



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 D1: Stakeholder Values and Ecosystems

Author(s): Thordis Sveinsdottir, Bridgette Wessels & Rod Smallwood (University of Sheffield); Peter Linde (Blekinge Institute of Technology); Vasso Kala & Victoria Tsoukala (National Documentation Centre, Greece); Jeroen Sondervan (Amsterdam University Press)

Dissemination level: Public

Deliverable type: Final

Version: 1

Submission date: Due 30 September 2013

Table of Contents

List of acroynms	4
Executive Summary	6
1 Introduction.....	12
1.1 Background to Open Access to Data.....	13
1.2 Defining Research Data and Open Access.....	14
1.3 Research Values, Motivations and Barriers: Outlining the Five Case Studies	16
2 Methodology	18
3 The stakeholder taxonomy	21
3.1 Background	21
3.2 The Functional Taxonomy	21
4 Document Review.....	32
4.1 Introduction	32
4.2 Synthesis of High Level Stakeholder Literature: Values, Motivations and Barriers in Open Access to Research Data	33
4.2.1 Scientific Values and the Value of Data	34
4.2.2 Motivations	41
4.2.3 Barriers.....	42
4.3 Stakeholder motivations for open access	45
4.3.1 Libraries and Repositories	45
4.3.2 Funders.....	47
4.3.3 Publishers.....	52
4.3.4 Advocacy Groups, Professional Organisations and CSOs	55
4.3.5 A Study on Subject-Specific Requirements for Open Access Infrastructure – the Case of OpenAIRE	56
4.4 Conclusions and Discussion.....	58
5 Case Study Research: Values, Motivations and Barriers to Open Data – the View from Scientists within Five Scientific Disciplines.....	60
5.1 Particle Physics and Particle Astrophysics: The PPPA Group, The University of Sheffield.....	60
5.1.1 Research practices within Physics	60
5.1.2 Values and motivations.....	60
5.1.3 Barriers to implementing open data.....	61
5.2 Health and Clinical Research: The FP7 Project EVA.....	63
5.2.1 Research practices in Health and Clinical Research.....	63
5.2.2 Values and Motivations	63
5.2.3 Barriers to implementing open data.....	64
5.3 Bioengineering: Auckland Bioengineering Institute and The VPH Community.....	65

5.3.1	Research practices in Bioengineering	65
5.3.2	Values and Motivations	65
5.3.3	Barriers to implementing open data	66
5.4	Environmental Research: JRC, EuroGEOSS and INSPIRE	67
5.4.1	Research practices in Environmental Research	67
5.4.2	Values and Motivations	68
5.4.3	Barriers to implementing open data	69
5.5	Archaeology: Open Context	70
5.5.1	Research practices in archaeology	70
5.5.2	Values and Motivations	70
5.5.3	Barriers to implementing open data	71
5.6	Overall Values, Motivations and Barriers Emerging from the Case Studies	73
5.6.1	Potential Barriers to Implementing Open Data within Science	73
6	Findings from the Validation Workshop	76
7	International Advisory Board Comments	78
8	Discussion	80
8.1	Overview of Research Practices in Making Data Open	82
9	Conclusion	84
10	References	86
	Appendix 1- List of Workshop Attendees' Institutions	94
	Appendix 2 – RECODE WP1 Workshop Agenda	95
	Appendix 3 – Interview protocols	97

LIST OF ACROYNMS

AHRC - Arts and Humanities Research Council
APARSEN – Alliance for Permanent Access
ARL - Association of Research Libraries
AUP - Amsterdam University Press
BBSRC - Biotechnology and Biological Sciences Research Council
CERN - European Organization for Nuclear Research
CESSDA - Council of European Social Science Data Archives
COPD - Chronic Obstructive Pulmonary Disease
CSOs - Civil Society Organisations
DARIAH - DigitalAI Research Infrastructure for the Arts and Humanities
DFG - German Research Foundation
DNFR - The Danish National Research Foundation
DOI - Digital Object Identifiers
DRIVER – Digital Repository Infrastructure Vision for European Research
DRIVER-II - Digital Repository Infrastructure Vision for European Research II
DSA – Data Seal of Approval
EC - European Commission
EDNA - e-Depot Netherlands Archaeology
ERC – European Research Council
ESF - European Science Foundation
ESRC - Economic and Social Research Council, UK
EU - European Union
EU JRC – European Union Joint Research Centre
EUDAT – European Data Infrastructure
EuroCRIS – The European Organisation for International Research Information
EvA - Emphysema versus Airways disease
FNRS - Fonds de la Recherche Scientifique, Belgium
FOIA - Freedom of Information Act
FWF - Austrian Science Fund
GEOSS - Global Earth Observation System of Systems
GIS – Geographic Information Systems
HE – Higher Education
HEI - Higher Education Institutes
ICORDI – International Collaboration on Research Data Infrastructure
ICT - Information and Communication Technology
IFLA - International Federation of Library Associations
IGO – Intergovernmental Organisations
INSPIRE - Infrastructure for Spatial Information in the European Community
IPR - Intellectual Property Rights
IT - Information Technology
JISC – Joint Information Systems Committee
LERU - The League of European Research Universities
LHC - Large Hadron Collider
LIBER - Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries
NEH - National Endowment for the Humanities
NERC - Natural and Environmental Research Council
OAPEN - Open Access Publishing in European Networks

OAR - Open Access Repositories
OECD - Organisation for Economic Co-operation and Development
OpenAIREplus - 2nd Generation of Open Access Infrastructure for Research in Europe
OpenDOAR - The Directory of Open Access Repositories
PARSE – Permanent Access to the Records of Science in Europe
PEER - Publishing and the Ecology of European Research
PDB – Protein Data Bank
PPPA - Particle Physics and Particle Astrophysics
PREPARDE - Peer Review for Publication & Accreditation of Research Data in the Earth Sciences
RCUK - Research Councils United Kingdom
RFO - European Research Funding Organisations
RI - Research Institutes
RPO - Research Performing Organisations
SHERPA - Securing a Hybrid Environment for Research Preservation and Access
SHERPA/JULIET - Securing a Hybrid Environment for Research Preservation and Access (Research funders' open access policies)
SHERPA/ROMEO - Securing a Hybrid Environment for Research Preservation and Access (Publisher copyright policies & self-archiving)
SiS - Science in Society
SOAP - Study of Open Access Publishing
SPARC - Scholarly Publishing and Academic Resources Coalition
SSH - Socioeconomic Sciences and Humanities
SSHRC - Social Sciences and Humanities Research Council of Canada
SURF - collaborative organisation for ICT in Dutch higher education and research
UK – United Kingdom
UKDA - UK Data Archive
USFD – University of Sheffield
VPH - Virtual Physiological Human

EXECUTIVE SUMMARY

This report is the deliverable for Work Package 1 (WP1), Stakeholder Values and Ecosystems, 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. WP1 focuses on understanding stakeholder values and ecosystems in Open Access, dissemination and preservation in the area of scientific and scholarly data (thus not government data). The objectives of this WP are as follows:

- Identify and map the diverse range of stakeholder values in Open Access data and data dissemination and preservation.
- Map stakeholder values on to research ecosystems using case studies from different disciplinary perspectives.
- Conduct a workshop to evaluate and identify good practice in addressing conflicting value chains and stakeholder fragmentation.

In order to reach these objectives WP1 conducted document analysis of policy and related documents and protocols to map the formal expression of values and motivations. Secondly, it conducted five case studies to understand the values, motivations and barriers to Open Access to data in different disciplines. The case studies were as follows:

- **Case study 1** addressed particle physics in relation to the data management issues of large volumes of data.
- **Case study 2** addressed health sciences in relation to the issue of quality control, ethics and data security.
- **Case study 3** addressed bioengineering, specifically complex modelling that may prove difficult to replicate and test in models for heterogeneous datasets.
- **Case study 4** addressed environmental research, in particular addressing multidisciplinary interoperability models for heterogeneous datasets.
- **Case study 5** addressed archaeology, including procedures for evaluating the quality of Open Access data and the technical approaches to preserving diverse types of data.

After the document review and case study analysis, WP1 held a validation and dissemination workshop that sought to better understand how to match policies with stakeholder drivers and motivations to increase their effectiveness in promoting Open Access to research data. The review, case study analysis and validation workshop are aligned with a stakeholder taxonomy that was undertaken in WP6 (Stakeholder Engagement and Mobilisation), which is integrated into this report.

WP 1 uses the following definitions of Open Access to Data and publications in its document review, case study review, and taxonomy development.

- First, it draws on the European Commission definition of “Open Access” as “free ... access to and use of publicly-funded scientific publications and data”¹. This definition holds that research data is an integral part of the Open Access paradigm and it allows for a range of different research processes.

¹ European Commission, Commission Recommendation on access to and preservation of scientific information, C(2012) 4890 final, Brussels, 17 July 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

- Second, the Berlin Declaration is drawn on in relation to its statement that Open Access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material. The remit of the Declaration for Open Access contributions includes two key points: a) authors and rights holders must grant users free access to the materials including a license to copy, use, distribute and display material subject to proper attribution of authorship and responsible use; b) a complete version of the work should be in an appropriate standard format and submitted in an online repository with suitable technical standards that seek to enable Open Access, unrestricted distribution, interoperability, and long-term archiving.
- Third, WP1 draws on the Berlin Declaration’s vision of Open Access, which is that Open Access to data has the potential to create “a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community”². The Berlin Declaration is used to frame the definition of scientific and scholarly research WP1 addresses in its review and case study analysis. WP1 addresses science in its broadest term to cover the physical, biological and engineering sciences, the social sciences, health and medicine, the environmental sciences, and the arts and humanities.

WP1 takes a broad definition of research data, which is: that data is any material used as a foundation for research, and thus data can be published texts, artefacts or raw unprocessed data. This is supported by the Organization of Economic Co-operation and Development (OECD) wide definition of research data, which is that research data is any kind of resource that is useful to researchers.³ The broad approach is also supported by the European Commission (EC), which defines research data as data which “may be numerical/quantitative, descriptive/qualitative or visual, raw or analysed, experimental or observational, examples include digitized primary research data, photographs and images, films, etc”.⁴ Other definitions of data include datasets, which are collections of factual information, and linked data, where data is described by a unique identifier that enables the linking of data. More directly in terms of defining open data, the Royal Society defines open data as data that is accessible, usable, assessable, and able to be evaluated.⁵

To understand the Open Access ecosystem, the RECODE stakeholder taxonomy identifies the key stakeholder groups as: research councils; universities and academies; libraries, archives and repositories; data centres; information aggregators and foundations; research institutes; EU funded projects; policy makers, scholarly societies and national funded projects; inter-governmental organisations; civil society organisations; standards organisations; service providers, professional associations; media; and publishers. The RECODE functional stakeholder taxonomy categorizes these stakeholders (who perform different activities) in the five basic functions in the Open Access ecosystem. These are Funders and Initiators; Creators; Disseminators; Curators, and Users. The functions and stakeholders interact within functions and across functions in an interactive manner. For the

² Max Planck Society, *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003. http://oa.mpg.de/files/2010/04/berlin_declaration.pdf

³ Organization for Economic Cooperation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD, Paris, 2007. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

⁴ European Commission, op. cit., 2012, p.45.

