



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

Deliverable D2.1: Infrastructure and technology challenges

Author(s): Lorenzo Bigagli (National Research Council of Italy); Thordis Sveinsdottir, Bridgette Wessels & Rod Smallwood (University of Sheffield); Peter Linde (Blekinge Institute of Technology); Jeroen Sondervan (Amsterdam University Press)

Dissemination level: Public

Deliverable type: Final

Version: 2.0

Submission date: Due 31 March 2014 (extension from 28 February 2014 agreed with PO)

Table of Contents

List of acroynms	4
Executive Summary.....	7
1 Introduction	9
1.1 Scope and definitions	9
1.2 Background: the RECODE stakeholder taxonomy	9
1.3 Document structure	10
2 Methodology.....	12
2.1 WP2 Stakeholder taxonomy	15
3 Framing infrastructure and technology challenges: a review of the literature	18
3.1 Reference initiatives in Open Access to Research Data in Europe	18
3.2 Main challenges in infrastructure and technology.....	22
3.2.1 Heterogeneity and interoperability	23
3.2.2 Accessibility and discoverability.....	28
3.2.3 Preservation and curation	30
3.2.4 Quality and assessability	32
3.2.5 Security.....	34
3.3 Findings from the literature review	35
4 Scoping infrastructure and technology challenges: an online survey.....	37
4.1 Key issues for all the stakeholders	38
4.2 Stakeholder specific issues and concerns	41
4.2.1 Key issues for data Disseminators/Curators.....	41
4.2.2 Key issues for data Producers.....	43
4.3 Findings from the online survey	44
5 Case study Research: Infrastructure and technology challenges and recommendations in Open access to research Data – the View from Scientists within Five Scientific Disciplines	46
5.1 Particle Physics and Particle Astrophysics: The PPPA Group at the University of Sheffield and the CMS experiment at CERN.....	46
5.1.1 Infrastructure and technology challenges and recommendations within the Physics case study	47
5.2 Health and Clinical Research: The FP7 Project EVA and Open Health.....	50
5.2.1 Infrastructure and technology challenges and recommendations within the Health and Clinical Research case study	50
5.3 Bioengineering: Auckland Bioengineering Institute and The VPH Community.....	53
5.3.1 Infrastructure and technology challenges and recommendations within the Bioengineering case study.....	53
5.4 Environmental Sciences: the EC Joint Research Centre	55
5.4.1 Infrastructure and technology challenges and recommendations within the Environmental Sciences case study.....	56

5.5	Archaeology: Open Context and the Mappa project	58
5.5.1	Infrastructure and technology challenges and recommendations within the Archaeology case study	59
5.6	Supplementary interviews	62
5.7	Findings from the interviews	63
6	Validation Workshop.....	66
6.1	Findings from the workshop.....	70
7	International Advisory Board Comments.....	74
8	Discussion.....	76
8.1	Recommendations on infrastructure and technology for Open Access to research data 79	
9	Conclusions	82
	Appendix 1 – Survey questionnaire	83
	Appendix 2 – Interview protocols	96
	Introductory Section	96
	Producer of research data	96
	Disseminator/Curator of research data	97
	Funder.....	97
	End user	98
	Appendix 3 – List of Workshop Attendees’ Institutions.....	99
	Appendix 4 – RECODE WP2 Workshop Agenda	100

LIST OF ACROYNMS

ABI – Auckland Bioengineering Institute
API – Application Programming Interface
ARK – Archival Resource Key
CAS – Chemical Abstracts Registry Service
CC0 – Creative Commons license "No Rights Reserved"
CC-BY – Creative Commons license "Attribution"
CC-BY-SA – Creative Commons license "Attribution-ShareAlike"
CEA – Commissariat à l'énergie atomique et aux énergies alternatives
CERN – European Organization for Nuclear Research
CESSDA – Council of European Social Science Data Archives
CICG – Centre International de Conférences de Genève
CIDOC – International Committee on Documentation
CMS – Compact Muon Solenoid, or also Content Management System
CNG – Centre National de Génotypage
CPU – Central Processing Unit
CRM – Conceptual Reference Model
CRUI – Conference of Italian University Rectors
DINI – Deutsche Initiative für Netwerkinformation
DMP – Data Management Plan
DOI – Digital Object Identifier
DoW – Description of Work
DRAMBORA – Digital Repository Audit Method Based on Risk Assessment
DVCS – Distributed Version Control System
EC – European Community
DG CONNECT – Directorate General for Communications Networks, Content and Technology
EGIDA – Coordinating Earth and Environmental Cross-Disciplinary Projects to Promote GEOSS
EHR – Electronic Health Record
ESA – European Space Agency
ESO – European Southern Observatory
EU – European Union
EVA – Emphysema versus Airways disease
FP7 – EU Seventh Framework Programme for Research and Technological Development
GCI – GEOSS Common Infrastructure
GEO – Group on Earth Observations
GEOSS – Global Earth Observation System of Systems
GPR – Ground Penetrating Radar
H2020 – Horizon 2020
HHS – Health and Human Services
ICT – Information and Communication Technology
INSPIRE – Infrastructure for Spatial Information in the European Community
IPR – Intellectual Property Rights
iRODS – integrated Rule Oriented Data Systems
ISO – International Organization for Standardization
IUPS – International Union of Physiological Sciences
JOAD – Journal of Open Archaeological Data
JRC – Joint Research Centre

JSON – JavaScript Object Notation
JSON-LD – JavaScript Object Notation - Linked Data
LEP – Large Electron Positron
LHC – Large Hadron Collider
LSID – Large Structure Identifier
MITA – Medicaid Information Technology Architecture
NASA – National Aeronautics and Space Administration
NERC – Natural Environment Research Council
NSF – National Science Foundation
OAI – Open Archives Initiative
OAIS – Open Archival Information System
OBO – Open Biological and Biomedical Ontologies
ODE project – Opportunities for Data Exchange
OGC – Open Geospatial Consortium
OID – Object Identifier
ORCID – Open Researcher and Contributor ID
OWL – Web Ontology Language
PARSE.Insight – Permanent Access to the Records of Science in Europe
PID – Persistent Identifier
PLOS – Public Library of Science
PMR – Physiome Model Repository
PPPA – Particle Physics and Particle Astrophysics
PSI – Public Sector Information
PURL – Permanent Universal Resource Locator
RDF – Resource Description Framework
RECODE – Policy RECommendations for Open access to research Data in Europe
SAFE – Standard Archive Format for Europe
SCIDIP-ES – SCIENCE Data Infrastructure for Preservation with focus on Earth Science
SFTP – Secured File Transfer Protocol
SOS – Sensor Observation Service
SOSE – System-of-Systems Engineering
TRAC – Trustworthy Repositories Audit and Certification
UK – United Kingdom
URI – Uniform Resource Identifier
URL – Uniform Resource Locator
URN – Uniform Resource Name
USA – United States of America
UUID – Unique User Identifier
WCS – Web Coverage Service
WebDAV – Web-based Distributed Authoring and Versioning
WHO – World Health Organization
WP – Work Package
WP1 – RECODE Work Package 1, Stakeholder Values and Ecosystems
WP2 – RECODE Work Package 2, Infrastructure and technology
WP5 – RECODE Work Package 5, Policy guidelines for open access and data preservation and dissemination
WP6 – RECODE Work Package 6, Stakeholder Engagement and Mobilisation
VPH – Virtual Physiological Human
W3C – World Wide Web Consortium
WPS – Web Processing Service

XML – Extensible Markup Language
XRI – Extensible Resource Identifier

EXECUTIVE SUMMARY

In this deliverable, we report on our work on infrastructural and technological barriers to Open Access and preservation of research data as identified by key stakeholder groups. Through a mix of qualitative, quantitative and document review methods, we identified five key barriers to successfully implementing Open Access to research data in Europe: data heterogeneity and issues of standardisation; accessibility and discoverability issues; data preservation and curation; data quality and assessability; and data security. We explore these issues in detail and present existing good practice, and technical and infrastructural solutions used to mitigate such barriers.

This work was conducted within the EU FP7 funded project RECODE, which focuses on developing policy recommendations for Open Access to Research Data in Europe. In particular, this work is coordinated by RECODE Work Package 2 (WP2), Infrastructure and technology. It distinguishes between different categories of stakeholders in terms of how the experience and respond to these challenges. Specifically, we distinguish between:

- **Producers** of research data
 - E.g. researchers elaborating raw data
- **Disseminators/Curators** of research data
 - In charge of the distribution and preservation infrastructure (information systems, e-infrastructure) for storage, access, and maintenance of data
 - E.g. publisher, library
- **Funders**
 - Providing financial and policy support to research
- **End users** of research data at large
 - Including researchers, the industry, governmental agencies, ecc.

WP2 takes a broad definition of infrastructure, including: technological assets (hardware and software); human resources involved; all the procedures for management, training and support to its continuous operation and evolution.

The WP team conducted a literature review and consulted and analysed a large number of sources to scope the known technological and infrastructural challenges to Open Access and preservation of research data, and the possible existing solutions for their mitigation. The literature review tells us that technical issues are being discussed in a relatively small grid of reoccurring problems. If we talk about open research data, questions around standardization, interoperability, reuse and preservation are prevalent. In contrast, relatively few issues arose in relation to bandwidth, storage capacity or usability. On the basis of the document review, infrastructure and technology challenges are not considered as the most important obstacles to Open Access to research data, compared to financial, legal, and policy challenges.

Further to the literature review, the WP team sent out a scoping questionnaire to the broader stakeholder communities, to further explore the prevalence of the key issues that had been identified in the literature review, i.e. areas of data heterogeneity, accessibility and discoverability, preservation and curation, quality and assessability, and security. Although the survey was completed by a small number of people, and hence we do not intend to generalise over a whole population, the findings do give an insight into what are considered major barriers to implementing Open Access to research data (heterogeneity and interoperability, data documentation and quality assessment) and also which stakeholders are

considered most reliable when it comes to preserving data and storing it (national institutional repositories or digital libraries). Interestingly, in light of the scope of this research work, technological issues seem to be rather low on the priorities of those who replied to the survey.

The WP team obtained further inputs by means of targeted interviews with key individuals from each of the five RECODE case studies (physics, health, bioengineering, earth sciences, and archaeology), in order to elaborate on the infrastructural and technological issues they encounter in their research practice. Again, one interesting finding emerging from our interviews is that technological barriers are not reported as of high concern to implementing Open Access to research data, whereas financial, cultural and legal challenges are higher on the list of concerns. Overall, respondents reported more experience with Open Access publications rather than Open Access to research data, and data preservation. In most instances we found data management plans at an early stage. Technical solutions for data management and preservation are fragmented, often designed for a narrow purpose, rather than centralised. As we found in the literature review and the survey, most respondents mentioned issues of documentation and metadata as a key challenge to enable retrieval, re-use and preservation of research data. However, the technological challenges mentioned by respondents in the case study interview differ somewhat between disciplines.

The WP team held a validation and dissemination workshop as an official side event of the 10th Plenary Session of the Group on Earth Observations & 2014 Ministerial Summit. The workshop attracted over 40 attendees from 14 countries, including policy makers, funding bodies, libraries, data management organisations and researchers, along with representatives from the RECODE case studies and RECODE team members. The workshop sought to validate and discuss the research findings, as well as to obtain additional feedback and insights from representatives of the RECODE case studies and major international initiatives, to share their perspective in understanding Open Access to research data, in relation to infrastructure and technology challenges. The workshop discussion overall validated our survey and case study results. Data heterogeneity was picked up as a very important challenge, and options for making the data accessible and useable are deemed as somewhat lacking. With regard to accessibility, the workshop attendees agreed with our findings from the survey, in that the preference expressed is for the enhancing of digital libraries, and specialised repositories to store and curate research data. Data preservation, in terms of long-term storage solutions and curation options, remains a key challenge. Quality and security have a prominent importance in Open Access. This was highlighted especially during the workshop, especially for some scientific communities, such as health and archaeology.

It is clear from the combined results of the survey, the literature review, the case study interviews and the workshop that stakeholders in general have a limited knowledge about research data management and how to make data openly available in a multidisciplinary way. To reiterate, technological barriers were not reported to be of high concern in implementing Open Access to research data, when compared to financial, cultural and legal challenges. We maintain that Open Access to research data is still at an early stage within Europe and internationally.

On the basis of our research, we indicate possible recommendations on infrastructure and technology for Open Access to research data. These recommendations are intended as an 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.

1 INTRODUCTION

This report is the deliverable for Work Package 2 (WP2), Infrastructure and technology, of the EU FP7 funded project RECODE (Grant Agreement No: 321463), which focuses on developing Policy Recommendations for Open Access to Research Data in Europe. WP2 focuses overall on identifying the existing technological barriers to Open Access and preservation of research data, as well as any existing solutions that are being used to mitigate these barriers. The objectives of WP2 are as follows:

- Identify and report on current and emerging technologies being used in Open Access repositories to provide access to scientific information and research data.
- Identify the perceived technological barriers to Open Access to information and research data.
- Identify and report on possibilities for developing solutions to increase the interoperability and interconnection of Open Access repositories: common standards, mediation technologies.
- Conduct a workshop with stakeholders to produce recommendations for improved technologies and infrastructures.

1.1 SCOPE AND DEFINITIONS

In the context of this research work, with the term “*infrastructure*”, we mean:

- Technological assets (hardware and software);
- Human resources;
- Procedures for management, training and support to its continuous operation and evolution.

Examples of infrastructural and technological factors that may hinder Open Access and preservation of research data include: interoperability issues, functional gaps, lack of training and/or expertise on IT and semantics aspects, data quality and fitness for use, discoverability, access management, data selection, heterogeneous formats, structural complexity, lack of automatic mechanisms for policy enforcement, lack of metadata and data models, obsolescence of infrastructures, scarce awareness about new technological solutions, communication issues. In our research we further explored these issues in order to determine their immediacy and importance for key stakeholders. We found that while all these issues are recognised, they can be considered as specific aspects of five main challenges: heterogeneity; accessibility; preservation and curation; quality and assessability; and security.

The first RECODE Deliverable¹ provides more information on the definitions of Research Data and Open Access in the general context of RECODE.

1.2 BACKGROUND: THE RECODE STAKEHOLDER TAXONOMY

In order to examine the above issues, in relation to the implementation of Open Access to research data, we recognise that different stakeholders will play different roles and take on different responsibilities in the overall process. We therefore briefly introduce the RECODE

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

model of the Open Access stakeholder ecosystem, as a functional taxonomy that consists of five entities or functions within which performers are interconnected through flows.²

As depicted in Figure 1, the categories are not mutually exclusive and at any given time stakeholders may operate and interact from within different functional categories. Stakeholders have one primary function (PF) and can have several secondary functions (SF), hence we acknowledge that data creators can also act as users, disseminators and/or curators within the open data ecosystem.

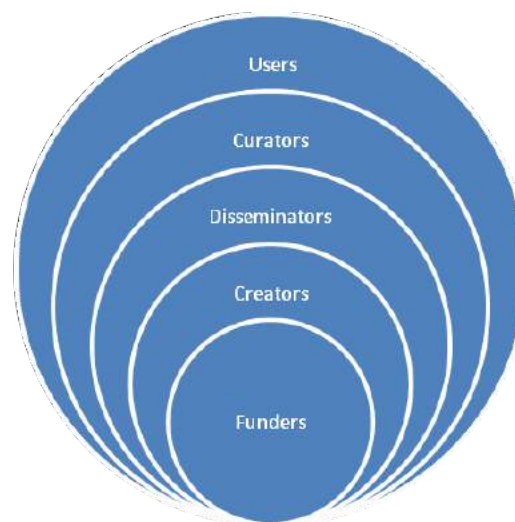


Figure 1 - The RECODE stakeholder functions

In addition to focusing on the five key issues outlined above, we also explore these issues as experienced and interpreted from within different stakeholder groups.

1.3 DOCUMENT STRUCTURE

This deliverable is organized as follows:

- **Chapter 2** introduces the methodology employed by the WP team and further details the stakeholder taxonomy and its relevance to the WP2 aims and objectives.
- **Chapter 3** reports on our literature review and establishes the infrastructure and technology challenges in Open Access to Research Data considered most relevant to our scope.
- **Chapter 4** presents our analysis of the online survey, elaborating on the key issues regarding technology and infrastructure as perceived by stakeholders across Europe.
- **Chapter 5** outlines our findings from the case study research and presents a detailed overview of the key issues as presented to us by the different stakeholders within various scientific disciplines.
- **Chapter 6** reports on the activities and discussion of the stakeholder workshop, in which we sought to validate our findings from the above research.
- **Chapter 7** presents our discussions with the Advisory Board Members and outlines comments and feedback received on the draft report.

² Sveinsdottir, et al., op. cit., 2013, pp. 21-31.

- **Chapter 8** presents a discussion of all the findings of the above research activities and seeks to consolidate, presenting overall draft recommendations based on our work.
- **Chapter 9** concludes our work, outlining our key objectives and introducing RECODE future research on policy guidelines for the implementation of Open Access to research data in Europe.

2 METHODOLOGY

In order to reach the objectives of WP2, the WP team performed several tasks, including a literature review, a survey of the existing practice, a case study research and a stakeholder validation workshop, as presented and agreed in the RECODE Description of Work³.

We have carried out a scoping document and literature review, gathering information about existing solutions, good practice, and initiatives, to identify existing barriers to Open Access and preservation of research data (Task 2.1 – Technological infrastructural requirements: technological barriers). We have analysed the possible solutions for mitigating the identified technological barriers, to identify possible recommendations on Open Access to scientific information and research data (Task 2.2 – Technological issues: recommendations).

Apart from technical reports, guidelines and other grey literature, industry material, software documentation, results from previous projects, scholarly articles, and other generic web information, we have paid particular attention to the lessons learned from wide-scale data-sharing initiatives in environmental sciences, like GEOSS⁴ and INSPIRE⁵. In fact, such domain is emblematic for research data sharing, access, dissemination and preservation.

GEOSS, a global initiative grouping around 80 nations and other international organisations coordinating and sharing information on nine societal benefit areas, by means of Earth Observation, is building a System-of-Systems based on a brokering/mediation infrastructure, which has proven able to provide harmonized discovery and access to heterogeneous multi-disciplinary data, according to a scalable approach. GEOSS 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. GEOSS 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. The European Directive INSPIRE establishes an infrastructure for spatial information in Europe, to support EU environmental policies and activities that may have an impact on the environment. INSPIRE aims to deliver integrated spatial information services to its target users, which include policy-makers at European, national and local level and the citizen.

We have been able to build on our expertise from previous EU-funded projects that coordinated cross-disciplinary efforts to promote GEOSS, such as EuroGEOSS⁶ and EGIDA⁷. EGIDA has been particularly useful to this work, since it has produced a general methodological approach⁸ for implementing a (re-) engineering process of the existing Science and Technology infrastructures and systems, to be adopted at the national/regional

³ RECODE Project, Annex I – “Description of work”, 15 July 2012.

⁴ Global Earth Observation System of Systems, “GEO Group on Earth Observations”, 2014. <https://www.earthobservations.org/index.shtml>

⁵ European Commission, “INSPIRE: Infrastructure for Spatial Information in the European Community”, no date. <http://inspire.ec.europa.eu/>

⁶ EuroGEOSS, “Welcome to EuroGEOSS the European Approach to GEOSS”, no date. <http://www.eurogeoss.eu/default.aspx>

⁷ EGIDA: Coordinating Earth and Environmental Cross- Disciplinary Project to Promote GEOSS, “Welcome to the EGIDA Project Website”, no date. <http://www.egida-project.eu/>

⁸ Mazzetti, Paolo, and Stefano Nativi, *D4.8 The EGIDA Methodology - Final version*, January 2013.

level for a sustainable contribution to GEOSS and other relevant European initiatives. The EGIDA Methodology is based on a System of Systems approach, through the mobilization of resources made available from the participation in national, European and international initiatives and projects, hence it seemed applicable in the context of infrastructural and technological recommendations for Open Access to research data.

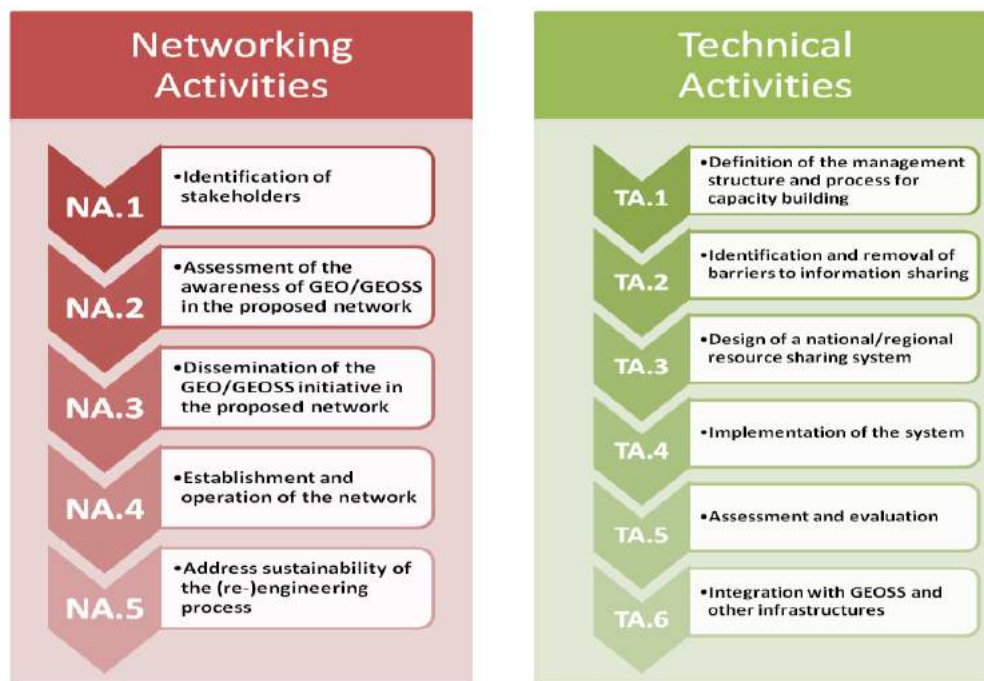


Figure 2 - EGIDA Methodology

As shown in Figure 2, the EGIDA Methodology defines two sets of activities running in parallel:

- Networking Activities: to identify and address the relevant Science and Technology community and actors (Community Engagement);
- Technical Activities: to guide the infrastructure development and align it with the GEO/GEOSS interoperability principles (Capacity Building).

For each activity several actions and sub-actions are defined, with related practices and guidelines derived from the design phase. Technical Activity TA.2 concerns “*Identification and removal of barriers to information sharing*”, considering behavioural, economical, legal, and technical barriers.

As defined by the EGIDA Methodology, “*technical barriers are related not to the will or possibility to share resources, but to the capability to do it. Some participants may be willing and authorized to share resources but are not able to do it.*”⁹ We applied Guideline TA.2.1:

“*The existence and nature of obstacles to data sharing can be discovered and analysed through surveys and interviews. Members of the Stakeholders Network can provide information about behavioural, legal, technical and financial barriers to data sharing.*”¹⁰

⁹ Ibid., p. 41.

¹⁰ Ibid., p. 42.

Hence, we have identified the stakeholder categories of interest for our scope, and explored existing practices and views, by designing and distributing an online survey questionnaire, with the aim of gaining an overview of current data practices and stakeholder views within the European scientific community. The questionnaire can be found in Appendix 1.

We have distributed the questionnaire by a number of means, including the RECODE mailing list and Twitter account, with the support of WP6, Stakeholder engagement and mobilization. We have also been supported by the EC DG CONNECT, who kindly offered to advertise the questionnaire via its Twitter account, to distribute it to a wide audience, including the participants in the EC Public Consultation on Open Research Data.

The online survey was open from November 2013 to January 2014; we have received around 50 responses, mostly from disseminators (62%) and producers (29%) in natural and computer sciences. Because the questionnaire was fully accessible without restrictions, we could not assess the statistical relevance of the sample, however we think the results provide interesting insights that steered our subsequent investigation. Furthermore, the information functioned as stimuli for the workshop discussions.

To gain a more detailed and holistic view of the identified technological and infrastructural barriers and solutions, a case study method was employed within five scientific fields, detailed below. Semi-structured interviews (Task 2.3 – Case study interviews) were conducted with selected technical staff within each of the five RECODE case studies, defined as a discrete research field that has its own ontology, epistemology and methodology. The aim was to get an overview of barriers and issues, but also to explore any discipline specific issues that may arise regarding access and preservation of research data.

- **Case study 1, Physics**, addressed particle physics in relation to the data management issues of large volumes of data.
- **Case study 2, Health**, addressed health sciences in relation to the issue of quality control, ethics and data security.
- **Case study 3, Bioengineering**, addressed complex modelling that may prove difficult to replicate and test in models for heterogeneous datasets.
- **Case study 4, Environmental Sciences**, addressed environmental research in relation to multidisciplinary interoperability models for heterogeneous datasets.
- **Case study 5, Archaeology**, addressed procedures for evaluating the quality of open data and the technical approach to preserving diverse types of data.

We have asked the Directors of projects and Research Units that agreed to take part in the project at the proposal stage to suggest case study participants for the interviews. Each interviewee was asked to respond to a semi-structured interview based on an expansion of our online questionnaire. All sections featured open questions, so that the respondent could elaborate on technical and infrastructural challenges. The interview protocols were structured in two parts:

- An introductory section to gather basic information on the respondent profile and his/her perspective on the scope of discourse; this section also contained generic questions based on the material and the discussion at the EC Public Consultation on Open Research Data, held in Brussels on the 2nd July, 2013 (see chapter 3.1);
- A profile-specific section depending on the respondent's perspective in the WP2 taxonomy that has been introduced in chapter 2.1.

We also distributed the interview protocols to other interested stakeholders in the RECODE contact list, not necessarily associated with the RECODE case studies, including several funders and disseminators/curators, involving around thirty individuals in the interviews. The interview protocols can be found in Appendix 2.

