft/

[103] Bicarregui, Juan, Norman Gray, Rob Henderson, Roger Jones, Simon Lambert, and Brian Matthews, *DMP Planning for Big Science Projects*, MaRDI-Gross Project, v1.0, 17 August 2012. http://purl.org/nxg/projects/mardi-gross/repor

[104] Ibid., p. 14.

[105] Ibid.

[106] Research Councils UK. "RCUK Common Principles on Data Policy", no date. http://www.rcuk.ac.uk/research/datapolicy/

[107] Digital Curation Centre. "NERC - Natural Environment Research Council", no date. http://www.dcc.ac.uk/resources/policy-and-legal/research-funding-policies/nerc

[108] UK Data Service. "Our purpose", no date. http://ukdataservice.ac.uk/about-us/purpose.aspx

[109] UK Data Service, "Collections Development Policy", 28 January 2014.

Thus, the value of the research data is first assessed for the ingestion. The criteria for selection, according to the archive, are the following: relevance to the remit; scientific and historical value; new sources or types of data; international value; uniqueness/risk of loss; usability/redistribution/operational benefits; replication data and resources.[110]

According to DCC, appraisal and selection criteria used for traditional (paper-based) material would need to be modified to be applied to data, placing more emphasis on the technical capability to preserve data, the on-going cost of maintaining data ('digital mortgage') and making preservation decisions early on in the data life cycle.[111] As DCC notes, a universally applicable appraisal and selection framework is not realistic; different kinds of material, produced in different contexts for different stakeholders will require different approaches. While the DCC acknowledges that the criteria for assessing a dataset or a resource's value will be discipline-specific, it nonetheless provides seven general criteria: relevance to the mission, scientific or historic value, uniqueness, potential for redistribution, non-replicability, economic case, and full documentation.[112] It also maintains that an appraisal and selection policy "needs to ensure consistent, transparent and accountable decision-making so that commitments can be tracked and accounted for".[113]

An example of a policy for quality control can be found at the SeaDataNet. This is a European service for Ocean and Marine Data Management that provides access to ocean and marine research data of the major national ocean and marine research centres of European member states. It has developed detailed procedures for data quality control.[114] As stated in the its 'Quality Control Procedures' Manual, "data resources are quality controlled and managed at distributed data centres that are interconnected by the SeaDataNet infrastructure and accessible for users through an integrated portal. The data centres are mostly National Oceanographic Data Centres (NODCs), which are part of major marine research institutes that are developing/ operating national marine data networks and international organisations such as IOC/IODE and ICES. The data sets come from various sources and time periods. This imposes strong requirements towards ensuring quality, elimination of duplicate data and overall coherence of the integrated data set".[115]

Another emerging strategy is the certification of data repositories or data sets. Certification initiatives aim to ensure the quality of data repositories as well as to influence the quality assurance of data. One example is the Data Seal of Approval (DAS). DAS was established by a number of institutions "committed to the long-term archiving of research data".[116] The Seal "is granted to repositories that are committed to archiving and providing access to scholarly research data in a sustainable way. It is assigned by the DSA Board and renewed every year

http://ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf

[110] Ibid, p. 6.

[111] Digital Curation Centre, "Instalment on 'Appraisal and Selection'", 2007. http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/appraisal-and-selection/appraisal-and-selection.pdf

[112] Ibid.

[113] White Angus and Andrew Wilson, "How to Appraise and Select Research Data for Curation", *DCC How-to-Guides. Edinburgh: Digital Curation Centre*, 2010, p. 4. http://www.dcc.ac.uk/resources/how-guides

[114] SeaDataNet, "Data Quality Control Procedures", 2010. http://www.seadatanet.org/Standards-Software/Data-Quality-Control

[115] Ibid.

[116] Data Seal of Approval, "About the Data Seal of Approval", no date. http://www.datasealofapproval.org/

through a modification procedure".[117] Other reported sources of accreditation are the Deutsche Initiative fűr Netwerkinformation (DINI) Certificate[118], Trustworthy Repositories Audit and Certification (TRAC)[119] and ISO 16363:2012.

Developing measures for the assessment of the technical quality of data have been a central focus in the strategies discussed above, while such measures for the scientific quality seem to be less developed. Such measures would evaluate data and the content of the metadata in terms of, for example, whether appropriate methods were used to collect the data or whether the data accurately reflect actual observations or responses. Evaluating data on that level usually requires expert knowledge and can only be achieved through peer review or review by a dedicated subject specialist.

In part, the lack of quality assessment measures that focus on scientific quality is the result of data repositories not feeling responsible or capable of evaluating data on that level. One of our case study respondents pointed out:

> *As an infrastructure provider, I think there are only very few things you can do in terms of data quality. One of the things you can do is, set up an automated checking system so that at least you can pass the data or check that its schema makes sense and that certain simple tasks can be performed with the data that can be automated into a workflow. […] Almost everything above that, requires people who know about the domain*
> (Interview 2, manager data centre, bioengineering).

Similarly, another respondent noted that their role as data management coordinator of a health science project was focused on one particular aspect of quality assurance:

> *Yes. I mean we are involved in quality control. I mean during integration of data from different partners, there's always a quality control step. Very classical for building this kind of integration, so I mean you would enforce explicit description of data attributes, so what kind of values they should contain, what kind of extreme values are allowed, what does an expected range mean, what is the expected missing values and so on. So classical criteria for data quality.*
> (Interview 1, project manager at software company, health).

He distinguishes this kind of quality control from the evaluation of the scientific value of data as a quality aspect that the researchers are responsible for:

> *That's another quality aspect. Like the way you produce the data, is that meaningful? And in that sense, we are not really involved because we are sort of not judging the way an experiment has been designed and executed. That's still traditional sort of peer review. You have a description and then basically everyone has to judge on their own. If you have real open access, you could of course start to enable feedback*
> (Ibid.)

[117] Data Seal of Approval, "Data Seal of Approval: an overview", no date, http://www.data-archive.ac.uk/media/57322/dsa_overview.pdf

[118] Deutsche Initiative für Netzwerkinformation, *DINI-Certificate Document and Publication Services 2007*, Göttingen, September 2006. http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf

[119] The Center for Research Libraries, *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, Chicago, IL, February 2007. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Nonetheless, appropriate peer-review methods for research data can increase the trustworthiness and value of individual datasets and strengthen research findings. As highlighted in the APARSEN project: "peer-review processes have different functions depending on the respective participant; the filter function is of priority with regard to the potential reader; the concern of a discipline is to improve the publication; the most important aspect for an author in the case of successful publication is his reputation".[120] Therefore the review process of research data has important implication for scientific disciplines, their e-infrastructures (data repositories) and institutions.[121]

Although it holds a central place within the scientific communication system, thus far, peer-review and editorial processes generally exclude an assessment of research data.[122] Nevertheless, several publishers and journals contribute to the review of data. Various journals require data to be made available, and some also require their reviewers to check the data.[123] An increasing number of publishers demand research data supporting publications to be made openly accessible in certified repositories, and some also require their reviewers to assess the data. Linking to data repositories offers some advantages over including data with the publication in non-actionable appendices. It is not always possible for journals to ensure that data sets adhere to accepted standards, have adequate metadata and are largely error free. The use of online publications appendices is also not ideal because related data are usually not peer reviewed and much of what is regarded as 'supplemental material' is not permanently archived and thus can become inaccessible over time. In order to address this particular problem, the digital repository Dryad directly links to journals and where appropriate to and from select specialised data repositories (e.g. GenBank).[124]

Data peer review is a key focus of a relatively new type of publication: the data journal. Data journals focus on publishing articles that discuss datasets, which are openly available in (certified) repositories, in terms of acquisition, methods, processing etc. Articles published in such journals describe the data acquisition process and include a discussion of the considerations regarding the experimental design.[125] They do not provide analysis or results. Nevertheless, the articles undergo peer review, as do the underlying data. These types of publications draw attention to the significance of research data as independent publication objects, as well as to their quality and ability for reuse.

According to Mayernik et al. data peer-review and the processes used can vary by the kind of publications or resource being reviewed.[126] They make a distinction between: 1) data analysed in traditional scientific articles, 2) data articles in traditional scientific journals, 3)

[120] APARSEN, op. cit. 2012, p. 13.

[121] APARSEN, op. cit., 2012

[122] Exceptions include Data Papers and Ecological Monographs of the Ecological Society of America, the Earth System Science Data Journal, BioInvasions Records and Data Sets in Ecology.

[123] See for example: Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, David Remsen, Vincent Smith, David Shotton, *Pensoft Data Publishing Policies and Guidelines for Biodiversity Data*, Pensoft Publishers, 2011.