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

Open Access ecosystem to function in a secure and rigorous way the relevant interdependencies between the functions need to be ensured and formalised.

The WP1 meta-synthesis analysis of the review of policy and related documents indicates that there is a general consensus about the benefits of open data access. In terms of the functional taxonomy there is support for developing Open Access to data from the Funders and Creators. Further, key stakeholders such as libraries, repositories, and universities are acting on the overall value consensus and are undertaking various types of preparatory development to support the development of Open Access to research data. The benefits are seen in the value-context that science is of great value to society, and the way society benefits from science is through an on-going dialogue in which knowledge emerges through science as a cumulative process. This view results in an approach that advocates that scientific results should be further scrutinised, re-analysed and tested. These points are seen to be the founding principles of science and it is within these core values that many of the review documents situate open data access. In terms of motivations, we found that the over-arching motivations for implementing open data are derived from the above values. Furthering access to data is seen to be able to deliver faster progress in science, by minimizing duplication of effort, and offering scientists a wider range of data to use for re-analysis, comparison, integration and testing. Further, there is a belief that Open Access will yield economic benefit, including new industries such as ‘re-use industries’. Evidence for this is mainly drawn from government-owned data, which might not directly transfer to scientific data. Further, there is ambiguity about the costs involved in developing and sustaining Open Access to data in relation to any realised benefits.

In terms of values, there are two broad themes regarding value that run through the RECODE stakeholder documents on open data access and it is within these values that references to data and openness are contextualised and justified.

- First, the stakeholder groups in the Funders and Creators categories and in the Curating and Disseminating categories of the taxonomy hold the view that science is valuable for broader societal goals, such as improved medicine and health care. Given this stance they argue that that Open Access to research data will enhance and improve the way science and scientific data can be used in relation to social goals, and thus enhance the value of the contribution science makes to society. The broad focus is that science, knowledge and education is a public good, and that digitally supported Open Access will ensure the maintenance of the rigour of science. There is also a strong notion that open data will deliver ground-breaking research at a faster rate, driving a more expedient scientific discovery, innovation and economic growth.
- Second, the above mentioned stakeholder groups extend the value of science and the role of data into economic value systems. The role of data in economic terms is seen as a type of currency or capital asset, which can be obtained, and re-used. In societal terms, data is valued as a public good in the sense that its production is funded by public money and thus should be accessible to the general public. These themes in the policy analysis point to the perception that institutions and government should seek the highest possible return from data, as a public investment. The European Commission, for example, refers to open data as ‘an engine for innovation, growth

and transparent governance'⁶ indicating that the prospect of openness can bring about various socio-economic benefits.

There are questions that have to be addressed in relation to these high level values and the various assumptions inherent in such high level approaches. These are:

- The links between research data and the practice of science are presented as self-evident and consequently the move towards Open Data Access is automatically aligned with the notion of advancing rigorous science. However, there are detailed research practices and specific types of data that frame the way in which the rigour of science is conducted. Developments in Open Access to Data, therefore, have to be sensitive to the specific processes of various types of scientific practice to ensure that existing research rigour is maintained as well as facilitating Open Access.
- More concerted effort needs to be made on the overall strategic agenda to ensure that Open Access can be implemented in a coherent way. This means ensuring that links between infrastructure, legal and ethical issues, and institutional frameworks are made in the development of Open Access to Data so that the Open Access ecosystem can support an appropriate approach Open Access to all types of data within their research areas.
- Anonymity and privacy of research participants needs to be fully safeguarded, and all Open Data referenced and attributed correctly as part of ethical research practice. However, access to sensitive datasets needs to be carefully managed without putting up a barrier to openness.
- Attention needs to be paid to technological issues, such as the way technology drives the collection of vast datasets, the lack of technical infrastructure to store data and interoperability issues.
- Cultural barriers are significant, especially issues such as competition within science for reward and reputation, the lack of trust between scientists and the lack of career related rewards and prestige resulting from publishing and sharing data.

The case study analysis identifies some important details that need to be addressed so that Open Access to Data does deliver some of the perceived benefits. In particular there needs to be greater clarity in defining key concepts in Open Access and research data in order to further drive implementation of Open Access to Data. The main and overarching recommendation is that there is a need to form a consensus around the definition of data, openness and access.

The assumptions about the science and data relationship in the high level documents and the barriers identified in various stakeholder documents are highlighted in the case study analysis. Throughout the case studies there was a clear echo of the values and motivations expressed in the document review. Overall, researchers within all fields are positive about openness to data within their respective fields but remain sceptical about the practicalities and details of developing Open Access. Research fields vary greatly in terms of how open they are, how they share within the research community and their attitudes towards Open Access. While archaeology is presented as a field with relatively limited sharing of data, bioengineering is very open in terms of sharing models and methods. The scientists we

⁶ European Commission, Open Data, an engine for innovation, growth and transparent governance, COM(2011) 882 final, Brussels, 12 December 2011.
<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF>

interviewed recognise the benefit of having access to increased amounts of data as it has the potential of driving faster advancement of science and discovery within their fields, as well as reduce duplication of effort. It is recognised that the robustness of research outputs is increased by access to larger datasets (health and clinical studies) and by access to data from other experiments (physics). Scientists in larger research collaborations recognise the importance of collaboration and sharing of data and methods when it comes to solving increasingly complex problems (environment studies, health and clinical research and bioengineering) and is a necessity for the funding of large-scale equipment for experiments (Particle Physics and Particle Astrophysics (PPPA)). In some instances collaborations are seen to foster sharing (bioengineering). Because of competition within the field, the collaborations create memorandums around the sharing of data.

Although positive about Open Data, scientists express a variety of concerns deriving from their own personal experiences and knowledge of scientific practices within their respective fields. These are:

- Competition for prestige and funding makes scientists reluctant to publish data openly due to the fear of being scooped. Scientists suggest that a certain time limit on openness needs to be established so that the scientist can to publish his or her findings first.
- The amount of work and the time needed to make data meaningful and useful if made openly available. For instance, the time needed to annotate, create and apply metadata and document context. This extra work would take up time from other research activities such as data collection, analysis, publications and applications for funding, all of which bring clear and demonstrable rewards and benefits to scientists and their careers.
- Publishing data still remains largely unrecognised as a valuable scientific activity. The publication of data does not count in terms of a scientist's career progression and merit. This has consequences when trying to incentivise scientists to publish their data.
- Although many funding bodies request data management plans and insist that data be made openly available, they are not seen to be fully funding the actual activities that are needed for making data available through Open Access. This is especially seen as a detriment to the move towards Open Data in archaeology, as the field is both grappling with vast amounts of un-digitised data and dwindling funds overall. Other fields, such as Particle Physics, also report that in order to fully comply with the move towards Open Data there would need to be an increase in funding for storage and staff costs.
- Scientists who work in interdisciplinary collaborations recognise that different disciplines may have different attitudes to data. This may include ethical considerations with the use of data derived from human subjects. For example, environmental research grapples with combining research data with data from public authorities, which come with a set of legality issues surrounding the use and integration of data.
- Working with sensitive data causes hesitation in implementing Open Access. The types of data that cause concern are: a) data derived from human subjects; b) culturally sensitive data, such as data related to religious belief; b) human data, as used in health and clinical research and bioengineering; c) sensitive location data, as can be the case with archaeology and environment research.

A key issue is that any data that is openly available needs to be meaningful so that it can be used and re-used. Therefore, it is not sufficient to simply make access to data open by putting raw datasets in a repository or posting them online. The full benefits of making research data openly available requires access to the context in which the data is collected and analysed if it is to be meaningful to others. In many instances this might require a vast amount of work for which there is no funding and for which benefits and rewards are currently lacking.

In summary, WP1 clearly found an overall drive for Open Data Access within the policy documents, which is part of a wider drive for open science in general. The values underpinning this move are the view of science as an open enterprise, where knowledge is sought and where discovery rests on scientists working together to solve specific challenges, which increasingly are becoming interdisciplinary in nature. The argument for publicly funded science to be made open to the public is also strong, although it is not often clear how this openness should be operationalised. 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 which might be discrete or interlinked. Some disciplines deal with sensitive data while others deal with data that may have IPR or legal issues. It is important that these differences be acknowledged in further policy for Open Data as it will inform the debate about whether we require subject specific requirements, or common infrastructure for Open Data Access.

1 INTRODUCTION

This report is the deliverable for Work Package 1 (WP1), Stakeholder Values and Ecosystems, 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. WP1 focuses overall on understanding stakeholder values and ecosystems in Open Access and data dissemination and preservation. The objectives of this WP are as follows:

- Identify and map the diverse range of stakeholder values in Open Access and data dissemination and preservation.
- Map stakeholder values to research ecosystems using case studies from different disciplinary perspectives.
- Conduct a workshop to evaluate and identify good practice in addressing conflicting value chains and stakeholder fragmentation.

In order to reach these objectives the RECODE team conducted document analysis of policy, and related documents, publication protocols, data management protocols and ethical protocols to map the formal expression of values and motivations. Secondly, this WP examined five case studies in order to map the scientific ecosystems in different disciplines, including the physical sciences, medicine, bio-sciences, environmental sciences, and the humanities in different national contexts.

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

As part of the case studies the WP team conducted 29 interviews with key personnel to investigate values, motivations and barriers in the different disciplines as well as map their relationships with different stakeholders and organisations.

This report also draws on a consultation and validation workshop which was held at the University of Sheffield on the 4th September 2013. The workshop attracted 38 attendees from policy makers, funding bodies, libraries, data management organisations and HEI researchers, along with participants from the case studies. A complete list of Institutions which were represented at the Workshop can be found in appendix 1⁷. The agenda of the workshop sought to validate and discuss findings from WP1 with relevant stakeholders. The agenda for the workshop can be found in appendix 2 and slides used by the RECODE team are accessible on the project website at: <http://recodeproject.eu/events/recode-workshops/>.

⁷ Due to issues of privacy, a full list of names will not be made public.