In order to validate and expand the findings from the above activities and obtain additional feedback and insights, we held a stakeholder consultation workshop (Task 2.4 – Workshop). The workshop was organised on the 14th of January 2014 in Geneva, as a side event of 10th Plenary Session of the Group on Earth Observations & 2014 Ministerial Summit. Workshop invites were sent to key stakeholders identified in the document review and the survey, as well as to case study participants and major international initiatives, to share their perspective in understanding Open Access to research data, in relation to infrastructure and technology challenges. The workshop was furthermore advertised on the RECODE and conference website, as well as on the RECODE email list, which holds contact details for various stakeholders throughout Europe.

The workshop attracted over 40 attendees from 14 countries, including policy makers, funding bodies, libraries, data management organisations and researchers, along with representatives from the RECODE case studies and RECODE team members. A complete list of institutions represented at the Workshop can be found in Appendix 3¹¹ and the full workshop agenda can be found in Appendix 4. The Workshop presentations and minutes are accessible from the RECODE website¹².

This document has also been provided to the RECODE international advisory board members to solicit their feedback, which has then been incorporated in a final revised version of the deliverable.

2.1 WP2 STAKEHOLDER TAXONOMY

As different actors in research data management perceive infrastructure and technology challenges differently, we have identified the stakeholder categories of interest for our scope, based on the functional categories elaborated in the framework of WP1 (Stakeholder Values and Ecosystems) and WP6 (Stakeholder Engagement and Mobilisation)¹³, which gives a balanced picture of the stakeholder ecosystem for Open Data, including diverse groups from government, industry, the public and mass media.

We have modified the RECODE stakeholder taxonomy congregating the Disseminator and Curator roles, as we can assume they share similar concerns, with respect to infrastructure and technology. In fact, as shown in this excerpt from the RECODE functional taxonomy, outlining functions, performers, activities and records, the performers of the two functions overlap significantly, differing mainly only for the primary function (PF) and secondary function (SF).

¹¹ Due to issues of privacy, a full list of names will not be made public.

¹² Policy RECommendations for Open Access to Research Data in Europe, “Recode Workshops”, no date. <http://recodeproject.eu/events/recode-workshops/>

¹³ Sveinsdottir, et al., op. cit., 2013, pp. 21-31.

C. Disseminating

Libraries/Archives (SF)

Activity: Disseminates research publications and data

Records: a) Manuals

Universities/Academy (SF)

Activity: Disseminates research publications and data

Records: a) Open Access policy

Data Centres (PF)

Activity: Disseminate, procure and preserve research data

Records: a) Research Management protocols

b) Open Access policy

Publishers (PF)

Activity: Offers publication, recognition and distribution platforms

Records: a) Open Access policy

b) Rights agreement

D. Curating

Libraries/Archives (PF)

Activity: Disseminate, procure and preserve research publications and data

Records: a) Manuals

Universities/Academy (SF)

Activity: Curate and preserve publications and data

Records: a) Open Access policy

Data Centres (SF)

Activity: Disseminate, procure and preserve research data

Records: a) Research Management protocols

b) Open Access policy

Publishers (SF)

Activity: Offer publication and limited preservation

Records: a) Open Access policy

b) Rights agreement

Generic citizens may be considered as potential users of research data (though most likely with a secondary function), for example performing the activity of reuse of publication and data, or social interaction. Moreover, citizens may be viewed as research data producers, in crowdsourcing scenarios (cf. Citizen Science). However, we assume that their involvement in data use and production is mediated by appropriate supporting applications (e.g., mobile apps) that practically isolate them from the implied technological and infrastructural issues, to overcome usability issues. Hence, we decided not to focus specifically on generic citizens

in the scope of our work. This is in line with the overall RECODE stakeholder taxonomy¹⁴, which does not include the generic citizen in its scope.

In conclusion, in the context of WP2 we distinguish between:

- **Producers**
 - The source of research data
 - E.g. researchers elaborating raw sensor datasets
- **Disseminators/Curators**
 - The actors in charge of the distribution and preservation infrastructure (information systems, e-infrastructure) for storage, access, and maintenance of research data
 - E.g. publisher, library
- **Funders**
 - The parties providing financial and policy support to data collection activities in research
 - E.g. research councils, funding agencies
- **End users**
 - The generic final recipient of research data
 - E.g. researchers, the industry, governmental agencies, data users at large

We have used these categories to structure our survey on infrastructure and technology challenges for Open Access to research data, as well as to organize our findings.

¹⁴ Sveinsdottir, et al., op. cit., 2013, pp. 21-31.

3 FRAMING INFRASTRUCTURE AND TECHNOLOGY CHALLENGES: A REVIEW OF THE LITERATURE

In our literature review we consulted and analysed a large number of sources to scope the known technological and infrastructural challenges to Open Access and preservation of research data, and the possible existing solutions for their mitigation. This chapter reports on our main findings from the review. Our scope included technical reports, guidelines and other grey literature, industry material, software documentation, results from previous projects, scholarly articles, and other information to identify initiatives, good practice, and existing solutions relevant to Open Access to research data.

Definitions of research data vary, with some contributions defining research data as potentially all data (including public sector information), and some limiting it to data that is the product of research.

From the perspective of researchers, research data includes all data from an experiment, study or measurement, including metadata and details on processing data. Researchers seem open to generalized Open Access, including even negative/discarded data, with few to no restrictions (except for privacy reasons). For publishers, data linked to publications is part of the publication. Several participants reiterated that data sharing should be recognised as a scientific product, just like a publication. According to some, data sharers should receive incentives.

We have first addressed the initiatives that we consider most relevant to scope the primary challenges on infrastructure and technology, which are the five key issues of data heterogeneity, accessibility and discoverability, preservation and curation, quality and assessability, and security. We have then organized our research in terms of such challenges, according to the stakeholder categories introduced above.

3.1 REFERENCE INITIATIVES IN OPEN ACCESS TO RESEARCH DATA IN EUROPE

In the early stages of working on the literature review we recognised the following initiatives as the most relevant for scoping the main infrastructure and technology challenges pertaining Open Access to research data:

- **EC Consultation on Open Access to Research Data**
The European Commission held a public consultation on open research data on the 2nd July 2013 in Brussels, which was attended by a variety of stakeholders from the research community, industry, funders, libraries, publishers, infrastructure developers and others (around 130 persons)¹⁵. The discussion focused on questions posed by the Commission to structure the debate.
- **Horizon 2020 (H2020) Pilot on Open Access to Research Data**
H2020 features an Open Research Data Pilot aiming to improve and maximize access to and re-use of research data generated by projects; the scope of the Pilot is quite large, covering 20% of H2020-funded projects, which will be required to define a detailed DMP covering individual datasets and deposit the research data, preferably into a research data repository; they shall also take measures to enable third parties to

¹⁵ Information on the consultation, including the agenda, the list of participants, the list of contributions and the final report are available at: <http://ec.europa.eu/digital-agenda/node/67533>

access, mine, exploit, reproduce and disseminate their research data free of charge for any user (e.g., under a Creative Commons License like CC-BY, CC0) as well as provide information about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (such as specialized software, algorithms, analysis protocols, etc.), or the tools and instruments themselves, where possible¹⁶.

- **Big Data**

Although originating from the enterprise sector and usually related to Business Intelligence, many of the infrastructure and technology issues of Big Data are relevant also in the context of Open Access to research data; we anticipate that the future ubiquity of sensors and the uptake of Citizen Science approaches will imply an increasing overlapping of Open Access and Big Data issues, particularly as concerns sharing, preservation and curation of research data; Big Data is data that is impractical to manage (capture, curate, process, share, analyse, visualise) with the traditional tools, given the limitations of the hardware and software infrastructure at a given time; this is a relative definition, whose practical significance changes with the advancement of the technological baseline, currently in the order of the zettabyte (billions of terabytes); Big Data is characterised with the classic 3 V's model¹⁷: Volume, Variety, and Velocity; some definitions add a fourth one: Veracity¹⁸.

- **INSPIRE and GEOSS**

Respectively at the European and at the global level, INSPIRE and GEOSS are both aiming at implementing data sharing across many different scientific disciplines and have recommended a set of specific principles and technologies for data discovery, access, and use; they are recognised as significant initiatives also in the RECODE DoW¹⁹; INSPIRE is a legal framework to ensure the interoperability of spatial datasets and services needed to support environmental policy and policies that affect the environment; GEOSS is a global effort of a voluntary nature; in particular, the GEOSS 10-Year Implementation Plan explicitly acknowledges the importance of data sharing in achieving the GEOSS vision and anticipated societal benefits: *"The societal benefits of Earth observations cannot be achieved without data sharing"*²⁰. To achieve this, GEOSS promotes a set of Data Sharing Principles^{21 22} for full and open exchange of data. Besides, GEO Members are invited to encourage their data-providing organizations to make available datasets as GEOSS Data Collection of Open Resources for Everyone (GEOSS Data-CORE), a distributed pool of documented datasets with full, open and unrestricted access at no more than the cost of

¹⁶ European Commission, Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 1.0, 11 December 2013.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf and European Commission, Guidelines on Data Management in Horizon 2020, Version 1.0, 11 December 2013.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

¹⁷ Genovese, Yvonne, and Stephen Prentice, *Pattern-Based Strategy: Getting Value From Big Data*, Gartner Special Report, 17 June 2011.

¹⁸ The growing debate on Big Data has spurred a parallel proliferation of V's: Validity (data that is correct), Visualization (data in patterns), Vulnerability (data at risk), Value (data that is meaningful), and yet more (Verisimilitude, Variability, etc.)

¹⁹ RECODE Project, Annex I, op. cit., 15 July 2012, pp. 6, 10.

²⁰ Group on Earth Observations, *10-Year Implementation Plan Reference Document*, ESA Publications Division, Noordwijk (The Netherlands), February 2005, p. 205.

²¹ Group on Earth Observations, White Paper on the GEOSS Data Sharing Principles, subsequently published concurrently as *Toward Implementation of the GEOSS Data Sharing Principles*, *Journal of Space Law*, Vol. 35, No. 1, 2009, and *Data Science Journal*, Vol. 8, 2009.

²² GEOSS Data Sharing Working Group, *GEOSS Data Quality Guidelines*, 19 June 2013.

reproduction and distribution. Data CORE has been a key mechanism to address the limitations identified in implementing the Sharing Principles and there has been a big push last year to increase the stock of the CORE, leveraging the voluntary nature of GEOSS.

- **High-Level Expert Group on Scientific Data**

In their final report²³, the High Level Expert Group on Scientific Data identified the benefits and costs of accelerating the development of a fully functional e-infrastructure for scientific data, which already partially exists, but needs a more structured approach and global framework. The working group has developed a far-seeing vision on the issues of a technical e-infrastructure on a European level that is ready for the future:

*“Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.”*²⁴

The Group envisions that, by 2030:

- All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.
- Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted. Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.
- 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.
- The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.
- The public has access to and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it. All can be adequately educated and prepared to benefit from this abundance of information.
- Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Governments become trustworthy.
- Global governance promotes international trust and interoperability.²⁵

The report is a call for action to build an infrastructure to realise this vision, overcoming all the related issues and barriers. They present the so-called Collaborative Data Infrastructure (see Figure 3), which suggests, in the broadest possible terms, how different actors, data types and services should interrelate in a global e-infrastructure for science:

²³ High-Level Expert Group on Scientific Data, *Riding the Wave: How Europe can gain from the rising tide of scientific data*, European Union, October 2010. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

²⁴ Ibid., p. 4.

²⁵ Ibid., pp. 4-5.

“Data generators and users gather, capture, transfer and process data - often, across the globe, in virtual research environments. They draw upon support services in their specific scientific communities - tools to help them find remote data, work with it, annotate it or interpret it. The support services, specific to each scientific domain and provided by institutes or companies, draw on a broad set of common data services that cut across the global system; these include systems to store and identify data, authenticate it, execute tasks, and mine it for unexpected insights. At every layer in the system, there are appropriate provisions to curate data - and to ensure its trustworthiness.”²⁶

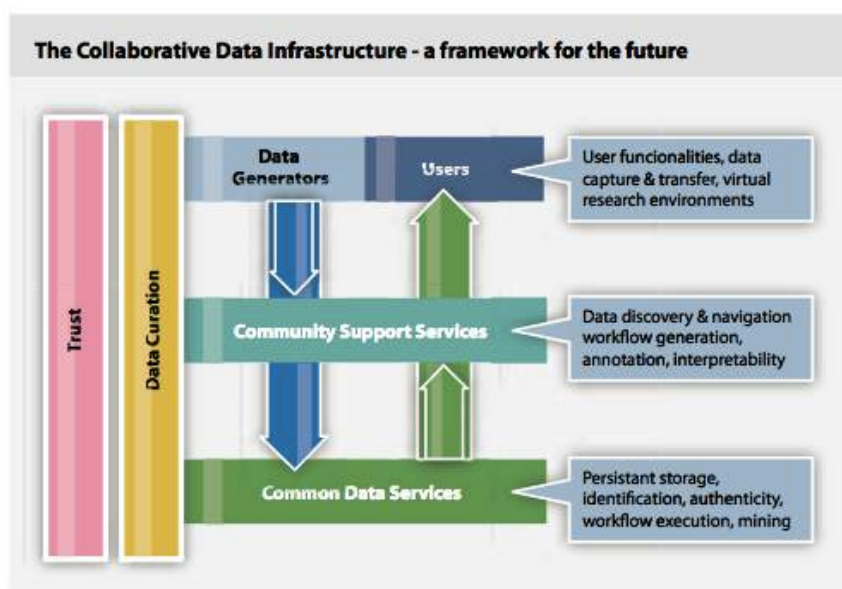


Figure 3 - diagram of the Collaborative Data Infrastructure

The Group identifies the following requirements for technical solutions for the e-infrastructure:

- Open deposit, allowing user-community centres to store data easily;
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years;
- Format and content migration, executing CPU-intensive transformations on large datasets at the command of the communities;
- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information;
- Metadata support to allow effective management, use and understanding;
- Maintaining proper access rights as the basis of all trust;
- A variety of access and curation services that will vary between scientific disciplines and over time;
- Execution services that allow a large group of researchers to operate on the stored data;
- High reliability, so researchers can count on its availability;
- Regular quality assessment to ensure adherence to all agreements;
- Distributed and collaborative authentication, authorization and accounting;
- A high degree of interoperability at format and semantic level.

²⁶ Ibid., p. 31.

3.2 MAIN CHALLENGES IN INFRASTRUCTURE AND TECHNOLOGY

In summary, based on an analysis of the above reference initiatives, we can identify the following broad categories of concern for Open Access to research data:

- **Heterogeneity and interoperability**
As Variety is defined in the Big Data context, decision-makers have always had an issue translating large volumes and types of transactional information into decisions, mainly coming from social media and mobile (context-aware). Variety includes tabular data (databases), hierarchical data, documents, e-mail, metering data, video, still images, audio, stock ticker data, financial transactions and more. This challenge covers the issues that occur because of different ways of formatting, storing, operating, and standardizing the data.
- **Accessibility and discoverability**
This challenge is related to the Big Data aspects of Volume and Velocity, as it involves streams of data, structured record creation, and availability for access and delivery. Velocity means both how fast data is being produced and how fast the data must be processed to meet demand. The increase in data volumes within enterprise systems is caused by transaction volumes and other traditional data types, as well as by new types of data. Too much volume is a storage issue, but too much data is also a massive analysis issue. With the huge amounts of data being stored and accessed issues may arise around bandwidth. Metadata is important for discoverability and therefore accessibility of data.
- **Preservation and curation**
With the ever-growing amounts of data, a pertinent question is what data should be stored indefinitely and what can be purged. Furthermore, when a selection has been made with regard to data being stored, stakeholders must consider the length of time that it will be stored and the method of storage. Decisions and judgements regarding the selection of online storage, with instant and Open Access possibilities for recent and more relevant data, and offline storage for older and less relevant data will need to be made. These decisions will be context specific and in some cases may be subjective.
- **Quality and assessability**
According to some definitions, a fourth V characterises Big Data: Veracity, an indication of data integrity, including trustworthiness, provenance, lineage, quality, and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions. Are there ways of reviewing data? Do we need Data Management Plans in order to increase the quality of the data? Researchers and users need to know if the available data is of good quality. If we want data sharing to be more effective, we need to look into ways of reviewing data by developing and implementing tools to assess their quality.
- **Security**
With regard to protecting sensitive research data, e.g. data from human subjects, a consideration for data security needs to be demonstrated. Security issues incorporate any restrictions on the usage, access, and consultation of data and metadata, and their enforcement from a technical viewpoint, e.g. protocol for authentication, authorization and auditing/accounting, privacy issues, policy enforcement, licensing.

When reviewing literature which refers to infrastructure and technology issues regarding Open Access to research data, it is evident that there is an underlying worry about the vast amounts of data are produced each day, which are neither discoverable, accessible nor re-useable due to lack of curation, storage, and overall management. There is a concern that all this data is now feeding into a “*virtual reservoir*” and is not stored uniformly in one place, but in various formats in scattered disparate repositories of varying sizes across the globe²⁷, see also Pearlman, et al²⁸, and Bermudez²⁹.

The role of technical infrastructure is seen to be the provision of uniform and equal access to a broad variety of research outputs, i.e., making data understandable, searchable, retrievable, available, assessable, and secure. The following sub-chapters are organized by these categories and elaborate on relevant technologies, initiatives, good practice, and existing solutions for each challenge.

3.2.1 *Heterogeneity and interoperability*

From reviewing the literature it becomes clear that seamless Open Access to data is a complex technological undertaking, especially due to the heterogeneity of scientific data practices. The H2020 Pilot will enforce that data is interoperable to specific quality standards by asking this DMP question: is the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organizations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins)?³⁰ Added benefit comes from being able to linking datasets to produce deeper and better-integrated understanding.

“The vocabulary used in the semantic description of data – i.e. in the metadata- can vary so greatly between heterogeneous linked datasets that the whole lacks a shared vocabulary capable of revealing the underlying meaning”³¹

Work on standardisation of language and data practices will be needed to allow for seamless search and location of re-usable data. The Linked Data initiative³² and Open Knowledge Foundation³³ have defined guidelines for publishing structured data in standardized and queryable format.

In the reviewed literature there is an overall consideration of how to successfully implement Open Access to different types of heterogeneous data.

²⁷ Thomson Reuters Industry Forum, *Unlocking the Value of Research Data*, July 2013, p. 3.

<http://researchanalytics.thomsonreuters.com/m/pdfs/1003903-1.pdf>

²⁸ Pearlman, Jay, Albert Williams and Pauline Simpson (eds.), James Gallagher, John A. Orcutt, Peter Pissierssens, Lisa Raymond and Pauline Simpson (Authors), “Report of the Research Coordination Network RCN: OceanObsNetwork. Facilitating Open Exchange of Data and Information”, NSF/Ocean Research Coordination Network, May 2013.

²⁹ Bermudez, Luis, *Making Sense of Millions of Observations Using Open Standards*. Air Sensors, Session III Big Data: Management and Analysis, presentation slides, 2013.

³⁰ 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.

³¹ The Royal Society, “Science as an open Enterprise”, 2012, p. 34.

http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

³² Linked Data – Connect Distributed Data Across the Web, “Linked Data”, no date. <http://linkeddata.org/>

³³ Open Knowledge Foundation, “The Open Knowledge Foundation”, no date. <http://www.okfn.org/>

“As for the data deluge: managing the flood of new data and information is a daunting task, and one that no single organisation – or indeed nation – can manage it alone. Now more than ever, there is a need to integrate diverse life science data from many different databases and make it discoverable. We must respond to the needs of researchers and build usable interfaces that facilitate easy re-use of the material.”³⁴

The focus here is on interoperability and how that can be achieved in datasets that contain many different formats of data. This is also important as End user abilities to search and re-use data need to be borne in mind when accessibility is concerned. The report of the High-Level Expert Group of Scientific Data states that there needs to be a focus on interoperability in order to search through and work with relevant data files anywhere in the world. Researches can jointly work on projects and share data alike. That is possible, but with great effort, skill, cost and time. Focusing and taking big steps in interoperability can make that much more efficient.³⁵

During the EC Consultation the question of how to ensure that data can be re-used led to discussions about technical aspects of heterogeneity of open research data. The discussion centred not just on whether and how data should be re-used, but also on the adequacy of e-infrastructures for data re-use. Some participants suggested avoiding huge centralized repositories, and advocated solutions based on interoperable distributed systems, to leverage on the existing infrastructures. Solutions should also take into account the specificities and attitudes in the different fields of science, which imply very heterogeneous requirements and features.

In the context of re-use, the Directive on the re-use of Public Sector Information (2003/98/EC, currently under revision) was referred to several times. While PSI is distinct from research data and governed by a specific directive, it is important to remember that data from public administrations, where there is a lack of culture for Open Access, can be very valuable to research. Hence, it is worth keeping an eye on technical developments within the field of Open Access to PSI, in case these are useful and can be replicated in the implementation of Open Access to research data.

The need for good standards, with respect to EU legal frameworks, and the importance of metadata has been stressed several times, particularly provenance metadata, to guarantee repeatability. Some comments reinforced the need to promote a culture of standards also in education, and to educate researchers on open data. Simple templates with the approach of Creative-Commons were suggested to the EC as a most effective contribution to convince researchers in adopting Open Access to research data.³⁶

The re3data.org³⁷ focuses on the problems of the heterogeneous research data repository landscape. Data repositories need to serve different academic and disciplinary communities

³⁴ Cameron, Graham, “ODE Project: 10 Tales of Drivers and Barriers in Data Sharing”, European Bioinformatics Institute in Alliance for Permanent Access, 2011, p. 7. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/7836_ODE_brochure_final.pdf

³⁵ High-Level Expert Group on Scientific Data, op. cit., 2010, p 19.

³⁶ European Commission, “Results of the consultation on Open Research Data – Digital Agenda for Europe”, no date. <http://ec.europa.eu/digital-agenda/node/67533>

³⁷ Re3data.org, “Registry of Research Data Repositories”, no date. <http://www.re3data.org/>

with their respective concepts of research data. Information infrastructure requirements arise from these different concepts and user requirements.

These repositories are making their data openly accessible, usually by publication of research data as an independent information object or data with textual documentation (so-called data paper) or publication of research data as enrichment of an interpretive text publication (so-called enriched publication).

These strategies have in common that a technical infrastructure is required that ensures safe storage and accurate accessibility by a data archive, data centre, library, archive and the like. These different kinds of data repositories are all lacking in standardization. The OAI PMH³⁸ was defined to standardize interchange and discovery of research papers, but data repositories so far lack a similar solution, as even on a disciplinary level the diversity is great.

Re3data.org studies have shown that a majority of scientists are willing to place their data, or some of their data, into a central data repository with no restrictions. One of the obstacles to this willingness is the lack of knowledge by scholars on suitable existing repositories. The re3data.org project is attempting to close this gap by developing and operating a directory of research data repositories. To finalize an indexed and structured description of research data repositories of all domains in a web-based registry is the target of the project. In the summer of 2012 the first version of a vocabulary for metadata description of repositories was published. Now repositories can be indexed in re3data.org if only requirements and details on access to and licensing of the data are met.³⁹

Interoperability is key to any data system. As pointed out in the literature⁴⁰, there are many levels of interoperability, “*from basic machine interactions to human exchanges to human rewards and motivations*”. On the machine side, two extremes have been identified and between them are a variety of approaches that mix varying degrees of each of them:

- A brokering approach, which provides an intermediary information system layer that translates between different domain information infrastructures.
- A Federated approach, which mandates certain standards that must be followed by each domain system so that the different systems will be interoperable.

Both of these approaches “*must address the issues of semantics, metadata, workflows, and so on. The brokering approach reduces the workload on discipline repositories by centralizing the interoperability developments into the middleware layer. This encourages greater participation on the part of the discipline information infrastructures by reducing local efforts.*”⁴¹ The authors argue for the brokering method rather than strict standardisation, which they maintain is a distant dream. The former allows the domain system to maintain its independence while enabling full interoperability, and “*provides an intermediary information system layer that translates between different domain information infrastructures allowing the domain system to maintain its independence while enabling full interoperability.*”⁴².

³⁸ Open Archives Initiative, “Protocol for Metadata Harvesting”, no date. <http://www.openarchives.org/pmh/>

³⁹ Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirnbacher and Uwe Dierolf, *Making Research Data Repositories Visible: The re3data.org Registry*, PLOSOne, November 2013, Volume 8, Issue 11.

⁴⁰ Pearlman, et al., op. cit., 2013, p. 10.

⁴¹ Ibid.

⁴² Ibid.

There are well known instances of the brokering method, one of which is GEOSS, discussed in detail in this report.

An example of a more federated approach is the PANGAEA Data Publisher for Earth and Environmental Science⁴³, which curates and store scientific data from all kinds of multidisciplinary scientific programs and publications. In their work PANGAEA became “*an agent for homogenization of analytical measurements assigned (by the scientific community) to define accepted parameters.*” These parameter definitions are seen as imperative for data management and storage. In the case of PANGAEA data repositories with trained scientific data curators are needed to assure true scientific parameter homogenization. The data submitted originally is not simply taken in without question, but assembled datasets are sent back and forth between PANGAEA data curators and the authors until it is finally quality assured and validated by the responsible author. This is described as a long-term process and a time consuming task for the authors involved.⁴⁴

Brokering technology can facilitate interoperability at all levels. It integrates and supplements the standardization approach building effective systems of (autonomous) systems. Ultimately, interoperability solutions of a global nature will be a combination of middleware, standards and a compendium of good practice.