[124] Dryad Digital Repository, no date. http://datadryad.org/

[125] Gorgolewski, Krzysztof J., Daniel S. Margulies, and Michael P. Milham, "Making Data Sharing Count: A Publication-Based Solution", *Frontiers in Neuroscience,* Vol. 7, No. 9, 2014. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3565154/

[126] Mayernik, Mathew S., Sara Callaghan, Roland Leigh, Jonathan Tedds, Steven Worley, "Peer Review of Datasets: When, Why and How", *Bulletin of the American Meteorological Society*, Early Online Release, May 7 2014.

open access repositories and 4) datasets published via articles in data journals. The authors highlight a difference between journal-based data publications and data repositories with the latter focusing on the technical aspects of datasets to secure management and curation. Yet, the authors identified some commonalities in relation to data peer-review and quality assurance processes of data. Such common elements include:

- the need for data accessibility (via data centers or repositories);
- the provision of adequate information for the dataset to be reviewed;
- clear guidelines for data peer-reviewers on how to perform the data review and what characteristics need to be examined.[127]

Open Context, the organization featured in our archaeology case study, publishes data contributed by researchers through a process that includes reviewing, editing and aligning data to standards. Open Context publications are thus expected to "complement and enhance conventional publications through the dissemination and preservation of rich digital data and media".[128] A professional editorial member of staff monitors contributions to make sure that contributions stem from field research programmes. It also offers additional services to enhance the quality of the data published, including version control.[129]

Large-scale consortia often have established their own internal data review mechanisms, often consisting of multiple layers of automated checks and peer review. Before data are published or shared they have to go through several quality assessment steps. One example of such an internal multi-layered review process comes from the particle physics case study, in particular from the ATLAS experiment. "At about 140 million electronic channels and an event rate of 10^5 Herz it is essential to monitor the ATLAS hardware and the quality of data in an effective manner".[130] The ATLAS Data Quality Monitoring Framework (DQMF) involves automated analysis of monitoring data through user-friendly algorithms. In the online environment the information can be used to overcome problems, while in the offline environment the results are used to assess the quality of the reconstructed data.[131] In one of our interviews a data base coordinator in the particle physics case study described the multiple layers of quality assessment:

> *Already from the online system, at the time of acquiring the data. There are checks, tasks which run on the data as they come in or samples of the data as they come in and check the quality. They look if all the data preparation is normal for example. Measure the temperature of the electronics and detector and check if there is a spike or something going wrong. They measure the gas flows for the gas systems. They measure the currents in the power supplies and things like this. A very low level technological checks. Then there are further checks which means the events are reconstructed partially on line and then fully off line, just after being taken, within 24 hours usually. […] Everything is checked to find out if there is any anomaly in the behaviour of the detectors, if any part has switched off, you never know. […] In general we analysis between 95 and 98 percent of the data that we collect, one or the other. […] So at the end data are processed through calibration procedures, which*

[127] See also APARSEN, op. cit., 2012

[128] Open Context, no date. http://opencontext.org/about/publishing

[129] Ibid.

[130] Corso-Radu, A., S. Kolos, H. Hadavand, R. Kehoe, M. Hauschild, "Data Quality Monitoring Framework for the Atlas Experiment at the LHC", no date. http://www.physics.smu.edu/web/research/preprints/SMU-HEP-07-15.pdf

[131] Ibid.

> *apply conversions of constants, because what we measure are currents and positions from the detector when particles cross them. […] The calibration parameters are applied to what we call the raw data, the initial data and we end up with what we call the analysis data. […] This is the input for the physics analysis. And then for each analysis people still can't check the quality of the data they get, because it can be formally correct from the point of view of the detector hardware but make no sense from the point of view of physics. […] So this data are cleaned further and another few percent are like cleaned away.*
> (Interview 2, database coordinator, particle physics)

Another layer of quality assessment when the data are analysed:

> *First of all there are guidelines and general rules of how to deal with data; how to apply calibration corrections and what to do. There is a general analysis framework in which all the software is run so that these checks, at least formally, all the procedures are followed. When it comes to the analysis itself, there is an analysis group which is formed to analyse, to do a given analysis, search for a certain particle or to measure a certain effect. Then when this analysis is close to the end, so close to writing the paper, the collaboration forms what is called an editorial board. The editorial board are different people from a different analysis group. Still from within the collaboration. The editorial board does a review of the analysis procedure, checks all the results, reviews the paper while it is written, the publication in preparation. So there are many interactions between the editorial board and the analysis team. In the end they will give the go ahead for the publication. At this point the paper […] goes through the whole collaboration for comments. […] At the end when everybody is happy everybody, everybody means the editorial board and the analysis team, the paper is sent to a journal for publication.*
> (Ibid.)

Data review is thus a part of the structure of the ATLAS experiment. Data is not openly accessible until various checks, corrections and processing steps have been performed. When asked whether procedures would change if the data the experiment produced would be made more openly available, he responded that there should be no major changes, because there had to be a scientific check. Internal quality assessment procedures are an advantage of longer running collaboration, in which such procedures can be established and fine-tuned.

In sum, peer review can take place a various stages of the data production and curation process and can involve several stakeholders. It can take place after the data has been processed and analysis are published, or at an earlier stages through internal review processes or through openly accessible data repositories or data journals.

A final strategy that we highlight here is the adoption of Data Management Plans (DMP). DMPs have become a more commonly used means to encourage researchers and research groups to address the integrity and quality of data. Funders, universities and data centres are increasingly encouraging or even mandating researchers to develop a DMP at the beginning of their research projects.[132] Such a plan should specify, among other conditions, how the researchers intend to address the quality of research data.

Consistent progress is observed in the United Kingdom, the United States of America and Australia across national funders, agencies and Research Producing Organisations (RPOs) in terms of data management policies. UK research funders, namely the Research Councils (RCUK) and the Wellcome Trust have formulated policies as well as a range of tools to help researchers and institutions give shape to their data management.[133] While not all Research Councils in the UK require DMPs for the research they fund, they nonetheless require applicants and institutions to state explicitly how they will address the issue of access to and management of the data generated by publicly funded research. In fact, the EPSRC requires institutions it funds to have roadmaps and be able to comply fully with its requirements by May 2015.[134] The ESRC requires that individual applicants provide in their DMPs information on planned quality assurance and back-up procedures, responsibilities for data management and curation.[135] Moreover, the council expects "research data to be accompanied by high quality metadata to provide secondary users with important additional information, for example, the origin, circumstances, processing/analysis and/or the researcher's management of the data".[136] Effective in 2011, the US National Science Foundation requires grant applicants to submit a DMP, which also undergoes peer-review during the proposal evaluation. The DMP is a supplementary document to the NSF proposal that should in general provide information on the data generated by the project and the plan for managing these data. If no such plan is included the NSF will return the proposal without review. Most research councils and funders in the US and the UK tend to provide general guidelines about how to ensure the quality of data for re-use, rather than providing specific standards.

The European Commission has recently published the guidelines for the Open Access to Research Data Pilot, which applies in seven areas in the Horizon 2020 funding scheme.[137] In its "Guidelines on Data Management in Horizon 2020", the European Commission emphasizes the significance of data sharing, security and quality, as factors that enable intelligent reuse of research data.[138] A DMP is required with the grant applications for the specific areas and is evaluated under the section "Impact". The DMP should be an evolving document "outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open and how it [sic] will be curated and preserved".[139] Researchers are expected

[132] E.g. in: University of Edinburgh, "Research Data Management Policy", 24 January 2014. http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy

[133] The Digital Curation Centre (DCC) in the UK provides an overview of both funders' data policies (stating requirement for data plan, expectations on data sharing and available support such as guidance and data centres) and institutional ones, Available at: http://www.dcc.ac.uk/resources/policy-and-legal

[134] EPSRC, "EPSRC policy framework on research data", no date. http://www.epsrc.ac.uk/about/standards/researchdata/

[135] ESRC, ESRC Research Data Policy, 2010. http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

[136] Ibid., p. 3

[137] These areas are Future and Emerging Technologies; Research Infrastructures (e-Infrastructures); Leadership in enabling and industrial technologies-Information and Communication Technologies; Societal Challenges: Secure, Clean and Efficient Energy- smart cities and communities; Climate Action, Environment, Resource Efficiency and Raw materials-except raw materials; Europe in a changing world: inclusive, innovative and reflective societies; and Science with and for Society.

[138] European Commission, op. cit. 2011, (p.2, and annex 1 and 2, pp.5-6)

[139] Ibid, p. 3.

to provide reference to existing suitable standards of the discipline and, in the absence of such, how and what metadata will be created. Additional information is expected in relation to how research data can be discoverable, accessible, assessable and intelligent, useable beyond the original purpose for which it was collected and interoperable to specific quality standards. The document should be part of the deliverables of projects early in the project process.

From the above it is clear most disciplines quality assurance practices, with regard to the technical quality of digital data in particular, have been established. Researchers, data repositories, data centres, and publishers play actives roles in ensuring the quality of data through manual and automated mechanisms. Yet, several aspects of these practices are still in the early stages of development. The majority of stakeholders involved in the quality assurance process, for example, acknowledge the positive contribution of a peer-review process for research data, yet its implementation is still in embryonic stage. Quality aspects related to data re-use and retention is examined more on the basis of the technical quality rather than on content. Moreover, data quality is increasingly addressed in the context of DMPs, but so far few institutions have developed strategies to evaluate compliance. As the next section shows, there still remain some barriers in further developing these aspects.

3.2 Remaining Barriers and Challenges

The most important barrier for institutions in further developing data quality assurance processes probably lies in the blurred distribution of responsibilities among stakeholders. As Pearlman et al. succinctly point out "data stewardship is critical to the longevity and sustainability of data sharing and management throughout the data lifecycle, but it is unclear where the responsibilities for this effort lie".[140]

Of equal importance are issues of fragmentation in the scientific landscape. The difficulties presented by the heterogeneity and interoperability of data, such as the different ways of formatting, storing, operating and standardizing data, have been discussed in the second RECODE report.[141] Fragmentation, however, presents additional challenges for institutions in terms of the distribution of scarce resources. Data centres, institutional repositories and publishers that serve research groups from multiple disciplines have to make decisions about the extent to which they invest in ensuring the quality and integrity of data sets from various disciplines. For many institutions it is too expensive to employ several data librarians or data scientists who are specialised in particular subjects and therefore capable of quality assessment. Moreover, as reviewing practices are community-specific and dependent on the form of data, it is impractical for repositories or publishers to formulate recommendation about data quality for each discipline and each data type.

The role of researchers is central in relation to data quality as the responsibility during the first stage of the data life cycle rests with them as part of their overall responsibility for research that is valid, accurate and ethical. Yet, engaging researchers in developing quality assurance practices poses a challenge. In particular, issues related to data management might seem counter-productive to researchers as they feel it will require significant work to get data

[140] Pearlman, Jay, Albert Williams III, and Pauline Simpson (eds.) "Report of the Research Coordination Network RCN: OceanObsNetwork. Facilitating Open Exchange of Data and Information", NSF/Ocean Research Foundation, May 2013.
https://darchive.mblwhoilibrary.org/bitstream/handle/1912/5937/RCN_Open_data_report_final.pdf?sequence=1

[141] cf. Bigali et al., op. cit., 2014.

into a format that could be used by others. Furthermore, while there is an abundance of statements, guidelines and frameworks regarding 'good scientific conduct' they usually do not refer to quality of data for re-use.