1.1 BACKGROUND TO OPEN ACCESS TO DATA

The drive to provide Open Access to research data, especially research data produced as a result of public funding, is often justified by reference to the public interest. Since research is publicly funded, so the argument goes, making it available to stakeholders is seen as a logical and socially responsible step to widening access to research.⁸ The perceived benefits of Open Access include the ability of researchers to subsequently make use of the data for further research, as data re-use prevents costly duplication, and Open Access may permit more data to be brought into complex, interdisciplinary areas of enquiry. Open Access to research data would also enable the validation of research results, for example by assisting reproducibility and ensuring quality control. Policy makers could use the data to inform decision making and the private sector could use it in the development of new products and services. Civil society organisations (CSOs) and citizens would have access to data to better inform themselves about important scientific developments and to participate in public debates.⁹ Such benefits are laudable; however, Open Access raises concerns for the functioning of research ecosystems that serve to underpin the rigour of research.¹⁰

Another aspect of the development of Open Access to Data is the growth of ‘e-research’¹¹, which refers to the use of digital technologies to support new and existing forms of research and research practice. This is fostering a reconsideration of the way scientific and scholarly knowledge is produced and shared.¹² The practice of research rests on open inquiry that traditionally worked through the publication of research results via peer review in which primary data was not always openly shared. However, the development of digital means of producing, storing and manipulating data is creating a focus on ‘data-led science’¹³, which requires that the way data can be shared and made openly available is addressed. In broad terms, Open Access to research data refers to making various types of data openly available to public and private stakeholders, user communities and citizens. This initiative, however, involves more than simply providing easier and wider access to data for potential user groups. The development of Open Access involves a reconsideration of the whole system of the production and dissemination of knowledge, which WPI terms as ‘research ecosystems’.

The move to Open Access and Open Access to Data involves significant changes in research practices and in research ecosystems. Open Access to Data extends across the life cycle of the production of knowledge, including data collection, data analysis, data management, and publication of findings, as well as the legal and ethical frameworks guiding research. Although some developments are shared across research practices, these are adapted within specific disciplines in the physical sciences, social sciences, and humanities. This means that the development of Open Access to Data may vary across research disciplines and within interdisciplinary research collaborations. The practices and values of specific disciplinary research are embedded in wider ecosystems and stakeholder values of knowledge production. Although there is interest in Open Access by policy makers, institutions, and academics, the

⁸ Organization for Economic Cooperation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD, Paris, 2007. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

⁹ The Royal Society, *Science as an open Enterprise*, London, 2012. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

¹⁰ Ibid.

¹¹ Beaulieu, Anne and Paul Wouters, “E-research as intervention”, in Jankowski, N. (ed.) *E-research: Transformations in Scholarly Practice*, Routledge, New York, 2009, pp. 54-69.

¹² Jankowski, Nicholas W. (ed.) *E-Research: Transformations in Scholarly Practice*, Routledge, New York, 2009.

¹³ The Royal Society, op. cit., 2012, p. 7.

range of various research practices across scientific disciplines means that the development of Open Access is fragmented and not fully understood within the research community. Furthermore, several studies of researchers' attitudes point to the fact that although a majority of researchers support free access to research data¹⁴, many lack the tools, standards, incentives and information to make their data publicly available. Furthermore, significant fragmentation, insufficient strategies and funding have also been identified as significant barriers to enhancing access to research data.¹⁵

To summarise, currently the development of Open Access is fragmented and insufficient strategies and funding are major barriers to enhancing coherent Open Access to research data.¹⁶

1.2 DEFINING RESEARCH DATA AND OPEN ACCESS

There are a number of definitions of Open Access, which vary in specificity. The European Commission, for example, defines "Open Access" as "free internet access to and use of publicly-funded scientific publications and data".¹⁷ This definition sees that research data is an integral part of the Open Access paradigm and it allows for a range of different research processes. The Berlin Declaration's vision of Open Access is that it has the potential to create "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community".¹⁸ The Berlin Declaration states that Open Access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material. The Declaration's criteria for Open Access contributions also include two key points. First, authors and rights holders must grant users free access to the materials including a license to copy, use, distribute and display material subject to proper attribution of authorship and responsible use. Second, a complete version of the work should be in an appropriate standard format and submitted in an online repository with suitable technical standards that seeks to enable Open Access, unrestricted distribution, interoperability, and long-term archiving.^{19 20}

Defining research data is similarly problematic, since any material used as a foundation for research can be classified as research data, whether that is published texts, artefacts or raw unprocessed data. The Organization for Economic Co-operation and Development (OECD), for example, uses a wide definition of research data that includes any kind of resource that is

¹⁴ Repositories support project, *Survey of academic attitudes to Open Access and institutional repositories – an RSP and UKCoRR initiative*, 2011. <http://rspproject.files.wordpress.com/2011/12/attitudes-to-oa-basic-summary-report.doc>

¹⁵ Directorate-General Research and Innovation, *Online survey on scientific information in the digital age*, European Commission, Brussels, 2012, p. 28.

¹⁶ Ibid.

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

¹⁸ Max Planck Society, *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003. http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf

¹⁹ Ibid.

²⁰ For another widely-used, similar definition of Open Access see the Budapest Open Access Initiative, "Read the Budapest Open Access Initiative", 2002. <http://www.budapestopenaccessinitiative.org/read>

useful to researchers.²¹ In the most recent survey on information in the digital age, the European Commission (EC) defines research data as data which “may be numerical/quantitative, descriptive/qualitative or visual, raw or analysed, experimental or observational. Examples are digitized primary research data, photographs and images, films, etc.”²² Other definitions of data include datasets, which are collections of factual information, and linked data, where data is described by a unique identifier that enables the linking of data. Open Data refers to data that is accessible, usable, assessable, and able to be evaluated.²³

Despite the above mentioned variability and difficulties of definition, there is a strong policy push to develop Open Access at the national level and within certain world regions (e.g., Europe, North America, Asia-Pacific, etc.). The European Community context provides an exemplar of the issues involved in developing Open Access. There are a number of policies, initiatives, and projects in the European research community that seek to support the development of Open Access to research data, the linkages between research data and publications, and the preservation of scientific data [e.g. FP7 OPEN ACCESS pilot, APARSEN²⁴, DRIVER and DRIVER-II²⁵, DARIAH²⁶, OpenAIREplus²⁷, etc.]. Many of these projects and initiatives seek to address the barriers associated with making data more accessible, for example intellectual property issues, ethical considerations, conflicting stakeholder values, and disciplinary differences. However, each initiative focuses on a specific aspect of Open Access in general terms without necessarily addressing how that aspect links to the wider Open Access ecosystem. One notable exception is the newly issued European Commission Recommendation of July 2012, which integrates Open Access to research data (alongside Open Access to scientific publications, development of e-infrastructures and improved stakeholder collaboration) within a larger Open Access to scientific information fields.²⁸ Given the current fragmentation of the Open Access to research data sector, a comprehensive Open Access policy framework is needed to span various disciplinary, stakeholder and research practices. This requires examining, in the first instance, stakeholder values, motivations and barriers to Open Access. In this deliverable we address one specific stakeholder group, which is the research community. This community is made up a many different disciplines, different multi- and interdisciplinary research collaborations and research that has different rationales and relationships with broader societal concerns.

²¹ Organization for Economic Cooperation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, op. cit., 2007.

²² European Commission, Commission Recommendation on access to and preservation of scientific information, op. cit., 2012, p.45.

²³ The Royal Society, op. cit., 2012, p. 12.

²⁴ Alliance Permanent Access, “About APARSEN”, no date. <http://www.alliancepermanentaccess.org/index.php/aparsen/>

²⁵ Digital Repository Infrastructure Vision for European Research, “DRIVER”, 2013. <http://www.driver-repository.eu/>

²⁶ Digital Research Infrastructure for the Arts and Humanities, “DARIAH-EU”, 2013. <http://www.dariah.eu/>

²⁷ Open Access Infrastructure for Research in Europe, OpenAIRE, no date. <http://www.openaire.eu/>

²⁸ European Commission, Commission Recommendation on access to and preservation of scientific information, op. cit., 2012.

1.3 RESEARCH VALUES, MOTIVATIONS AND BARRIERS: OUTLINING THE FIVE CASE STUDIES

Researchers often have different values, drivers and interests. Furthermore, even within research categories, disciplinary fragmentation impacts upon Open Access and data sharing. These values and motivations are embedded within the disciplinary frameworks of scientific research and in the practice of doing research. Very often the specifics of collecting and managing data within research communities means that barriers emerge in how to share specific data as well as also barriers to making that data openly available. Further, increasingly, research questions demand access to data from different disciplines, yet disciplines differ in their approach to data sharing and reuse. It can be difficult to use those datasets without sufficiently descriptive and understandable metadata. Some of the ways in which research is conducted and managed, and research values and motivations embedded within those practices is seen in our case studies. The cases studies range from Open Access to Data that has already been generated to the collection processes of primary data. They cover a range of data managing issues by addressing human data, clinical data, complex modelling data and theory building across text-based, visual, audio and numerical data.

For example, Particle Physics produces extremely large volumes of data - the Large Hadron Collider (LHC) at CERN produces about 15 petabytes of data per annum. The LHC Computing Grid is the world's largest computing grid, and the Particle Physics and Particle Astrophysics (PPPA) Group at USFD²⁹, led by Neil Spooner, is a member of one of four regional Computing Grid Groups in the UK. This case provides insights into values and motivations that shape the collecting, disseminating, storing and processing large quantities of numerical data from experiments which have hundreds of academic partners. Before recording the raw data, it is pre-processed to reduce the number of events from around 40 million per second to 200 per second. It highlights some of the barriers because even with this reduction, making the data publicly available is questionable - the resources necessary for storing and processing the data are only available to very large consortia. What issues does this raise for the acceptability of the conclusions from the analysis of the data?

Another context is clinical and social science research. The values and motivations shaping the collection and validation of personal data in clinical and social science contexts and its use in research raises barriers in terms of the problems of quality control, ethics, and security, in addition to the possibly more straightforward problems of accuracy and measurement noise (as a result of both individual variability and process variability). The FP7 project, EVA (Project number 200605, Markers for emphysema versus airway disease in COPD³⁰, led by Loems Ziegler-Heitbrock at Helmholtz Zentrum München) provides insight into the use of Standard Operating Procedures and other tools to optimize the quality of the collected data, and to ensure the ethical treatment of personal data.

Yet another area where we see a complex interplay of values, motivations and barriers is bioengineering. In this research field, there is a perception that the data used for developing computational models of human physiology is, in a sense, fragile, and that the outputs of computational models of extremely complex systems may not be repeatable in the manner that is expected for acceptance in the current scientific paradigm. There are many levels at which these issues can be raised: how is the initial reduction in complexity (which is essential

²⁹ University of Sheffield, "Research in Particle Physics and Particle Astrophysics", no date. <http://www.hep.shef.ac.uk/research/>

³⁰ Emphysema vs Airways Disease: The EvA Project, "Welcome to EvA", 2008. <http://www.eva-copd.eu>

in order to make the problem computationally tractable) validated; how can a complex model be described in a manner which enables reproducibility; what is the origin of the biological data used to build the model; is reproducibility of results an impossible condition to meet when the results may be the end product of tens of person-years work?

Work in the environmental sciences provides other insights into the values and motivations in making environmental data open but it also highlights the barriers to making that data open. The mission of the Digital Earth and Reference Data Unit, at the DG Joint Research Centre in Italy, is to address sustainability and competitiveness challenges by developing and promoting wide access to the reference data and systems needed for robust policy making. The Centre coordinates the scientific and technical development of the European Directive INSPIRE (Infrastructure for Spatial Information in the European Community), supports its implementation, and leads research towards the next generation environmental information infrastructures at European and Global level. INSPIRE aims to deliver integrated spatial information services to its users. These services should allow the users to identify and access spatial information from a wide range of sources, from the local to the global level, in an inter-operable way and for a variety of uses. The target users of INSPIRE include policy-makers at European, national and local level and the citizen. What are the issues regarding technical and multidisciplinary interoperability in Open Data access, what challenges are there in sharing, and providing Open Access to research data from a variety of sources, and in a variety of formats?