Pearlman, et. al. cite an EC communication, which states that interoperability encompasses at least three overarching and different aspects:

1. *“Semantics, which ensures that exchanged information is understandable and usable by any application or user involved;*
2. *Technology, which concerns the technical issues of linking up computer and information systems, the definition of open interfaces, data formats and protocols.*
3. *Organization, which deals with modelling organizational processes, aligning information architectures with organizational goals, and helping these processes to co-operate. This category can also include important interoperability challenges, like: data policy, legal, cultural, and people harmonization.”*⁴⁵

Interoperability is not an on-off capability; there are various levels of interoperability. Different models for levels of interoperability already exist and are used successfully to determine the degree of interoperability implemented by a disciplinary infrastructure. One of them, the Level of Conceptual Interoperability Model (LCIM), applies well to assess the Earth Sciences infrastructure levels of interoperability. This goes beyond the technical interoperability addressing conceptual/semantic models interoperability.

Pearlman, et.al. (2013) consider universal standards for data a distant hope, and similar to the hope of a universal language. The solution here is middleware development and information brokering. Ultimately, the broadly inclusive collaborations across scientific disciplines need a more formal way to make data generally available. Translators for formats must develop as a middleware market. Recent developments in information brokering have been quite encouraging and demonstrations with selected user scenarios and communities have pointed to significant benefits. Further development, implementation and uptake of brokering middleware are recommended as an important step forward.

⁴³ PANGAEA, “Data Publisher for Earth & Environmental Science”, no date. <http://www.pangaea.de/>

⁴⁴ Cameron, op. cit., 2011, p. 9.

⁴⁵ European Commission, Communication from the Commission to the Council and the European Parliament: interoperability for Pan-European eGovernment Services, COM (2006) 45 final, Brussels, 2006.

The OpenAIRE⁴⁶ and its follow-up OpenAIRE plus are EC funded projects whose goals are to deliver an Open Access scholarly communication data infrastructure for Europe capable of collecting and monitoring Framework Programme article output. OpenAIRE has the goal to provide a domain agnostic metadata schema and provide interoperability through a small number of properties, making interoperability possible in the simplest manner possible and as a result keep the technical barriers for implementation as low as possible.⁴⁷ Among OpenAIRE plus major objectives are experiments to interlink datasets and publications across different disciplines, by automatically inferring semantic relationships between them, by enabling end users to construct enhanced publications, and by interoperating with existing infrastructures.⁴⁸

The Orbital project at the School of Engineering at University of Lincoln, UK is developing a university research data management infrastructure. The project has produced a literature review⁴⁹ on research data management and, based on this, has proposed a set of recommendations in support of further development of their research data management structure. Some of the technical recommendations are to incorporate the functionality of DataCite⁵⁰ via its API into the local application, in order to allow Lincoln researchers to secure a DOI for their data objects.

In the roadmap and policy document “*Open Overheid*”⁵¹ the Dutch government is making steps towards full open governmental information. These developments follow the Open Government Partnership project.⁵² We can learn some interesting things from opening up citizen and governmental data. To realize fast and easy access to public information it is important that open standards are included wherever possible in the design of the information systems. New systems will not directly provide a change in approach and culture, but they will only facilitate change. In the document there is a passage on the development and installation of the big information database with governmental data. This involves both structured information (data) and unstructured information (documents). Although the approach may be different, it is both the design of databases and the design of information systems, which needs to take into account important aspects that contribute to openness, the degree of openness and sustainable digital access (archiving). Open standards, open formats for reusability, metadata and linked data for traceability and consistency, but also privacy, level of security and accessibility are aspects that already need to be taken into account during the process of document creation.

⁴⁶ OpenAIRE, “Home”, no date. <https://www.openaire.eu/>

⁴⁷ Elbaek, Mikael, and Lars Holm Nielsen, “OpenAIRE Guidelines for Data Archive Managers 1.0”, June 2013. http://openaire-dev.cern.ch/record/6918/comments#.UzACC8g_wmc

⁴⁸ Manghi, Paolo, Łukasz Bolikowski, Natalia Manola, Jochen Schirrwagen and Tim Smith, “OpenAIRE plus: the European Scholarly Communication Data Infrastructure”, *D-Lib Magazine*, Vol. 18, n. 9/10, September/October 2012. DOI 10.1045/september2012-manghi

⁴⁹ Stainthorp, Paul, *An Engineering Research Data Management (RDM) Literature Review*, Orbital project, University of Lincoln, 2012. <http://orbital.blogs.lincoln.ac.uk/files/2012/04/Literature-review.pdf>

⁵⁰ DataCite, “Helping you to find, access, and reuse data”, no date. <https://www.datacite.org/>

⁵¹ Ministerie von Binnenlandse Zaken en Koninkrijksrelaties, *Actie plan: Open Overheid*, September 2013, p. 21. <https://data.overheid.nl/sites/data.overheid.nl/files/actieplan-open-overheid.pdf>

⁵² Open Government Partnership, “What is the Open Government Partnership?”, 2013. <http://www.opengovpartnership.org/>

3.2.2 Accessibility and discoverability

The importance of accessible formats and sharing of code, software and hardware so as to facilitate access and re-use of data is key when it comes to the different needs, demands and capabilities of potential users. H2020 provides the following definitions that help scoping this fundamental challenge:

- Accessible – DMP question: are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licensing framework for research and education, embargo periods, commercial exploitation)?
- Discoverable – DMP question: are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?⁵³

In the ODE project⁵⁴ the partners have set out to describe the factors that motivate, inhibit and enable the sharing of research data. These may be variously defined in terms of individual-psychological, social, organisational, technical, legal and political components. A recognised technical and infrastructural barrier is data discovery, where the ODE project concludes that there is no supportive infrastructure for international, cross-disciplinary data discovery. Three possible enablers here are Open Linked, Persistent, unique data identifiers with search engines (e.g. DataCite⁵⁵) and Interoperating Data Centres in specific disciplines (e.g. CESSDA in Social Science).

Metadata standards are essential for discovery and identification. In conjunction with this, standard data formatting is very important in order to have content prepared for machine processing. DOIs, as described above, are mentioned as a good example of a usable standard. Data citation and description for discovery and use are mentioned as a very important factor. Researchers are recommended to “*create and describe citable datasets using appropriate disciplinary metadata*”.⁵⁶

In the document “*Information Economy Strategy. Industrial Strategy: government and industry in partnership*”, produced by the UK government initiative “*The new Information Economy Council*”, an excellent digital infrastructure is mentioned as a pre-condition for a thriving information economy sector. The UK government digital strategy is focused on making government data more accessible, relevant and updated using the platform GOV.UK.⁵⁷ The consumer expectations on the availability, speed and reliability of the Internet make the government ambition for UK digital infrastructure to increase speed and standards in the broadband and to support the development of 5G technologies.⁵⁸

⁵³ 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.

⁵⁴ Dallmeier-Tiessen, Sunje, Robert Darby, Kathrin Gitmans, Simon Lambert, Jari Suhonen and Michael Wilson, *Compilation of Results on Drivers and Barriers and New Opportunities*, Opportunities for Data Exchange (ODE), 2012.

⁵⁵ DataCite, “Helping you to find, access, and reuse data”, no date. <https://www.datacite.org/>

⁵⁶ Dallmeier-Tiessen, et al., op. cit., 2012, p. 52.

⁵⁷ GOV.UK, “Welcome to GOV.UK.”, no date. <https://www.gov.uk/>

⁵⁸ Information Economy Strategy. Industrial Strategy: government and industry in partnership, HM Government, June 2013.

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206944/13-901-information-economy-strategy.pdf

The Global Biodiversity Information Facility (GBIF) is a major international open data infrastructure, funded by governments. The GBIF representatives Chavan and Ingwersen, concludes that scientific data is neither easily discoverable nor accessible. Another impediment is the lack of professional recognition of scientific data publishing efforts. Their solution to these and other barriers is the establishment of a data publishing framework. One of the five components of this framework is technical-infrastructure. Three of the major components of the technical infrastructure are:

- Persistent Identifiers to data publishers, datasets, the data record itself, as well as to data versioning, and data citation;
- Data Usage Index (DUI) at every access point;
- An effective Data Citation mechanism.

These elements are interdependent. About persistent identifiers to data it is said that: “*Unique global identifiers should be assigned not only to datasets, but also to its publishers, every individual datum and its author(s), data versioning, and data citation. Further, simplified mechanisms are needed that make it easy for individuals to assign these identifiers to their data. Given the options available, the choice of choosing the suitable unique global identifier should reside with data publishers*”.⁵⁹

The authors furthermore include the above components in 24 recommendations for GBIF. Recommendation 13 emphasizes “*the need for GBIF to adopt a stable and proven persistent identifier such as the digital object identifier (DOI), rather than unstable persistent identifiers*”. Recommendation 14 states that GBIF must develop a prototype of a data usage index. Recommendation 15 mandates that GBIF institutionalize a data citation mechanism and establish a data citation service facilitating deep-data citation, registration and resolving of citations. Other technical infrastructure recommendations regard the development of a tool to convert tabular data into resource description framework (RDF) formats conforming to a standard ontology, and the facilitation of discovery and mobilization of all streams/types of relevant biodiversity data.⁶⁰

In summary, the following technologies and resources can be considered good practice to address the problems related to data discovery and access, such as identification:

- DOI (Digital Object Identifier) and other persistent identifiers – in recent years several technologies have been developed to assign persistent identifiers to specifically digital data. To name a few: ARKs, DOIs, XRIs, Handles, LSIDs, OIDs, PURLs, URIs/URNs/URLs, and UUIDs. DOIs are the identification scheme currently adopted by most (commercial) publishers for their online publications. Not that many publishers yet are experimenting with data and data publications. Only just recently a non-profit publisher as PLOS announced⁶¹ an updated data policy⁶² supporting public access to data. Persistent identification allows data centres to register a huge amount of markers to track the origins and characteristics of the information. With the fast

⁵⁹ Chavan, Vishvas, and Peter Ingwersen, “Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community”, *BMC Bioinformatics*, vol. 10 (Suppl14):S2, 2009.

⁶⁰ Moritz, Tom, S Krishnan, Dave Roberts, Peter Ingwersen, Donat Agosti, Lyubomir Penev, Matthew Cockerill and Vishvas Chavan, “Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing framework task group”, *BMC Bioinformatics*, vol. 12 (Suppl. 15):S1, 2011.

⁶¹ PLOS BLOGS, “PLOS’ New Data Policy: Public Access to Data - EveryONE”, 24 February 2014. <http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data/>

⁶² PLOS, “Update on PLOS Data Policy”, 23 January 2014. <http://www.plos.org/update-on-plos-data-policy/>

growing amount of data it is essential to come up with solutions and standardization. Metadata support allows effective management, use and understanding.

- ORCID⁶³ (Open Researcher and Contributor ID) – provides a persistent digital identifier that distinguishes a researcher from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between the researcher and his professional activities ensuring that your work is recognised. The services became live in 2012.
- DataCite⁶⁴ – an international organisation which aims to establish easier access to research data, increase acceptance of research data as legitimate contributions in the scholarly record, and to support data archiving to permit results to be verified and re-purposed for future study; provides schema, API, services for data citation. Founded 2009.

3.2.3 *Preservation and curation*

The H2020 Pilot requires that data is useable beyond the original purpose for which it was collected. The proposed DMP question is: are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?⁶⁵

The already mentioned ODE project⁶⁶ recognises preservation as a technical and infrastructural barrier to the sharing of research data. Researchers will not prepare and submit data until there is an acceptable infrastructure for preservation. Some specific issues under this challenge have been identified:

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

The project suggests solutions for these impediments:

- Archives supported by journal publishers (e.g. Nature) sustained by a business model;
- Archives supported by learned societies (e.g. the CAS Registry of the American Chemical Society) sustained by a business model;
- Archives funded by funding bodies (e.g. UK Economic and Social Data Service);
- Institutional archives (e.g. ESO archive of astronomical images, university archives proposed by USA NSF and UK Research Councils).
- E-infrastructure to support/share the effort of creating the metadata needed to enable the re-use and combination of data from multiple sources, e.g. the SCIDIP-ES project.

⁶³ ORCID, "Distinguish yourself in three easy steps", no date. <http://orcid.org/>

⁶⁴ DataCite, "Helping you to find, access, and reuse data", no date. <https://www.datacite.org/>

⁶⁵ 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.

⁶⁶ Dallmeier-Tiessen, et al., op. cit., 2012.

It is clear that an investment in long-term preservation must be undertaken in order to maximise the exploitation of stored data. For example, in the case study of particle physics, experiments cannot be re-run without a major revival effort. Likewise, rapidly changing technology can be a challenge for data preservation. In general, hardware and software soon becomes out-dated or unreadable. Migration to new platforms and virtualization of the software are some of the efforts that have to be invested in for long-term preservation and re-use.⁶⁷

A part of the preservation and curation task is to make sure hardware and software is kept up to date, and also make sure that older hardware and software needed for reading older data is curated alongside datasets, in order to ensure these are still accessible. A report by the EC-funded PARSE.Insight project, focusing on preservation of digital information in science, suggests that the responsibility for preservation issues could be supported by a government level organisation, for example a component of the EU. An alternative approach is to move the responsibility to a “*consortium-based organisational structure, such as the Alliance for Permanent Access. This structure brings together key stakeholders in many sectors and can play a key role in developing standards for preservation and curation. However, even this may need to be underpinned by governmental guarantee in order to provide real confidence in the infrastructure’s longevity*”⁶⁸, as infrastructure run on short term grants may not be trusted by scientists to store and curate data long term.

Not only does this top-down approach result in better interoperable and sustainable networks, but it also draws a clear scenery of the European science landscape, allowing new stakeholders to build a business model on top of the infrastructure. Researchers are assured that their data is compatible and safe because of certification and legislation while new businesses can offer new services on top of this secure layer of the infrastructure. A good example is the OAIS Reference Model (ISO 14721:2003), which has become a worldwide-adopted standard for building a sustainable digital archive.

The EC Consultation on Open Research Data⁶⁹ recognised the need for improved data management practices as a key concern, closely linked with data preservation and the sustainability of data repositories. The readiness of professionals to engage in data curation was also highlighted. Some participants wondered whether a new profession is emerging, like an “*embedded informationalist*”, specifically for data documentation. All stakeholders agreed that any funding body policy on open research data must call on researchers to take the issue of data management seriously by developing Data Management Plans for their research projects.

Also related to sustainability, namely to the governance of versioning, is a data citation mechanism with standards “*flexible enough to accommodate deep citations, versioning, as well as any amount of additional information of interest to archivers, producers, distributors, publishers, or others without losing functionality. An issue around citations of versions of same datasets is critical and needs to be resolved in such a way that links between prior and new versions are functional and consistent*”.⁷⁰

⁶⁷ Cameron, op. cit., 2011, p. 15.

⁶⁸ PARSE.Insight, “Deliverable D2.2 – Science Data Infrastructure Roadmap”, 2010.
http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

⁶⁹ European Commission, “Results of the consultation on Open Research Data – Digital Agenda for Europe”, no date. <http://ec.europa.eu/digital-agenda/node/67533>

⁷⁰ Chavan, et al., op. cit., 2009.

The GEOSS Common Infrastructure (GCI), as the primary tool where the interaction between data providers and users are materialized, must provide efficient and effective support to the implementation of the Data Sharing Principles. This requires the long-term sustained operation of the GCI itself. Until now, the GCI has been maintained on a voluntary basis, in accordance with the GEOSS implementation methodology. The GEOSS Action Plan calls for the GEO Members and Participating Organisations to make resources available for the sustained operation of the GCI and the other initiatives set out.

3.2.4 *Quality and assessability*

The H2020 Pilot asks to ascertain if data are assessable and intelligible with the following DMP question: are the data and associated software produced and/or used in the project 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)?⁷¹

‘With the increasing size and complexity of the data being produced, a major bottleneck today is the contextualization and integration of data. A researcher who is interested in a particular topic might want to look beyond one specific analysis to other research that might be related. How can we integrate and display this information?’⁷²

A report from the ODE project⁷³ gives a very good assessment of the quality issues at stake. For instance, as regards the technical infrastructure, trustworthiness of the data as related to usability, quality control and metadata are problematic. Major issues here are:

- Not “*feeling safe*” in dealing with unfamiliar data;
- Impossibility of data centre staff having detailed technical knowledge of all data (e.g. museum curators);
- Lack of clear definition of the level of data quality that the potential data users will require;
- Interdisciplinary data requires a unifying factor for data to make reuse easier (e.g. data maps to a common geographical co-ordinate system);
- Datasets not meaningful in themselves; need algorithms and software to interpret them;
- Lack of clear definition of the metadata that the potential data users will require to interpret the data;
- Lack of a process to ensure quality standards and ensure acquisition of metadata;
- Lack of data management training for staff;
- Cost of providing the effort to ensure the quality standards are enforced, and the metadata gathered.

⁷¹ 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.

⁷² Cameron, op. cit., 2011, p. 7.

⁷³ Dallmeier-Tiessen, et al., op. cit., 2012, p. 19.

Solutions for these impediments could be a combination of:

- Agreeing auditable standards for publishable data quality and metadata within disciplines;
- Certification of data centres for data quality and usability by a trustworthy body;
- Peer reviewing of data supporting academic research publications to certify its quality;
- The development of education and training materials for these data quality standards;
- The training of data producers with these materials;
- Implementation of automated data quality and metadata content tools to test pre-archive data;
- Providing the rewards to lead to the contribution of producer effort required (see the incentives barrier below);
- Inclusion of a mandatory data management plan in research project proposals;
- Introducing specific job profiles with career paths for data preparation and quality assurance staff – such staff may be embedded in research groups or hosted in data centres;
- Overcoming the financial barrier to pre-archive activities.

Data sharing is something a minority of researchers are engaging in. It has been repeatedly reported that one of the great infrastructural reasons for this is the lack of academic credit for sharing data. Data citation is a major issue that has to be addressed, especially by the Disseminators/Curators, to promote data assessability. According to Dallmeier-Tiessen, et al.⁷⁴, “*data centres have a role to play in validation and quality assurance of metadata. But publishers are essential if universal standards of effective citation are to be embedded in the system as whole*”. The minimal requirements for this should be:

- Persistent resolvable identifiers (e.g. DataCite⁷⁵ DOIs) and a stable architecture for resolving them;
- Consistent citation formats;
- Universal data citation rules applied by publishers;
- Appropriate descriptive metadata associated with the dataset, so that users can understand the data and assess their relevance;
- Provenance metadata (creator(s), source organisation, holding organisation), so that users can assess the value of the dataset, its authority and trustworthiness

There is really no standard format for data metrics available. The reason for this is that the development of metrics for data is dependent on that researchers are sharing data on a bigger scale. On the other hand proper metrics would help researchers making their data more visible. In a report for Knowledge Exchange⁷⁶ the authors explored the possibilities of creating metrics for data and they point out one very important thing that must be in place in order to increase data sharing and that is a reward system for scientists that considers data metrics. Such a system could be approached from two perspectives:

- *“Instead of formally citing datasets, the users typically acknowledge data use in the text of the document or in the acknowledgements section. This needs to be changed by promoting data publication and data citation among researchers, and particularly*

⁷⁴ Dallmeier-Tiessen, et al., op. cit., 2012, p. 52.

⁷⁵ DataCite, “Helping you to find, access, and reuse data”, no date. <https://www.datacite.org/>

⁷⁶ Costas, Rodrigo, Ingeborg Meijer, Zohreh Zahedi and Paul Wouters, “The Value of Research Data - Metrics for datasets from a cultural and technical point of view”, Knowledge Exchange, 2013. <http://www.knowledge-exchange.info/datametrics>

developing a publication model where they can see the general advantages (both from a general scientific point of view, but also from an individual point of view) that data sharing can bring to their scientific careers and the development of their work”⁷⁷

- By both research institutions and funding organizations track and use data metrics for funding, hiring, tenure and decisions about promotion. Another need to be developed is a standard for data citation since there are no guidelines for the description of datasets. The basis for developing a reward and a citation system is of course that institutions must fund and maintain a data-sharing infrastructure, which means investing in repositories, data management, curation etc.

Another important aspect of quality and assessability of data is a repository accreditation. The quality of data repositories must be increased in order to make them trustworthy for researchers, users, publishers and other service providers, being able to handle data for the long term. Standards such as the Data Seal of Approval⁷⁸, the Deutsche Initiative für Netzwerkinformation (DINI) Certificate⁷⁹, Trustworthy Repositories Audit and Certification (TRAC)⁸⁰ and ISO 16363:2012 are reported as a source for accreditation.

3.2.5 Security

In the results of the EC Consultation, potential barriers to Open Access are connected with issues of public security, privacy and data protection, as well as IPR protection and possible commercialisation. Concerning public security, the potential use of data for terrorism was mentioned. Privacy and data protection are seen as particularly relevant in areas like health, in particular for clinical trials and the issue of opening up negative results. For IPR and possible commercialisation of research results, representatives from industry expressed the view that data resulting from projects that are close to market should not be open, as a rule, but may be opened on an individual case-by-case basis.⁸¹

One of the factors in a functioning data infrastructure is a trustworthy data repository where curated records can be found and reused in a safe way. In order to facilitate this a number of certification schemas and accreditation procedures have been developed around the world. Now a working group with representatives from two of the certification communities aims to develop a common framework for certification and a service of trusted data repositories. Doing this they hope to simplify the certification options with a procedure that requires less time and effort, which will greatly benefit researchers, funders, data repository services and science publishers. The long-term goal is the development of a global network of trusted, certified digital repositories.⁸²

A framework for privacy and security is another important factor for the information economy. Here the government will continue to engage in the European Union to ensure

⁷⁷ Costas, et al., op. cit., p. 29.

⁷⁸ Data Seal of Approval, “Towards sustainable and trusted data repositories”, no date. <http://www.datasealofapproval.org/en/>

⁷⁹ 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>

⁸⁰ 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

⁸¹ European Commission, “Results of the consultation on Open Research Data – Digital Agenda for Europe”, no date. <http://ec.europa.eu/digital-agenda/node/67533>

⁸² Repository audit and certification interest group DSA_WDS partnership working group case statement, Draft case statement to the RDA Council, January 2014.

guidance on data protection and data sharing. It promises to lead the way in using an Identity Assurance (IDA) approach for securing public digital services. This identity authentication system will promote international interoperability and enable better services. 2011 a National Cyber Security Strategy was published and its first priority was to block cyber crime and secure digital trading places. This strategy will be further expanded through partnerships with industry and academia.⁸³

In order to make researchers publish their data there must be mechanisms that give them credit and academic prestige. *“The Data usage Index is intended to demonstrate to data publishers that their efforts creating primary datasets do have Impact by being accessed and viewed or downloaded by fellow scientists. By visiting (searching or retrieval) and viewing dataset records one may assume interest in the dataset, whilst the volume of downloading volume may demonstrate usage.”*⁸⁴

3.3 FINDINGS FROM THE LITERATURE REVIEW

The literature review tells us that technical issues are being discussed in a relatively small grid of reoccurring problems. If we talk about open research data, questions around standardization, interoperability, reuse and preservation are prevalent. In contrast, relatively few issues arose in relation to bandwidth, storage capacity or usability. Those issues often have a practical solution, such as placing extra servers, using online tools, such as community software, cloud-services, etc.

Arguably, there is a big role for the Disseminators/Curators (libraries, publishers, institutional repositories) to develop solutions for the issues and challenges, like heterogeneity of data and ways of accessing and discovering the data through dedicated repositories or databases run by publishers. These are the stakeholders that deal directly with most of the technical developments in order to store, preserve and access the data.

A big issue, which can be found in the literature and also was discussed during the workshop (see chapter 6), is the vast amount of data and the problem of interpreting all this data. The data itself is meaningless. We need context, algorithms, software and documentation to interpret, re-use and discover the data. This is related to the overall challenges of heterogeneity and assessability.

In a survey on scientific information in the digital age, sent out to Open Access stakeholders all over Europe, over 1100 responses were received. About 90% of the respondents disagreed with the statement: *“Generally speaking, there is NO access problem to research data in Europe”*. There was no major discrepancy among stakeholders on this question. In the follow up question on potential barriers to research data, *“Lack of funding to develop and maintain*

⁸³ Information Economy Strategy. Industrial Strategy: government and industry in partnership, HM Government, June 2013.

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206944/13-901-information-economy-strategy.pdf

⁸⁴ Chavan, et al., op. cit., 2009.

the necessary data infrastructures” was rated by more than 80% of the respondents as an important or very important barrier.⁸⁵

In GEOSS, an Action Plan⁸⁶ was developed, noting that the Data Sharing Principles may remain an abstract goal until all parties (members, contributors, users) can appreciate how they take form concretely. The Action Plan identifies some of the barriers to implementing the GEOSS vision. Much of the perceived barriers are of financial nature. Various data providers have the perception that the implementation of the full and open exchange of data, metadata and products in GEOSS could pose challenges to their development, resulting in limited revenue, in particular as payments for reuse are not consistent with the accepted Implementation Guidelines for the GEOSS Data Sharing Principles. Further, many providers cannot see a clear articulation of a business model linked to the adoption of the principle of full and open exchange.

A Dutch study on open data government strategies⁸⁷ mentions technical barriers such as: limited quality of data; complicated dataset structures; lack of standardization of open data policy and standards; network overload. However, the study concludes that a crucial barrier for implementation of Open Access is not the technical infrastructure as such, but the closed culture within government based on fear of disclosure of government failures, together with the lack of understanding of the precise effects of open data strategies, which restrains the drive for opening up government data.

In summary, we can say that, on the basis of the literature review, infrastructure and technology challenges are not considered the most important obstacles to Open Access to research data, compared to financial, legal, and policy challenges. This finding is furthermore echoed in our online survey results and in our case study findings, which are presented in chapters 4 and 5 respectively.

⁸⁵ European Commission, Online survey on scientific information in the digital age, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

⁸⁶ Group on Earth Observations, *GEOSS Data Sharing Action Plan*, GEO-VII Plenary, 3-4 November 2010.