An additional issue concerning the role of researchers relates to the data peer-review process. While the data peer-review process is expected to benefit the researcher by increasing the visibility of a scientist's work and his/her citation rate, the main barrier lies in the availability and willingness of referees to undertake such tasks. This is a non-negligible problem given that it is already an issue for scientific publications and is further accentuated by the existence of few incentives for attracting referees.[142] Scientists have even expressed their concern about the possibility to review data for all its quality.[143]

Another remaining challenge is the integration of the integrity and quality of data in existing evaluation systems at institutions. Although data management plans often require researchers to address the quality of data, funding bodies have yet to develop strategies on evaluating and rewarding researchers and research groups for their efforts on data management. The UK Data Service offers guidelines on data documentation, formatting, submission and data quality, which state 'We apply various quality control checks to all research data whilst we process them for archiving into the data collection. The level of quality control depends on how much additional value is to be added to the data, based on anticipated future usage.'[144] It is however not clear whether data quality is reported back to funders, or whether funders do monitor data, aside from merely monitoring whether data has been submitted.

Finally, the selection and retention of data continues to pose a challenge for institutions. As the number of data sets grows data centres and repositories have to develop strategies to maintain and curate their diverse data collection, which requires domain knowledge and specific expertise. The Research Information Network (RIN) maintains that the curatorial role of data centres is twofold, 'first, ensuring that individual datasets are academically "good" (as much as it can) and second, ensuring that it creates and preserves collections which can be a useful starting point for new research'[145]. This may require additional staff that has the required disciplinary knowledge to evaluate the value and quality of data sets. At the institutional level, research data producing institutions will need to make overall provisions and earmark funds for the cost of managing research data, including the need to secure their quality and integrity. However, the deep knowledge of archive or data centre staff cannot always be ensured, placing the responsibility for quality back on to the data producer.

This section outlined some of the remaining barriers that institutions face in further developing quality assurance processes, including the fragmentation of the scientific landscape, engaging researchers, making data quality part of existing evaluation systems and the lack of strategies for data retention and selection. The next section will discuss some developments that contribute to overcoming some of these barriers.

[142] Costello, Mark J., William K. Michener, Mark Gahegan, Zhi-Qiang Zhang and Philip Bourne, "Biodiversity data should be published, cited and peer-reviewed"*, Trends in Ecology and Evolution*, Vol. 28, No. 4, August 2013, pp. 454-461.

[143] APARSEN, op. cit., 2012

[144] UK Data Service. "Quality assurance", no date. http://ukdataservice.ac.uk/manage-data/format/quality.aspx

[145] Research Information Network, *Data centres: their use, value and impact*, JISC, 2011. http://www.jisc.ac.uk/media/documents/publications/general/2011/datacentres.pdf

3.3 Working towards solutions

A promising solution to some of the remaining barriers is the further development of research cultures in which data becomes an integral part of the scientific and scholarly evaluation system. The various initiatives that support data publications and citations are good examples of this. Institutions, however, can play a bigger role in facilitating such practices. Research institutions and scholarly societies can support researchers in producing and maintaining high quality data sets through training and education, as well as acknowledging researchers' data management and review efforts in promotions and awards. Journals can promote and enforce data citation practices and incorporating data quality requirement in their editorial policies.

Other stakeholders have also begun to explore ways of ensuring the quality of data sets and repositories. Some journals contribute to quality assurances of research data by developing standards, methods and criteria for reviewing data effectively. They formulate requirements regarding, for instance, the documentation of data and incorporate such requirement in their editorial policies. Furthermore, in recent developments, increasingly more journals require that research data supporting their publications should be openly accessible. A recent example is the new PLOS policy, effective March 2014, which requires research data that supports publications to be openly available through an appropriate repository.[146] The policy discusses the fact that research data should be recorded and deposited according to disciplinary standards, and provides to this end extensive references to discipline-specific bodies and links to their requirements for data documentation.

Data journals, publishing information on data acquisition, methods, processing of specific datasets and other data related issues, help to increase the standards for the quality of research data. They also help to establish good practices, such as referencing data and making them available through accredited repositories. Publishers in both STEM sciences and the social sciences and the humanities are increasingly turning towards developing this new kind of publication.[147] Data papers, allowing researchers to publish their datasets as citable scientific publications, have become a way to give credit to researchers sharing their data.

Various libraries and data centres have experimented with new mechanisms to enhance data quality, for instance through providing researchers platforms to discuss data sets or offering tools for Altmetrics. Altmetrics is the study and use of scholarly impact measures based on activity in online tools and environments. According to the Altmetrics manifesto 'altmetrics expands our view of what impact looks like, but also of what's making the impact'.[148] Recent developments include the partnership between Scopus and Altmetric.com whereby altmetrics data will be included alongside the traditional bibliometrics in the Scopus interface and the merger between Elsevier and Mendeley.[149]

With regard to distributing responsibility, some institutions have taken the lead in explicitly addressing who is owner or steward of particular data sets and at what stage. University College London's (UCL) Research Data Policy, for example, has highlighted the different lines of responsibility (data creators, UCL research data and network service executive,

[146] Bloom, Theo, "Data Access for the Open Access Literature: PLOS's Data Policy", Public Library of Science, 12 December 2013. http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/

[147] Ubiquity Press is well-known publisher of data journals in the Humanities http://www.ubiquitypress.com/

[148] Roemer, Chin, Robin and Rachel Borchardt, "Institutional Altmetrics and Academic Libraries", *Information Standards Quarterly*, Vol. 25, No. 2, Summer 2013.

[149] Ibid.

director of UCL library service and UCL records manager, research information IT services group, vice provost and provost) in its data policy.[150]

3.4 CONCLUSION

This Chapter looked at the issue of maintaining and evaluating the quality and integrity of research data. Ensuring the quality of research data is a prerequisite to achieving the promises of open access to research data. In many disciplines, formal and informal mechanisms are already in place to check the quality of research data produced. Research communities may perform several review processes, manually and automatically validating data at various stages in the data life cycle. Open access to digital research data often requires additional mechanisms, for instance, to ensure that data are re-usable and interpretable. Several stakeholders play an active role in these processes, including data repositories and centres, consortia, and publishers.

The above analysis shows that institutions have focused primarily on developing strategies to ensure the technical quality of data deposited (e.g. are the correct formats used, is the metadata complete, etc.). Less effort has gone into establishing review practice that focus on the scientific value of data, partly because it is a time consuming and difficult task. An important barrier that has to be overcome in order to move forward is the lack of incentives for researchers to engage in data review processes. Few institutions acknowledge the time and effort that these processes require. The Chapter also showed that long-term perspectives have yet to be further developed. Issues such as how to deal with increasing volumes of heterogeneous data or how to deal with data selection and retention have been less of a priority for many institutions. The distribution of responsibility between various stakeholders is also an area that requires further attention. Nevertheless, recent developments offer some promising solutions. Data journals; publishers implementing peer review processes and developing standards; methods and criteria for the review of research data; and the development of new mechanism to assist researchers in evaluating openly accessible data are some examples that contribute to the quality of open access research data.

[150] Paul Ayris, “UCL Research Data Policy”, version 4.0, 2 August 2013.

4 EDUCATING AND TRAINING RESEARCHERS AND OTHER RELEVANT STAKEHOLDERS

One reason why unrestricted data sharing is still not the norm in most disciplines is that many researchers and other professionals lack the knowledge and skill to make their data publicly accessible or to use existing data sets.[151] They may not know where to find the data they are looking for, or how to access and use the data. It requires considerable skill to do these things. A researcher would have to know how to work with various formats and software tools and how to efficiently search for the data they need. Researchers also need to become skilled in preparing their data for digital publishing and re-use. They have to know about mark-up languages, standards, metadata, storage possibilities, and other technical requirements. If they wish to add value to their own data, then they also have to learn how to link their datasets to other data sets and publications.

Besides researchers, other stakeholders need to develop and maintain their knowledge and skills to in order to make and keep research data openly accessible. The technical staff, (data) librarians and data scientists have to acquire technical skills and stay abreast of technological and policy developments to support researchers effectively in their data management activities. Moreover, professionals working with research data, as well as researchers, may encounter legal and ethical issues regarding open access that requires a certain expertise. Finally, the leadership and management of institutions have to stay up-to-data with policy and infrastructural developments in the area of open data to properly perform their role.

In order for open access to take root in more disciplines, institutions therefore must invest in education, training and skills development. Several reports and studies have underlined the need for the development of skills in data management and re-use and the role institutions have in this.[152] LERU calls it a “key-enabler of open data”.[153] The Royal Society holds that “principles of data management should be an integral part of the training of scientists in the future”.[154] In one of its reports the Opportunities for Data Exchange (ODE) project argues, “Improving the skills and understanding of researchers in data management is essential. Training should begin in the institutions that train researchers, at the outset of postgraduate study at the latest, possibly even earlier”.[155] These reports all point out that a solid education in data management will ensure the trustworthiness and usability of research data.

The reports also make clear that education and training are considered to be the responsibility of the various stakeholders: governments should set new policies for data management skills to be taught at university and secondary school level[156], funders should educate their grantees on data management and institutions, with the help of data centres, libraries and IT departments should provide training and educate their researchers and other staff on data

[151] Lyon, op cit. 2007.

[152] Dallmeier-Tiessen, et al., op. cit., 2012; Pryor, Graham and Sarah Jones and Angus Whyte. *Delivering Research Data Managment Services: Fundamentals of good practice*, Facet Publishing, 2013.; Lyon, Liz, “The Informatics transform: Re-engineering libraries for the data decade”, *The International Journal of Digital Curation*, Vol.7, No.1, 2012.

[153] Achard, Pablo, Paul Ayris, Serge Fdida, Stefan Gradmann,Wolfram Horstmann, Ignasi Labastida, Liz Lyon, Katrien Maes, Susan Reilly, Anja Smit, *LERU roadmap for Research Data*, League of European Research Universities, 2013, p. 29.

[154] The Royal Society, op. cit., 2012, p. 63.

[155] Dallmeier-Tiessen, et al., op. cit., 2012.