In the Open Context project we see some of the values and motivations for making a free, Open Access resource for the electronic publication of diverse types of research datasets from archaeology and related disciplines. It also highlights how it deals with some of the barriers to Open Access, such as the way it enforces editorial control through its editorial board, utilizes open licensing frameworks and focuses on data portability. Open Context is maintained and administered by the Alexandria Archive Institute³¹, a not-for-profit organisation³², based in Berkeley, California, while IT development is carried out in collaboration with the Berkeley School of Information. Open Context provides information regarding researcher values and motivations in making their data open, as well as the barriers, within the ecosystem of archaeology.

These case studies provide a research practice and framework base that provides us with a way to understand the values, motivations and barriers to make data open from the point of view of the research community, which is valuable for thinking ahead and towards a common European policy for implementing Open Access to research data.

³¹ The Alexandria Archive Institute, “The Alexandria Archive Institute”, no date. <http://www.alexandriaarchive.org/>

³² Open Context is financially supported by The William and Flora Hewlett Foundation, The National Endowment for the Humanities and The Institute of Museum and Library Services.

2 METHODOLOGY

The methodology in WP1 consists of document review, case study research and a stakeholder validation workshop. The purpose of these three approaches in WP1 methodology is twofold. First, it seeks to support the development of a coherent approach to Open Access to Data by the research community. Second, it seeks to ensure that policy recommendations are informed by research practice. The development of Open Access to research data is complex and the research environment is diverse, which requires an approach that can address both the emergent character of Open Access and the specific research contexts and practices in which it is to be implemented. We therefore undertook a meta-synthesis review of the literature, case study research and held a stakeholder validation workshop.

The diversity of research practice, the complexity of research and data ecosystems and the early stage of developing Open Access to research data requires a review process that can draw on a range of sources. To this end the project undertook a meta-synthesis literature review. This is a non-statistical technique that integrates, evaluates and interprets findings from multiple sources (predominantly qualitative studies) and identifies core elements and themes. The review process in RECODE addressed the different domains of Open Access to research data from across the research and data ecosystem in a range of policy guidelines, policy reports and grey literature. To undertake the review the RECODE project drew on its experts in the WP1 project team. Each expert addressed two areas from the following research domains: research disciplines of social sciences, humanities, sciences, and health and medicine including ethics; policy makers; libraries and repositories; and publishing. The first phase of the review involved the experts undertaking a broad scoping study of each area to identify sources and main areas of documentation. The second phase involved taking a sample of those documents for deeper analysis. The sample was based on the stakeholder taxonomy created in WP6 (Stakeholder Engagement and Mobilisation). The review of the final selection of documents focused on values, motivations and barriers in the development of Open Access to Data. The review process produced an outline scoping of the main literature and an analysis of the main themes found in the literature (see sections below).

The review of documents addresses high level motivations and values in Open Access to Data. In order to address the way these high level statements might be operationalised, WP1 conducted case studies in five areas of research practice. The aim of case study research is to yield close and in-depth understanding of a case or small number of cases in their real world context.³³ The in-depth focus on a small number of cases enables research to gain insights into the context and complex conditions that make up particular case study contexts and data is drawn from multiple sources of evidence to address that complexity. The methodology is based on an interpretive approach that seeks to understand the meaning and practice of different research fields in order to understand how Open Access to research data is being discussed and appraised within different research communities. The main foci are the values and motivations for taking up Open Access to research data and the barriers in the development of Open Access in specific areas of research.

The project selected five embedded case studies from within the broad case study of Open Access to Data (these were briefly described above) that cover a range of research disciplines that raise important issues about data - types of data, and their collection, management,

³³ Bromley, Dennis B., *The case-study method in psychology and related disciplines*, John Wiley & Sons, Chichester, 1986 and Lin, Thomas, "Cracking Open the Scientific Process", *The New York Times*, 16 January 2012, p. D1.

interpretation, curatorship, and the sharing and dissemination of research results. The case studies as a cohort show how particular disciplinary knowledge and research practice is important in creating data that is meaningful to specific knowledge communities. The case studies also address the four points that the Royal Society³⁴ notes as important in developing Open Access to Data: each study explores the complex relationships involved in making data accessible; intelligible; assessable and usable. 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. Recruitment of case studies was done at the proposal stage in which Directors of projects and Research Units agreed to take part. The recruitment of case study participants was purposive with the Director and core team in each case suggesting participants for interview. The sample in each case consisted of five research scientists who were at different levels of seniority and hence with different ranges of roles and responsibilities. The levels were: Director; senior researcher; junior researcher; and technical support staff. We used semi-structured interviews in the case studies so that we could explore the values, motivations and barriers to Open Access to Data in each embedded case study. The interview topic guide enabled the research participants to discuss the key themes identified in the review process and it also allowed each participant scope to discuss the issues that were important from his or her perspective and experience (see Appendix 3 for interview guidelines). In total we conducted 29 interviews across all the cases and between 5-7 interviews in each case. All interviews were fully transcribed and we undertook a thematic framework analysis of the data.

The third aspect of WP1 is a validation workshop. The purpose of this workshop was to gain stakeholder feedback about the findings from the literature review and the case studies (see Appendices 1 and 2). The workshop addressed the perspectives in understanding Open Access to research data in relation to stakeholder values and motivations in research ecosystems. The aim of the workshop was as follows:

- Present key findings from a review of literature regarding Open Access to Data and from the five case studies of research practice.
- Solicit participants' and case study partners' feedback on the effectiveness of these policies as well as identify any gaps to assess their significance for policy development and for research practice.
- Better understand how to match policies with stakeholder drivers and motivations to increase their effectiveness in promoting Open Access to research data.

The workshop was attended by 38 people from the following stakeholder groups: research funders; university researchers; libraries; repositories; data archives; and publishers. The first panel discussion explored the values of scientific practice and the motivations and barriers that shaped the way researchers may or may not support Open Access to Data. The panel involved case representatives from three of the case studies, namely: physics, environmental science; and archaeology. The overall conclusion of the workshop was that the stakeholders validated many of the WP1 findings and that the case study approach was useful in identifying many of the values, motivations as well as barriers to Open Data access. The

³⁴ The Royal Society, op. cit., 2012.

workshop participants noted that we had addressed high level aspects of Open Data access in terms of the document reviews as well as practice-based issues from research practice. The middle ground, meso-level needs further exploration, which will be covered in WP2 Infrastructure and Technology, WP3 Legal and Ethical Issues in Open Access and Data Dissemination and WP4 Institutional Evaluation and Support for Open Access Data. A detailed overview of the workshop discussion and findings is in the sections ‘Case Study Research: Values, Motivations and Barriers to Open Data – the View from Scientists within Five Scientific Disciplines’ and ‘Findings from Validation Workshop’.

3 THE STAKEHOLDER TAXONOMY

3.1 BACKGROUND

Stakeholders in the open data access ecosystem are a diverse group of players from government, industry, the public and mass media. Due to the financing structure of science the primary stakeholders are affiliated with government organisations. One of the tasks in the RECODE project was to identify the stakeholders of the Open Access ecosystem with special focus on stakeholders of open research data. Therefore we have created a functional stakeholder taxonomy in which stakeholders have been categorised. Alongside this we have tried to map the values and interests associated with each group.

During the first months of WP1 we listed possible stakeholders and undertook a review of policy documents on data management and ethics. It soon became clear that there is an enormous amount of material on Open Data handling. There are also a large of stakeholders involved. In order to create a workable model of the Open Data stakeholder ecosystem we saw the need to narrow the stakeholder list down to a few broad functions with a limited number of stakeholders. This resulted in a functional stakeholder taxonomy which we believe gives a balanced picture of the stakeholder ecosystem for Open Data.

3.2 THE FUNCTIONAL TAXONOMY

Our model of the Open Access stakeholder ecosystem is a functional taxonomy that consists of five entities or functions with performers interconnected through flows. It can be visualised as a layered cake where stakeholders operate and interact in different layers at the same time. Creators can also be Users and Disseminators. Curators can be Disseminators and Users and so on. The flows of the ecosystem go in many directions and can involve the same performers in more than one function. In order to focus on the relevant stakeholder with the relevant function we have constructed the taxonomy so that stakeholders can have several secondary functions (SF) but only one primary function (PF). A primary function performer is a performer with an essential importance to the function. Secondary stakeholders are performers not essential to the function.

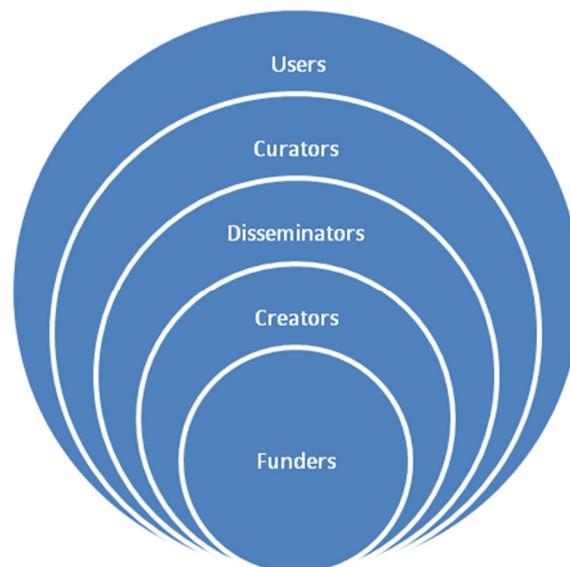


Figure 1: The Five Stakeholder Functions

We have identified five basic functions in the Open Access ecosystem: 1) Funders & Initiators, 2) Creators, 3) Disseminators, 4) Curators and 5) Users. These functions are represented by different performers (stakeholders). Each performer undertakes activities and produces records in relation to Open Access Data. We are aware of different stakeholder affiliations (government, public, industry and government institutions etc.) but these are not included in the taxonomy in the interests of clarity.

Below we outline the functions, performers, activities and records of the functional taxonomy. Each performer is listed as having either a primary function (PF) or a secondary function (SF)

A. Funding & Initiating

Research Councils (PF)

Activity: Distributing funds

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Foundations (PF)

Activity: Distributing funds

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Policy Makers (PF)

Activity: Influencing decisions, Initiating processes

Records: a) Propositions

b) Recommendations

c) Open Access policy

Advocacy Groups (PF)

Activity: Influencing decisions, Initiating processes

Records: a) Recommendations

b) Open Access Policy

CSOs (SF)

Activity: Influencing decisions, Initiating processes

Records: a) Recommendations

b) Open Access Policy

B. Creating

Universities/Academy (PF)

Activity: Produce research publications and data

Records: a) Open Access policy

b) Ethical protocols

c) Research Management protocols

Research Institutes (PF)

Activity: Produce research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Scholarly Societies (PF)

Activity: Produce research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

IGOs (PF)

Activity: Produce research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Standards Organisations (PF)

Activity: Produce standards documents

Records: a) Standards

b) Protocols

Service Providers (PF)

Activity: Produce infrastructure services

Records: a) Standards

b) Protocols

Data Centres (SF)

Activity: Produce research data

Records: a) Research Management protocols

b) Open Access policy

EU Funded Projects (SF)

Activity: Testing ideas, infrastructures etc.