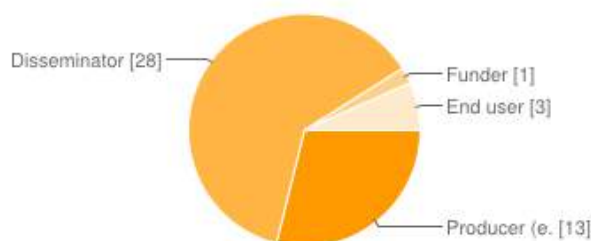
⁸⁷ Huijboom, Noor, Tijs van den Broek, Valerie Frissen, Linda Kool, Bas Kotterink, Morten Meyerhoff Nielsen and Jeremy Millard (Authors), Punie, Yves, Gianluca Misuraca and David Osimo (Editors), *Public Services 2.0: The Impact of Social Computing on Public Services*, Institute for Prospective Technological Studies, 2009. <http://ftp.jrc.es/EURdoc/JRC54203.pdf>

4 SCOPING INFRASTRUCTURE AND TECHNOLOGY CHALLENGES: AN ONLINE SURVEY

Further to reviewing key literature, we sent out a scoping questionnaire to the broader stakeholder communities, to further explore the prevalence of the key issues that had been identified in the literature review, i.e. areas of data heterogeneity, accessibility and discoverability, preservation and curation, quality and assessability, and security.

We designed an online survey using Google Drive Forms⁸⁸ and distributed it widely among key stakeholder groups. For this purpose the RECODE stakeholder mailing list was used, alongside advertising the survey on the EC DG Connect and the RECODE Twitter accounts.

The survey was open between November 2013 and January 2014 and yielded 45 responses, 62% came from stakeholders self-identifying as Disseminators/Curators, and 29% as Producers – from natural and computer sciences. Only three respondents identified themselves as End users, which could early bias the replies. However, we could assume that most of the Producers act also as Users, so that their concerns could still be represented in the survey results. However, we have not assessed this assumption. As only one respondent identifies as a Funder, we can neglect the survey results for this category.



Producer (e.g. researcher)	13	29%
Disseminator/Curator (e.g. publisher, librarian)	28	62%
Funder	1	2%
End user	3	7%

Further to general questions directed at all stakeholder groups, the survey asked questions relevant to each stakeholder group specifically, in order to explore the prominence and importance of the issues within each group.

Due to the distribution method of survey links and the openness and accessibility, it is difficult to calculate a response rate. It is evident that the sample is rather small, and we do not intend here to generalise over a whole population, we do however feel the answers present interesting results that helped steer our subsequent research. The survey results were also used to stimulate a workshop discussion amongst the RECODE workshop participants.

In this chapter, we present descriptive statistics for the most relevant and significant questions, in order to shine a light on technological and infrastructure issues as the group respondents see them. As described above, the questionnaire contained mostly closed questions and was divided into two parts, in order to ascertain:

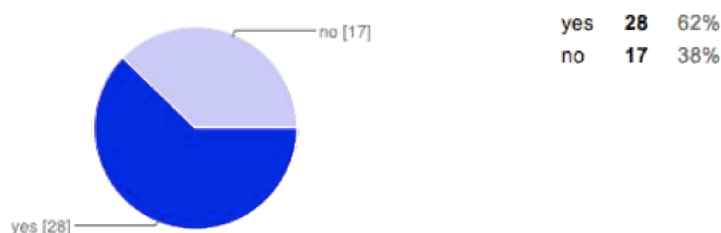
⁸⁸ Google Drive, “Tools to help you get your stuff done”, no date. <http://www.google.com/drive/apps.html>

- The respondent profile and his/her perspective on Open Access to research data; this section is for all the stakeholders and also contains generic questions based on the material and the discussion at the EC Public Consultation on Open Research Data (see chapter 3.1);
- The respondent's perspective according to a specific stakeholder category in the WP2 taxonomy introduced in chapter 2.1; this section is profile-specific.

4.1 KEY ISSUES FOR ALL THE STAKEHOLDERS

In the introductory question, to which all respondents replied, we asked whether they felt that data should be preserved indefinitely, in principle and, if not, who should decide on what to purge. It is interesting to note that a majority (62%) of respondents feel that all data should be kept indefinitely and indicates that they see value in data, irrespective of its type and age.

Do you think research data should all be preserved indefinitely, in principle?



If not, who should decide what to preserve and until when?

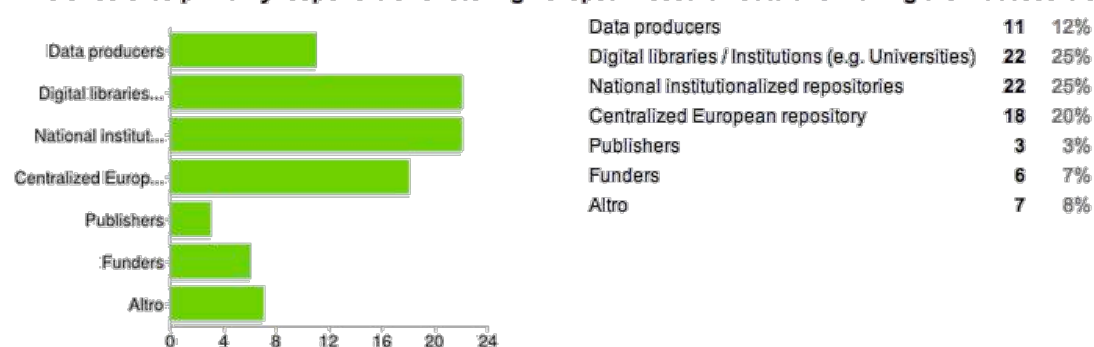


When asked about which group should decide which data to preserve and for how long, interestingly the great minority (4%) chose Funders and none chose Publishers. Here it needs bearing in mind that the groups that took the survey were mostly either ‘disseminators’ or ‘producers’ of data. This does however raise interesting questions regarding trust when it comes to discussing responsibility for data and its curation. For example, questions regarding the trend of enriched publications where publishers are responsible for providing data in relation to published scholarly work. It is also worthy of note that funders, many of which have set up repositories for data from their funded projects, are not seen as a group which should have a deciding power of how long data is preserved. The message here could be that, decisions regarding data preservation should not be financially motivated but rather made on the basis of data value, other than financial.

The respondents’ preference for data preservation is for the data producers themselves or disciplinary associations, closely followed by end users and then librarians/repository managers. The response to this question does raise some interesting issues regarding trust and which stakeholder groups are seen as a reliable choice when it comes to preserving data.

When it comes to selecting the stakeholder group that should be responsible for storing European research data and making it accessible, the responses were as follows:

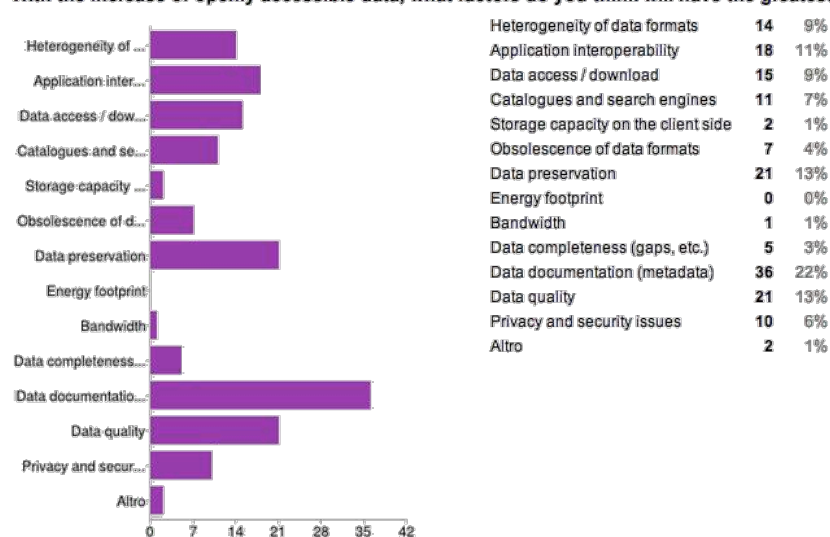
Who should be primarily responsible for storing European research data and making them accessible?



The response to the question posed above indicates that again publishers and funders are the least favoured groups to take on the responsibility for storing research data. For example, 25% of the respondents thought that Digital libraries and national institutionalised repositories should be primarily responsible for data storage. There was even less support for, only 20% of the responses thought that a centralised European repository should hold primary responsibility for data storage and accessibility. The responses indicate that stakeholders prefer that data is stored and made accessible by institutions that are specialised in the field of data and data preservation (70%). This does suggest, as has been detailed in literature on Open Access to research data, that specialised infrastructure and skilled staff will be imperative for a successful implementation of Open Access data policies.

The following question was designed to specifically address the factors that can be considered a barrier to the successful implementation of Open Access to research data. The question was open in the sense that respondents could tick as many issues as they felt were relevant, it was however not a ranking exercise.

With the increase of openly accessible data, what factors do you think will have the greatest impact?



It is interesting to see here that data practices and processes that could be described as non-technical in nature (e.g. data documentation, preservation and quality) are the three barriers

which score highest of the thirteen barriers selected from the literature review and the WP1 RECODE report. This indicates that the perception is that data work might be lacking, which will hinder re-use of Open Access research data. It is however interesting to note that data completeness is not a key concern to respondents, but adequate metadata is. In view of the preferences made for the preservation and storage of data, perhaps the message here is that lack of metadata and quality might in some way be minimized if preservation and storage is in the hands of skilled data specialists, such as digital libraries or national/European data repositories.

With regard to technical and infrastructural issues application interoperability followed by concerns over data access/download and heterogeneity of data formats are also of some concern to respondents, which does chime with the review of documents, in which interoperability remains a key concern.

Noticeably, the technological issues of bandwidth and storage capacity issues are considered rather a low priority, which echoes findings from both WP1 work as well as interviews conducted within the case studies for this report. When considering the group of respondents come from the groups ‘disseminators’ and ‘data producers’ this can indicate that these groups do not deal with issues of bandwidth on a daily basis and therefore do not see this as a potential barrier. It could also indicate that their data practices are by no means heavy on bandwidth. This could imply that the respondents are considering a batch/offline model of computation, possibly overlooking more dynamic computational models, such as workflow-oriented approaches (e.g. Model Web). We have not investigated this further.

We have further analysed the responses to this question, partitioning them by stakeholder group (data are not shown). It becomes clear that data documentation is considered the most important issue for both producers (60%) and disseminators (80%). Both categories also consider prominently data preservation and data quality (46%). The third most pressing issue for producers is heterogeneity of data formats (46%), possibly their primary issue when generating data. Disseminators instead focus on application interoperability (43%), which is indeed a big issue when you want to serve disparate users.

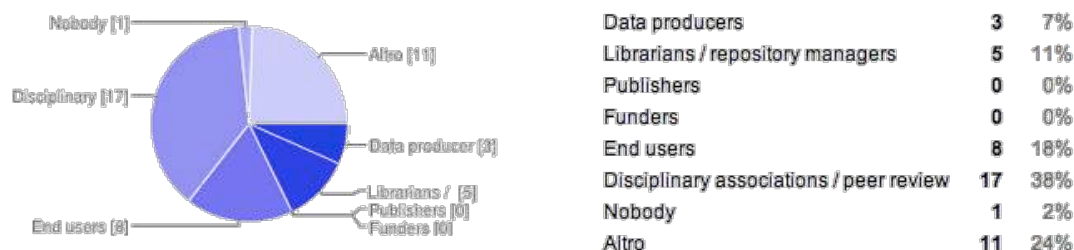
None of the respondents considers the energy footprint a big priority, which could indicate a certain underestimation of digital economy sustainability, as alleged in literature.⁸⁹ The digital infrastructure and economy uses a tenth of the world’s electricity supplies and it is likely that we will see an extensive growth in the upcoming years, following the explosion in data production. It is estimated that by 2020 we will produce 35 zettabytes (1 trillion Gb) annually.⁹⁰ This not only implies issues concerning storage and bandwidth, but also has a tremendous influence on the use of energy sources. Most respondents from the survey seem not really engaged on in this matter. A study by JISC in 2010 shows that the energy footprint of the data collection of the University of Cardiff Developing can be by reduced by 70-80%. So national/international strategies for sustainable data storage therefore seem very

⁸⁹ Walsh, Bryan, “The Surprisingly Large Energy Footprint of the Digital Economy”, *Time Magazine*, 14 August 2014. <http://science.time.com/2013/08/14/power-drain-the-digital-cloud-is-using-more-energy-than-you-think/>

⁹⁰ The Best Computer Science Schools, “Big Data, Small Footprint?”, no date. <http://www.bestcomputerscienceschools.net/big-data/>

relevant⁹¹. In 2010, in the Netherlands, SURF launched its Green ICT and Sustainable Development pilot programme in an effort to stimulate and promote the sustainability of ICT solutions.⁹²

Who should evaluate the quality of research data?



Some of the questions touched on the issue of quality in the context of Open Access, where 38% of respondents felt that quality of data should be evaluated by disciplinary associations or peer review, indicating that this is a process that would fall outside of the responsibilities of those preferred to store and preserve data. However, it is worth noting the quite high number of “other” (“*Altro*”, in the figure), confirming quality as a usual hot topic in science-related discussions. The 11 “other” responses received describe a lack of decisiveness in this area as they did in most instances not nominate a group but were a combination of phrases like “*difficult*” or “*undecided*”. Some respondents pointed out that traditional peer review is impractical for data, and new approaches are needed for this purpose. Possible options mentioned are continuous quality review, a la Wikipedia, or the inclusion of data reviewing in the publication process by scientific journals.

4.2 STAKEHOLDER SPECIFIC ISSUES AND CONCERNS

As explained above, the survey was designed to capture issues respective to each stakeholder group, in order to determine whether perceived issues differed in importance according to which stakeholder group respondents self-identified as. As we explain above, the two largest groups of respondents were Disseminators/Curators and Producers of research data. Hence we present relevant findings from these two groups only, firstly discussing Disseminators/Curators relevant findings and secondly those of Producers.

4.2.1 Key issues for data Disseminators/Curators

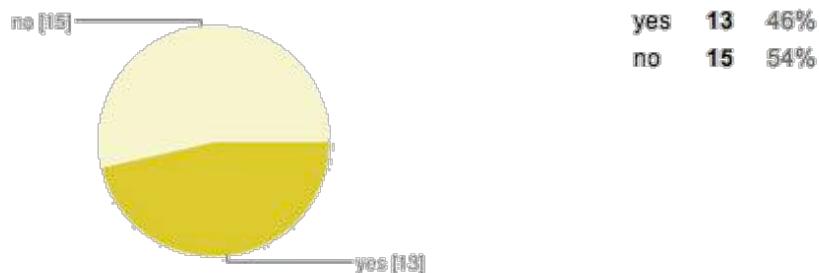
Just under half of the Disseminators/Curators who replied to the survey reported having direct experience of implementing Open Access to research data. Presumably this is because most research data is not Open Access, but restricted to research use only. This is important to bear in mind when results from the overall survey, as well as the Disseminators/Curators specific statistics are read. It suggests that a slight majority of the respondents are reporting on issues they do not have first-hand experience of, and in that case it would be interesting to know on which sources/experiences they base their perceptions of Open Access to research data.

⁹¹ Dickson, Chris, and Paul Rock, “Project Final Report”, Planet Filestore project, December 2010, p.12. <http://www.jisc.ac.uk/media/documents/programmes/greeningict/planetfilestorepp/PlanetFilestoreFinalReport.pdf>

⁹² See: <http://www.surf.nl/en/themes/green-ict-and-sustainability>

Disseminator/Curator of research data

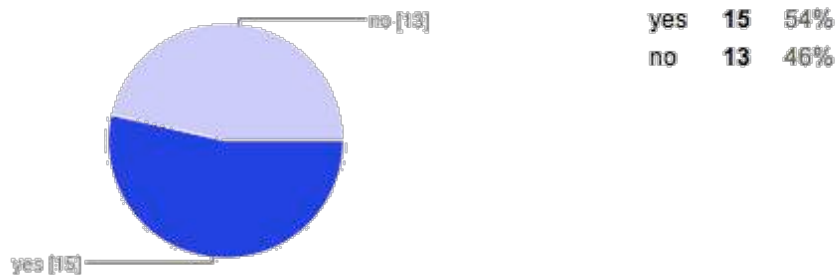
Do you have direct experience of implementing Open Access to research data?



In the specific survey section for the Disseminator/Curator role, we note that almost half of the respondents do not consider user feedback for quality assessment and improvement. This is coherent with the result (not shown in the picture) the half of the respondents do not audit data usage at all. However, a slight majority of respondents consider user feedback for quality assessment and improvement and the same applies to the question regarding tools which allow for linking research data to scientific publications. This may indicate that we are at a turning point in developments towards more bottom-up engagement in science (cf. the GEOSS user feedback brokering⁹³). It is therefore imperative that this opportunity is taken to highlight and consequently replicate overall good practice in this area.

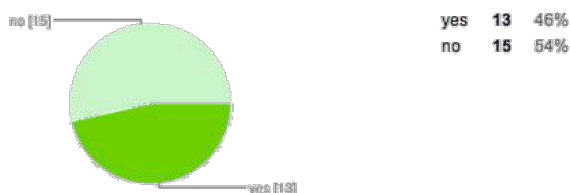
Disseminator/Curator of research data

Do you take user feedback into account for improving the quality of your data?

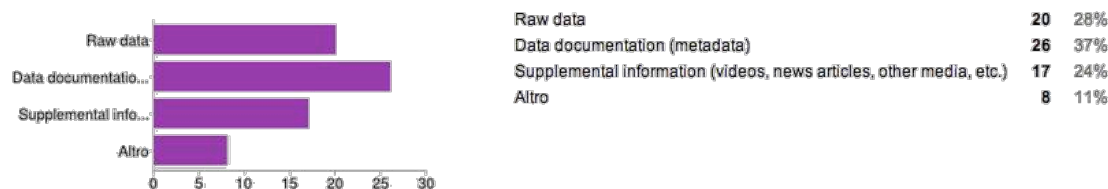


⁹³ GeoViQua, “Quality Aware Visualization For The Global Earth Observation System Of Systems”, no date. <http://www.geoviqua.org/>

As a disseminator/curator of research data, are you offering tools in order to associate research data to scientific publications?



What kind of additional information would you consider relevant to complement scientific publications?



Questions on possible additional information to complement scientific publications highlighted the importance assigned to data documentation (metadata). The issue of metadata, as can be seen above is one of key importance and in addition to feature prominently in our survey it is echoed in the WP1 findings, as well as in the case study interviews conducted for this WP, and featured during the workshop discussions as well (see chapter 6).

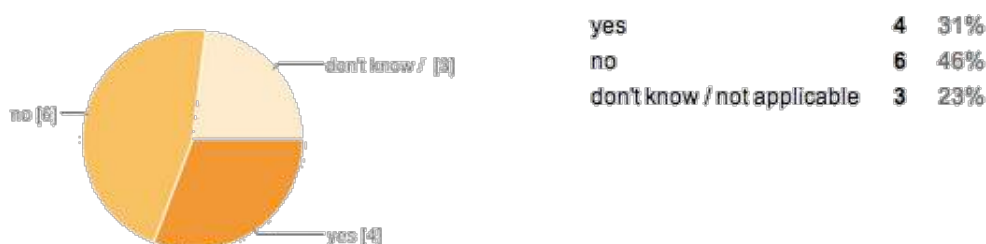
4.2.2 Key issues for data Producers

From our analysis of the data Producers responses concerning their direct experience of Open Access, a similar picture emerges to that of the disseminator/curator role, in that just over half of the respondents report that they have direct experience of Open Access to research data. Only a third of the respondents are aware of a Data Management Plan in use at their institution. However, the majority feel that the level of Open Access adoption within their field is high.

As Open Access with regard to both publications and research data is a high profile issue, it is likely that the prevalence of discussions regarding its overall development and implementation does affect the perception of level of adoption.

Producer of research data

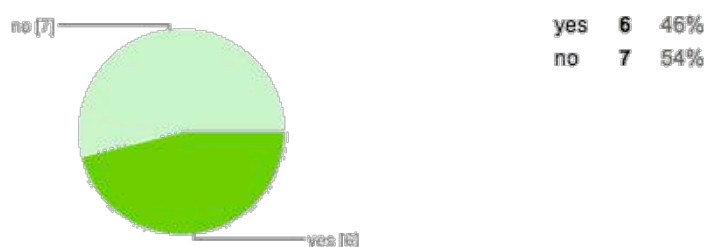
Does your institution/organization have a data management plan?



How would you categorize the level of adoption of Open Access in your field?



Do you have direct experience of releasing your data according to an Open Access policy?



Note: respondents were asked to give their opinion on a scale from 1 (very low) to 5 (very high). Horizontal axis represents the values in that scale. Vertical axis represents the number of respondents who marked that value.

4.3 FINDINGS FROM THE ONLINE SURVEY

Although the survey was completed by a small number of people from largely two stakeholder groups (Disseminators/Curators and Producers), the findings do give an insight into what are considered major barriers to implementing Open Access to research data and also which stakeholders are considered most reliable when it comes to preserving data and storing it. The survey results were also used to stimulate a workshop discussion amongst the RECODE workshop participants. Some key points emerge in relation to infrastructural and technical challenges.

Heterogeneity and interoperability emerge as important issues overall. This is where effort should be put into further exploring the methods of brokering and standardisation of research data formats, and the impact of these two approaches on workload, technology and infrastructure. Furthermore, these aspects need to be explored further in a disciplinary context.

The responsibility for data storage and accessibility is clearly delegated to neutral institutions (national institutional repositories or digital libraries), which may be perceived as ‘specialist’ entities in data storage and curation. Further developing these infrastructures may thus mitigate some of the concerns expressed over issues of accessibility, preservation and curation.

The need for data documentation and quality assessment emerge again as key issues. With regard to technology developments for Open Access to Research Data this is a complex issue as much of it is regards data preparation at the source, i.e. from the side of the data producer. However, implementing tools for user feedback as part of any Open Access platforms, as

well as ensuring that users have a way of publically rating datasets could be a way to mitigate some of the worries around data quality. However, in light of how data quality can be a subjective issue, this may be a contested way of ensuring better quality of research data. This could suggest the future need of specific data documentation skills, as mentioned in the literature. User feedback mechanisms could be explored further as a possible good practice.

With regard to security and privacy issues, only 6% of the respondents selected these as a potential barrier to further implementing Open Access to research data. However, this result needs to be viewed in light of respondent backgrounds, as security and privacy issues are more prominent in certain disciplines dealing with data from human subjects. In our respondent group only 15% of respondents came from Health and Medical Sciences and no respondent self-identified as being from Social Sciences. We are therefore not able, based on these findings, to accurately depict how prominent security and privacy issues feature as a barrier, in the eyes of Producers and Disseminators/Curators of data overall.

Interestingly, in light of the scope of this research work, technological issues seem to be rather low on the priorities of those who replied to the survey. Technical issues concerning storage and bandwidth are almost not questioned. It looks like the perception is that these issues can be solved easily and pragmatically. Again, the types of stakeholders that replied may explain this and we might have seen different results from other respondents. What they do flag up however, are well-known issues, also emerging from the literature review, such as heterogeneity and accessibility issues.

5 CASE STUDY RESEARCH: THE VIEW FROM SCIENTISTS WITHIN FIVE SCIENTIFIC DISCIPLINES

As per Task 2.3, we interviewed key individuals from each of the five RECODE case studies, in order to elaborate on the infrastructural and technological issues they encounter in their research practice. Given our focus on technical issues, we interviewed people within technologically oriented roles such as system administrator, infrastructure managers, etc. In addition to the interviews, we also broadened our coverage and gathered contributions from other sources, outside the specific cases outlined in the description of the RECODE case studies. This allowed for a deeper understanding of challenges faced by each discipline, whilst at the same time adhering to the definition of a case study outlined in WP1:

*"In our study a case is defined as being a discrete research field that each has its own ontology, epistemology and methodology. This allows the RECODE project to address fields of study that include various related disciplines that combine to address a major research question. The overall context of the case study is the development of Open Access to research data in academic and policy research, which forms a single case. We developed a single case design based on the issues of Open Access to Data and conducted five case studies within single case design framework."*⁹⁴

We asked the Directors of projects and Research Units, who agreed to take part in the project at the proposal stage, to suggest case study participants for interviews. The interviews were semi-structured, and questions were based on an expansion of our online questionnaire (see chapter 4). All sections featured open questions, so that the respondent could elaborate on technical and infrastructural challenges. The interview protocols were structured in two parts:

- An introductory section to gather basic information on the respondent profile and his/her perspective on the scope of discourse; this section also contained generic questions based on the material and the discussion at the EC Public Consultation on Open Research Data, held in Brussels on the 2nd July, 2013 (see chapter 3.1);
- A profile-specific section depending on the respondent's perspective in the WP2 taxonomy that has been introduced in chapter 2.1.

We also distributed the interview protocols to other stakeholders in the RECODE contact list and obtained further responses. Although this input is not associated with the RECODE case studies, we consider it interesting material, namely because it includes several responses from a funder's perspective. This material is presented in section 5.6.

In total, we have involved around thirty individuals in the interview phase. The interview protocols can be found in Appendix 2.

5.1 PARTICLE PHYSICS AND PARTICLE ASTROPHYSICS: THE PPPA GROUP AT THE UNIVERSITY OF SHEFFIELD AND THE CMS EXPERIMENT AT CERN

The case study on particle physics was conducted within the Particle Physics and Particle Astrophysics (PPPA) Group of the Department of Physics and Astronomy at the University of Sheffield (UK)⁹⁵. The PPPA Group is a member of one of four regional Computing Grid

⁹⁴ Sveinsdottir, et al., op. cit., p. 19.

⁹⁵ The University of Sheffield, "Department of Physics and Astronomy", 2014. <http://www.sheffield.ac.uk/physics/research/pppa>

Groups in the UK for the CERN Large Hadron Collider (LHC) Computing Grid, the world's largest computing grid. We extended our case study research to the LHC Compact Muon Solenoid (CMS) experiment, one of the largest international scientific collaboration in history, involving 4.300 particle physicists, engineers, technicians, students and support staff from 179 universities and institutes in 41 countries, and one of the two general-purpose experiments at CERN's LHC that have been built to search for new physics.