[156] European Commission, op. cit., 2010

management[157]. Researchers should also serve as mentors to early investigators and students who are interested in pursuing data sciences.[158] Even within institutions responsibilities are distributed. Providing training for open research data within universities, for example, is a team effort divided between IT departments educating researchers and librarians about the technical requirements for open research data; the library supporting data management and discovery; university departments and research groups developing new courses within postgraduate education as a core academic competency; faculty administration raising awareness; and data scientists collecting and making data ready for reuse.

Providing training for researchers and other stakeholders can be challenging. Different disciplines have their own methods, standards, repositories and data sharing norms that require different forms of training. Standardized educational programmes may not be sufficient even if they are targeted at particular disciplines. The various other stakeholders involved also have different training needs. Moreover, open research data is a new area in which policy and technological developments occur at a rapid pace. In the following section we will again look at some of the policies and strategies that institutions have developed to address some of the issues described.

4.1 Policies and strategies

In some disciplines, such as particle physics, genetics or social geography, digital data management training is already an integral part of the (post-graduate) curriculum, but for many disciplines this is a relatively new area. Disciplinary-focused training programs in data curation at universities are scarce.[159] Universities, departments, research groups and research institutes are only just beginning to gain experience in providing the appropriate courses, workshops and tools to support researchers as well as librarians, information specialists and other staff in their data management activities. Important developments in this respect are the increasing number of training programmes and materials that data centres, libraries and research consortia offer researchers as well as the establishment of professional training programmes for data curators and information specialists.

The available training for data management and curation is mainly given by dedicated national bodies, libraries, information science schools or by data centers in the UK, The Netherlands and the in the US. The UK Digital Curation Centre (DCC) plays a leading role in offering training for practitioners in need of resources on data management. The DCC offers workshops in data management and also short intensive three-day courses or half-day courses for absolute beginners. The courses are structured around the DCC curation lifecycle model[160]. They also offer information and a range of tools to help researchers prepare their data management plans. The Data Archiving and Networking Services (DANS) institute in

[157] Jones, Sarah, Grahamn Pryor, and Angus Whyte, *Developing Research Data Management Capability: the View from a National Support Service,* iPres Conference 2012.
http://www.dcc.ac.uk/sites/default/files/documents/institutional-engagements/Institutional-engagements-iPres.pdf

[158] National Science Board, op. cit., 2005.

[159] Creamer, Andrew T., Myrna E. Morales, Donna Kafel, Javier Crespo, Elaine R. Martin, "A sample of Research Data Curation and management Courses", *Journal of eScience Librarianship*, Vol. 1, No. 2, article 4, 2012; Walters, Tyler, "Data curation program Development in U.S. Universities: The Georgia Institute of Technology Example", *The International Journal of Digital Curation*, Vol. 4., No. 3, 2009; Lyon, Liz, op. cit. 2012, p.132-135

[160] Keralis, Spencer D. C., *Data Curation Education: A Snapshot*, CLIR Publication No. 154, Council on Library and Information Resources (CLIR), Washington DC, 2012.

the Netherlands provides various workshops, training courses and guest lectures for researchers and students in the humanities and social sciences at various Dutch universities and research institutions.[161] In the US, organizations like the Inter-University Consortium for Political and Social Research (ICPSR), a unit of the Institute for Social Research at the University of Michigan, is also offering its members short courses on how to manage and curate research data.[162] In Australia, ANDS has acquired considerable experience in providing support and training courses for researchers and students across all disciplines.[163]

Libraries are also seen to be organisations well-suited to play a greater role in guiding researchers in their data management practices. Libraries have a long tradition of subject liaisons who work closely with researchers to understand their needs.[164] In that capacity they could be the “last mile” of research data infrastructures – “the part of the network that will provide connections between the systems and the researchers, and ultimately, to new users of the data”.[165] The Canadian Association of Research Libraries[166] also sees an important opportunity for libraries in developing new and valuable services. Librarians can take on a role as data stewards for various stakeholders and especially for researchers. This can take several forms, including organizing conferences, passing out literature, develop courses, web tutorials, advocacy programs tailored for specific research communities.

Indeed, several libraries and digital libraries, such as the Edinburgh University Library or the California Digital library, have started to develop this new intermediary role. As liaisons, they help researchers deposit their data at the point of data creation. They advise about standards applicable to the needs and create curation plans for the whole life cycle of the data in compliance with funder mandates. They offer seminars and workshops or individual tuition to research- and professional staff.

Training for using data and data management also takes place within research consortia. A consortium may organize workshops to instruct participating researchers about how data needs to be stored and what formats to use, and how datasets can be accessed. Within the Health case study one of the partners in the EvA project organizes workshops about data management and the technical requirements of the project:

> *[W]hat we do is we provide training or try to create awareness in the sense that we have dedicated sessions on data knowledge management within the projects, a little bit depending on the importance of the issue that might go into the direction of a separate workshop, or it might be just a smaller part of standard workshops.*
> (Interview 1, project manager at software company, health)

[161] Data Archiving and Networked Services, op. cit., no date.
[162] Inter-University Consortium for Political and Social Research, ICPSR, no date. https://www.icpsr.umich.edu/icpsrweb/landing.jsp
[163] Australian National Data Service, “Training and tutorials”, no date. http://www.ands.org.au/training/index.html
[164] Gabridge, Tracy, The *last mile: Liason roles in curating science and engineering research data, Research Library Issues,* A bimonthly report from ARL CNL and SPARC, August 2009. http://old.arl.org/bm~doc/rli-265-gabridge.pdf
[165] Ibid., p. 15.
[166] Shearer, Kathleen and Diego Argáez, *Addressing the Research Data Gap: A Review of Novel Services for Libraries*, Canadian Association of Research Libraries, 2010. http://carl-abrc.ca/uploads/pdfs/library_roles-final.pdf

Training is an integral part of the ATLAS experiment at CERN, and researchers are educated about how to work with the data as part of their general post-graduate training. Various tutorials and workshops are organized both centrally and by the local physics groups. One interviewee explains:

> *There are tutorials, which are organised centrally and distributed. The central tutorials provide information about the software framework. How to use it, how to plug in your own algorithmic code to do the analysis. How to move the data for the analysis. Also how to get to access to all the data of the experiment and do more complicated things. [...] Then different countries, each country tends to organise their tutorials in their own language. For their people, in addition, so people have a choice whether to take the global tutorial or a more specific one. And some physics analysis groups organise tutorials on their analysis tools. There are some very sophisticated statistical analysis tools which are now developed and used for some of the analysis, to check the statistical and systematic errors, cross correlate them to some different channels to compare simulated events and so on. Some physics groups organise some specific tutorials on this type of the very specialised software. Students that come in used to do a PhD for example will take probably a few months to get familiar with the whole environment.*
> (Interview 2, computing coordinator, Particle Physics).

He also notes that the experiment has been running since 2009 and start-up problems have mostly been solved.

Another target audience, besides researchers, for development of training programmes are librarians and information specialists. Institutions have recognized the need to invest in skill development of staff involved in research data services and several have taken first steps in this direction. In Europe, the recently established Research Data Netherlands, a collaboration between the 3TU data centre and DANS institute, provides a course for data supporters and information resources.[167] In the UK the Digital Curation Centre (DCC) lists several courses in "data management and curation education and training".[168] Several UK and Swedish universities offer face-to-face or distance courses for librarians and other information specialists. In the US there were at least a dozen institutions teaching courses on data management and curation at ALA accredited library and information science (LIS) programs in 2012.[169] There are more digital curation programs emerging and the trend is towards allowing open enrolment for scholars and professionals outside the library circuit using a pedagogic model based on a collaborative model of teaching between librarians, LIS educators, research faculty and data centre specialists. Yet, in 2012, only five Library and information science schools offered graduate certificates explicitly in data curation. All but one restricted its enrolment to LIS students.

As various reports and studies have pointed out, open access to research data generates a need for skilled data scientists. For example the US National Science Board observed that:

> *New jobs and areas of expertise are emerging in response to the evolving role of data in science and engineering, yet opportunities for education, training, and workforce*

[167] Research data Netherlands, "RDNL", no date. http://www.researchdata.nl

[168] DCC, "Data management and curation education and training", no date. http://www.dcc.ac.uk/training/data-management-courses-and-training

[169] Creamer, op. cit., 2012

> *development are not fully recognized and supported. The proliferation of shared, interoperable data creates new computational and data -enabled science and engineering research opportunities that require the support of trained experts and researchers.*[170]

Data specialists are described as a cross between informaticians and librarians and are currently "often informatics trained scientists expert in the tools and processes of data management".[171] They have to be skilled in such things as migration of data, building ontologies, metadata production etc. Data scientists or curators can fulfil a role in collecting, describing and connecting data, as well as developing standards in collaboration with researchers and based on understanding the ontology of a domain[172]. The Royal Society acknowledges that data scientists are crucial in supporting researchers and institutions with data management issues. As an example of this they refer to how the National Science Foundation in USA allocated funding for undergraduate training in complex data and encouraged universities to develop graduate programs in Big Data[173].

In their yearly review of top trends in academic libraries the Association of College & Research Libraries (ACRL) finds that increased focus on open data and data management is the new top trend. Therefore universities like North Carolina State and Stanford are rolling out graduate and certificate programs to prepare professionals for careers related to the analysis and manipulation of Big Data. This, in turn, will place new demands on the skill and training of library staff in the use of complex data[174].

4.2 REMAINING BARRIERS AND CHALLENGES

Institutions are giving increasingly more attention to education and training, as the previous section shows, but some barriers and challenges remain. We found during our case study interviews and the RECODE workshop that one significant problem that institutions face is that only a minority of researchers are interested in the data management courses offered, unless they are an integral part of their research activities. Another remaining barrier is the unclear distribution of responsibility. There is confusion about who is responsible for what, especially in academia, in part because the various institutions may not be sufficiently equipped to provide education and training. This is hampering the development of necessary training programmes for different professional players.

According to Halbert there is a "daunting array of barriers that hamper the prospects for effective research data management practices and programs".[175] One of the most striking barriers, according to him, is the lack of professional preparation. "Yet, almost no one within the academic community receives systematic professional training and certification in the

[170] National Science Board. *NSB Digital Research Data Sharing and Management: Statement of Principles,* 2011. http://www.nsf.gov/nsb/committees/dp/principles.pdf

[171] The Royal Society, op. cit., 2012, p. 64.