Records: Recommendations

National Funded Projects (SF)

Activity: Testing ideas, infrastructures etc.

Records: Recommendations

Information Aggregators (SF)

Activity: Produce data retrieval services

Libraries/Archives (SF)

Activity: Procure and distribute research publications

Records: Manuals

C. Disseminating

Data Centres (PF)

Activity: Disseminate, procure and preserve research data

Records: a) Research Management protocols

b) Open Access policy

EU Funded Projects (PF)

Activity: Testing ideas, infrastructures etc.

Records: Recommendations

National Funded Projects (PF)

Activity: Testing ideas, infrastructures etc.

Records: Recommendations

CSOs (PF)

Activity: Disseminate research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Publishers (PF)

Activity: Offers publication, recognition and distribution platforms

Records: a) Open Access policy

b) Rights agreement

Professional Associations (PF)

Activity: Influences researcher behaviour

Records: a) Ethical protocols

b) Open Access policy

Libraries/Archives (SF)

Activity: Disseminates research publications and data

Records: a) Manuals

Scholarly Societies (SF)

Activity: Disseminates research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

IGOs (SF)

Activity: Disseminates research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Universities/Academy (SF)

Activity: Disseminates research publications and data

Records: a) Open Access policy

Research Institutes (SF)

Activity: Disseminate research publications and data

Records: a) Ethical protocols
b) Research Management protocols
c) Open Access policy

Information Aggregators (SF)

Activity: Disseminate research publications and data

Media (SF)

Activity: Mass dissemination and marketing of research

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

E. Using

Information Aggregators (PF)

Activity: Commercial possibilities

Media (PF)

Activity: a) Creates news and features,
b) Enriching possibilities for social interaction

Universities/Academy (SF)

Activity: a) Reuse research publications and data
b) Commercial possibilities

Records: a) Open Access policy
b) Ethical protocols

c) Research Management protocols

Research Institutes (SF)

Activity: Reuse research publications and data,

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Scholarly Societies (SF)

Activity: Reuse research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Data Centres (SF)

Activity: New access to research data

Records: a) Research Management protocols

b) Open Access policy

IGOs (SF)

Activity: a) Reuse research publications and data

b) Commercial possibilities

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

CSOs (SF)

Activity: Reuse research publications and data

Records: a) Ethical protocols

b) Research Management protocols

c) Open Access policy

Standards Organisations (SF)

Activity: Validate standards usability, commercial possibilities

Records: a) Standards

b) Protocols

Service Providers (SF)

Activity: Commercial possibilities

Records: a) Standards

b) Protocols

Publishers (SF)

Activity: Reuse publications and data, commercial possibilities

Records: a) Open Access policy

b) Rights agreement

Libraries/Archives (SF)

Activity: Procure and distribute research publications

Records: Manuals

To provide context for the above taxonomy a short definition of the stakeholders and their respective position in the open data ecosystem is now given.

Research Councils (PF=Funding and Initiating), which include, for example, national research funding councils like Research Councils UK, Swedish Research Council together with the European Association for Research Councils. Other examples are Science Europe and all major government organized funding agencies like JISC in the UK and SURF in the Netherlands. Research councils and Foundations, no matter if they are organized by governments or publicly sponsored, are prioritised stakeholders since they fund research and can set up mandates and policies for how their money is used.

Foundations (PF=Funding and Initiating) are public charitable funding initiatives such as the Wellcome Trust, Research Foundation Flanders, and the Leverhulme Trust and so on.

Policy makers (PF=Funding and Initiating) are those policy makers within the EU and national government authorities who produce important Open Access policy documents. EU and governments in the UK, Australia, the Netherlands and other countries are at the forefront promoting the issue of Open Data and therefore are continually revising policies in the area.

Advocacy groups (PF=Funding and Initiating) are organized interest or pressure groups in the Open Access arena, such as Research Data Alliance, The World Confederation of Open Access Repositories, iCORDI, SPARC etc. These groups engage members in different working groups to propel Open Access forward by proposing standards, infrastructure changes, organizing conferences and so on.

Civil Society Organisations (CSO) (PF=Disseminating; SF= Funding and Initiating, Using) cover organisations independent of government and industry that have a stake in the establishment of Open Access. One example of a CSO is Open Society Institute. CSOs concerned with Open Access and Open Data are an important lobby factor especially in organizing professionals and scientists on a broader scale for causes outside professional associations.

Professional Associations (PF=Dissemination) are organisations that gather members from a particular profession seeking to further the profession. Examples are British Sociological Association and Association of Internet Researchers, and so on. Professional associations are advocates of their profession and they are a strong lobbying group with great insights into the needs and motivations and ethical problems of different areas of scientific research. In the UK, professional associations are usually charities, and therefore have a legal obligation to further the public good, not the interests of their members³⁵.

Universities/Academy (PF= Creating; SF= Disseminating, Curating, Using) covers universities and higher education organisations including researchers and other staff engaged in primary or secondary research. Universities and University library repositories have been in the vanguard advocating and implementing Open Access policies and disseminating research documents in a more or less standard way using the Green Road of Open Access.

³⁵ The Charity Commission, “The regulator for charities in England and Wales”, no date. <http://www.charitycommission.gov.uk/>

Research Institutes (PF= Creating; SF= Disseminating, Using) can file under government or industry and produce research in national or international constellations. Examples are CERN, UK Biobank, and the Max-Planck Institute. Research Institutes represents the creator role of research data stakeholders but also the user role. Research institutes represent different research areas with different needs, different data and different research cultures and doing so gives a broad picture of needs and motives.

Scholarly Societies (PF= Creating; SF= Disseminating, Using) or learned societies are usually formed to promote certain academic disciplines. They are organized by scholars usually via a university, and create and distribute scientific results. Examples are Massachusetts Medical Society, Leopoldina German Academy of Science, and The International Federation of Library Associations. Included in this category are National Academies such as the Royal Society.

Data Centers (PF= Disseminating; SF= Creating, Curating, Using) are key players in the development of Open Access to Data. They are one of the stakeholders that have implemented and tried out infrastructures and created solutions that work for distributing Open Data. They are organised both nationally and internationally. They are mainly government financed operations for making datasets available like the Swedish National Data Service or Gesis German Social Science Infrastructure Services, both are organized under the Council of European Social Science Data Archives (CESSDA) working to improve access to data for researchers and students.

Inter-Governmental Organisations (IGO) (PF= Creating; SF= Disseminating, Using). An IGO is an entity created by treaty, involving two or more nations, working in good faith, on issues of common interest. Organisations of this kind in the Open Access ecology are OECD, EUDAT, The World Bank, UNESCO. By reason of their international capacity and the strong involvement of government, IGOs are important stakeholders.

EU funded Projects (PF= Disseminating; SF=, Creating) that address areas in Open Access such as Open Access Infrastructure for Research in Europe (OpenAIRE), Digital Research Infrastructure for the Arts and Humanities (DARIAH), Publishing and the Ecology of European Research (PEER), Study of Open Access Publishing (SOAP) and others are included as a stakeholder group because they inform policy decisions in the field of Open Access.

National funded Projects (PF= Disseminating; SF=, Creating), which are created and run on a national basis like CLARIN-NL in the Netherlands, OpenAccess.se run by the National library in Sweden and OAPEN-UK make major contributions to the development of Open Access in terms of knowledge, practice and tools.

Standards organisations (PF= Creating; SF= Using) in Open Access are WC3, Open Geospatial Consortium, International Organisation for Standardization. In order for Open Data to have a global impact it is necessary to take into account opinions and implementations by standards organisations.

Service Providers (PF= Creating; SF=Using) are usually companies that provide necessary services like telecommunications, software applications, storage providers for data and computers etc. To distribute big datasets it is important to implement an infrastructure that supports Open Access.

Information Aggregators (PF= Using; SF= Creating, Disseminating) are companies that use mined or otherwise collected data to create information services - data services, news services, search services. Companies like Google, Springer and ProQuest are examples of this. These stakeholders are an important channel for the non-scientific as well as for the science user of an Open Data system.

Libraries/Archives (PF= Curating; SF= Creating, Disseminating, Using) refers to university libraries, special libraries and national libraries and national and local archival institutions that maintain collections of content. Many libraries already have infrastructures adapted to disseminate Open Access publications and are starting to also disseminate Open Data.

Publishers (PF= Disseminating; SF= Curating, Using) are adopting to Open Access as a business model. We include all kinds of publishers, commercial, society, university etc., of research publications and datasets. The publishing industry has not taken a very strong position concerning Open Data. They have been quite slow to embrace Open Access to journal papers etc. and the journal based dissemination structure of science give publishers a strong position as stakeholders. The traditional publishing culture, and the influence this has on academics career structures are important reasons to account for publishers' values and drivers when it comes to future Open Data policies.

Media (PF= Using; SF= Disseminating), which includes traditional television, radio and newspapers together with new digital and social media and video producers. These stakeholders are big potential users of government and research data but are also producer of new and refined data.

Below is a figure which illustrates how the different stakeholders operate and connect within the Open Data ecosystem.