We interviewed key personnel in charge of managing the provision of research computing infrastructure and services, supporting the LHC and academic work in the area of solar astrophysics and related data lifecycle.

5.1.1 Infrastructure and technology challenges and recommendations within the Physics case study

As known, Particle Physics research produces extremely large volumes of data and is commonly mentioned in the context of Big Data, where data volume is a specific, prominent issue. Much of the practice consists of large-scale, collaborative efforts, with tens or even hundreds of international partners, in some instances over a long period of time, which generate massive data that is stored for further processing. Analysis of the vast quantities of data cannot be undertaken with a single desktop computer or a single large supercomputer, consequently Grid or Cloud technologies are used for analytical purposes

The LHC produces about 15 petabytes of data per annum. The CMS⁹⁶ experiment alone has collected so far around 64 petabytes of raw data from the collisions taking place every second, at peak performance, inside its detector.

CMS is now taking its first steps in making up to half of its data accessible to the public in accordance with its policy for data preservation, re-use and Open Access. The first release will be made available in the second half of 2014, and will comprise of a portion of the data collected in 2010. Along with the many published papers, this data constitutes the scientific legacy of the CMS collaboration, and preserving the data for future generations is of paramount importance. As the head of the CMS Data Preservation and Open Access project states: “*We want to be able to re-analyse our data, even decades from now*”.⁹⁷

Our respondents at PPPA define data as “*all outputs from a research process utilising e-infrastructure*”. Asked what type of research data should be openly accessible, the answer was any data that is deemed to be of an acceptable quality. Such data should have associated metadata and a visible record of peer review, although it could be as simple as a blog comment. However, the feeling is that data openness should be limited when ethical issues surface, e.g. impacts on personal privacy, personal data, intellectual property and commercial sensitivity.

Regarding who should be responsible for storing and providing access to data, among the different actors in the Open Access ecosystem (data producers, digital libraries/repositories, national institutionalized repositories, centralized European repository, publishers, funders) the respondents expressed that there is no single solution to this issues, so they mentioned all

⁹⁶ European Organization for Nuclear Research, “Compact Muon Solenoid experiment at CERN’s LHC”, no date. <http://cms.web.cern.ch/>

⁹⁷ Rao, Achintya, “LHC data to be made public via open-access initiative”, *International Science Grid This Week*, 27 November 2013. <http://www.isgtw.org/feature/lhc-data-be-made-public-open-access-initiative>

of the above institutions as possible solutions. However, it was noted that “*what is missing is linked access, discovery mechanisms and suitable provenance tracking systems.*”

The most impacting factor with the growth of Open Access in Physics research is network bandwidth access to computational resource required to process datasets and applications. Another anticipated issue is that software used to process datasets changes quickly and becomes incompatible with systems:

“We must make sure that we preserve not only the data but also the information on how to use them. To achieve this, we intend to make available through Open Access our data that are no longer under active analysis. This helps record the basic ingredients needed to guarantee that these data remain usable even when we are no longer working on them.”⁹⁸

Although, in principle, providing open scientific data will allow potentially anyone to sift through and perform analyses of their own, in practice doing so is very difficult: it takes CMS scientists working in groups many months or even years to perform a single analysis that must then be scrutinized by the whole collaboration before a scientific paper can be published. A first-time analysis typically takes about a year from the start of preparation to publication, not taking into account the six months it takes newcomers to learn how to use the analysis software.⁹⁹

However, technical aspects are not perceived as crucial obstacles to the adoption of Open Access, when compared to ethical, legal, financial, and political factors. Of least importance are cultural aspects, what seems compatible with the quite common practice of sharing data in Physics. In fact, from the point of view of a scientist, there is a good sharing of data, though at several levels: published results are shared without problems, whereas raw data are not immediately shared. This is justified by the need to ensure quality control, and to provide protection for intellectual assets.

CMS data is classified into four levels in increasing order of complexity of information:

- Level 1 encompasses data included in CMS publications. In keeping with CERN’s commitment to Open Access publishing, all the data contained in these documents and any additional numerical data provided by CMS are, by definition, open;
- Level 2 data are small samples that are carefully selected for education programs. They are limited in scope: while students get a feel for how physics analyses work, they cannot do any in-depth studies;
- Level 3 is made of data that CMS scientists use for their analyses. It includes meaningful representations of the data, along with simulations, documentation needed to understand the data, and software tools for analysis. CMS is making this analysable level-3 data available publicly, in a first for high-energy physics;
- Level 4 consists of the so-called ‘raw’ data — all the original collision data, without any physics objects, such as electrons and particle jets, identified. CMS will not make this data public.

In line with the fact that CMS is now taking its first steps in making up to half of its data accessible to the public, our PPPA respondents reported no experience of Open Access,

⁹⁸ Rao, op. cit., 2013.

⁹⁹ European Organization for Nuclear Research, “How does CMS publish and analysis?”, 14 December 2011. <http://cms.web.cern.ch/content/how-does-cms-publish-analysis>

though sharing and reuse are seen positively, as “*new tools and research procedures reveal new insights*”. A Data Management Plan is in use: research outputs can be discussed and presented at local seminars, community conferences, prior to publication in a topically selected journal.

From the point of view of a data disseminator/curator, data storage and maintenance is reported as a challenge. Ideally, data should be preserved indefinitely, for example in slow dense storage (e.g. tape), with appropriate metadata, in particular quality and usage guidelines (in the form of accreditation/peer review). However, although the importance of quality is stressed, reportedly there is no documentation of quality description with the data. Even user evaluation of quality, such as user feedback, is considered just as an aspect of the quality of service, of interest for librarians, etc.

Data usage is audited and reviewed periodically, tracking by departments, projects, user groups and individuals. The Berkeley Restricted Data Management application¹⁰⁰ is used for policy enforcement.

Several resources and technologies for data storage, preservation and curation are in use by the community of Particle Physics research, as reported by the respondents. Among the referred technologies are: digital libraries such as the NASA Astrophysics Data System¹⁰¹, the Virtual Solar Observatory¹⁰², the arXiv repository¹⁰³, where almost all scientific papers in many fields of mathematics and physics are self-archived; Mendeley¹⁰⁴, a free reference manager and academic social network; the integrated Rule Oriented Data Systems (iRODS), a community-driven, open source, data grid software solution that helps researchers, archivists and others to manage (organize, share, protect, and preserve) large sets of computer files; the ESA Standard Archive Format for Europe (SAFE), a common format for archiving and conveying data within ESA Earth Observation archiving facilities; the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)¹⁰⁵, a digital preservation and curation software platform; the UK Digital Curation Centre¹⁰⁶.

Key technological issues in implementing Open Access to physics data are:

- As physics deals with extremely large and heavy data sets, bandwidth is perceived to be a future challenge for implementing Open Access to data within the field, especially for Disseminators/Curators and End users;
- Disseminators/Curators may have to be very selective about which data to keep and for how long, also there might be instances where some data would be kept offline and access to that might therefore have to be delayed;
- With regard to End users, technological barriers to access, analysis and use of physics data are perceived to be very high due to the size, complexity and format of the data. Currently, grid computing is used to store and analyse this data. In order to make it available by Open Access, selection may need to be made in terms of organising data

¹⁰⁰ University of California Berkley Security, “RDM What is Restricted Data?”, no date. <https://security.berkeley.edu/taxonomy/term/69>

¹⁰¹ The SAO/NASA Astrophysics Data System, “Welcome to the Digital Library for Physics and Astronomy”, no date. <http://adsabs.harvard.edu/index.html>

¹⁰² The Virtual Solar Observatory, “News”, 20 May 2005. <http://umbra.nascom.nasa.gov/vso/>

¹⁰³ Cornell University Library, “arXiv.org”, no date. <http://arxiv.org/>

¹⁰⁴ Mendeley, “Your personal library of research”, no date. <http://www.mendeley.com/>

¹⁰⁵ DCC, “Drambora interactive”, no date. <http://repositoryaudit.eu/>

¹⁰⁶ DCC, “Because good research needs good data”, no date. <http://www.dcc.ac.uk/>

into smaller and more manageable datasets for users with less computing power and/or specialist know-how.

5.2 HEALTH AND CLINICAL RESEARCH: THE FP7 PROJECT EVA AND OPEN HEALTH

The case study on health and clinical research was conducted within the FP7 funded project EVA (Markers for emphysema versus airway disease in Chronic Obstructive Pulmonary Disease; project number 200605).¹⁰⁷ The EVA consortium aims at bringing together clinical medicine, radiology, image analysis and genetics (including gene expression analysis), laboratory diagnostics and bioinformatics, to improve the use of Standard Operating Procedures and other tools to optimise the quality of the collected data, and to ensure the ethical treatment of personal data. The consortium consists of thirteen partners, including ten clinical partners, who provide cases and samples and two partners, to analyse the samples.

We have interviewed key personnel of the Bioinformatics Group at the Centre National de Génotypage (CNG), Institut de Génomique, CEA (France), one of EVA partners. In addition, we gathered further information researching the healthcare community, particularly Open Health News¹⁰⁸, on Information Technology solutions and activities in the “*Open Health*” sector.

5.2.1 *Infrastructure and technology challenges and recommendations within the Health and Clinical Research case study*

Heterogeneity of data is a major issue in health research, which is increasingly interdisciplinary, as can be seen by the mix of disciplines involved in the EVA project. Health and clinical research benefits most from being able to link patient data from many sources, overcoming different research practices and traditions, and different views of health, disease and patients, relevant to each discipline.

There are several Open Access resources on population health and healthcare around the world. One of the most visited is the U.S. Department of Health and Human Services (HHS) providing Open Access to a wide range of health information and datasets that are generated and/or held by the U.S. Government.¹⁰⁹ Others include: the World Health Organization (WHO) Global Health Observatory, which provides data and analyses on global health priorities;¹¹⁰ also institutions such as the World Bank collect data on national health systems, disease prevention, reproductive health, nutrition, population, and more;¹¹¹ EuroStat provides data and statistics about the public and private sectors of the European Union, including the healthcare industry.¹¹² Even Google, with its Public Data Explorer, makes large public datasets on healthcare, economics, and other subjects readily accessible and easy to explore and visualize.¹¹³

¹⁰⁷ Emphysema versus Airways diseases, “Welcome to EvA”, no date. <http://www.eva-copd.eu>

¹⁰⁸ Open Health News, “The voice for the open health community”, no date. <http://www.openhealthnews.com/>

¹⁰⁹ Data.gov, “Health”, no date. <http://www.data.gov/health>

¹¹⁰ World Health Organization, “Global Health Observatory (GHO)”, no date. <http://www.who.int/gho/en/>

¹¹¹ The World Bank, “Health”, no date. <http://data.worldbank.org/topic/health>

¹¹² Open Health News, “EuroStat”, no date. <http://www.openhealthnews.com/resources/Eurostat>

¹¹³ Google public data, “Health expenditure per capita, PPP (constant 2005 international \$)”, no date. http://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&met_y=sh_xpd_pcap_pp_kd

According to CNG respondents, application interoperability, cataloguing, and data documentation (metadata) will have the greatest impact on Open Access to research data. In relation to research data storage and maintenance, difficulties are reported “*to maintain a performing environment to use (access, query...) the data in an efficient way*”.

Interoperability and data sharing between key internal and external systems are seen as primary objectives of sound IT Architectural roadmap for a business organization in healthcare (e.g. clinic, hospital). To help achieve such an objective, several architectural processes have emerged, among which the Medicaid Information Technology Architecture (MITA)¹¹⁴ Framework may be the most relevant to healthcare organizations.

The MITA Framework provides guidelines for the creation of a roadmap that ought to contain the following key sections:

- The Business Architecture section of the IT Architecture roadmap should present a collectively agreed upon vision for the future of the business organization.
- The Information Architecture section should identify the major types of information systems needed to support the business functions of the organization.
- The Technical Architecture section should describe current and planned technical services, their connectivity, and standards that the organization should use when they plan and specify new IT systems to be acquired and implemented.

A prominent challenge is data security, due to legal and privacy issues arising when sharing genome and patient data. To complicate matters further there are a number of different stakeholders involved in various projects, e.g., pharmaceutical companies, patient organisations, ICT staff, journals and research institutions, each of which have a different take on the use and sharing of data. The pharmaceutical industry might be reluctant to share data, which is commercially sensitive, whilst patient groups would be unwilling to opening up data due to privacy concerns. Some respondents mentioned the danger of making some data open, referring to the specific example of genome sequencing. The question of patents is also a concern in the Health community.

While research data and open data are closely connected, the different nature of types of research data needs careful consideration, and data awareness and a culture of sharing need to be created. Suggestions from the interviewees include the provision of tools that ease data sharing and reuse; adapting copyright legislation to the needs of research and the use of common standards for resource identification; methods of citation that are helpful to curators and acknowledge authors and institutions; interlinking of datasets, metadata and publications.

Open Health News reports open grid technologies coupled with cloud computing as an increasingly adopted good practice, especially in the fields of healthcare and bioinformatics.¹¹⁵ Grid computing provides the ability to perform higher throughput computing by taking advantage of many networked computers to form the equivalent of a virtual 'super computer'. Grid computing utilizes the unused capacity of many separate computers connected by a network to solve large-scale computational problems. With grid computing, organizations can collaborate and pool both internal and trusted external

¹¹⁴ CMS.gov, “Medicaid Information Technology Architecture (MITA)”, no date.

<http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidInfoTechArch/index.html?redirect=/MedicaidInfoTechArch/>

¹¹⁵ Open Health News, ““Open” Grid Solutions in healthcare Continue to Make News, December 3, 2012. <http://www.openhealthnews.com/hotnews/open-grid-solutions-healthcare-continue-make-news>

computer resources to tackle projects that require an extremely large capacity of computing power.

Grids are being used in healthcare in at least three ways:

- *Computational grids* are being used to solve large-scale computation problems in healthcare research;
- *Data grids* do not share computing power, but instead provide a standardized way to swap data internally and externally for data mining and decision support;
- *Collaborative grids* let dispersed users share information and work together on extremely large datasets. Examples of collaborative 'open' grid projects in the field of bioinformatics and healthcare include: BioSimGrid¹¹⁶, Folding@Home¹¹⁷, FightAIDS@Home¹¹⁸.

A very specific challenge related to Open Access in Health and Clinical Research is security. Privacy is absolutely crucial when it comes to genomic information. Integrated longitudinal health records that include personal, clinical, and genomic information will provide unprecedented access to details about an individual in a manner that was previously inconceivable. To protect individuals and their genomic information will require better access controls, data encryption, system audit tools, release of information procedures, training, etc. Next generation Electronic Health Record (EHR)¹¹⁹ will contain genomic information modules and provide predictive care capabilities supporting the continued movement towards more personalized medicine. Much of the work on genomic information systems being done involves extensive collaboration between public and private sector organizations with a heavy emphasis on standards and open source solutions.

Key technological issues in implementing Open Access to health and clinical data are:

- Data heterogeneity remains a key issue for health and clinical research, due to the interdisciplinarity within current and future health research projects; this remains an issue for all stakeholder groups, in various ways. Producers may find that in order to produce data, which is reliant on other data, heterogeneity in data sets and locations may significantly delay their work. For End users the discoverability and integration of heterogeneous datasets may be ill possible due to technical limitations. For Disseminators/Curators, in order to preserve and Open Access to heterogeneous data sets the decision as to whether to use a brokering approach or a federated approach will need to be made. Funders may need to become involved and provide funds for added data work in the case the federated approach and standardisation of data becomes a solution for heterogeneity;
- Different disciplines (producers, users and curators/disseminators) have different views of data, as well as methods of providing access, data preservation and curation;
- Security and privacy are very important issues, as health and clinical research deals with individual and patient data; there are already robust security measures within health and clinical research, but Disseminators/Curators will need to become involved at this stage to provide secure access and secondary data use;

¹¹⁶ BioSimGrid, no date. <http://www.biosimgrid.org/>

¹¹⁷ Stanford University, "What if...", no date. <http://folding.stanford.edu>

¹¹⁸ Fight AIDS@home, "The Olsen laboratory", no date. <http://fightaidsathome.scripps.edu>

¹¹⁹ Often used interchangeably with EPR (Electronic Patient Record) and EMR (Electronic Medical Record), although differences between them can be defined.

- It is perceived that healthcare organizations, as End users and Producers, need to be more proactive in collaborating on the construction of the unified clinical and genomic health information systems of the future.

5.3 BIOENGINEERING: AUCKLAND BIOENGINEERING INSTITUTE AND THE VPH COMMUNITY

The case study on bioengineering was carried out within the Auckland Bioengineering Institute (ABI)¹²⁰, in New Zealand, and the Virtual Physiological Human (VPH)¹²¹ Community. Their objective is to develop a systemic approach that avoids a reductionist approach and seeks not to subdivide biological systems in any particular way by dimensional scale (body, organ, tissue, cells, molecules), by scientific discipline (biology, physiology, biophysics, biochemistry, molecular biology, bioengineering) or anatomical sub-system (cardiovascular, musculoskeletal, gastrointestinal, etc.)

We interviewed key bioengineering scientists in the field of computational modelling of the human body and its functions, from tissue structure to mechanical, electrical and cellular activity. The institute's work also encompasses computational physiology, instrumental medical devices and medical informatics.

5.3.1 *Infrastructure and technology challenges and recommendations within the Bioengineering case study*

Over the past decade, clinical medicine and biomedical science have been transformed by the emergence of bioengineering, which has enabled the development of a characterisation of structure and function in cells, tissues, organs and the living body in detail. For this information to be integrated, comprehensive models of human biology based on quantitative descriptions of anatomic structure and biophysical processes, which reach down to the genetic level, are under development.

The VPH is a methodological and technological framework that will enable collaborative investigation of the human body as a single complex system. This collective framework will make it possible to share resources and observations formed by institutions and organizations creating disparate, but integrated computer models of the mechanical, physical and biochemical functions of a living human body. The VPH framework is formed by large collections of anatomical, physiological, and pathological data stored in digital format, by predictive simulations developed from these collections, and by services intended to support researchers in the creation and maintenance of these models, as well as in the creation of end user technologies to be used in the clinical practice. VPH models aim to integrate physiological processes across different length and time scales (multi-scale modelling) and enable the combination of patient-specific data with population-based representations.

The paradigm of VPH is the “*physiome*” concept, that is the quantitative and integrated description of the functional behaviour of the physiological state of an individual or species. The concept was introduced in 1997 by the IUPS Physiome Project, the first worldwide effort to define the physiome through the development of databases and models that facilitated the understanding of the integrative function of cells, organs, and organisms. The project focused

¹²⁰ The University of Auckland, New Zealand, “Auckland Bioengineering Institute”, no date. <http://www.abi.auckland.ac.nz/en.html>

¹²¹ VPH Institute, “Welcome to the VPH Institute”, no date. <http://www.vph-institute.org/>

on compiling and providing a central repository of databases, linking experimental information and computational models from many laboratories into a single, self-consistent framework.

In terms of sharing models and code, bioengineering is described by respondents as a fundamentally open field, but less so in the sharing of experimental data. Nevertheless, the 200-people ABI has a direct experience of releasing data with an Open Access policy for 30-40 years. Namely, experimental measurements and models are available online just completely Open Access.

One interviewee recalls: *“It was a single database for the group, lot of people from all over the world were accessing the data for many years. Now things have moved on and there are minimal sources of that data, but the open group still has a policy of making sure that experimental data are made available to anyone that wants it.”*

Although a Data Management Plan is not in place yet, the ABI is beginning to provide resources for people to centralize where they store data, via the library. The access policy is not uniform, though, and in particular there is no agreed, official policy on data publication.

Being based on computational models of extremely complex biological systems, a prominent challenge of bioengineering research pertains the reproducibility of the outputs. In fact, experiments may not be repeatable in the manner that is expected for acceptance in the current scientific paradigm. Uncertainty and random factors can be introduced at many levels, such as initial approximations applied to reduce complexity, essential in order to make the problem computationally tractable; lack of validation; lack of expressiveness in the descriptive ancillary information (metadata) associated to complex models; provenance of the biological data used to design and validate the model; size and duration of the experiments, whose results may be the end product of tens of person-years work.

Asked specifically on metadata technologies, for data identification and discovery, the respondents indicated Web 2.0 standards (RDF, OWL, XML). It is generally assumed that the data files have metadata associated with them, as semantic annotations based on ontologies, namely RDF-encoded information. As RDF stores are likely to become large, scalability issues may arise. ABI provides web services for searching against ontology-based metadata, for retrieving particular models or datasets. RDF trees are seen as an effective technology to this end. In fact, RDF Tree is an approach (and a Java library, in development), to producing developer-friendly serialisations of RDF graphs, based on JSON and XML. Unlike other full-fledged Linked Data serializations (e.g. JSON-LD), it is a simple approach to building predictable, stable JSON and XML representations of graph data. RDF triplets are typically consolidated into an OWL structure, in order to support logical reasoning.

The development of ontologies is seen as a critical point. As a respondent underlines: *“We avoid developing ontologies ourselves, as you really need to be an expert in the area”*. A good practice on this aspect is to rely on collaborations with others, particularly recognised stakeholders, such as the European Bioinformatics Institute, and to adhere to shared standards, such as the ontologies defined by the Open Biological and Biomedical Ontologies (OBO) Foundry.¹²²

¹²² The open Biological and Biomedical Ontologies, “OBO Foundry ontologies”, no date. <http://www.obofoundry.org/>

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies, who are establishing a set of principles for ontology development, with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain.

A specific question regarded technical difficulties related to research data storage and maintenance. In that regard, difficulties are mainly seen as associated with human resources, e.g. lack of technical expertise for understanding, designing, and implementing appropriate services based on RDF, Web Services, etc. These are addressed by training activities for individual capacity building.

However, the general feeling is that technological aspects, in terms of how storage and maintenance are handled, are quite routine and do not present significant, specific issues requiring unanticipated solutions.

As for repository, a technology in use is the Physiome Model Repository (PMR)¹²³, which in fact is used to store data as well. The PMR is an open source software suite that integrates a Content Management System (CMS) with a Distributed Version Control System (DVCS). It enables scientists to easily upload, store and manage models, without reliance on a central repository and keeping full control over their data.

PMR supports the concept of building a complex model by modules that are imported from lower levels in a hierarchy, maintaining the full provenance information. The whole development history can be tracked and easily merged back into the shared database.

Access control is implemented by the Plone-based CMS. That also supports curation, distribution, and collaborative development, allowing project teams to share data within the team and to make them publicly available, when calibration reaches the desired point, just by flagging a switch.

Key technological issues in implementing Open Access to bioengineering data are:

- Outputs reproducibility is a prominent challenge for End users in bioengineering research, namely for the limitations of the metadata associated to complex models, such as the provenance of the biological data used to design and validate the model;
- The development of ontologies is seen as a critical point, because of the deep knowledge involved for defining interoperable ontologies in the biomedical domain, this impact on Producers, who have to annotate their data appropriately to support discovery and access;
- With regard to storage and maintenance, the key issues for Disseminators/Curators are presented as being human resource related, where a lack of technical expertise and skill to work with data is presented as a challenge to future implementation.

5.4 ENVIRONMENTAL SCIENCES: THE EC JOINT RESEARCH CENTRE

The case study on environmental research was conducted at the Digital Earth and Reference Data Unit of the Institute for Environment and Sustainability, at the EC Joint Research

¹²³ CellML, “Physiome Model Repository”, no date. <http://www.cellml.org/tools/pmr>

Centre.¹²⁴ The mission of the unit is to address sustainability and competitiveness challenges by developing information systems and promoting wide access to the reference data and services needed for robust policy making.

Interviews were carried out with key technical researchers leading the unit internal project on open data access and interoperability. The results of the case study reflect a mix of the current situation and the planned infrastructure setup for acquisition and dissemination of research data.

5.4.1 Infrastructure and technology challenges and recommendations within the Environmental Sciences case study

The JRC is an active contributor to GEOSS: among other contributions, it led the EuroGEOSS¹²⁵ EU-funded project, which implemented multidisciplinary interoperability across the thematic areas of Drought, Forestry and Biodiversity within GEOSS. The JRC also coordinates the scientific and technical development of INSPIRE, investigating issues regarding technical and multidisciplinary interoperability of spatial datasets and services needed to support environmental policy and policies that affect the environment, and what are the challenges in sharing and providing access to data from a variety of sources, and in a variety of formats. As a consequence, the JRC has a specific interest in Public Sector Information, which is in the main scope of INSPIRE.

As regards Open Access, it must be born in mind that most of the data held by the JRC is not owned by the JRC, but by the Member States that provide access to the JRC for specific projects, or as part of legal requirements. Therefore the JRC would have to respect data ownership on the matter of Open Access, as constrained by third party IPR.

Being part of the European Commission, JRC refers to the official EC practice, so the uptake of Open Access to research data is in a very initial phase, and no direct experience is reported. Next year an open data project will be initiated. One aim of the project is to increase access to JRC data (including research data) following the commission decision of 12 December 2011 on the reuse of Commission documents.¹²⁶

Currently, JRC has no common policy for data management. That would have to be defined in the context of the open data project. Likewise, there is no auditing in place for data utilization. A first approach to that is to establish a common data inventory.

Until now, the JRC has not adopted a precise definition of research data, and possibly will not in the future. According to our respondents: *“At the moment we operate with a broad definition specifying all data produced within JRC are within scope”*. Noteworthy, this definition includes PSI managed and organized by the JRC in the framework of the INSPIRE Directive.

¹²⁴ European Commission, “Joint Research Centre”, no date. <http://ies.jrc.ec.europa.eu/>

¹²⁵ EuroGEOSS, “Welcome to EuroGEOSS the European Approach to GEOSS”, no date. <http://www.eurogeoss.eu/default.aspx>

¹²⁶ European Commission, Commission Decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU), OJ L 330, 14 December 2011. <http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF>

According to the Commission decision of 12 December 2011 on the reuse of Commission documents, all data shall be made openly accessible, unless other restrictions apply, e.g., from third parties, or for sensitive information. In that context, “documents” are defined in a broad sense and also include data.