[172] Lyon, op. cit., 2007, p.54.

[173] The Royal society, op. cit. 2012, p. 64.

[174] ACRL Research Planning and Review Committee, "Top trends in academic libraries – a review of the trends and issues affecting academic libraries in higher education", *College & Research Libraries News,* Vol. 75, No. 6 June 2014, pp. 294-302.

[175] Halbert, Martin, *Prospects for research data management, in Research Data Management, Principles, practices and prospects*, Council on Library and Information Resources, November 2013.

management of research data. Still worse [...] virtually no one in academia perceives that they have a professional responsibility or mandate for research data management functions”.[176]

In so far as libraries have taken up the responsibility; they are not fully equipped to address all data management issues. As data stewards or liaisons, libraries are facing new roles and responsibilities in the current research and technology culture, but they will have to overcome the current skill gap. A survey by Cox and Pinfield of libraries in Higher Education Institutes in the UK revealed that library staff’s skill gaps were identified by respondents as one of the key barriers for fulfilling RDM obligations. The specific skills needed were data curation skills (mentioned by nearly 90%), technical IT skills and knowledge of research methods. About 40% also recognised the need for disciplinary knowledge.[177] In earlier work, Auckland et al had identified the following skill gaps, based on a survey of library staff in the UK:

- Advising on preserving research outputs;
- Advising on data management and curation;
- Support complying with the various mandates of funders;
- Metadata advice and advocacy;
- Assisting locating sources of research funding;
- Developing metadata schema.[178]

One of our case study interviewees with significant experience in developing and implementing open software and open data infrastructures argued that because of the significant skill gaps, in particular with regard to the technical aspects of data management, libraries can only fulfil certain aspects of data management support. The lack of sufficiently technically skilled staff to advise and educate researchers in how to digitally publish their data such that they can be easily found, interpreted and reused forms a barrier. She holds that there are important differences between open publications and open research data that are often overlooked.

> *Our library information management people, we force them to go out into the centres and to the labs and group and they discovered there is all this data, all different formats […] These people are coming from a place where it is a book. I know it might be an electronic book. The metadata associated with a paper is like some key words, the author and things like that and it is in this particular journal but it is actually not that much you are interpreting by reading it. Whereas with data, you have to interpret it by knowing what the heading meant and calling it n1 n2 n3 doesn’t help. It is a different level and so that’s education.*
> (Interview 4, professor computer science, bioengineering)

She notes in this regard:

> *The idea of open access data just being treated as open access to papers is just a flawed argument. It is not true, data should be treated more like open access to software, it’s much more complicated. Much more evolving, much more multi-dependent kind of item. […] Libraries have typically taken on board the research data management activity, as indeed has ours. I personally think that the libraries are not the people who can do it, these things are not a one size fit all.*

[176] Halbert, op. cit., 2013, p. 6.
[177] Cox, op. cit, 2013, p. 8.
[178] Auckland et al., op. cit., 2012, p.3.

(Ibid.)

Another problem is making the connection between the relatively new and quickly evolving field of digital data management and the everyday practice or doing research. Cox and Pinfield point out that 'while librarians' information management skills may be relevant, it could be challenging to translate them to research data contexts (including metadata creation or good data housekeeping). They argue that it is like any area of specialist activity, complex and jargon ridden; there is a whole social world of organisations, projects, thought-leaders and key influencers, technologies, discourses, concepts and terminology that have to be mastered in order to be 'taken seriously'.[179] Another study showed that researchers are not familiar with terms like 'digital curation' and 'digital repository' and suspicious of policies that issue all sorts of requirements and mandates. They prefer advice that conveys a sense of purpose and assistance.[180]

Further progress in the area of training and skill development is also hampered by the balance of power. Librarians introduce and administer the institutional repository and the idea about open access with a great knowledge about scholarly communication issues but since they do not bring any funding into the university the library is mostly perceived as a service based unit without much influence.

4.3 Working towards solutions

Several institutions have taken initial steps to bring education and training in the area of open access and data management further. As illustrated above, training and skill development has become a priority in the push for open access and several initiatives have set out to address this challenge. In 2012 the EU-funded project Digital Curator Vocational Education Europe interviewed and made a survey targeting cultural heritage staff, librarians and researchers asking about the need for training in the field of digital preservation and curation. The survey showed that the training methods considered most suitable were small workshops a few days up to a week in length. The most pressing need for training was in digital preservation-specific and technical skills.[181] Now the ongoing project Facilitate Open Science Training for European Research (FOSTER)[182] will act on this knowledge and support different stakeholders in complying with the open access policies set out for Horizon 2020. It aims to strengthen training capacities, also in managing open data. The project will thus identify already existing contents and training activities that can be reused and repacked. It intends to offer workshops and training of trainers who can carry on further training and dissemination activities.

Besides the various initiatives to provide training by dedicated projects and organizations such as the FOSTER project and the DCC, good practices are emerging within universities. Brown and White describe how the University of Southampton, through collaboration with UK Research Data Service and involvement in projects like the Institutional Data

[179] Cox and Pinfield, op. cit., 2013.

[180] Freimna, Lesley, Catharine Ward, Sarah Jones, Laura Molloy, Kellie Snow, *Incremental. Scoping study and Implementation plan: A pilot project for supporting research data management,* University of Cambridge, University of Glasgow, July 2010. http://www.lib.cam.ac.uk/preservation/incremental/documents/Incremental_Scoping_Report_170910.pdf

[181] Engelhardt, Claudia, *The DigCurV review of training needs in the field of digital preservation and curation: An overview of the main findings*, DigCurV, 2013. http://www.digcur-education.org/ita/Risorse/DigCurV-2013-proceedings/Engelhardt-paper10

[182] FOSTER, "homepage", no date. http://www.fosteropenscience.eu/

Management Blueprint Project (IDMB), started to improve and formalize initiatives to support researchers at the university in managing their research data.[183] The decision was made to take a bottom-up approach based on researchers needs and a top-down approach to design requirements for an institutional policy and infrastructure. The approach was to develop an understanding of different disciplinary needs, partnership and cooperation with the researchers and their workflow, implement simple and low cost technical solutions and applications and focus on training and support. Six objectives were formulated to blend policy and infrastructure with local disciplinary perspectives:

- "implement the draft institutional research data management policy with an associated one-stop-shop of web guidance and data management planning advice;
- develop flexible support services and guidance for researchers extending across the research lifecycle;
- create and embed a range of training materials and workshops for postgraduates and early career researchers;
- enhance repository infrastructure to create comprehensive records of data outputs;
- scope options for storage and archiving including institutional structures, locally managed storage of small-scale outputs and a platform for sharing data;
- develop a suite of case studies to investigate multidisciplinary issues in depth including gathering granular evidence for cost analysis".[184]

Brown and White found that "researchers were open to new practice as long as it was researcher led, integrated into research workflow, reflective of discipline distinctions and supported by advice and training. Clarity over policy and responsive service support were essential".[185] It was also very important that the institutions at the university felt that they were in command of the investments and service support regarding data management without feeling compelled by a set of requirements. For data management planning service for researchers, a training programme was developed to engage with various groups from postgraduate researchers to senior scientists. Planning and realization of these courses, lectures, workshops and seminars were always done together with the researchers themselves. An evolving training programme for the services staff has also been designed and is under constant revision as the level of engagement by researchers and their expectations rise. Additional, automated and web tools have been developed for training purposes. For example automated tools to support minting of DataCite DOIs and web-based guidance to help interpret funders' requirements. Finally, Brown and White consider the university policy on research data management the most important component of the project.

Another example is the Orbital project at the School of Engineering at the University of Lincoln, UK. The project has proposed a set of recommendations in support of further development of their research data management structure.[186] The project underlines that researchers are very heterogeneous not only in terms of discipline, but also between individuals in the same team. It is, therefore, important to gain an understanding of the "culture" within any given set of researchers before considering how to influence their

[183] Brown, Mark and Wendy White, "University of Southampton: A Partnership Approach to Research Data Management", In Graham Pryor, Sarah Jones and Angus Whyte (Eds.), *Delivering Research Data Management Services: Fundamentals of good practice*, Facet Publishing, 2013.

[184] Ibid, p. 11.

[185] Ibid.

[186] Stainthorp, Paul, *An Engineering Research Data Management (RDM) Literature Review,* University of Lincoln, 2012. http://orbital.blogs.lincoln.ac.uk/files/2012/04/Literature-review.pdf

research data management behaviour. One of the recommendations is that for advocacy and training purposes interviews and surveys are developed to understand researcher requirements and behaviour. Such interviews and surveys should provide a basis for developing advocacy/training materials that will motivate researchers, as well as make them understand the obligations to institutions, funders and the public. Another recommendation is that guidance and advocacy materials concentrate on the academic benefits of research data management and institutional policies and that support material makes researchers aware of the importance of reflecting on appropriate research data management costs into funding bids.

Examples, such as these, show that development of training programmes and materials for open research data requires a top-down/bottom-up approach. Institutions should reach out to research communities to gain insights into the needs and practices of these communities, while such efforts need to be backed by clear institutional policies.

From our workshop in Riga it became clear that a focus on training alone might not be sufficient to persuade researchers to partake in training programmes on offer. A more successful strategy would be to concentrate on incentivising and enabling the open access practices and supporting researchers in making those happen. There is no reason to provide training if there is no demand for it. The focus should be on increasing the demand and, preferably, the need for training should emerge from the research culture itself. Physics, genetics, environmental sciences are ahead in managing open data because they have a culture and a history of working collaboratively and sharing data.