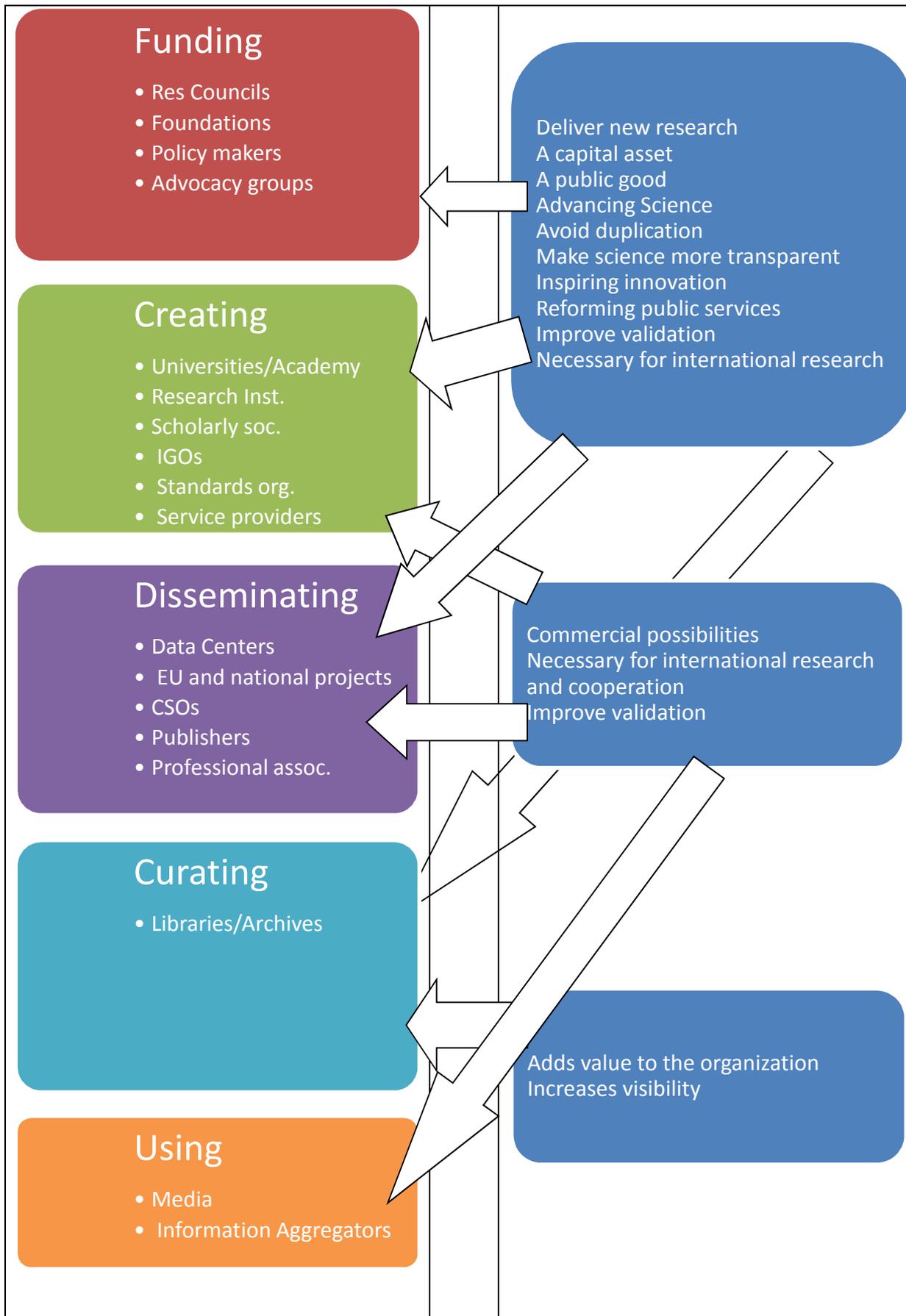


Figure 2: Stakeholder Values and Function

The functional taxonomy of Open Access stakeholders was constructed in parallel to the review work and mapping of stakeholder values in WP1. The construction work of the taxonomy and the wide search for data management protocols, Open Access policies, and ethical protocols have fertilised and re-fertilised both tasks to ensure that we cover all aspects of the Open Access ecosystem.

4 DOCUMENT REVIEW

4.1 INTRODUCTION

The objective of this task is to identify and map the diverse range of stakeholder values in Open Access and data dissemination and preservation policies. The RECODE partners conducted documentary analysis of national, European and international policy, and related documents, publication protocols, data management protocols and ethical protocols to map the formal expression of values within the stakeholder groups identified in the stakeholder review.

This was followed by a meta-synthesis analysis of policy material from the identified stakeholders. The team relied on the members' expertise within these specific fields. Due to the vastness of the field in question and the great quantity of material on Open Access and Open Data Access the team firstly undertook a broad scoping of material before applying a sampling matrix to further narrow the material. A thematic analysis was then undertaken on the sample in question before identifying high level horizontal themes and any cases of best practice.

What the team found was that, overall, there seems to be a general consensus around the benefits of Open Data Access, and a drive towards participation at both high level (EC and national science policy) and mid-level (libraries, repositories, funding bodies, publishers, advocacy groups, professional organisations and CSOs)

The two broad themes emerging from analysis were (these will be discussed in more detail below):

Values: Science is presented as of great value to society. It is seen to be based on an on-going dialogue, and knowledge emerging from research is seen as cumulative. Scientific results should be further scrutinised, re-analysed and tested; these are seen to be the founding principles of science and it is within these core values that much of the stakeholder literature situates Open Data Access. Throughout the high level policy literature there is also a clear message that openness in science will deliver economic benefit.

Motivations: What we found to be the over-arching motivations for implementing Open Data was derived from the above values. Furthering access to data is seen to be able to deliver faster progress in science, by minimising duplication of effort, and offering scientists a wider range of data to use for re-analysis, comparison, integration and testing. Furthermore, the promise of social and economic benefit and furthering public access to science is a strong argument throughout.

In many instances we found that these are aligned with a set of very broad definitions that encompass the complexity of the field but they are so broad that they do not sufficiently address and guide the operationalisation of open data access.

Whilst the document review focuses primarily on the stakeholder values and motivations of users, curators, disseminators, and funders, the case studies explore values, motivations and barriers at the grounded practice level of creators (i.e. researchers, research groups). These values are further explored using case studies of five disciplines: Bio-engineering, Physics, Archaeology, Health and Environmental research. The aim was to explore if and how the

stakeholder values, motivations and barriers identified in the document review map on to values, motivations and barriers as expressed by practicing researchers.

4.2 SYNTHESIS OF HIGH LEVEL STAKEHOLDER LITERATURE: VALUES, MOTIVATIONS AND BARRIERS IN OPEN ACCESS TO RESEARCH DATA

‘The fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge.’³⁶

Through the process of reviewing policy material, and case study research, the team found that definitions of data and Open Access can vary between disciplines, institutions and policy documents. Many of the policy documents acknowledge the complexity of definitions and set to set up glossaries and lists of definitions. Three examples will be outlined here, each from a different stakeholder type.

The OECD³⁷ defines research data as ‘factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.’

The Royal Society Report, Science as an Open Enterprise³⁸, distinguishes between ‘raw’ data collected/generated by the researcher and ‘derived’ data, which indicates that the data has been processed in some way. In addition there is metadata, which is defined as ‘data about data’, which contains information about how the data was generated, and/or a description regarding its structure, licensing terms and standards.

More specifically, the University of Edinburgh, Information Services divides metadata into three broad categories:

- **‘Descriptive** - common fields such as title, author, abstract, keywords which help users to discover online sources through searching and browsing.
- **Administrative** - preservation, rights management, and technical metadata about formats.
- **Structural** - how different components of a set of associated data relate to one another, such as tables in a database.’³⁹

The division between raw and generated data is not always clear-cut and the Royal Society Report discusses this especially with reference to simulation data, derived from simulation experiments performed on virtual representations of the real world where it is not feasible to experiment with the real system due to reasons such as cost and inherent limitations, e.g. in

³⁶ European Commission, *Riding the wave: How Europe can gain from the rising tide of scientific data*, final report of the High level Expert Group on Scientific Data, Brussels, October 2010, p. 4. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

³⁷ Organization for Economic Cooperation and Development, *Report on Cross-Border Enforcement of Privacy Laws*, OECD, Paris, 2007, p.13 www.oecd.org/dataoecd/17/43/37558845.pdf

³⁸ The Royal Society, op. cit., 2012.

³⁹ University of Edinburgh. “Research data management guidance”. 20 September 2013. <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/documenting-data>

environmental research climate change needs to be simulated. In these instances, making the data open and available whilst meeting obligations of accessibility, intelligibility, assessability and usability, the simulator would need to be included along with a full description of how it was used, the values of the parameters, the starting conditions, the software platforms that it ran on etc. In order to make a generic statement, a full description of the formal system that generated the data would be needed. It is therefore clear that in some instances it is not only data that will need to be opened up for access but also models, equipment, and systems, which may further complicate implementation of Open Data policy.

Open Access is defined by the Royal Society⁴⁰ as data that is ‘accessible, intelligible, assessable and usable’. The OECD further defines openness as ‘access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open Access to research data from public funding should be easy, timely, user friendly and preferably Internet based.’ Furthermore, Open Access should respect legal rights and adhere to research ethics regarding personal information, privacy and confidentiality.⁴¹

It is clear that different types of stakeholders will have different perceptions and definitions of data and openness according to what their role in the research process and data journey (from collection to publishing) they fulfil. The higher level stakeholders (Initiators and Funders) thus present an overall broad policy, with wider definitions, which the medium level stakeholders (Users, Disseminators, Creators and Curators) appropriate further to fit their practices. To further the drive for implementation of Open Access to research data, these definitions may need further work around forming a consensus on the definition of data, openness and access. The case study further supports this finding, as researchers in many instances felt that a further clarification was needed so that practical issues around data sharing and preparation could be tackled.

4.2.1 Scientific Values and the Value of Data

There are two broad themes regarding value that run through stakeholder literature on Open Data Access and it is within these values that references to data and openness are contextualised and justified.

Firstly, we refer to the values held and presented by stakeholder groups on both high and medium levels, regarding science as an undertaking and how Open Access to research data is seen to further strengthen these underpinning values. These broadly outline the notion of doing good science, science, knowledge and education as a public good, and maintenance of the rigour of science through current technological and cultural changes. There is also a strong notion that Open Data will deliver ground-breaking research at a faster rate, driving a more expedient scientific discovery, innovation and economic growth.⁴²

Secondly we refer to the value of science as noted by the above groups, thus meaning the societal and economic value of scientific practice and the role of data as currency/capital

⁴⁰ The Royal Society, op. cit., 2012.

⁴¹ OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, op. cit., 2007, p.15.

⁴² European Commission, “Digital Agenda for Europe: Open Data”, 2013. <https://ec.europa.eu/digital-agenda/node/70>

asset, which can be obtained, and re-used; a public good in the sense that its production is funded by public money and thus should be accessible to the general public.

Scientific Values

*‘The Internet has fundamentally changed the world of science and research. (...) Research and funding policies need to adapt to this new environment’*⁴³

*‘There are huge benefits to society in making the outputs of publicly funded research publicly available and thus facilitating the free exchange of knowledge’.*⁴⁴

There is broad acknowledgement and agreement, in guidelines and recommendations for Open Data Access, that science is changing, and rapid development of information and communication technologies is frequently mentioned to be the key driver for this change. As much data is now transformed in to digital form, storage, sharing, amending is made easier. As bandwidth increases and processors become more powerful, this data is easily shared over vast distances, and between various groups of researchers.

Throughout, the links between research data and the practice of science are presented as self-evident and consequently, the move towards Open Data Access is automatically aligned with the notion of advancing rigorous science. Greater sharing and access to data is seen to strengthen the already underpinning values of science allowing ‘effective self-correction of research’ and for ‘secondary analysts (to) verify, refute, or refine original results’ (CESSDA).⁴⁵ Furthermore, policies on Open Access are seen to improve conditions for researchers by reducing duplication of effort and minimising the effort spent searching for information and data.