The cases when openness must be excluded are stated in the Commission decision of 12 December 2011 on the reuse of Commission documents:

- a) Software or to documents covered by industrial property rights such as patents, trademarks, registered designs, logos and names;
- b) Documents for which the Commission is not in a position to allow their reuse in view of intellectual property rights of third parties;
- c) Documents which pursuant to the rules established in Regulation (EC) No 1049/2001 are excluded from access or only made accessible to a party under specific rules governing privileged access;
- d) Confidential data, as defined in Regulation (EC) No 223/2009 of the European Parliament and of the Council (1);
- e) Documents resulting from on-going research projects conducted by the staff of the Commission which are not published or available in a published database, and whose reuse would interfere with the validation of provisional research results or where reuse would constitute a reason to refuse registration of industrial property rights in the Commission’s favour.

According to our respondents, the primary responsibility for storing and providing access to European research data should lie with the data producers, “*but with support from centralized repositories being it regional, national, or European.*”

In general, our respondents judge financial and legal issues the main key obstacles that need to be removed for promoting the adoption Open Access in environmental sciences. After those, technical barriers are considered to be more a hindrance than cultural ones. This suggests that the acceptance of Open Access in environmental sciences could be more limited by technology than by the willingness of stakeholders to share their data.

The JRC covers the role of data disseminator, as part of the European Commission (due to above Commission decision) and as the technical coordinator of INSPIRE and maintainer of the INSPIRE geoportal, which provides access to the datasets and services made available by the Member States. From this point of view, there is a concern that preservation of data is too costly. This would make it unpractical to preserve research data indefinitely, though desirable in principle. When evaluating what data to discard, priority needs to be given to raw data that are not preserved anywhere else: “*Derived data needs to be documented, but not necessarily kept, meaning that it can be reproduced.*”

A common approach and agreement within the research community (producers) would be best to decide what to preserve and until when: “*usage should determine whether data need to be accessible online or not*”.

Support for associating research data to scientific publications is planned. In particular, metadata are considered relevant additional information to complement scientific publications, particularly with pointers to the data used in that particular research work.

The respondents underline the importance of standard metadata, especially as concerns Public Sector Information. However, no common approach on metadata is taken across JRC. For geographic data, the INSPIRE regulations and guidelines on metadata and network services are followed.

On the quality of data, user feedback is recognised as a valuable resource, which should be accounted for, particularly in multi-disciplinary contexts, where the concept of fitness for use is emphasized, more than theoretical, objective expressions of quality. However, there is no consistent mechanism for soliciting, retrieving, and leveraging user quality assessment, at the moment. This is one of the aspects that some respondents would like to improve.

Among the referred technologies are: the European Data Infrastructure¹²⁷, which provides solutions for preservation and curation of data, and the Dryad digital repository¹²⁸, which supports associating research data to scientific publications.

Key technological issues in implementing Open Access to environmental data are:

- The definition of data by the JRC includes PSI managed and organized by the centre in the framework of the INSPIRE Directive, hence legal aspects and institutional arrangements are prominent over technical ones;
- The importance of standard metadata is underlined, especially for Public Sector Information; this is a cross-cutting aspect through all the challenges and stakeholder categories, especially the End users;
- Data preservation is perceived as too costly, for Funders and Disseminators/Curators, which makes it impractical to preserve research data indefinitely, though desirable in principle; a shared approach and agreement within the research community (Producers and End users) should be taken to decide what to preserve and until when, giving priority to raw data that are not preserved anywhere else, over derived data that can be reproduced;
- User feedback is perceived to be valuable for the on-going development of datasets, however a solid mechanism for gathering and implementing user assessment does not exist at the moment; it is worth noting that the topic of quality in geospatial information is a hot topic in GEOSS, also explored by FP7-funded projects like GeoViQua¹²⁹, which combined geospatial data together with information on their quality and provenance within GEOSS catalogues. Disseminators/Curators of data should consider if and how user feedback should be gathered and acted on, whilst End users and Producers may need to adapt their practices to include leaving data feedback, and acting on feedback received.

5.5 ARCHAEOLOGY: OPEN CONTEXT AND THE MAPPA PROJECT

The case study on archaeology addressed Open Context¹³⁰, a free, Open Access resource for web-based publication of diverse types of research datasets from archaeology and related disciplines. Open Context is developed by the Alexandria Archive Institute¹³¹ and backed by

¹²⁷ EUDAT, “European Data Infrastructure”, no date. <http://eudat.eu/>

¹²⁸ Dryad, “Homepage”, no date. <http://datadryad.org/>

¹²⁹ GeoViQua, “Quality Aware Visualization For The Global Earth Observation System Of Systems”, no date. <http://www.geoviqua.org/>

¹³⁰ Open Context, “Welcome to Open Context”, no date. <http://opencontext.org/>

¹³¹ The Alexandria Archive Institute, “The Alexandria Archive Institute”, no date. <http://alexandriaarchive.org/>

the California Digital Library¹³², USA. Open Context takes the model of data sharing as a form of publishing. It promotes a more value-added way of sharing data than the typical database, excel, media files (pictures), and field notes. Open Context relates all these to global schema, ontology, controlled vocabulary; besides, it provides web services and API, with a specific focus on visualisation.

We also included in the case study the University of Pisa (Italy) MAPPa Open Data repository project¹³³, which aims at creating the first Italian open digital archaeological archive, as a network of systems and standardised procedures for managing archaeological data, and implement Open Access to all public data relating to archaeological investigations.

We interviewed key technical personnel, who work on leading the development of services, integration with other Web-based research collections, and preservation of data through digital library services.

5.5.1 Infrastructure and technology challenges and recommendations within the Archaeology case study

Traditionally, archaeologists work on site excavations, which yield data in many different formats, e.g., diaries, artefacts, spatial data, images and statistical data. Given the possible long duration of the excavation phase, for convenience, data are typically digitized and managed electronically for a substantial part of their life cycle. Resources like Open Context and MAPPa focus on data publishing and adding value to existing datasets by means of ICT technologies, which allows for integrating datasets and linking of different data formats to present a holistic view of archaeological data.

Research data are all kind of data, from raw measurements to linked open data produced during the research process, and any form of data documentation, usually structured in some form. Nowadays, no digitization is applied, as most field documentation is born digital.

Respondents report an increasing use of ontology (e.g. the ISO CIDOC CRM cultural heritage ontology) to enrich the description of resources with semantics. However, this is sometimes controversial, as any ontology is a subjective theoretical construct, always biased by the culture of the designer. As multiple meanings are usually conceivable, it is important to accommodate multiple ontological hierarchies, without restricting the data to a particular one.

According to the MAPPa respondents, access to data should be unlimited in archaeology, but in practice Open Access is very seldom adopted. Currently, the Conference of Italian University Rectors (CRUI), joined by many Italian Universities, including the University of Pisa, is promoting a policy for Open Access publishing of research data¹³⁴, including electronic theses and dissertations, registered with the Open Archives Initiative¹³⁵.

There is a specific important security issue in archaeology, in that the location can be regarded as sensitive information, and in the instances of valuable sites, access to location

¹³² University of California, "California Digital Library", no date. <http://www.cdlib.org/>

¹³³ University of Pisa, "Mappa Project", no date. <http://mappaproject.arch.unipi.it/?lang=en>

¹³⁴ University of Pisa, "Home ETD (Electronic Theses and Dissertations): presentazione, conservazione e disponibilità in forma elettronica delle tesi discusse nell'Ateneo", no date. <http://etd.adm.unipi.it>

¹³⁵ Open Archives Initiative, "Home", no date. <http://www.openarchives.org/>

data may even be legally protected (e.g. in cases of national heritage). In fact, apart from the due protection of excavation locations from other competing groups, there is a pressing need to limit the risk of looting and vandalism. Moreover, in the USA and other countries characterized by colonialism and genocide, respect for indigenous and local communities and their sacred locations and artefacts, further advocates the need for restricted access to location data. Open Context takes the approach of lowering the resolution of location information, to mitigate these problems.

Access control solutions are described as:

- Authorisation from the data owner to access data;
- Embargo for a given period of time;
- Access to approved researchers only;
- Providing secure access to data;
- Excluding the ability to download data.

No auditing is reported for the actual use of data.

Responsibility for storing and providing access to European research data is debated. According to MAPPA, the best solution would be national repositories connected to publications on data. For example, the MAPPA repository publishes all kinds of data, but is connected with the Journal of Open Archaeological Data for good quality data. According to our respondents from Open Context, there is “*no single solution, there will always be winners and losers*”. Centralization may be seen as undesirable, on the other hand decentralization is more expensive, is prone to fights over standards, etc. However, it may also bring more innovation.

Future factors to face are digital data preservation, and documentation and standards alignment, from a more technical point of view. From a researcher point of view, big data (such as originating from GPR and photogrammetry) is seen as a primary issue. As one of the interviewee puts it: “*we do not understand well enough what data are more interesting for future reuse*”.

When ranking barriers hindering Open Access to research data, the respondents agree that infrastructure and technology challenges are not priority, with respect to cultural, political/legal, and financial factors. Only traditional publication is valued. Researcher are overloaded, risk taking is difficult for them, as data sharing has no clear reward. Open Access is probably related to academic freedom and autonomy, and the current “*Taylorism in university management*”, causing some disciplines to be underfinanced, is a major hurdle for its adoption.

From the point of view of producers of research data, the adoption of Open Access is perceived as very scarce. Yet archaeological data are considered quite shareable and reusable, for example for predictive modelling, preventive archaeology, quantitative analysis, etc.

MAPPA has a policy for publications: gold Open Access publishing on Open Access journals or on the MAPPA repository, with DOI and CC-BY/CC-BY-SA license, or green Open Access self-archiving (e.g. on Academia.edu¹³⁶ or ResearchGate¹³⁷) of papers published in

¹³⁶ Academia.edu, “Academia.edu is a place to share and follow research”, no date. <http://www.academia.edu/>

¹³⁷ ResearchGate, “Homepage”, no date. <https://www.researchgate.net/home.Home.html>

journals or other types of publications. MAPPA is linked with the Journal of Open Archaeological Data (JOAD) and included in their list of recommended repositories.

As a disseminator/curator of research data, MAPPA respondents advocate that research data should be preserved indefinitely, which seems a reasonable demand in archaeology. Preservation should ensure the authenticity, reliability and logical integrity of data in perpetuity through:

- Data migration achieved by monitoring hardware and software developments and updating the data accordingly:
 - Version migration: conversion to newer versions of a format; often the only option for preserving data in proprietary formats, although only practical for *de facto* standard, i.e. proprietary formats widely used in the research practice;
 - Format migration: conversion to formats that optimise dissemination;
- Media refreshment: migration between media which leave data unchanged;
- Appropriate metadata standards;
- OAIS ISO 14721:2003;
- Normalisation: migration to widely supported open standards;
- Reuse (reuse itself aids preservation).

Likewise, all research data should be accessible online, in open repositories through which researchers disseminate the data they decide to open. All datasets should be licensed with CC-BY or CC-BY-SA licenses and published with a DOI, to provide a persistent, global, standardised identification of research data datasets.

When asked about their policy for data management, respondents from MAPPA describe it as follows:

“Currently, our main purpose is to persuade the archaeological community of the importance of open data, so we use a basic policy for data management: we acquire raw data from archaeologists and we validate it from a legal point of view. Once data are validated, we embed metadata to each dataset, describing all the information regarding the dataset itself; then we store the data in our repository, with appropriate security.”

Format validation and possible migration is planned for the near future. Currently, there are no standard formats, nor standard conversion practices.

Legal aspects seem to be a primary concern, for MAPPA. They have published a detailed guide on the procedures to prepare and provide the material for publication. In compliance with the law, published documents must not contain personal data of persons who have not previously agreed to their publication. They provide specific disclaimers for the authors to download, to help them collecting the authorisations needed to put their material online. In the absence of clearance, the personal data reported in the original documents are obfuscated. The publication of particular resources, such as images, must be authorised by the Ministry of Culture.

The most important type of data is raw data and metadata; besides that, all kinds of supplemental information are desirable. As for data documentation, MAPPA has defined a recommended metadata profile based on Dublin Core and ISO19115 core metadata, for the

geographical section¹³⁸. The profile describes: the archaeographic production, the structure and format of the digital data, the history of the archaeological investigation, the sources used, the method and the relationship with the physical data.

Quality is perceived as depending primarily on the data type: for example, superintendencies on archaeology guarantee on raw data about excavations. However, data quality is not covered by the MAPPA profile, as, above all, researchers must evaluate the quality of research data: *“we firmly believe that the quality of research data must be the responsibility of researchers in a sort of open peer review method.”*

Open Context reports on the use of OpenRefine (formerly known as Google Refine), to improve the quality of data (e.g. entity reconciliation). Besides, the Git¹³⁹ DVCS, commonly used to support teamwork in software production, is used for storing, publishing and curating textual resources.

Key technological issues in implementing Open Access to archaeology data are:

- Heterogeneity of data formats, due to different types of data yielded from any one excavation; this remains an issue for End users (discoverability and integration of different datasets) as well as Disseminators/Curators, which will need to make a decision on how data heterogeneity should be tackled;
- Access control, needed for sensitive data, namely the location of archaeological sites; technologically sound solutions may be needed by Disseminators/Curators to enforce data security and appropriate access limitations;
- Regarding responsibility for storage and preservation of data, respondents felt that perhaps a distributed approach would lead to more innovation in the field, rather than a centralised approach; a technological challenge would then be to ensure interoperability between different software solutions.

5.6 SUPPLEMENTARY INTERVIEWS

We have gathered additional interview material from a number of other stakeholders in the RECODE contact list. Although this input is not strictly associated with the RECODE case studies, we consider it interesting material, namely because it includes several responses from the perspective of a funder.

Awareness of solutions for long-term research data preservation is very low among all stakeholder groups. However, everyone agrees that, if the right tools were being offered, raw data, data documentation (metadata), and supplemental information (videos, news articles, other media) should be stored.

When asked what the technical steps for making research data openly accessible would be, quite many respondents answered that, besides having good and solid technical equipment (servers, database management and back-up systems), accurate metadata and good metadata standards are crucial for making data accessible for (academic) re-use. Standardization of metadata (e.g. within OGC) is emphasised as a pressing need.

¹³⁸ University of Pisa, "MAPPA Project, Proposta di Metadati", 2012. <http://mappaproject.arch.unipi.it/mod/metadati.php>

¹³⁹ Git, "Fast version control", no date. <http://git-scm.com/>

There are standardization efforts concerning metadata, and filing/database systems. However, almost no respondent is aware of a generic data policy. One respondent reported on the adoption of a quite clear five-steps Data Management Plan:

- 1) Setup hardware, server, database management and backup system;
- 2) Describe metadata, ownership, access control, add DOIs;
- 3) Use OGC-compliant services (WPS, WCS, SOS) to support data discoverability and accessibility;
- 4) Audit data downloading activities; send related information to the data producers and the data manager;
- 5) Provide alternative tools for downloading data, e.g. console based application, SFTP, WebDAV, etc.

This list seems to capture the requirements of a data repository, although at a quite basic level. For example, it does not address application interoperability issues, which may require data conversion and transcoding.

5.7 FINDINGS FROM THE INTERVIEWS

An interesting finding emerging from our interviews is that technological barriers are not reported as of high priority or concern for implementing Open Access to research data, whereas financial, cultural and legal challenges are higher on the list of concerns. This confirms a similar finding from the literature review and the online survey.

Overall, respondents reported more experience with Open Access publications rather than Open Access to research data, and data preservation. In most instances we found data management plans at an early stage. Technical solutions for data management and preservation are fragmented, often designed for a narrow purpose, rather than centralised.

As we found in the literature review and the survey, most respondents mentioned issues of documentation and metadata as a key enabling factor for retrieval, re-use and preservation of research data. However, the technological challenges mentioned by respondents in the case study interview differ somewhat between disciplines. In the following paragraphs we elaborate on the areas of convergence and divergence in the case studies concerning the main infrastructure and technology challenges, to identify possible common solutions to evaluate for suggesting potential recommendation.

Heterogeneity and interoperability

The environmental research, archaeology, health and clinical research case study respondents reported on a common issue in which technology either posed a challenge or was seen as a possible solution to a current issue: data heterogeneity was presented as an issue linked firstly to interdisciplinarity (health and clinical research) and secondly, as an issue arising from different methods employed on any one excavation (archaeology). Developing a suitable technological solution that allows for linking of different data would enhance the possibilities of re-use. Also, interoperability between different disciplines, in terms of standards and terminology, as well as between repositories, in terms of software solutions, were discussed as a technological challenge.

Health researchers reported that to help interoperability between key internal and external systems the Medicaid Information Technology Architecture Framework (MITA) was the most relevant to healthcare organizations. In bioengineering a well used open source Content Management System software used for creation and managing of repositories is the Physiome Model Repository (PMR) which helps researchers with storage and keeping control of their data without using central repositories solutions. In Italy, archaeologists within MAPPA use metadata schemas based on Dublin Core and ISO19115.

Another facet of interoperability is searching against ontology-based metadata for finding particular datasets. In bioengineering RDF trees in combination with an OWL structure are seen as a usable technology for this. In Bioengineering the development of ontologies are seen as critical for retrieval of data. The best practice on this aspect is said to be to collaborate with recognised stakeholders and adhere to shared standards, such as ontologies defined by the Open Biological and Biomedical Ontologies Foundry (OBO).

In archaeology there is an increasing use of ontologies especially the ISO CIDOC CRM but ontologies are also considered controversial and biased products based on subjective theoretical constructs. Therefore it is considered important to accommodate multiple ontologies. There is evidence that different open community standards are used to relieve some of the problems of heterogeneity. Using available standards in a collaborative way, for linking and interoperability is obvious a solution to many of the technical infrastructural challenges discussed.

Accessibility

In physics, the sheer size and volume of the datasets is seen to pose problems with bandwidth from the client side, i.e. public access is seen as almost impossible, as the processing power of any single computer is not enough to work with data of this size.

Another big problem reported across all the research fields in the study, is the documentation of the data and therefore the importance of standard metadata. There is no common approach to metadata from the EU. In the environmental sciences case study, the INSPIRE guidelines on metadata are used and for preservation the European Data Infrastructure¹⁴⁰ provides solutions, as well as the Dryad¹⁴¹ repository for storage. In the case of the archaeology case study, there are definitions and recommended metadata profiles based on Dublin Core and ISO19115 core metadata structures. In the fields of science of the other case studies, there are few initiatives, and good practice for metadata is unavailable.

There is a need for sorting out what metadata schemas are available and could be recommended for use in different areas of research depending on granularity and interoperability specifications.

Preservation and curation

Preservation and curation of data can be very costly. A practical take on this issue comes from the environmental sciences where a shared approach is suggested. Researchers agree on what to preserve, giving priority to raw data, since derived data can be recreated (however, on

¹⁴⁰ EUDAT, “European Data Infrastructure”, no date. <http://eudat.eu/>

¹⁴¹ Dryad, “Homepage”, no date. <http://datadryad.org/>

the other hand, derived data may be more frequently needed, so this approach may be inefficient).

In the field of archaeology, a distributed approach of data storing and preservation could lead to more innovation in the field, rather than a centralised approach. At this moment in the UK, Netherlands and other countries you have centralised repositories for archaeological data since it is obliged in the EU to store and preserve excavation data.¹⁴²

Quality and assessability

Several of the researchers interviewed pointed to the fact that it was not the technical issues that were the main problems talking about infrastructure. It was mainly cultural or financial or human resource barriers that slowed things down. One such thing was the lack of technical expertise for understanding, designing and implementing appropriate services based on technologies such as RDF, CMSs, etc. There is a lack of competence on all levels and massive amount of training for researchers, curators, librarians and other related university staff is badly needed to give momentum to open research data. This problem is addressed by training activities for individual capacity building using expertise from data centres, research institutes and library schools.

There is no solid mechanism for gathering and implementing user assessments of open datasets. User feedback is a valuable asset for both the development of datasets as well as an indicator of quality. Some journals and publishers, like BioMedCentral and the Public Library of Science, are doing work in the area of user feedback.

Security

Data security and access control was mentioned in the case of archaeology and health and clinical research, to protect sensitive data. Technological solutions for ensuring levels of access are needed, as well as robust data security measures. In archaeology one measure to deal with sensitive data have been lowering the resolution of the data.

Data obfuscation may be a viable good practice for protecting sensitive data while making them available to the general public. To this end, the data infrastructure must be instrumented with appropriate security services for user authentication, authorisation, and auditing.

¹⁴² European Convention on the Protection of the Archaeological Heritage (Revised), <http://conventions.coe.int/Treaty/en/Treaties/Html/143.htm>

6 VALIDATION WORKSHOP

A consultation and validation workshop was held as an official side event of the 10th Plenary Session of the Group on Earth Observations & 2014 Ministerial Summit. The workshop attracted over 40 attendees from 14 countries, including policy makers, funding bodies, libraries, data management organisations and researchers, along with representatives from the RECODE case studies and RECODE team members. A complete list of institutions represented at the Workshop can be found in Appendix 3¹⁴³.

The agenda of the workshop sought to validate and discuss our findings, as well as to obtain additional feedback and insights from representatives of the RECODE case studies and major international initiatives, to share their perspective in understanding Open Access to research data, in relation to infrastructure and technology challenges.

We recognised that many networks, initiatives, projects and communities are addressing the key technical and infrastructural barriers to Open Access to research data. However, these organizations are often heterogeneous and fragmented by discipline, geography, stakeholder category (publishers, academics, repositories, etc.) as well as other boundaries, and often work in isolation or with limited contact with one another.

Hence, we took advantage of the workshop venue, the GEO Plenary Session, to offer a multi-disciplinary space for GEOSS and other stakeholders in Open Access to research data, to identify good practices, areas where further support is required, and possible common solutions to overcome or mitigate what is hindering Open Access and data preservation in science.

The workshop explored the infrastructure and technology challenges by undertaking a review of the existing state of the art and by examining the five RECODE case studies in different scientific disciplines: particle physics and particle astrophysics, clinical research, medicine and technical physiology (bioengineering), the humanities (archaeology), and the environmental sciences (GEOSS).

The aims of the workshop were to:

- Present key findings from a survey of the existing solutions and good practice for Open Data Access, from the five RECODE case studies and other initiatives;
- Gather feedback on the effectiveness of the existing solutions, as well as any gaps, to assess their significance in practice;
- Better understand how to increase the effectiveness of the current technological baseline in supporting Open Access to research data.

The workshop agenda comprised a morning session with several presentations and an afternoon session featuring plenary open discussions. The full agenda for the workshop can be found in Appendix 4. The workshop presentations and minutes are accessible on the RECODE website.¹⁴⁴

The morning presentations covered the EC perspective and the GEOSS Data Sharing Principles, and the Collaborative Research Action on e-Infrastructure and Data Management

¹⁴³ Due to issues of privacy, a full list of names will not be made public.

¹⁴⁴ European Commission, Policy RECommendations for Open Access to Research Data in Europe, “Recode Workshops”, op. cit., no date.

promoted by the Belmont Forum, a high level group of the world's major and emerging funders of global environmental change research and international science councils, co-founded by UK NERC and the US National Science Foundation in 2009. The key findings from RECODE activity were also presented. The afternoon plenary open discussions scoped gaps, and practical significance of the existing technological baseline for Open Research Data Access, and recommendations on how to increase its effectiveness. They were introduced by the following key questions:

- Where should open research data be stored and made accessible?
- How can we mitigate the technological barriers to Open Access to research Data?
- How can we cope with technological sustainability and obsolescence?
- What emerging technologies could be optimized to promote ease of deposit and retrieval of research data?

Both the morning and afternoon sessions of the workshop engaged a lot of discussions. Especially the open plenary of the afternoon session evoked many and often associative discussions. But the focus was mainly on four specific areas that seem to be at the heart of the infrastructure and technology challenges of Open Access to research data: preservation, including thoughts on repositories and data management plans; metadata; citing and academic credits; standards/formats. Running through all these themes were discussions on quality.

On preservation and repositories

The presentation by the Directorate General Research and Innovation representative Michel Schoupe raised the first comments, all about long-term preservation sustainability. The suggested solution to this was an increase of community-run and certified repositories covering several fields of science. There are already some good examples of such repositories for specific subjects, like Dryad¹⁴⁵, DataONE¹⁴⁶, GenBank¹⁴⁷, and even the GitHub¹⁴⁸ code repository is reportedly used for scientific data. But still here is a growing concern about certified repositories for data preservation and dissemination. As mandates for making datasets openly available will increase, so will the need for preservation solutions. As another speaker, Max Craglia from JRC, put it: provenance and persistence are the two big challenges.

This is also what our survey indicates - researchers have quite a lot of experience with Open Access publications, but not with data publications or data preservation. That the question of preservation is high on the agenda of stakeholders is also quite clear from the survey where data preservation is one of the three top factors that respondents thought could hinder the uptake of Open Access to research data in the near future. This is also emphasised in interviews with RECODE case studies. In this context the problem of storage was mentioned and cloud storage is another area that needs to be considered in more depth when we are talking about persistence, ownership and license agreement. According to the views expressed at the workshop, an important factor for researchers is a full integration of repositories and libraries and that libraries should be the facilitators for data access and management plans.