Research producing institutions can benefit from sharing experiences with other institutions that have already successfully implemented training programs. Similarly, libraries and data centres can benefit from collaborating in developing training programs for researchers and their own staff. Data centres tend to have considerable expertise in data management. Since 1976, CESSDA (Consortium of European Social Science Data Archives) has served as an informal umbrella organisation for the European national data archives. The CESSDA data archives and other similar subject data archives are in a good position to work with universities libraries and negotiate with archives on training. Libraries are generally institutions responsible for digital curation and preservation of print with a long experience of creating and applying metadata standards and retrieval services and with close contact with researchers, but their competence concerning research data still have to be proved. When the demand for provision of interoperable metadata, repository- and retrieval knowhow and training of researchers in data management starts to increase, researchers and data centres could also profit from library expertise in these areas[187].

4.4 Conclusion

Open access to research data requires specific skills and knowledge that have to be developed and maintained. As the Chapter showed, several institutions have taken up the challenge of educating and training researchers, librarians, information and data scientists and other professionals, building on existing and emerging digital data management practices. Libraries, data repositories, data centres and dedicated organizations play and important part in offering workshop, training materials and other kinds of support.

[187] Osswald, Achim and Stefan Strathmann, *The role of libraries in curation and preservation of research data in Germany: Findings of a survey*, IFLA conference, Helsinki, 2012. http://conference.ifla.org/ifla78

Nevertheless, several barriers have yet to be overcome, including distributing responsibility between stakeholders, engaging researchers and bridging the skill gap in libraries and data centres. Moreover, data management and curation skills have to be better and more commonly embedded in post-graduate education and new curricula and professional qualifications have to be developed. All the different stakeholders with their organizations will need to cooperate, as the barriers are multiple and complex. Funders and policy makers should clearly mandate data management and also earmark funds for training, infrastructure, data curation projects etc. Professional associations have to reflect on instigating new opportunities for training of professionals. Librarians, IT-specialists and research office staff from the universities need to collaborate with archivists and curators from data centres and vice versa. Finally, researchers must find new priorities regarding the importance of data management and need to find ways of recognising data management in appointment and promotion policies.

5 CREATING AWARENESS

A key issue for institutions in making data openly accessible is motivating researchers to publish and share their data. Many repositories, created to encourage open access publications as well as data sharing, remain largely empty.[188] Borgman points out that despite significant investments in and the promotion of data sharing, the "dirty little secret" is that not much sharing may be taking place. Relatively few studies, she notes, show consistent data release and data sharing seems to be concentrated in a few fields. "[L]ittle research data is [sic] circulated beyond the research teams that produce them, and few requests are made for these data".[189] Studies indicate that researchers are reluctant to share their data because they have various concerns ranging from being scooped to not being able to protect the privacy of their research subjects.[190]

As we have seen in previous RECODE reports, some of the concerns researchers have about sharing data are based on a partial understanding of what open access entails and what the possibilities and risks are, because technical skills and knowledge are lacking or because there are few good examples available.[191] For many disciplines it is a new development and there is limited experience to draw from when considering the advantages and disadvantages of making data openly available. One case study interviewee described it as follows:

> *Other than those early adopters, I think you have a lot of people who are concerned for various reasons about the open access movement, so they're worried that they're going to get scooped. They're worried that they're going to lose some sort of rights or privileges that come with collecting data. They are resistant to having to learn how to use new tools that make open data and reproducibility easier. They generally kind of just have their process and they feel like they're tested already in terms of their time and their commitment and they don't really want to add this to the list of things that they have to worry about.*
> (Interview 4, data curation specialist, archaeology)

She points out, though, that the willingness and the interest to open research data varies per discipline:

> *And so that varies a lot depending on the discipline that you're in. Some of the disciplines generate tons and terabytes of data all from one instrument. Well sure, you might as well share it. I mean there's tons of it. But if you're talking about penguin counts in the Antarctic, then it's like, 'Well, that was a pretty hard data set to collect. You really want to make that available right away? Maybe not.'*
> (Ibid.)

Another problem is the current lack of incentives for researchers in most disciplines to learn about openly sharing their data. Although in some disciplines researchers are rewarded for the effort they put into generating and maintaining data sets and making them available, this is not common practice in most disciplines. Researchers in these disciplines thus have few reasons to inform themselves about the possibilities of open research data.

[188] Nelson, op. cit., 2009.
[189] Borgman, op. cit., 2012, p. 1060.
[190] Kuipers & van der Hoeven, op. cit. 2009.
[191] European Commission, *Online survey on scientific information in the digital age*, op. cit., 2012.

Awareness about what open access to research data entails and what the possibilities and limitations is not just low amongst researchers; institutions too are not always up to date on what open access can bring. As we have seen in previous chapters, not all funding bodies, research institutes, research-producing organizations, publishers, etc. have taken steps to enable more open access to research data. An issue that adds to the difficulty of creating more awareness amongst these institutions is that they each have their own interests and viewpoints, which means that a one size fits all strategy is hard to find.

Creating awareness about open research also pertains to making institutions aware of all the activities necessary to make data findable, re-usable and interpretable. Open access is not just about storage of data, as we have seen. One of our respondents, with extensive experience in developing data sharing tools and open software, expressed concern that institutions are unaware of what open access really entails:

> *What they haven't grasped, you could be having data that isn't [University] data because of doing joint work. Doing collaborative work, we're working across institutional and national borders, and that completely freaks them out. [...] Another thing they are struggling with is the notion of you may be storing your data in an offsite community repository. [...] We have got the idea of the data management planning tool that we have produced at [the University], you have to say which database it is going into. And this is a drop down list, well there are roughly 2500 of these in biology alone, how are we going to...drop down list, not really going to crack it. [...] And then the third thing is this whole notion of cataloguing.[...] They kind of grasped the idea there might be DOIs for [the University] data but the fact there maybe DOI for somebody else's data that you are holding because you are part of a joint project... big freak out. Then the idea that these DOIs, you will have to have access, other people will have to access these things. It is a bit like an old fashioned, we just put in a shelf and we will admire it kind of approach to the data management.*
> (Interview 4, Professor Computer Science, bioegineering)

She points out that there is also a lack of awareness on the level of management in terms of resourcing:

> *They really have not grasped the resourcing I would say in our university anyway the resourcing is an inverted pyramid, loads and loads of people on committees over sighting and one and a half people and a dog to actually implement everything and get services out. So I think they haven't realised the scale of the activity and it is quite complicated, and the people who are at the heart of it are good people but they are just overwhelmed at the moment, by just what is going on.*
> (Ibid.)

The challenge in creating awareness is thus getting all the different stakeholders at all levels in the institutions, including researchers, information specialists, and management, to learn about and reflect on open access and what it entails.

5.1 Policies and strategies

Various institutions have developed advocacy policies and strategies aimed at making researchers aware of the possibilities of open access and to encourage changes in research cultures. Although there are only a few, institutional data management policies and funding mandates are probably the most forceful strategies to promote open access amongst researchers. One of the first universities in the UK to adopt a data management policy was the University of Edinburgh. Its approach in developing the policy was not to mandate open access, but to encourage the sharing and publication of data, by offering researchers the appropriate support and infrastructure. It integrated all research services into one department of Information Services, which include classical library functions but also divisions like IT-infrastructures, Digital Curation Center, the JISC-designated national data centre (EDINA) and the Data Library. The development of a Research Data Management Roadmap by the Information Services has resulted in investments into data storage and data management. The roadmap is an ongoing activity with an incremental approach where promotion, advocacy and training are integral.[192] In contrast, the University College of London (UCL) research data policy does explicitly push for open access as default.[193] An important aspect of the UCL research data policy is that it ascribes clear responsibilities to student researchers, researchers, Research Data and Network Services department, Library Services and Provosts. The UCL Library Services, for example, are responsible for providing guidance and advocacy for research data management, data deposition and related metadata description. In doing so, UCL aims to make data management legitimate and transparent.

Institutions have also used a variety of other strategies to encourage open access, ranging from distributing leaflets and brochures, to organizing workshops and pilot projects. Some university libraries have been particularly active in trying to engage researchers and departments. One of the digital libraries in our case studies, for instance, considers outreach and supporting campus libraries part of its responsibility, as one respondent mentioned:

> *The [digital library] at large and my group in particular really tries to help the campus librarians with their outreach and communication of these things. So we provide not only the tools and services, but we provide slide decks and webinars and posters and flyers and postcards and things that they can use within their community to really push out some of these topics to the researchers that they interact with every day.*
> (Interview 4, data curation specialist, Archaeology)

Nevertheless creating awareness had not been a priority for most institutions. Funding bodies policies about research data, such as those from the Wellcome Trust[194] and The Royal Society[195], will sometimes bring up the need for training but seldom mention advocacy or awareness. An exception is the League of European Research Universities (LERU). In its roadmap for research data it dedicates a full chapter to advocacy. LERU considers it to be a very important tool to promote the benefits of open data and make the shift towards a culture of open access happen. It argues that advocacy can equip funders and decision makers with good arguments for making research data open, such as transparency and validity of research results.[196] Moreover, it recommends that advocacy of open research data should take place at every level within research universities. Most important here is that the institutional

[192] Rice, Robin, Cuna Ekmekcioglu, Jeff Haywood, Sarah Jones, Stuart Lewis, Stuart Macdonald and Tony Weir, "Implementing the Research Data Management Policy: University of Edinburgh Roadmap", *International Journal of Digital Curation*, Vol. 8, No. 2, 2013.

[193] Ayris, Paul, UCL Research Data & Network Services, "UCL Research Data Policy", August 2013. http://www.ucl.ac.uk/isd/staff/research_services/research-data/researchdata/uclresearchdatapolicy

[194] Wellcome Trust, *Policy on data management and sharing*, 2010. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

[195] The Royal Society, op. cit., 2012.

[196] LERU, op. cit., 2013, pp. 10-12.

leadership agrees wholeheartedly on support for open data. In order to gain this support the senior management and all the appropriate stakeholders such as faculties, library and IT services, the research office and other support departments are to be informed about open research data. The overarching recommendations LERU gives its members concerning advocacy is that they should:

- Engage at an international level to build and collect evidence and advocate for the value of open access to research data;
- Foster a debate amongst stakeholders and disciplines around data sharing;
- Develop and clearly articulate incentives for researchers to make their data open;
- Promote best practice in data management, citation and interoperability to increase the visibility of data;
- Develop formal policies for promoting and rewarding those generating and sharing data of use to the scientific community.