An example of this discourse can be seen clearly in the report produced by JISC, which aimed to identify benefits arising from curation and open sharing of research data, produced by UK Higher Education and Research Institutes⁴⁶. The benefits were listed as follows:

- Wider access to data would enable greater cross sectorial collaboration as well as considerable opportunities within education and training as a consequence of access to researchers in industry, government and non-government organisations. This would allow for use and re-use of data, including reduced cost of collection and duplication, sharing the direct and indirect costs of collection (e.g. avoiding survey fatigue and thereby improving response rates), new uses unforeseen at the time of collection and data mining opportunities.
- Easier detection of fraud and plagiarism, easier assessment and peer review at the grant application and assessment, publication and research evaluation stages.

⁴³ European Commission, Commission Recommendation on access to and preservation of scientific information, op. cit., 2012.

⁴⁴ JISC. “Open Access for UK Research - JISC’s contributions”, 2 March 2012.

<http://www.jisc.ac.uk/publications/programmerelated/2010/openaccessmainbrochure.aspx>

⁴⁵ Council of European Social Science Data Archives, “Depositing Data – Benefits”, 2012. <http://www.CESSDA.org/sharing/depositing/1/>

⁴⁶ Fry, Jenny, Suzanne Lockyer, Charles Oppenheim, John Houghton and Bruce Rasmussen, *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes*, JISC, London, 2008. http://repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf

Consequently there will be an opportunity to create a more complete and transparent record of science.

- There will be increased opportunity to raise the visibility of researchers, repository host institutions and funders, by linking them to valued resources.

There is a high level understanding across stakeholder groups that the move towards Open Data is beneficial and thus should be put on the strategic agenda. Science is presented as a cumulative knowledge process, and a dialogue in which data plays an important role in improving earlier work. Open Access to Data is seen to significantly speed up this process through avoiding duplication of effort, fostering of collaboration and overall making science a more transparent process with the aim of driving public participation, inspiring innovation and reforming public services.⁴⁷

The European Organisation for International Research Information (EuroCRIS) declared in 2011 in their Rome Declaration⁴⁸ that there was a broad agreement on the need to coordinate development in Current Research Information Systems and Open Access Repositories (OAR) The declaration states that ‘high quality research information is critically important to research institutions, research funders, policymakers and society at large and that information on publicly-funded research should be available, shareable and integrated seamlessly.’⁴⁹

Science is presented as a universal endeavour and Open Access as a leveller which ensures a more equitable access to knowledge across an uneven world. Open Access to data is thus a great diffuser of knowledge which should create a better informed public, which can now be allowed a greater share in any scientific advancement.

Open data access is seen to hold an important role in maintaining the rigour of science as it allows for the re-use, testing, merging, and re-analysis of data. It is seen to open up data to scrutiny which is seen to lead to rigorous scientific practice.

It is interesting to note that several policy documents⁵⁰ take the Human Genome project as an example of successful data sharing and collaboration that has caused science to progress with considerable speed within this specific field. One these policy makers, The European Commission states that:

‘It is estimated that government investments of \$3.8 billion in the Human Genome Project, a US co-ordinated research endeavour including major European contributions, have had an economic impact worth \$796 billion, created 310 000 jobs and launched the genome revolution. This is an excellent illustration of the power that Open Access to scientific information can have.’⁵¹

⁴⁷ United Nations Educational, Scientific and Cultural Organisation, *Policy Guidelines for the Development and Promotion of Open Access*, Paris, 2012. <http://unesdoc.unesco.org/images/0021/002158/215863e.pdf>

⁴⁸ The European Organisation for International Research Information, *Rome Declaration on CRIS and OAR*, 2011. <http://www.eurocris.org/Documents/RomeDeclaration.pdf>

⁴⁹ Ibid

⁵⁰ E.g. United Nations Educational, Scientific and Cultural Organisation, op. cit., 2012, OECD, *Principles and Guidelines for Access to Research Data from Public Funding*, op. cit., 2007.

⁵¹ European Commission, Towards better access to scientific information: Boosting the benefits of public investments in research, Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions, COM(2012) 401 final, Brussels, 17 July 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

This indicates that the push for Open Access may be greatly influenced by this progress and there may be a lack of understanding of the diversity within the field of science, and especially what types of data they may collect, and how amenable this data is for making available for Open Access. It is evident that the Genome Project is seen as an example of the success that Open Data could bring to other disciplines and this example is used to drive the discussion around openness in science. There are, however, important questions about generalisability, and whether this success can be replicated within different fields of science remains an unanswered question.

Ethics

‘Legal and policy issues: national laws and international agreements, particularly in areas such as intellectual property rights and the protection of privacy, directly affect data access and sharing practices, and must be fully taken into account in the design of data access arrangements.’⁵²

Many policy documents reviewed touched briefly on ethics in relation to Open Access and the issues raised, which fall into two key themes. Firstly, that anonymity and privacy of research participants would be fully safeguarded, and secondly that all Open Data would be referenced and attributed correctly as part of ethical research practice. Consequently, access to sensitive datasets should be carefully managed, however without putting up a barrier against successful sharing or openness of data.

The ethical issues are most clearly seen in the context of health and clinical research, in which the value of datasets from very large numbers of individuals is well-recognised, but the traditional contract between medical practitioner and patient is one of complete confidentiality, and this contract carries over to large scale studies (e.g. biobanks)⁵³ in which the studied individuals are recruited as examples of a population, not as individuals with a medical history. Furthermore, the issue of genetic privacy is an important one, as detailed genome information can easily be traced back to an individual once medical history datasets are linked⁵⁴. The patients/subjects have the right that the collected data be not individually attributable, and the researchers have the duty to ensure that this is indeed the case. However, a medical practitioner, as researcher (as opposed to a scientist as researcher) would have a professional obligation to provide treatment if the collected data indicated that this as required by an individual. This suggests that Open Access is not an absolute – that different levels of access to data may be appropriate for different individuals or groups, and that some method of establishing credentials would be necessary for more detailed access to the data.

The Wellcome Trust is the UK’s largest charitable foundation dedicated to scientific research. The mission of the Trust is to support biomedical research and the medical humanities. With regards to data, The Wellcome Trust does not have a policy on Open Access but focuses clearly on data sharing within the research community. The Trust, in collaboration with the UK Economic and Social Research Council (ESRC), Medical Research Council and Cancer Research UK currently run The Expert Advisory Group on Data Access, which provides ‘strategic advice on the emerging scientific, legal and ethical

⁵² OECD. *Principles and Guidelines for Access to Research Data from Public Funding*, op. cit., 2007

⁵³ European Commission, *Biobanks for Europe: A Challenge for Governance*, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/biobanks-for-europe_en.pdf

⁵⁴ Angrist, Misha, “Genetic privacy needs a more nuanced approach”, *Nature*, Vol. 294, February 2013, p.7

issues associated with data access for human genetics research and cohort studies.’⁵⁵ In the Group’s minutes from meetings in 2012⁵⁶ there are references to discussions about the sensitivity of the Trusts’ biomedical and social data, and how that would be exempt from complete openness, by use of ‘data safe havens’, which provide secure linking between databases, without the need to create new ones, as referred to by the UK Cabinet Office’s Open Data White Paper.⁵⁷ This is seen to allow for greater data security as the information will not be all in one place, but distributed, yet linked.

The UK Data Archive (UKDA), which hosts digital research data in the social sciences and humanities presents, on their website, a coherent overview of the issues of ethics, privacy and sharing of data in social sciences and humanities research, whereby it identifies the perceived tension between data sharing on the one hand and data protection on the other. It also acknowledges the potential risk that researchers may believe that research data obtained from people cannot be shared, as this would violate data protection and research ethics. The UKDA presents guidelines and advice on how to share data ethically and highlights the role of Research Ethics Committees, which are situated within many institutions and organisations, and can play a mediating role in reconciling data sharing and data protection by advising researchers. The UKDA states it is important to distinguish between personal data collected and research data in general, and provides definitions on personal sensitive and confidential information.⁵⁸ The guidelines furthermore state that the UK data protection laws only apply to personal data where consent has not been given to disclose, but they do not apply to anonymised data and that identifiable information can and may be excluded from data. The UKDA maintains that a combination of gaining consent for data sharing, the anonymising of data and controlled access can enable the ethical and legal sharing of data.

A review of research funders’ ethical protocols within the humanities revealed in many instances the ethical codes refer to an overall responsible conduct of research in terms of protection and disclosure of personal information, especially when involving human subjects. The ethical protocols are intended to ensure excellence in science, transparency, and equal opportunities, and set a general conduct of ‘good research’. The ethical codes often refer to non-discrimination policies, equality and diversity. The equal treatment of all disciplines, the gender mainstreaming, equal opportunities, the equality between men and women in the scientific and academic community and ethical standards are the values on which many ethical codes are based. Those policy documents make clear that no one may be excluded from science or academia due to reasons not related to science, such as gender, ethnic background, age or state of health.

Administrative ethics is also an issue often to be encountered; anti-bribery and anti-fraud plans, data protection; transparency and fairness, budgeting and certification requirements, health and safety precautions, subject matter eligibility and public communications policies are included in the policy documents.

⁵⁵ The Wellcome Trust, “Expert Advisory Group on Data Access”, no date. <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/index.htm>

⁵⁶ The Wellcome Trust, *Minutes of the Second Meeting of the Expert Advisory Group on Data Access (EAGDA)*, 19 October 2012. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp041115.pdf

⁵⁷ HM Government, *Open Data White Paper: Unleashing the Potential*, Cabinet Office, London, 2012. <https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential>

⁵⁸ UK Data Archive, “Create & Manage Data: Consent and Ethics”, 2013. <http://www.data-archive.ac.uk/create-manage/consent-ethics/legal?index=4>

Our review overall did reveal that research ethics are not explicitly discussed in great detail within the remit of Open Access to Data, but are rather referred to in relation to discussions about the overall ethical conduct of research, and data management. Although the above guidelines and discussions refer to sharing data, and not explicitly to Open Access to research data, they do provide a starting point for thinking about how Open Data can be implemented without jeopardising robust ethical research procedures.

The second ethical issue is that all datasets are the result of the intellectual effort of an individual or group, and they have the right to have their contribution acknowledged when the dataset is re-used, in exactly the same manner as their intellectual effort as represented in their publications would be acknowledged. This can be done using Digital Object Identifiers (DOIs)⁵⁹, and DataCite provides a framework within which this can be done. The DataCite website states:

*'We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books. Data citation can help by: enabling easy reuse and verification of data; allowing the impact of data to be tracked, and creating a scholarly structure that recognises and rewards data producers'*⁶⁰

The Economic Value and Cost of Science and Data

The economic value of science and data is a prevalent theme in policy literature on Open Data Access. Two key sub-themes emerged from the policy analysis, firstly describing data as currency or a capital asset, and secondly the perception of data as a public investment, from which institutions and government should seek the highest possible return.