¹⁴⁵ Dryad, "Homepage", no date. <http://datadryad.org/>

¹⁴⁶ DataONE, "Data Observation Network for Earth", no date. <http://www.dataone.org/what-dataone>

¹⁴⁷ GenBank, "GenBank Overview", no date. <https://www.ncbi.nlm.nih.gov/genbank/>

¹⁴⁸ GitHub, "Build software better, together", no date. <https://github.com/>

On metadata

Connected to the issue of preservation is of course the need for good quality metadata. This is also stressed both in the survey and in the interviews – metadata are crucial for retrieval, re-use and preservation. In the afternoon in the first plenary open discussion metadata was again mentioned as something that now is mostly missing, but in the future has to be one of the cornerstones in the development, structuring and storing of data. Useful and accessible data must be well documented. Without proper documentation, it is difficult to replicate and validate the research based on the available data. This was said several times during the discussions. In the second plenary open discussion one of the participants specified the need for mandatory metadata standards. Researchers that submit datasets should be referred to accepted formats that can be used. For non-standard datasets, it would be the responsibility of the data owner and the data coordination team so see how best to make that data accessible. Generally simple solutions (software and services) were called for, in all aspects involving the researchers, especially for time-consuming work like adding good and metadata that conforms to standards. One of the open plenary discussion themes was quality. Good quality metadata, for example, must contain license information and security aspects so that systems for data management can allow for multi-level access to data.

Citation and academic credit

One aspect of quality brought up, was the way we can measure the usage of data. Data citation will become more and more important. As yet there are no standards for data citation. However, some participants in the open plenary discussion argued that there is no such thing as an inherent “*dataset quality*”. Data quality is in the eye of the beholder. But during the discussion it was argued that there is a relation between data, accessibility and quality assessment and that this means that there need to be metrics of the data level available for the user in order to assess what data to use. One participant emphasized that open research data must be supported by data citation policies as a way to protect the investment in infrastructure. To get researchers to submit data a scheme for citation has to be in place reassuring researcher proper academic credit. In connection with this, the PREPARDE¹⁴⁹ project work on data publishing workflow and peer-review was mentioned, advocating the need for a citation/linking network for data, similar to what CrossRef¹⁵⁰ is to publications. Data repositories like Dryad¹⁵¹ and Zenodo¹⁵² are in fact working on data citation principles. Publishers definitely have a role to play in creating citation/credit systems for researchers. The industry standard for identifying publications, DOI, is available also for datasets. Everybody agreed on using DOIs alternative URNs as unique identifiers also for datasets. And the ORCID¹⁵³ system for identifying researchers was mentioned in connection with this and the need to credit the original data creators.

On standards

In one of the presentations of the morning session, Stefano Nativi, whilst presenting the activities of the Belmont Forum, promoted the need to use standards although standards

¹⁴⁹ University of Leicester, “Peer Review for Publication & Accreditation of Research Data in the Earth Sciences”, no date. <http://www2.le.ac.uk/projects/preparde>

¹⁵⁰ CrossRef, “CrossRef.org”, no date. <http://www.crossref.org/>

¹⁵¹ Dryad, “Homepage”, no date. <http://datadryad.org/>

¹⁵² Zenodo, “Research. Shared”, no date. <http://zenodo.org/>

¹⁵³ ORCID, “Distinguish yourself in three easy steps”, no date. <http://orcid.org/>

sometimes are too numerous to be really effective. This is supported by the results in the survey, where 46% of the producers of data are concerned that the heterogeneity of data formats will be a major challenge. Owing to this the challenge of non-authoritative research data, so called crowdsourcing, was mentioned as a critical issue for the future. How can we deal with multi-disciplinary data such as crowdsourcing data? One of the answers to this was that we need to supplement the structural metadata that provides information about the organization and the structure of the data with semantic metadata that provides information about the data itself – what it is about and the available semantic relationships in the particular domain model in which the data is defined. Without proper documentation it will be difficult to replicate and validate research based on the available data. De-contextualisation was mentioned in association with crowdsourcing as a risk and also as a challenge for designing and enforcing security policies. On the other hand datasets are mostly the work of more than one individual so it follows that a technical infrastructure must support solutions for that. During the discussion on standards it was clear that the heterogeneity of data and data processes makes it very difficult to choose standards. Variables include different data practices, workflows and different ontologies.

Other infrastructure issues

One suggestion for identification of issues and challenges of open data research data infrastructures was to use a bottom-up perspective using pilot projects to probe issues like data curation, use of software, interoperability and then formulate policy guidelines based on the experiences in the field. The software used to create datasets should ideally be open source and the research community can ensure that issues with interoperability and accessibility will be dealt with. Third party software should be avoided. Another interesting idea that surfaced during the workshop was the suggestion for a technical infrastructure allowing different access levels, as different actors define quality at different levels. Scientists typically focus on raw data, whereas peer reviewers may be more interested in processed data. In conjunction with this the security aspect of data was mentioned as a very important aspect but most of the participants felt that many of the security issues could be addressed fairly well by technology already in existence. It is not security issues at the technological level that are a problem, rather it is implementing changes at the policy level that will take time and energy.

Finally, the basic requirement for any sort of technical infrastructure was discussed - Funding. Is there enough financial capacity to make data open over a longer period of time? Researchers have to pay up front for preservation, dissemination and access. There must be additional funding to hire and train staff managing these things. One possible way of providing more cost-effective services among institutions and organizations in neighbouring fields is to share services. Another suggested way to cover costs was to charge money for access to data and introduce a “*pay per use*” policy. Open source software tools could also be a way to reduce cost and keep the solutions inside the community. Virtual machines were said to have a role here in lowering costs and sustaining services executing security, access and license tasks.

6.1 FINDINGS FROM THE WORKSHOP

The workshop discussion overall validated our survey and case study results. Data heterogeneity was picked up as a very important challenge, which leads on to a discussion of standards in open data publishing. The growth of data is recognised and options for making the data accessible and useable are deemed as somewhat lacking. With regard to responsibility for storage, the workshop attendees agreed with our findings from the survey, in that the preference expressed is for the enhancing of digital libraries, and specialised repositories to store and curate research data. Data preservation, in terms of long-term storage solutions and curation options, remains a key challenge.

Several projects have been working on solutions for preservation, citation and description of research data. Technical components and standards are available, although we need to agree on their use: what formats, identifiers, citation standards for data, certification of repositories, storage, data management should be recommended. Besides, we need to agree on the issue of alignment between different standards in different domains, e.g. research data in different disciplines, INSPIRE Metadata for spatio-environmental data, DCAT¹⁵⁴ for PSI. This also means to align vocabularies, etc.

Heterogeneity and interoperability

Where you have heterogeneity you need interoperability. The open research data landscape is heterogeneous and fragmented by discipline, geography, stakeholder category (publishers, academics, repositories, etc.) as well as other boundaries. So apart from having financial capacity to make data open one of the important requirements are software systems that can exchange data using common standards for protocols and schemas. Sharing services is one possible way of providing more cost-effective services among institutions and organizations in neighbouring fields. Using open-source software tools could also be a way to reduce cost and keep the solutions inside the community for easier further developments. Using distributed repository systems administrated by university libraries to facilitate smaller datasets and access to those can be a way to accommodate the heterogeneity of data and research cultures.

The heterogeneity of data formats will be another major challenge. Owing to this the challenge of non-authoritative research data, so called crowdsourcing, can become a critical issue for the future. How can we deal with multi-disciplinary data such as crowdsourcing data? One of the answers is that we need to supplement the structural metadata that provides information about the organization and the structure of the data with semantic metadata that provides information about the data itself – what it is about and the available semantic relationships in the particular domain model in which the data is defined. It is clear, though, that the heterogeneity of data and data processes makes it very difficult to choose standards. Variables include different data practices, workflows and different ontologies.

Accessibility

Metadata is one issue that is always mentioned as crucial for accessibility. Researchers that submit datasets must be referred to standard metadata formats. In fact, mandatory standards

¹⁵⁴ W3C, “Data Catalog Vocabulary (DCAT)”, W3C Recommendation, 16 January 2014. <http://www.w3.org/TR/vocab-dcat/>

for metadata are much needed as well as standard formats for data files. For non-standard datasets, it would be the responsibility of the data owner and the data coordination team so see how best to make that data accessible.

Metadata creation can be time-consuming and there is a need for simple solutions. The software used to create datasets should ideally be open-source, so that the research community can ensure that issues with interoperability and accessibility will be dealt with appropriately. Projects like PRIME¹⁵⁵ and the already mentioned PREPARDE¹⁵⁶ have provided recommendations on metadata workflows for data repositories. Experiences from projects like these should be put to use.

Another thing touching the importance of repositories is the views expressed at the workshop that an important factor for researchers is a full integration of repositories with libraries, and that libraries should be the facilitators for data access and management. This fits with the survey, where a large number of respondents consider digital libraries and librarians as a first choice for storing European research data and making it accessible. Libraries together with national institutionalized repositories and centralized European repositories were the three favoured choices in the survey.

Preservation and curation

Long-term persistence of research data is obviously one of the subjects that concerns all stakeholders. Preservation is one of the three top factors in our survey that respondents thought could hinder the uptake of Open Access to research data in the near future. This is also emphasised in interviews with RECODE case studies. There is a lack of a culture of data management, governance, and infrastructure. Who shall curate and preserve the data? A suggested solution to this is community-run and certified repositories covering several fields of science. There are already some good examples of such repositories for specific subjects, like Dryad¹⁵⁷, DataONE¹⁵⁸, GenBank¹⁵⁹, and even the GitHub¹⁶⁰ code repository is reportedly used for scientific data. New certified repositories are available since some years back, but there is a need for more such subject-based solutions for data preservation and dissemination. As mandates for making datasets openly available will increase, so will the need for secure preservation solutions.

Stakeholders view national repositories together with European repositories as a solution to prevent preservation storage and curation gaps. This goes against ideas voiced, for example, in the EC Consultation on open research data, where huge centralized repositories were seen as misuse of resources for leveraging the existing infrastructure. At the Consultation, as well as at the workshop, distributed repositories administrated by university libraries to facilitate access to smaller datasets were recommended as a way to accommodate the heterogeneity of data and research cultures. The option of commercial, third party storage, such as cloud storage, is another area that needs to be considered with caution when talking about persistence of research data. Can cloud solutions hosted by commercial companies be

¹⁵⁵ PRIME, “Privacy and Identity Management for Europe”, no date. <https://www.prime-project.eu/>

¹⁵⁶ University of Leicester, “Peer Review for Publication & Accreditation of Research Data in the Earth Sciences”, no date. <http://www2.le.ac.uk/projects/preparde>

¹⁵⁷ Dryad, “Homepage”, no date. <http://datadryad.org/>

¹⁵⁸ DataONE, “Data Observation Network for Earth”, no date. <http://www.dataone.org/what-dataone>

¹⁵⁹ GenBank, “GenBank Overview”, no date. <https://www.ncbi.nlm.nih.gov/genbank/>

¹⁶⁰ GitHub, “Build software better, together”, no date. <https://github.com/>

trusted? The workshop attendants expressed scepticism on this, also for their cost and governance issues.

In general, funding is needed to make data open in a sustainable infrastructure. A majority of data management projects are bottom-up approaches simply because there are generally no organizational and financial policies outlining how to handle open research data. The risk is that we end up with structures and standards that will not support sustainable sharing and preservation. However, this is not a strictly technological issue.

Quality and assessability

Quality is a recurring issue, running through all the workshop discussions, as a sensitive topic that can spur wide debate. The workshop confirmed the fundamental reluctance of scientists, both in the role of Producers and in that of End users, to accept quality statements as a property inherent to the data, especially when expressed by third parties. Even the concept of data peer review, which was emerging from the survey as a possible option for improving data quality, was considered an unpractical solution.

The workshop attendants seemed more positive towards the more subjective and neutral concept of “*fitness for use*”, as advocated by several data sharing initiatives (e.g. GEOSS). The concept could be implemented supporting the collection of user feedback in data repositories, to be integrated with the metadata, to allow users to assess the data suitability for their specific purpose.

As reiterated in the interviews, complete and accurate metadata are of paramount importance for conveying quality of scientific data. In particular, provenance information (creator(s), source organisation, holding organisation) is considered of primary importance, so that users can assess the value of the dataset, its authority and trustworthiness, as well as to ensure the repeatability of processes. Without proper documentation, it is difficult to replicate and validate any research result, based on the available data. Mandatory metadata standards are seen as a good practice for enforcing and assessing quality in scientific data. Good quality metadata should also contain license information and information on security aspects, so that systems for data management can allow for multi-level access to data.

Security

One issue, which did not feature in our survey, nor in our interviews, was academic citations and academic credit and the role of technology in identifying individual academics and their contribution. Workshop attendees stressed that this was of great importance as academic credit was instrumental in furthering the implementation of Open Access to research data and the quest for publishing good quality data. If datasets are attributed to specific individuals or consortia, both these issues could be mitigated. The workshop attendees had knowledge of technology addressing this particular issue, e.g. ORCID¹⁶¹. PRIME¹⁶² and the already mentioned PREPARDE¹⁶³ have provided recommendations on accreditation requirements. Other projects like Dryad¹⁶⁴ and Zenodo¹⁶⁵ are working on data citation principles. Getting

¹⁶¹ ORCID, “Distinguish yourself in three easy steps”, no date. <http://orcid.org/>

¹⁶² PRIME, “Privacy and Identity Management for Europe”, no date. <https://www.prime-project.eu/>

¹⁶³ University of Leicester, “Peer Review for Publication & Accreditation of Research Data in the Earth Sciences”, no date. <http://www2.le.ac.uk/projects/preparde>

¹⁶⁴ Dryad, “Homepage”, no date. <http://datadryad.org/>

credit by making research data citable is a requirement that comes across especially in the creator stakeholder group. Using the available DOI or URN schema as unique identifiers is a solution that is often implemented when data is made citable.

To get researchers to submit data a scheme for citation has to be in place reassuring researcher proper academic credit. The PREPARDE project work on data publishing workflow and peer-review is exploring and advocating the need for a citation/linking network for data, similar to what CrossRef is to publications. Data repositories like Dryad and Zenodo are in fact working on data citation principles. Publishers definitely have a role to play in creating citation/credit systems for researchers.

The industry standard for identifying publications, DOI, is available also for datasets. Using URNs as unique identifiers also for datasets would make dissemination and reuse of datasets safer. The ORCID¹⁶⁶ system for identifying researchers is another initiative that can be used for reassuring that users are dealing with the correct records plus as a matter of giving credit to the original data creators.

Without proper documentation it will be difficult to replicate and validate research based on the available data. De-contextualisation is another matter mentioned in association with crowdsourcing as a risk and also as a challenge for designing and enforcing security policies. On the other hand datasets are mostly the work of more than one individual so it follows that a technical infrastructure must support solutions for that.

Another aspect of quality is how we can measure the usage of data. Data citation will become more and more important. As yet there are no standards for data citation. There is a relation between data, accessibility and quality assessment and this means that there need to be metrics of the data level available for the user in order to assess what data to use. There are projects like Dryad and Zenodo working on data citation principles. Getting credit by making research data citable is a requirement that comes across especially in the creator stakeholder group. Using the available DOI or URN schema as unique identifiers is a solution that is often implemented when data is made citable.

It seems that the security aspects of data are regarded as very important aspects but most of the participants felt that many of the security issues could be addressed fairly well by technology already in existence. It is not security issues at the technological level that are a problem, rather it is implementing changes at the policy level that will take time and energy.

¹⁶⁵ Zenodo, "Research. Shared", no date. <http://zenodo.org/>

¹⁶⁶ ORCID, "Distinguish yourself in three easy steps", no date. <http://orcid.org/>

7 INTERNATIONAL ADVISORY BOARD COMMENTS

The members of the International Advisory Board were sent a draft of this WP2 Deliverable prior to e-meetings to discuss it. The members of the Board are Max Craglia (Italy, EU JRC), Toby Burrows (University of Western Australia, Australia), Professor Jerome Reichman (Duke University, United States of America). The panel members initially included Boyong Wang, as representative for Dr Cao Jing (Handan City-EU Affairs representative; Dr Jing's representative for EU related projects in China), who kindly declined our request, as he recently moved from his previous position as ICT adviser for Yantai government, and is no longer updated on the research field of Open Access to data.

Max Craglia found the report generally well written and interesting. He suggested mentioning the GEOSS Data Collection of Open Resources for Everyone (GEOSS Data-CORE), which has been a key mechanism to address the limitations identified in implementing the GEOSS Sharing Principles. He underlined that INSPIRE is primarily concerned with interoperability, more than Open Data Access, and clarified that most of the data held by the JRC, as the maintainer of the INSPIRE geoportal, is not owned by the JRC, but by the Member States that provide access to the JRC for specific projects, or as part of legal requirements. Not being the data owner, the JRC is not in the position to disseminate data openly, as the ability to provide Open Access is constrained by third party IPR. Equally, the JRC view on data management is informed by that of the organizations on behalf of which JRC has the data in the first place. This may not necessarily be JRC's view, but they have to respect data ownership. He noted that there is a skills/knowledge challenge to understand the data fully, particularly in a multidisciplinary context, so that one can really use the data. A similar issue pertains data preservation and DMP, where the challenge is to define the criteria by virtue of which one can decide what to keep and what to throw away, especially in a Big Data world. In general, the real main issue is a lack of a culture of data management, governance, and infrastructure. With this regards, he found the recommendations somewhat lacking, as there is too much emphasis on metadata and standards and not enough on the culture of data management and infrastructure for data curation and preservation, which exists in some communities (e.g. university libraries, thematic areas like physics or social science), but is almost totally absent in the realm of administrative Public Sector Information, which should always be included into the definition of research data.

Jerome Reichman commented that our work has many points in common with several ongoing projects of the Belmont Forum and intended to refer their coordinators to RECODE. He was recently attending a Belmont Forum meeting where he brought up several issues that are relevant to Open Access to research data, especially from the perspective of funders and policy-makers, but with major implications for technical and infrastructural arrangements. He thought our work correctly frames the main infrastructure and technology challenges and would specifically consider the case studies research of interest. He points out some very important concerns that should not be overlooked. A concern is the fact that the increasing momentum of Open Access is spurring a great number of volunteer efforts for data sharing in very different contexts, which result in the implementation of data sharing solutions at very different scale, e.g. for a single community of practice, or a specific project, and in the fragmentation of data in a puzzle of individual pieces, which he referred to as “*semi-open data pools*”, or “*semi-commons*”¹⁶⁷. These, although in principle informed by the overall

¹⁶⁷ Reichman, Jerome H., Paul F Uhler and Tom Daederwerdere, *Governing Digitally Integrated Genetic Resources, Data, and Literature: Global Intellectual Property Strategies for the Microbial Research Commons*, Cambridge University Press, forthcoming 2015.

vision of data sharing, actually work in isolation from each other, as explicitly recognised in the RECODE DoW¹⁶⁸. The Open Access debate is largely unaware of the issue, which instead should be carefully considered (we have addressed this aspect in our workshop, for example). To mitigate the problem, funders may force publishers to publish data as Open Access together with scientific articles, or force that semi-commons transfer their data to Open Access repositories, when the cease to exist. Another concern is the combination of data that falls under different jurisdictions, e.g. EU and USA policies, especially when such combination of data is suitable to commercial exploitation. Funders and policy-makers should address this problem, for example by developing standard data transfer license that may be automatically enforced at the infrastructural and technological level. Another important concern is the governance of Open Access repositories as they become massive and Big Data and Open Data concerns overlap. As we have recognised (see chapter 3.1), this is a very current trend (e.g. in GEOSS), which may lead to a bureaucratisation of governance, where critical decisions are delegated to politician, typically not fully aware of the related scientific implications. It is instead necessary that the governance of big Open Access repositories primarily involve the scientists, which should ultimately implement the critical decisions.

Toby Burrows commented that it is important to recognise that there are multiple degrees of openness and that the definition of Open Access as indiscriminate is too simplistic. In fact, most researchers underline this and are comfortable with knowing that access to data is somehow regulated. Hence, an appropriate authorization infrastructure is crucial for Open Access to research data. He found the case studies very interesting and agreed with the identified challenges, which are broad enough to comprise the various issues. He also appreciated the overview table as a good way of summarizing, although approximation must be made when you classify such a complex matter. He noted that some of the findings recur in different sections, what reinforces the overall conclusions. With regards to recommendations on infrastructure and technology, he underlined that we should aim at stating the principles, and refrain from indicating specific technologies, as this could be a sensitive issue for many communities of practice. By recalling the work done in the framework of the Australian National Data Service, he pointed out that there is a mixture of solutions at different levels, such as community, national, disciplinary, or supranational. The real challenge is in making them work together as seamlessly as possible. In fact, there are also different levels or degrees of interoperability. At the simplest level, individual repositories make available flat-file versions of datasets, which can be downloaded and imported into other software environments. At the most complex level, a central service (usually discipline-based) harvests data exposed by various individual repositories, and enables researchers to work directly with the aggregated data in a uniform way. The UK Data Archive does this for social science quantitative data, for example.

¹⁶⁸ RECODE Project, Annex I, op. cit., 15 July 2012, p. 3.

8 DISCUSSION

It is clear from the combined results of the survey, the literature review, the case study interviews and the workshop that stakeholders in general have a limited knowledge about research data management and how to make data openly available in a multidisciplinary way. As reiterated in the literature review, in our online survey results and in our case study findings, technological barriers are not reported as of high concern to implementing Open Access to research data, when compared to financial, cultural and legal challenges. We maintain that Open Access to research data is still at an early stage within Europe and internationally. Hence there are significant shortcomings in the current technological framework, including lack of standards, policies, and best practice cases, which could help move on this research work. The technical infrastructure for Open Access to research data in most research communities is simply just not there.

However, the recent efforts to provide good infrastructure for sharing datasets, especially in the areas of biomedicine and physics, constitutes valuable experience and good practice to fall back on when discussing strategies for the future. Together with the lessons learnt from the more widespread Open Access to publication, as well as from global information sharing endeavours such as GEOSS, such findings provide useful insights and can contribute to the overarching set of recommendations for a European policy on Open Access to research data, which is the ultimate goal of RECODE.

Throughout we have identified a number of technological barriers, which need to be addressed to further the implementation of Open Access to research data in Europe. Here we attempt to further consolidate the barriers under five descriptive headings for the purpose of moving towards the overall RECODE objective of proposing policy recommendations, which should help to mitigate these categorical challenges. These headings do portray the overall grander technological challenges, but also take note of discipline specific challenges, mentioned in the survey and the case study research.

- **Heterogeneity** – relates to the variety of data at any level, e.g. format and encoding issues, data accessibility, protocol interoperability, but also high level issues, e.g. application interoperability, semantics mismatches, cross-disciplinarity, usability, internationalisation, discoverability of data in a growing number of disparate sources.
- **Accessibility** – relates to the volume of data and to its impact on the infrastructure capabilities and architecture, e.g. the organization of storage and processing resources, data discoverability, indexing and filtering, bandwidth issues.
- **Sustainability** – relates to the long-term impact of maintaining and operating an open infrastructure for research data, e.g. obsolescence, governance of updates/upgrades, data preservation and curation, persistence, scalability, energy footprint.
- **Quality** – relates to the technological support for the evaluation of data suitability and appropriateness, e.g. accuracy, completeness, documentation including metadata and other ancillary information, peer review, assessment and validation, usefulness, fitness for purpose.
- **Security** – relates to the restrictions on the usage, access, and consultation of data and metadata, and their enforcement from a technical viewpoint, e.g. protocol for authentication, authorization and auditing/accounting, privacy issues, policy enforcement, licensing.

Although the complex issues of Open Access are all interrelated, it can be noted that these challenges reflect orthogonal properties of data that are quite independent from one another.

For example, data heterogeneity would be a barrier to consider even in hypothetical case of a single dataset, of optimal quality, persistently stored in a secure infrastructure. In that case, Open Access would imply accessibility by many potential users and applications, so there would still be a prominent issue of *client-side* heterogeneity to address. Likewise, the accessibility barrier, which we relate to one of the orthogonal axes of Big Data, Volume, is independent from homogeneity, persistence, quality, and security of data, etc.

Quality and security have a prominent importance in Open Access: as data are made available to access and use by others, the issue of its quality and fitness for purpose becomes essential; as for the security challenge, we have found that many implications of Open Access in the current research practice are incompatible with the definition of Open Access sometimes found in literature, as completely unrestricted. This was highlighted especially during our case studies analysis and the workshop, especially for some scientific communities, such as health and archaeology.

We also acknowledge that the identified challenges differ in relevance depending on stakeholder group functions and perceptions. In fact, distinct actors of the Open Access ecosystem have a different perception of the technological barriers they confront with. For example, sustainability issues may be more relevant for Funders, than for data Creators. Likewise, heterogeneity issues have less impact on the Funders, than on the other stakeholder categories. Besides, these broadly identified challenges can be better qualified in specific issues for the distinct stakeholders. Over-arching solutions should take into account the different stakeholders' experience of infrastructure and technology issues to be effective. Table 1, below, provides a summary overview of the above challenges and some of their specific aspects, in relation to the stakeholder groups that are most affected by each of them.

Table 1 - infrastructure and technology challenges overview

	Creator	Disseminator / Curator	Funder	End user
Heterogeneity	standardization, encoding, semantics	interoperability, reuse, data cross-walk, internationalisation		usability, encoding, semantics
Accessibility	bandwidth	standardization, storage, scalability, distribution,		discoverability, storage, bandwidth
Sustainability	persistent identification	obsolescence, reuse, data migration, persistent identification	governance	obsolescence
Quality	provenance, training, fitness for use	provenance, training, completeness, peer review	certification	provenance, peer review, fitness for use

Security	authorization, attribution, licensing	authentication, authorization, accounting, privacy, obfuscation	licensing	authentication, privacy, trust
-----------------	---------------------------------------	---	-----------	--------------------------------

As highlighted by the overview table, the stakeholder categories that are most affected by infrastructure and technical issues are the Disseminator/Curator and the End user. This was to be expected, in light of levels of interaction between these groups and the infrastructure, as we present in our stakeholder taxonomy, where the Disseminator/Curator role is in charge of the hardware and software facilities (information systems, e-infrastructure) for data storage, access, distribution, maintenance, and preservation. With regard to the End user group, it is seen as a broad and diverse group, crossing a variety of disciplines, as well as including the general public. It is therefore to be expected that their different needs and capacity will pose a wide range of challenges. The interaction of the data Creator with the technological infrastructure is comparatively lower than that of the End user, as data creation can be regarded as a temporally limited event in the data life cycle, when compared to data exploitation, especially in the perspective of data preservation. Funders' concerns with infrastructure and technology aspects of Open Access to research data pertain mainly to sustainability issues and other policy aspects, such as licensing and certification.