Most institutions engage now and then in the enabling function either through providing training or creating awareness, but out of the many stakeholders there are some service providers and non-profit organizations behind them, that play a key role in creating awareness. Organizations like COAR, EUDAT, LIBER, RDA, KE and many more are advocating and informing researchers and institutions about the importance of open data management and they are a resource for libraries that are about to start data curation schemes. The Open Access Directory has listed about 200 organizations that make OA advocacy a significant part of their mission. Quite a few of these are also focusing on open data.[197] Some are calling for a proactive attitude for responsible data planning and support, suggesting that university libraries initiate discussions within the university to make this happen.[198] Service providers and the non-profits behind them also play a key role in creating awareness. These service providers, such as SherpaRomeo, DOAJ, SPARC, OAPEN, Registry of Research Data Repositories, DPC EIFL, and Dryad, maintain training and advocacy infrastructures and in so doing give access to training material, reference services, advocacy arguments etc. for all other stakeholders who engage in training or advocacy activities.

Several strategies have been developed to address some of the concerns about data sharing. Many data sharing initiatives, for example, allow for embargoes that allow authors to develop their publications and data and by doing so avoid the fears of unfair competition. Datasets can be considered citable entities and publishers and repositories are increasingly providing citable links and identifiers to datasets.[199]

5.2 Remaining barriers and challenges

The barriers for advocacy for open data are much the same as those mentioned above for training programs. There is generally a lack of incentives and recognition for sharing data, which makes it more difficult to interest researchers and research communities to invest in learning about the opportunities open research data offer. The pervasive perception that

[197] Open Access Directory, *Advocacy organizations for OA*, no date. http://oad.simmons.edu/oadwiki/Advocacy_organizations_for_OA

[198] Erway, Ricky, *Starting the Conversation: University-wide Research Data Management Policy*, OCLC Research, 2013. http://oclc.org/content/dam/research/publications/library/2013/2013-08.pdf

[199] White, Ethan P., Elita Baldridge, Zachar T. Brym, Kenneth J. Locey, Daniel J. McGlinn and Sarah R. Supp, "Nine simple ways to make it easier to (re)use your data", *PeerJ PrePrints*, Vol. 2, 5 Jul 2013. https://peerj.com/preprints/7v2/

sharing data is technically difficult and very time consuming adds to this barrier. Similarly, enabling open research data is one of many priorities for institutions' staff and it might not align well with other priorities.

Disciplinary fragmentation continues to present a challenge. Some disciplines are harder to reach than others. As open access still has some way to go in 'easier' disciplines, such as astronomy and genetics, disciplines in which outreach activities prove to be more difficult, such as the humanities, receive less attention in various promotion and education initiatives. A case study respondent, actively involved in creating awareness about open research data notes:

> *Well, the humanities is an interesting one because I found that as a researcher, a scientist coming in to the library community, this is something every librarian really likes to talk about and bring out, is digital humanities. It's kind of the go-to thing. Well, what about the digital humanities? So to me as a researcher, that's a really complicated problem. We have a lot of nuances that we can't even solve the easy problems yet, so I'm not excited about us jumping in to solve the hard problems. And so one specific…? No, we're kind of basically going for, I mean I think that in terms of the things we think about [...] is we think about things in terms of the long tail, so the little data sets that maybe don't have a good spot, that are individual researchers - I mean the large majority of researchers at the [University] at large fall into that category. And so if we can meet those needs first and then start thinking about the more specific groups, then that would be great.*
> (Interview 4, digital curation specialist, archaeology)

There is a need for information provision that should be targeted and tailored to the specific disciplines in the fragmented science landscape.

Challenges also remain in creating awareness amongst and within institutions, even amongst those enthusiastic about and advocating open access. Taking steps towards providing open access to research data often requires a change in culture within institutions as well. Yet, those advocating more open access should take into account that changing research cultures and reward systems can take some time. In some disciplines, such as particle physics, data management is already rewarded. However, opening the data for a wider public would undermine current reward systems, as one of our respondents noted:

> *Now, clearly, there is a mood amongst governments and funding agencies that takes a more open view of the data and we're not unaware of that, but what we can't do is an immediate transition that says, 'We take the data and then anybody that likes can suddenly analyse this stuff' because that would completely undermine the vast amounts of effort that are required by the people who analyse the data to collect the data and to maintain the detector and to upgrade the detector. So there is not a distinction between facilities operators and the physicists, the scientists that are accessing the data. There is a reward system that is built in.*
> (Interview 3, physicists involved with ATLAS data preservation, particle physics)

5.3 WORKING TOWARDS SOLUTIONS

Data citation and data papers are emerging solutions to some of the concerns about data sharing. A case study respondent pointed out that a joint effort has to be made to create incentives for researchers to open their data:

> *So how do we make this happen? And I think education is an important aspect and funding is an important aspect, but the question is also which groups need to provide what kind of framework to make this happen? And I think so funders is one group, but if we don't come up with a good way of recognising the importance of the data sharing and making it available, that has an impact on career planning, then it will be extremely difficult, regardless of the amount of education that's pouring into it because I mean the priority for a researcher is to have a position that they can work from and it's the impact factor of your publication ensures that and the data access does not ensure it.*
> (Interview 1, project manager at software company, health)

As mentioned before, funders, journals and their publishers can play an important role in providing incentives for data sharing. In the words of one respondent:

> *Journals aren't famous for actually following up on their mandates but there are two pressure points for all academics, publishing and money. Money going in and fame coming out. Those are the only points of any kind of influence over their behaviour. If you are going to do an intervention you have to do it with those two points really. Because those are the things that count toward their promotion and their progress.*
> (Interview 4, professor computer science, bioengineering).

A major task in creating awareness will be to focus and mobilize the energy of particularly open access advocacy organizations that are pushing the frontiers of open data management so that they are meeting the needs of different stakeholders. Besides raising awareness among researchers, there is also a need for policy makers of funding organizations to be ahead, or at least abreast, of the organizations they are funding, when it comes to developing policy and engaging in discussions about open research data. Another interesting and important target to raise awareness about open research data are the professional organizations. They need to hear from their members what open access to research data could work for their particular field and start issuing their own policies.

5.4 CONCLUSION

In this Chapter we looked at some of the initiatives that institutions have taken to create awareness about open access to research data within their own organization and among research communities. Libraries, in particular, often consider it their responsibility to encourage researchers and university departments to make their data openly accessible. In addition, there are a number of professional organisations that play an active role in creating more awareness about open access.

Yet, our review of the literature and the feedback during the interviews also indicate that creating awareness is not a priority in many institutions. Moreover, top-down approaches and advocacy may have adverse effects and it can be difficult to reach particular stakeholders. Different stakeholders have different needs and interests, which require tailored approaches to creating awareness about the possibilities and limitations of open research data in their particular area. Another key barrier is the lack of incentives for researchers and institutions to take an interest in open research data.

An important first step in addressing these issues it to create incentives for researchers and institutions to embrace open research data, by for example issuing data management mandates as well as establishing rewards and professional recognition for publishing, maintaining and using open research data. In addition, institutions will have to work together to create an environment in which the various stakeholders can discuss what open access to research data should look like and what that would entail for them.

6 CONCLUSION AND RECOMMENDATIONS

Institutions, such as universities, libraries, data centres, publishers, professional associations and funding bodies, play an important role in making open research data possible. They support researchers, provide infrastructure and funding, and set guidelines. In this report we have been concerned with the policies and strategies that these kinds of institutions have developed to enable open research data and with the challenges they face.

Open access to research data offers many benefits, but the analysis in this deliverable shows that there remain some challenges to overcome in order to realize them. Although in the long-term open access may bring significant cost savings, it also generates considerable costs in the short-term. Open access requires significant and continuous effort to make sure that data can be found, interpreted, evaluated and used. Evaluating and maintaining the quality of data, both in terms of bits and bytes as well as scientific value, requires considerable work and significant changes in organizational and cultural practices. Peer review strategies have to be developed; data citation practices have to be actively encouraged and made part of institutional evaluation and reward systems. Another challenge is the relatively low level of data management skills and awareness about the opportunities and limitations of open research data both within institutions and within many research communities. Yet, providing training to and creating awareness among researchers and other stakeholders is difficult because of the significant fragmentation within and between research communities and the rapid pace of technological developments.

We can make several observations based on the analysis in the previous Chapters that cut across the various challenges. The first observation is that, although open research data is a common practice in some research communities, in many other fields developments towards more open research cultures are still in the early stages. Current efforts to stimulate open access to research data focus in particular on supporting and motivating the early adopters among researchers and institutions to open up their data, by providing funding, infrastructure and data management services as well as developing policies and issuing mandates. Various promising initiatives have achieved significant successes in achieving this. Institutions have developed and offer a range of data services, often in collaboration with other institutions, and are learning as they go along. As the previous Chapters show, various reports and studies provide good starting points, guidelines and examples for the initial steps that institutions can take.

Yet, solving the harder problems, such as funding long-term preservation of data, evaluating the scientific quality of data or getting the more reluctant researchers to experiment with open access is put on hold. For such problems institutions have few reports or good practices to draw on. Although there are some good examples in a select set of disciplines, such as bioengineering and astronomy, successful practices in one discipline may not translate well to other disciplines. A more diverse set of discipline-specific as well as overarching examples and good practices are needed. Moreover, there is a lack of appropriate cost models to help institutions plan for future open access and data management needs. It is not clear for many institutions what kinds of investments are needed to develop, maintain and effectively use open data infrastructures.

The early stage developments also mean that various institutions gain new roles and responsibilities, as they begin to offer data services, establish infrastructures and issue policies. Whereas researchers in many disciplines used to be responsible for their data, even after their projects were finished, this responsibility is now partly delegated to data centres and repositories. In developing data management policies and services, institutions will have to consider how to give shape to these responsibilities.

The analysis in this report also shows that in terms of financial resources, knowledge and expertise, institutions will have a hard time addressing some of the challenges on their own. They will have to engage in collaborative efforts to develop data repositories, data management services, training programmes, etc. Libraries will have to work with data centres and other libraries to offer long-term preservation of and access to research data as well as skill development programmes. Universities, research institutes, and funding bodies will have to participate in international collaborations, overcoming their tendency to hold on to their research data.

Another observation pertains to the value and role of data management in scientific practices. The awareness about open research data and the incentives to make data open will increase when research communities start valuing data produced as much as they value publications, and when research institutions, universities, funding bodies and scholarly societies start evaluating and rewarding researchers and research groups based on their data management efforts. As various reports have pointed out, stimulating the recognition and acknowledgement of the value of sound data practices is an important step towards more open research communities. Efforts to, for instance, encourage and further develop data citation practices, by issuing persistent identifiers, developing and using citation standards, and supporting peer review, can contribute considerably to resolving the challenges described in this report. Moreover, the growing number of mandates that require open access to data, in this regard, provide a strong incentive for researchers and institutions to publish their data: funding grants are increasingly conditional on DMPs and the availability of research data; the number of journals that will not accept an article unless the data are accessible in certified repositories is growing; and progressively more governments are demanding that public funded data is openly available to the public. However, in order for these mandates to be effective institutions need to ensure compliance to mandates and stimulate the evaluation of DMPs and open research data. While various governments and institutions have been busy developing policies and mandates, little attention has so far been paid to how and when DMPs and open research data are to be evaluated.

The final observation is that a primarily top-down and centralized move towards open research data will only be effective to a certain extent. In order to have a vibrant open research ecosystem, institutions will have to acknowledge disciplinary heterogeneity and autonomy. Research communities will have to get some room to specify their own terms and conditions for open research data and experiment with various approaches. Some of our case study respondents also noted that there might be a generational dimension to the developments as well. As young researchers, more comfortable with new technologies, come into the field, open access to research data may become less of a problem.

6.1 RECOMMENDATIONS

Taking the above observations into account and based on the feedback we have received during the fourth RECODE workshop, we make the following preliminary recommendations. These recommendations are intended as input to be further discussed in the framework of RECODE WP5, which, based on the findings of the other work packages, will develop a set of good practice policy guidelines targeted at significant stakeholders and key policy makers.

Where possible, institute data management mandates and policies with open research data as the default and clear lines of responsibility, while ensuring that the required resources are available.

- Funding bodies and research councils have a particular responsibility in addressing the funding challenges. Mandating open access comes with a responsibility to make sure that there is funding available that will allow researchers and institutions to adhere appropriately to the mandates.
- Both funding bodies and research institutes should ensure that they have enough resources to evaluate compliance with mandates.
- The distribution of responsibilities in the data management process is not always clear, or is still being negotiated. This causes confusion and a risk of slowing down the efficiency and speed of the curating processes. Data policies provide a good basis to explicitly assign responsibilities and appropriate resources.

Stimulate and ensure compliance with mandates and policies and make data practices part of the evaluation and reward systems

- Funding and research institutions can play a significant role in changing research cultures to be more supportive of data sharing and open research data, by recognizing the value of proper data management in the evaluation of researchers and their work. In order to do so, they will have to ensure compliance to mandates and policies not only amongst researchers, but also amongst their own staff and reviewers.
- Scholarly societies can issues guidelines and facilitate discussions within research communities about standards and good practices.
- Publishers can instruct and support reviewers and editorial boards to pay more attention to the evaluation and proper citation of data sets.
- Professional associations can give attention to and support good practices within their communities. For example, they can create awards for good data management and inform their members about good practices with regard to open research data.
- Institutions and libraries can provide advice and support for compliance with funder mandates and templates for data management plans.

Create incentives for researchers to publish their data and make use of available open research data.

- Institutions can encourage researchers to publish their data by developing mechanisms that, besides academic publications, also recognise and reward data management, data sharing and reuse as valuable scientific output. For example, they can further develop career progression paths for data scientists; institute prizes and awards for good data management practices or develop tools for data citation and metrics (see the above recommendation).
- Institutional advocates for open data, such as libraries, can collect and promote examples of data reuse in various disciplines.

- Funders can create incentives for publishing research data and data reuse by instituting mandates or offering funding opportunities for research projects that actively reuse data.

Create room for innovative ideas and bottom-up initiatives to further develop data management services and sustainable business models

- In order to engage more researchers in the move towards more open research ecosystems and ensure that research data are not only available, but also reusable and interpretable, funding bodies and research institutes should encourage research communities to develop their own standards, norms, data repositories and best practices. From our research and the previous RECODE deliverables it is clear that there is no one size fits all solutions for data curation. What data are, how they relate to other research products and how they should be made accessible differs between and within disciplines. Funding and research institutions should therefore enable research communities to transition to open access on their own terms, taking into account the peculiarities of their field of research and research culture, by providing funding for bottom-up data management initiatives and allowing for alternative solutions in mandates and policies.
- Funding smaller scale open access initiatives also allows for creative and innovative solutions in the development of sustainable business models.
- Creating room for bottom-up initiatives also entails that research institutes and scholarly societies should collaborate with research communities in developing data management policies and services. This can also contribute to engaging the more reluctant research communities.
- Libraries and data centres can facilitate bottom-up initiatives by offering data services, where possible, developed in collaboration with or targeted at particular research communities and informing them about alternative data curation possibilities when the data management needs of researchers are too specific.

Start planning for long-term preservation and curation of open research data

- Funding bodies and research institutes should develop policies on long-term preservation of data in interaction with research communities. They should clarify responsibilities between the various institutions and between institutions and researchers. Data preservation policies should also address selection and retention of data as well as long-term funding.
- Research institutes and funding bodies can make their researchers aware of the requirements for long-term data storage by requiring a sustainability plan to be part of the DMP. Researchers should explain how they or particular institutions would curate the data and for how long as well as indicate how much this will cost.
- There is a need for well-developed and specific cost-models. Research institutions, including research producing institutions and higher education institutions, can benefit from developing and sharing cost models. Such models should take into account the various differences between kinds of data and research communities.

Pursue collaborations between and within institutions

- Several good examples exist of institutions working together to achieve efficiencies of scale, such as libraries working together in developing and providing data management training programmes or data centres working with libraries to provide data services and support for researchers. Institutions can benefit from studying such federated and distributed data management networks.
- Funding bodies should promote and encourage collaborations between institutions that aim to provide open access to research data. This includes collaborations that strive to give access to both research data as well as public sector data.
- Institutional repositories and data centres may find cost savings in collaborating with publishers and companies in providing data services, but they should make sure that they are able to fulfil their obligations, such as long-term preservation of data.

Develop strategies that support the evaluation of the quality of data and data repositories both in terms of technical quality as well as scientific value.

- Various institutions can play a role in promoting, enabling and supporting the development of practices that contribute to enhancing the quality of research data. Established practices of other scientific publications, including editorial quality control, independent peer review, standard citation of data sets, citation tracking, permanent archiving and use of other metrics, provide good examples to draw on in developing these practices. Yet, as our research shows the wide variety of data and definitions of data quality require specifically targeted strategies that have yet to be developed. Institutions, such as libraries, scholarly societies, data centres, consortia and publishers, can take the lead here.
- Institutional repositories, data centres and data publishers have a responsibility in ensuring the quality of metadata, file readability and adherence to standards, the development of policies for long-term access and re-use of data. To fulfil this responsibility they have a range of strategies to use, including automating part of the ingestion and quality control process, offering support for research in preparing their data, manual quality assurance, etc.
- In addition to evaluating the technical quality of research (meta)data, these institutions can further shape and expand their new role in the research data ecosystem by providing support to research communities in developing strategies evaluating the scientific quality of data. They could, for instance, provide online platforms for research communities to discuss and rate data sets or link data to scientific publications.
- Data centres and libraries with data repositories should also develop policies for the selection and retention of data, preferably in interaction with research communities. As more data sets become public, these institutions will have to start making decisions about what data should be kept and what data can be discarded.
- Publishers also have a role to play in ascertaining data quality by promoting research data peer review for data accompanying publications, developing policies requiring the deposit

of research data in certified repositories and further developing the market for data journals.

Create environments that stimulate open access and provide support and training for researchers and other relevant staff in their specific practices

- As researchers can be regarded as professionals and experts in their respective fields of research, it may prove more effective to offer them data management support tailored to their specific practices or needs, than to offer them general data management training courses. Funding agencies should therefore encourage the establishment of national centers of expertise in research data management, where researchers can ask for help in data management. Researchers prefer to contact experts rather than engage in training themselves. The focus of research institutes should be on the postgraduate community. Our literature review reveals that young scientists are the most open to advocacy and training efforts. They are also often involved in creating data sets.
- Funding agencies, research councils and higher education institutes should encourage the establishment of research data management programs for librarians and data scientists. In the US library schools have been provided government money for training librarians. Because of this US libraries are ahead the rest of the world when it comes to management of open data. Certificates and programs are targeted at library staff. Courses are both one-semester courses and full programs.
- Funding agencies, research councils and higher education institutes should also encourage the creation of shorter vocational courses on data curation and preservation and data management for library staff, researchers, project office staff and publishers. There is a massive need to upgrade the skills of all staff dealing with making open research data available. Short courses and distance training is a good way to prepare staff for managing data using available resources, but, ultimately, the development of the required skills should be a central part of professional training programmes.
- Institutions should encourage the exchange of community skills especially between libraries and data centres. Data management practices cross several professional borders. Cooperation between different specialties as well as clear definitions of responsibilities are helpful making the management process smoother.
- Publishers can act as gatekeepers for making data available but need training in order to supply data to their publications in a useful way.
- Encourage funding and supporting service providers like Sherpa Romeo, re3data.org registry, SPARC, University of Edinburgh's research data management training tool "Mantra" etc. These resources are vital to the training and advocacy infrastructure of open access both to publications and to research data, providing institutions with tools that enable efficient advocacy and training.
- If researchers are encouraged through their usual channels (professional associations, senior colleagues, reviewers, funding mandates, journals) to share data there is a greater probability that they will do this. Such gatekeepers could be positively influenced to recommend good data management practices by being convinced of open data benefits. This will result in increased demand for training.