Table 2, below, provides a summary overview of initiatives where good practice seem to be in use, in relation to the above infrastructure and technology challenges.

Table 2 – good practice overview

Infrastructure and Technology issues	Good practice
Heterogeneity	<p>GEOSS on brokering and data cross-walk</p> <p>INSPIRE, Open Knowledge Foundation, Linked Data on standards for data models</p> <p>OBO Foundry, OpenAIREplus on ontology interoperability</p> <p>PANGAEA, OpenAIREplus on federated approach</p> <p>re3data.org, CESSDA on data repositories</p>
Accessibility	<p>DOI, DataCite on data citation</p> <p>INSPIRE, OpenAIRE, Open Knowledge Foundation, Linked Data on standards for discovery and access</p> <p>SCIDIP-ES, CAS Registry, ESO on archive solutions</p>
Sustainability	<p>DRAMBORA, PANGAEA on curation</p> <p>OAIS Reference Model (ISO 14721:2003 on sustainable digital archive</p>
Quality	<p>Data Seal of Approval, DINI Certificate , TRAC, ISO 16363:2012 on repository accreditation</p> <p>GEOSS on user feedback and fitness for use</p>

	PRIME, PREPARDE on metadata workflows
Security	<p>CreativeCommons CC-BY or CC-BY-SA on licensing</p> <p>DRAMBORA on auditing</p> <p>Dryad, Zenodo on data citation</p> <p>ORCID, PRIME, PREPARDE on accreditation of researchers</p>

8.1 RECOMMENDATIONS ON INFRASTRUCTURE AND TECHNOLOGY FOR OPEN ACCESS TO RESEARCH DATA

On the basis of our research and the above discussion, we attempt here to indicate possible recommendations on infrastructure and technology for Open Access to research data. As suggested by a member of the Advisory Board, we aim at stating quite general principles, instead of indicating specific technologies, as this could be a sensitive issue for many communities of practice. These recommendations are intended as an 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.

We propose that any policy response take account of the five infrastructural and technological challenges that we have identified, whilst also paying attention to the more specific issues that arise in relation to a specific stakeholder function. However, it is important to realise that, in the complex Open Access ecosystem, all these aspects are interrelated. To effectively allow the researchers to identify, evaluate, access, and use relevant scientific information extracted from an open data repository, it is necessary to adopt technical and infrastructural solutions the holistically address data harmonization, discovery and access, data preservation, technological obsolescence, as well as data documentation and metadata, quality and relevance indicators, and security aspects.

At the same time, a policy for Open Access to research data should take into account the different attitudes in different fields of science towards the issue, as well as the specificities of the different communities, that imply essential, inherent heterogeneity. We concur with WP1 that, when discussing Open Data, there is a clear tendency to refer to science as a whole sector, thus there is a danger that the differences between disciplines are ignored in further policy making. Each discipline has different methods for gathering and analysing data. Data may be images, numerical, narrative, statistical and presented in small, medium or large datasets that might be discrete or interlinked.

Moreover, it is important to remember that the definition of research data includes Public Sector Information, hence the public, institutional stakeholders that manage that information should always be addressed. It is necessary to be flexible applying extensible technological and organizational solutions, avoiding solutions that do not satisfy the specificity of the different disciplinary communities, and hence raise entry barriers that are already high.

To mitigate heterogeneity, accessibility and quality issues, it is necessary to adopt open, good standards, as reiterated by many participants in our research. The importance of metadata and data standardization should be reinforced (e.g. defining common models and encodings), to promote ease of deposit and retrieval for stakeholders including researchers, universities, libraries and the general public. A culture of standards should be promoted, both in education and in research practice.

Data variety should be acknowledged and accommodated, resorting to solutions based on distributed architectures and filling the gaps by implementing interoperability between the existing systems. Leveraging on the existing research infrastructures, supplementing and not supplanting them, is essential to guarantee a sustainable action and a valuable participation of the research community.

Heterogeneity and accessibility should be addressed building upon the experience and lesson learned from the ongoing international programmes and initiatives for data sharing (e.g., GEOSS and INSPIRE in environmental sciences and PSI). In particular, the good practice of System-of-Systems and brokering/mediation solutions should be considered, as currently adopted for example in GEOSS, where the infrastructure is able to provide harmonized discovery and access services to heterogeneous data by means of the brokering approach¹⁶⁹. As recommended by the EGIDA Methodology Guideline TA.3d.1¹⁷⁰, an infrastructure for Open Access to research data should be conceived as a System-of Systems, to build on the existing and protect previous investments. This would also help mitigating the issue of semi-commons, underlined by a member of the Advisory Board.

To address the accessibility and the sustainability challenges, data discoverability should be promoted by enforcing the use of persistent digital identifiers, e.g. by requiring the use of DOI, URN, or similar PID technologies. User identification should also be enforced, for example with ORCID or equivalent solutions. Interoperability of PID should be ensured.

As for sustainability, a fundamental necessity is to promote a culture of data management and infrastructure for data curation and preservation. This is well established in some contexts, such libraries, or some fields of science (e.g. physics or social science), but is almost totally absent in others, particularly in the realm of administrative Public Sector Information. As recommended by the EGIDA Methodology, *“technical barriers are removed through the availability of technical expertise and tools. These include the establishment of technical Task Forces, and training activities for individual capacity building such as workshops, summer schools, web lectures, etc.”*¹⁷¹ As noted by one of the members of the Advisory Board: *“of course metadata and standards are important, but if you do not have data repositories, and archives for public sector data, and staff with the skills to look after and maintain the data, then metadata and standards are clearly not enough.”* In relation to this, the new professional roles and profiles required by the implementation of Open Access to research data should be investigated, particularly with respect to the issue of data curation.

¹⁶⁹ Nativi, Stefano, Massimo Craglia and Jay Pearlman, “The Brokering Approach for Multidisciplinary Interoperability: A Position Paper”, *International Journal of Spatial Data Infrastructures Research*, Vol.7, 2012, pp. 1-15. <http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/viewFile/281/319>

¹⁷⁰ Mazzetti, et al., op. cit., 2013, p. 46.

¹⁷¹ Ibid., p. 41.

Technological obsolescence should be addressed resorting to virtualisation technologies. Periodical migrations to more recent technological solutions should also be considered. The migration policies should address format conversion, transcoding, etc.

To address the quality challenge, it is necessary to enforce the presence of complete and accurate metadata, by requiring data Producers and Disseminators/Curators to provide and maintain appropriate ancillary information when publishing and curating the data. In particular, the completeness of information about provenance should be enforced, to ensure the repeatability of processes). To mitigate the sensitive issues related to quality statements, the concept of fitness for use should be adopted, as advocated by several data sharing initiatives (e.g. GEOSS). This could be implemented supporting the collection of user feedback in data repositories, to be integrated with the metadata, to allow users to assess the data suitability for their specific purpose.

To address the security challenge, it is important to distinguish among different levels of Open Access and to implement appropriate access control. A security framework for authentication, authorization and auditing is an unavoidable requirement, advocated by many communities and use-cases. Further, it is important to recognise that sharing does not necessarily mean unrestricted and free access; hence appropriate licensing policies (including data transfer licenses) should be enforced, especially when the data is suitable to commercial exploitation. Some disciplines deal with sensitive data that should be obfuscated when necessary, while others deal with data that may have specific IPR or legal issues. It is important that these differences be acknowledged in any future Open Access policies, and that security policies are automatically enforced at the infrastructural and technological level, where possible.

9 CONCLUSIONS

The aim of RECODE WP2 was to identify the perceived technological barriers to Open Access to research data, to survey the current and emerging technologies being used in Open Access repositories to provide access to scientific information and research data, and to recommend possible solutions to increase the effectiveness of the current technological baseline.

Starting from a revisiting of RECODE functional stakeholder taxonomy, we have undertaken a literature survey based on document review, as well as gathered information from a general wider audience through an online questionnaire. We have also addressed the perspectives of stakeholders in the specific fields of science of the RECODE case studies, belonging to various categories, such as data producers, policy makers, infrastructure providers, and final users like the researchers.

We have classified the facets of the identified infrastructure and technology issues according to five main challenges and to the stakeholder categories considered relevant to our scope. The resulting overview enables to highlight and capture correlations and dependencies between the various infrastructural and technological aspects of the complex implementation of Open Access to research data, and to better focus on possible mitigations. We consider that a similar analytical approach could be applied to other aspects of Open Access to research data, such as the legal and policy issues addressed by subsequent RECODE activities.

This work will be input and further discussed in the framework of RECODE WP5, which, based on these results and on the results of the other work packages, will develop a set of good practice policy guidelines targeted at significant stakeholders and key policy makers. The guidelines will provide examples of good practice, as well as checklists and tools to help stakeholders meet Open Access and data dissemination and preservation objectives. There will be a particular focus on recommendations that could assist policy-makers, national research funding bodies, libraries, repositories, publishers, academics, and users of research such as the media, industry, civil society organizations and citizens. The guidelines will be validated at the WP5 policy recommendations workshop, presented at the RECODE final conference, and then disseminated to the wider public.

APPENDIX 1 – SURVEY QUESTIONNAIRE

Technological barriers to Open Research Data Access - questionnaire

This questionnaire aims at gathering information on the existing technological barriers to Open Access and Preservation of Research Data, and the possible solutions adopted to mitigate them.

It has been developed in the framework of the FP7 RECODE Project (<http://recodeproject.eu/>), particularly with the objective of identifying the technological infrastructural requirements for Open Access to Research Data in Europe.

In this context, "infrastructure" means: technological assets (hardware and software), personnel involved, and all the procedures for management, training and support to its continuous operation and evolution.

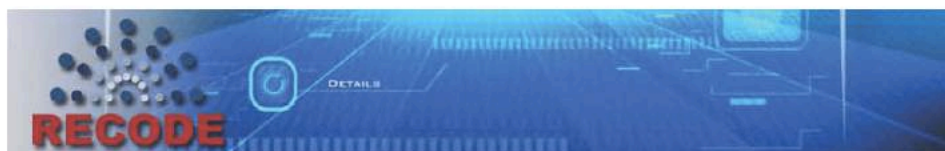
Possible technological barriers include lack of automatic mechanisms for policy enforcement, lack of metadata and data models supporting open access, obsolescence of infrastructures, scarce awareness about new technological solutions, lack of training and/or expertise on IT and semantics aspects.

The questionnaire is structured in two parts:

- An introductory section to gather basic (almost entirely optional) information on the respondent profile and his/her perspective on the scope of discourse; this section also contains generic questions based on the material and the discussion at the EC Public Consultation on Open Research Data, held in Brussels on July 2nd, 2013;

- A profile-specific section depending on the respondent's perspective. We distinguish between: the producers of research data (e.g. researchers elaborating raw data); the data disseminators/curators, who are in charge of the distribution infrastructure (information systems, e-infrastructure) for storage and access to research data (e.g. publisher, librarian); the funding bodies, providing financial and policy support to research; the end users of research data at large, including researchers, the industry, governmental agencies, etc.

* Required



Respondent identification

Identification information (almost entirely optional) will only be disclosed in aggregated form. If you would like to be updated on the progress of the RECODE project, please leave your e-mail address.

Name, Surname

E-mail address

Profession/Role *

Institution

Country *

If you are involved in specific initiatives to support open access to research data, please specify:

Do you think research data should all be preserved indefinitely, in principle? *

yes

no

If not, who should decide what to preserve and until when?

Data producers

Librarians / repository managers

Publishers

Funders

End users

Disciplinary associations / peer review

Do you think preserved research data should all be accessible online? *

yes

no

If not, who should decide what data should be accessible online?

Data producers

Librarians / repository managers

Publishers

Funders

End users

Disciplinary associations / peer review

Who should be primarily responsible for storing European research data and making them accessible? *

Data producers

Digital libraries / Institutions (e.g. Universities)

National institutionalized repositories

Centralized European repository

Publishers

Funders

Other:

With the increase of openly accessible data, what factors do you think will have the greatest impact? *

Please select the three aspects that in your opinion should be considered at the highest priority.

Heterogeneity of data formats

Application interoperability

Data access / download

Catalogues and search engines

Storage capacity on the client side

Obsolescence of data formats

Data preservation

Energy footprint

Bandwidth

Data completeness (gaps, etc.)

Data documentation (metadata)

Data quality

Privacy and security issues

Other:

Who should evaluate the quality of research data? *

Data producers

Librarians / repository managers

Publishers

Funders

End users

Disciplinary associations / peer review

Nobody

Other:

What role do you primarily play in relation to Open Access to Research Data? *

This questions leads to the specific questions for your profile. If multiple roles apply to your case, you can fill in the questionnaire more than once:


Producer (e.g. researcher)

Disseminator/Curator (e.g. publisher, librarian)

Funder

End user

16% completed

Powered by  This form was created inside of Essi Lab.

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Technological barriers to Open Research Data Access - questionnaire

* Required

Producer of research data

This section includes technical questions for producers of research data, e.g. researchers elaborating raw data.

What field do you work in? *

- Engineering
- Computer science
- Natural science (biology, chemistry, physics, etc.)
- Social science
- Health / Medicine
- Arts / Humanities
- Other:

How would you categorize the level of adoption of Open Access in your field? *

1 2 3 4 5

very low very high

Do you have direct experience of releasing your data according to an Open Access policy? *

- yes
- no

If so, please reference specific use-cases, repositories and technologies that you have used or are aware of:

How shareable are the data you work with? *

1 2 3 4 5

very little very much

How reusable are the data you work with? *

1 2 3 4 5

very little very much

Does your institution/organization have a data management plan? *

yes

no

don't know / not applicable

Can you briefly describe the policy you follow for data management (or provide a link to it)?

What solutions are you aware of for searching for and identifying relevant data?
Please reference specific standards and technologies for metadata management and cataloguing that you have evaluated:

Have you encountered technical difficulties related to research data storage and maintenance? *

yes

no

If so, please explain:

What solutions are you aware of for access control to research data?
Please reference specific standards and technologies for access authorization that you have evaluated:

The image shows a screenshot of a Google Drive form. At the top, there is a large, empty rectangular box for text input. Below this box, on the left side, are two buttons: « Back and Continue ». On the right side, there is a progress bar that is partially filled, with the text "33% completed" displayed below it. At the bottom left, it says "Powered by Google Drive". At the bottom center, it says "This form was created inside of Essi Lab:". At the bottom right, there are three links: "Report Abuse", "Terms of Service", and "Additional Terms".

Technological barriers to Open Research Data Access - questionnaire

* Required

Disseminator/Curator of research data

This section includes technical questions for disseminators and curators of research data, such as publishers, or those in charge of the distribution infrastructure (information systems, e-infrastructure).

Do you have direct experience of implementing Open Access to research data? *

yes

no

If so, please reference specific use-cases, repositories and technologies that you have used or are aware of:

What solutions are you aware of for preservation and curation of research data?

Please reference specific technologies for long-term storage and maintenance of research data:

What solutions are you aware of for associating research data to scientific publications?

Please reference specific standards and technologies for data linking and embedding:

As a disseminator/curator of research data, are you offering tools in order to associate research data to scientific publications? *

yes

no

What kind of additional information would you consider relevant to complement scientific

publications?

Raw data

Data documentation (metadata)

Supplemental information (videos, news articles, other media, etc.)

Other:

Can you briefly describe the policy you follow for data management (or provide a link to it)?

How do you describe and document the research data you manage, including their quality?
Please reference specific standards and technologies that you use or are aware of:

Do you take user feedback into account for improving the quality of your data? *

yes


no

Do you audit the use of the research data you manage? *

yes

no

50% completed

Powered by  This form was created inside of Essi Lab.
[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Technological barriers to Open Research Data Access - questionnaire

* Required

Funder

This section includes technical questions for funding bodies, providing financial and policy support to research.

Do you think data preservation activities should be eligible for funding at the institutional level? *

yes

no

Does your organization require research-funding applicants to describe a data management plan? *

yes

no

If yes, what would you require in a data management plan?

What solutions are you aware of for preservation and curation of research data?

Please reference specific technologies for long-term storage and maintenance of research data:

How does your organization plan to support the sustainability of the technical infrastructure for data preservation?

What solutions are you aware of for managing the energy footprint of research data handling?

Please reference specific initiatives and efforts on this topic:

66% completed

Powered by 

This form was created inside of Essi Lab:

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Technological barriers to Open Research Data Access - questionnaire

* Required

End user

This section includes technical questions for end users of research data, including researchers, the industry, governmental agencies, etc.

What field do you work in? *

- Engineering
- Computer science
- Natural science (biology, chemistry, physics, etc.)
- Social science
- Health / Medicine
- Arts / Humanities
- Other:

Do you have direct experience of using Open Access Data in your research work? *

- yes
- no

If so, please reference specific use-cases, repositories and technologies that you have used or are aware of:

Do you think the data you work with can be effectively shared and re-used in a cross-disciplinary fashion? *

- yes
- no
- don't know

Do you think the data you work with are properly organized and catalogued to support user queries? *

- yes
- no
- don't know

If not, what are the main limitations with respect to this aspect?

What specific standards and technologies for data cataloguing are you aware of?

Do you think enough research data are available to you to effectively carry out your research work? *

yes

no

don't know

Do you think the data you work with are sufficiently easy to access/download, when available? *

yes

no

don't know

Are you aware of specific standards and technologies for data access/download?

What solutions are you aware of for describing the quality of research data?
Please reference specific standards, tools and technologies that you have used or are aware of:

83% completed



APPENDIX 2 – INTERVIEW PROTOCOLS

The interview is structured in two parts:

- An introductory section to gather basic information on the respondent profile and his/her perspective on the scope of discourse; this section also contains generic questions based on the material and the discussion at the EC Public Consultation on Open Research Data, held in Brussels on July 2nd, 2013;
- A profile-specific section depending on the respondent's perspective. We distinguish between: the producers of research data (e.g. researchers elaborating raw data); the data disseminators/curators, who are in charge of the distribution infrastructure (information systems, e-infrastructure) for storage and access to research data (e.g. publisher, librarian); the funding bodies, providing financial and policy support to research; the end users of research data at large, including researchers, the industry, governmental agencies, etc.

INTRODUCTORY SECTION

The interviewer should confirm the identification data (name, surname, e-mail address, profession/role, institution, country). The interviewer should underline that identification information will only be disclosed in aggregated form.

1. Are you involved in specific initiatives to support open access to research data? If yes, specify.
2. What is your definition of research data?
3. What types of research data should be openly accessible?
 - E.g., raw data, metadata, cleaned data, good quality data
4. When and how does openness need to be limited?
5. Who should be primarily responsible for storing and provide access to European research data?
 - E.g., data producers, digital libraries/repositories, national institutionalized repositories, centralized European repository, publishers, funders
6. In your opinion, what factors will be most impacting, with the growing volume of openly accessible data?
7. How would you rank the following aspects that hinder Open Access to research data?
 - Cultural barriers, technical barriers, legal barriers, ethical barriers, financial barriers, political barriers

PRODUCER OF RESEARCH DATA

This section includes technical questions for producers of research data, e.g. researchers elaborating raw data.

1. What field do you work in?
2. What is the general adoption of Open Access in your field?
3. Do you have direct experience of releasing your data with an Open Access policy?
4. How shareable and reusable are the data you work with?
5. Does your institution/organization have a data management plan?
6. Can you briefly describe the policy you follow for data management, in particular publication?

7. What solutions are you aware of for searching for and identifying relevant data? Please reference specific standards and technologies for metadata management and cataloguing that you have evaluated.
8. Have you encountered technical difficulties related to research data storage and maintenance?
9. What solutions are you aware of for access control to research data?

DISSEMINATOR/CURATOR OF RESEARCH DATA

This section includes technical questions for disseminators and curators of research data, such as publishers, or those in charge of the distribution infrastructure (information systems, e-infrastructure).

1. Do you think research data should all be preserved indefinitely, in principle?
 - a. If yes, how would this work in practice?
 - b. If not, who should decide what to preserve and until when?
 - E.g., data producers, librarians/repository managers, publishers, funders, end users, disciplinary association/peer review
2. Do you think preserved research data should all be accessible online?
 - a. If yes, how would this work in practice?
 - b. If not, who should decide what should be accessible online?
 - E.g., data producers, librarians/repository managers, publishers, funders, end users, disciplinary association/peer review
3. What solutions are you aware of for preservation and curation of research data?
4. Do you have direct experience of implementing Open Access to research data?
5. What solutions are you aware of for associating research data to scientific publications?
6. As a disseminator/curator of research data, are you offering tools in order to associate research data to scientific publications?
7. What kind of additional information would you consider relevant to complement scientific publications?
 - E.g., raw data, data documentation (metadata), supplemental information (videos, news articles, other media, etc.)
8. Can you briefly describe the policy you follow for data management?
9. How do you describe and document the research data you manage, including their quality?
10. Do you take user feedback into account for improving the quality of your data?
11. How does your organization support the sustainability and availability of your data infrastructure?
12. How do you audit the use of the research data you manage?

FUNDER

This section includes technical questions for funding bodies, providing financial and policy support to research.

1. Do you think research data should all be preserved indefinitely, in principle?
 - a. If yes, how would this work in practice?
 - b. If not, who should decide what to preserve and until when?
 - E.g., data producers, librarians/repository managers, publishers, funders, end users, disciplinary association/peer review
2. Do you think preserved research data should all be accessible online?

- a. If yes, how would this work in practice?
- b. If not, who should decide what should be accessible online?
 - E.g., data producers, librarians/repository managers, publishers, funders, end users, disciplinary association/peer review
3. How does your organization plan to support the sustainability of the technical infrastructure for data preservation?
4. Do you think data preservation activities should be eligible for funding at the institutional level?
5. Does your organisation require research-funding applicants to describe a data management plan? If yes, what would you require in a data management plan?
6. What solutions are you aware of for preservation and curation of research data?
7. What solutions are you aware of for managing the energy footprint of research data handling?

END USER

This section includes technical questions for end users of research data, including researchers, the industry, governmental agencies, etc.

1. What field do you work in?
2. Do you have direct experience of using Open Access Data in your research work?
3. Do you think the data you work with can be effectively shared and re-used in a cross-disciplinary fashion?
4. Do you think the data you work with are properly organized and catalogued to support user queries?
 - a. If not, what are the main limitations with respect to this aspect?
5. What specific standards and technologies for data cataloguing are you aware of?
6. Do you think enough research data are available to you to effectively carry out your research work?
7. Do you think the data you work with are sufficiently easy to access/download, when available?
8. Are you aware of specific standards and technologies for data access/download?
9. What solutions are you aware of for describing the quality of research data?
10. Who should evaluate the quality of research data?
 - E.g., data producers, librarians/repository managers, publishers, funders, end users, disciplinary association/peer review

APPENDIX 3 – LIST OF WORKSHOP ATTENDEES’ INSTITUTIONS

Representative from the following institutions attended the RECODE WP2 Workshop:

Organisation	Country
Amsterdam University Press	The Netherlands
Blekinge Institute of Technology	Sweden
Centre for Research Communications	Poland
CERN	Switzerland
China University	China
CIMA Foundation	Italy
CODATA (ICSU Committee on Data for Science and Technology)	France
Consiglio Nazionale delle Ricerche	Italy
CSIR	South Africa
EDINA/UK Dataserve	UK
ENCES e.V.	Germany
Joint Research Centre	EC
Fisheries and oceans	Canada
Forschungszentrum Juelich	Germany
GEO	Switzerland
GIS Center, Feng Chia University	Taiwan
GitHub Inc	USA
GSDI	global
Institute of Geographic Sciences and Natural Resources Research of the Chinese Academy of Sciences	China
JRC	Italy
National Documentation Centre (EKT/NHRF)	Greece
NILU	Norway
Open Context / UC Berkeley	USA
OPeNDAP	USA
Royal Netherlands Academy of Arts and Sciences (KNAW)	The Netherlands
Software Mind	Poland
Stichting Liber Foundation	The Netherlands
Swedish National Data Service	Sweden
Terradue	Italy
Trilateral Research & Consulting	UK
University of Geneva	Switzerland
University of Pisa	Italy
University of Sheffield	UK
Xi'an University of Science and Technology	China
Zentral- und Hochschulbibliothek Luzern	Switzerland

APPENDIX 4 – RECODE WP2 WORKSHOP AGENDA

9:00 – Participant registration and coffee

9:30 – Introduction to RECODE project (Kush Wadhwa, Trilateral Research)

9:45 – Outline and purpose of today’s workshop (Lorenzo Bigagli, National Research Council of Italy)

10:00 – The EC perspective and the GEOSS Data Sharing Principles (Michel Schouppe, European Commission)

10:30 – Coffee break

11:00 – Belmont Forum Collaborative Research Action on e-Infrastructure and Data Management (Stefano Nativi, Belmont Forum Steering Committee)

11:30 – Key findings from a survey (questionnaire and literature review) and case studies interviews on the existing technological barriers, solutions and best practice for Open Research Data Access (Lorenzo Bigagli, National Research Council of Italy)

12:00 – Q&A

12:30 – Lunch break

13:30 – Plenary open discussion: Effectiveness, gaps, and practical significance of the existing technical solutions for Open Research Data Access (Jeroen Sondervan, Amsterdam University Press)

14.45 – Coffee break

15:15 – Plenary open discussion: Recommendations on how to increase the effectiveness of the current technological baseline in supporting Open Research Data Access (Bridgette Wessels, University of Sheffield)

16:30 – Next steps: legal and ethical issues in Open Research Data Access: Introduction to RECODE WP3 (Rachel Finn, Trilateral Research)

16:45 – Conclusions (Stefano Nativi, National Research Council of Italy)

17:00 – Finish