An aspect of this can be seen in the UK Arts and Humanities Research Council capital call for funding on 'Big Data' where research outputs such as 'open datasets' and 'new, linked and mixed' are referred to as a tangible asset and applicants are also expected to present plans for long term management and sustained impact of the investment.⁶¹

The consequence of seeing data as an asset or investment is that since the collection of data is financed by public funds, it is the right of the public to have access to said data. It is also seen as the responsibility of the governments, funding bodies and scientist to ensure that investment is maximised by opening access for the sake of re-use of data, and prevent costly duplication of effort.⁶²

*'Access to research data increases the returns from public investment in this area...'*⁶³

Open Access to research data is often justified by the argument that as much of research is publically funded the outcomes are a public investment/good in a sense and it is economically

⁵⁹ Digital Object Identifier System. "The DOI System", 2013. <http://www.doi.org>

⁶⁰ DataCite. "Helping you find, access and reuse data", 2009. <http://www.datacite.org/whycitedata>

⁶¹ Arts and Humanities Research Council, *Digital Transformations in the Arts and Humanities: Big Data Research Call for Proposals*, July 2013. <http://www.ahrc.ac.uk/Funding-Opportunities/Documents/Big-Data-Projects-call-document.pdf>

⁶² European Commission, *Towards better access to scientific information: Boosting the benefits of public investments in research*, op. cit., 2012.

⁶³ OECD, *Principles and Guidelines for Access to Research Data from Public Funding*, op. cit., 2007, p.3.

sensible to seek to the highest returns from scientific endeavour. It is also argued that on the basis of this the knowledge produced should be publically accessible. The public and researchers should thus not have to pay each time to access data. The focus here is also on the re-use of data, in this instance not necessarily for the purposes of scientific validity but to derive better value for money from re-using data already gathered, making new connections and discoveries without first spending on data collection. There is also a perception here that publically available data can provide economic benefits to businesses in the form of driving innovation and getting produce faster to the market.

It is a clear message, in much of the higher policy level literature, that data is seen as a growth opportunity in times of austerity and economic strife. For instance the UK Government Open Data White paper refers to data as ‘the 21st century’s new raw material’, implying that there is a perception that there is a vast amount of data in the world that is there ripe for mining and turning economic profit. Furthermore, data has been referred to as ‘the new oil’⁶⁴ and as ‘the currency of science’⁶⁵.

In addition to the resource discourse when discussing openness and data sharing, this development has also been seen to bring opportunities for the ‘emergence of re-use ‘industries’ in particular areas of research and observation, as has happened with geospatial, meteorological and oceanographic data, etc., [as well as] opportunities for the emergence of support and service ‘industries’, focusing on providing value adding products and services that enable easier storage, discovery and access to datasets’⁶⁶

In our review we were not able to ascertain fully whether the perceptions on economic value of research data were of the scale that the policy documents indicated. Economic figures that we found were mostly derived from estimates and calculations based on the value of government owned data, e.g., satellite and geological survey data. The example below is drawn from a presentation from Jean Parcher at the US Geological Survey:

NASA Landsat satellite imagery of the Earth’s surface environment, collected over the last 40 years, was sold through the US Geological Survey for US\$600 per scene until 2008, when it became freely available from the Survey over the internet. Usage leapt from sales of 19,000 scenes per year, to transmission of 2,100,000 scenes per year. Google Earth now uses the images. There has been great scientific benefit, not least to the Geological Survey, which has seen a huge increase in its influence and its involvement in international collaboration. It is estimated to have created value for the environmental management industry of \$935 million per year, with direct benefit of more than \$100 million per year to the US economy, and has stimulated the development of applications from a large number of companies worldwide.⁶⁷

While these figures give some indication to the value of data, however it is not clear whether this value fully translates over to calculating the value of research data derived from scientific

⁶⁴ European Commission. *Scientific data: open access to research results will boost Europe's innovation capacity*, IP/12/790, 17 July 2012. http://europa.eu/rapid/press-release_IP-12-790_en.htm

⁶⁵ Council of European Social Science Data Archives. *Manifesto of the New Global Data Generation*, 2012. <http://www.cessda.org/about/manifesto.html>

⁶⁶ Fry, et al., op. cit., 2008, p. 4. http://repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf

⁶⁷ Parcher, Jean, “Benefits of open availability of Landsat data”, Presentation, U.S, Geological Survey, 2012 www.oosa.unvienna.org/pdf/pres/stsc2012/2012ind-05E.pdf

research. As the example tells, Google Earth is using a substantial amount of this data, which does skew the picture somewhat, as for much data there may not be a huge demand or as large scale users as Google.

In addition to calculating the value of Open Data, it is important also to have an understanding of the costs involved, and the commitment required for long term data management and archiving, as these are integral aspects of making data shareable and Open Access. In 2008 JISC⁶⁸ commissioned a report on the costs of keeping data safe and archived, which provide a cost model and guidance to UK Universities. The report provided detailed descriptions of cost elements and an activity costing framework focused on costs relating to staff, equipment, travel, consumables, estate and indirect costs of establishing and operating a research data repository (at full economic costing). The major activity elements identified related to three phases, namely:

- **Pre-archive** – a phase primarily relating to research projects in universities creating research data for later transfer to a data archive, in which implications for repository costs are considered and data collection/creation is designed and implemented with curation and sharing in mind;
- **Archive** – a phase primarily relating to the acquisition/disposal, ingest, storage and management of data, but also expanding into the provision of access and user support; and
- **Support services** – covering administration, common services and estates.

This is further supported by the findings from the 2011 European Commission Survey on the Open Access pilot in projects within all FP7 framework funded projects (2008-2011).⁶⁹ Grant beneficiaries are expected to deposit peer-reviewed research articles or final manuscripts resulting from their projects into an online repository and make their best efforts to ensure Open Access to those articles within a set period of time after publication. The Survey findings reveal that the dissemination of research results in FP7, including self-archiving and costs related to Open Access, are often an under-estimated aspect. However, it does require specific measures and a sustained investment.

It is not within the scope of this report to provide detailed economic cost calculations, however we feel that the above could help with policy recommendations that aim to fuel a discussion and action regarding the calculation of costs of Open Data, as well as economic benefits, which seem to steer much of the discussion around Open Data in high level policy literature

4.2.2 Motivations

In their communication to The European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions, the European Commission refers to Open Data as ‘an engine for innovation, growth and transparent governance’⁷⁰ indicating that

⁶⁸Beagrie, Neil, Julia Chruszcz, and Brian Lovoie, *Keeping research data safe: a cost model and guidance for UK Universities*, Final Report to JISC, JISC, London, 2008.

<http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>

⁶⁹ Directorate General for Research and Innovation, *Survey on Open Access in FP7*, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-open-access-in-fp7_en.pdf

⁷⁰ European Commission, *Open data: An engine for innovation, growth and transparent governance*, op. cit., 2012.

the prospect of openness can bring about various socio-economic benefits. Motivations for facilitating Open Data access are rooted in this view as well as in the values underpinning scientific practices, as discussed above. There is a clear feeling among stakeholders that science is changing and for future successful participation in science, openness will be key.⁷¹ Not only are there references to technological change but also change in research subjects, which are becoming more complex and greater in scope i.e. societal challenges or grand challenges.⁷² These are seen to require interdisciplinary and international research efforts, and for the success of which Open Access is seen as imperative for a faster solution to these challenges.

Motivations were also found to be rooted in the notion of the public as a funder of scientific endeavour and consequently an important stakeholder in scientific discovery. Alongside calls for openness to research outputs for all citizens, there are further motivations for opening up public sector data. The European Commission supports the opening up of public sector data for four key reasons:

- Public data has significant potential for re-use in new products and services. Overall economic gains from opening up this resource could amount to € 40 billion a year in the EU;
- Addressing societal challenges – having more data openly available will help us discover new and innovative solutions;
- Achieving efficiency gains through sharing data inside and between public administrations;
- Fostering participation of citizens in political and social life and increasing transparency of government.⁷³

Although this focus is outside the remit of this project, this is an important policy matter and deserves mentioning here as the discourse of ‘public funding’ as a justification and a motivation for is used by policy makers, in both instances. These policy issues are thus likely to co-develop and issues with implementation are likely to be similar across the two topics.

Economic motivations for opening up access to research data are presented in strong terms in the policy literature. In times of austerity and economic crises, the motivation to be seen as cost effective and offering value for money is strong. As detailed above, this comes out in the discourses presenting data as a capital asset and an untapped resource.

4.2.3 Barriers

Barriers to furthering the Open Data agenda are discussed and referred to many of the policy/stakeholder documents reviewed. These range from being technological in nature, e.g., technology driving the collection of vast datasets to lack of technical infrastructure to

⁷¹ E.g. Research Information Network, The Finch Group Report, *Accessibility, sustainability, excellence: how to expand access to research publications : Report of the Working Group on Expanding Access to Published Research Findings*, London, 2012.

<http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf>

⁷² OECD, *Principles and Guidelines for Access to Research Data from Public Funding*, op. cit. 2007.

⁷³ European Commission, “Open data”, Digital Agenda for Europe, 2013. <https://ec.europa.eu/digital-agenda/node/70>

store the data in question and interoperability issues.⁷⁴ Also, cultural barriers are frequently discussed, and then first and foremost the competition within science, lack of trust between scientists and lack of career related rewards and prestige resulting from publishing and sharing data.

*'At all levels the amount of data created, maintained and shared by (research) infrastructures has been growing by orders of magnitude over a very short period of time, and the management of this data deluge has become a critical issue if the sheer volume is not to overwhelm researchers.'*⁷⁵

A key barrier, addressed in the policy documents reviewed, was the magnitude and speed with which data collection is being carried out. The technology, which is seen to drive Open Data access, is also seen to drive the possibility of collecting and storing far more data than ever before. Policy makers and stakeholders acknowledge this in their discussions about the need to build adequate infrastructure, norms, culture around the collecting, archiving and curating of data.⁷⁶

In terms of publishing datasets in relation to Open Access journal publications it is acknowledged by publishers that it will be both time and cost consuming to build infrastructure and sustainable business models around this practice.⁷⁷ Cultural barriers play a part here as it is pointed out that researchers are currently not rewarded for sharing or publishing datasets, only peer reviewed publications. The pressure among researchers to publish, to compete for and win grant funding, and to repeat the cycle, is strong and very persistent. Researchers' career trajectories largely depend on their success in these activities.⁷⁸ The career-related rewards for sharing or publishing datasets are still largely absent, which will make it even more difficult to develop a healthy publishing model based on the enrichment of publications.

CESSDA also raises the issue of the sustainability of research infrastructures. Now that a large proportion is run on short term funding, there is a danger of datasets getting lost, should an institution and consequently its repository not receive follow-on funding. Secure and permanent access to data will need to be taken into consideration as the movement for Open Data continues. Uncertainty over the future of datasets submitted to repositories could become a barrier to successfully implementing Open Access to publically funded research data.⁷⁹

The management of data is a core prerequisite in making data open. Good data management is a key feature in ensuring that research data is of a high quality and it therefore supports research excellence. However, and further to this, good data management is crucial for facilitating data sharing and ensuring the sustainability and accessibility of data in the long-

⁷⁴ European Commission, *Riding the wave: How Europe can gain from the rising tide of scientific data*, Final Report of the High level Expert Group on Scientific Data, Brussels, October 2010, p.4. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

⁷⁵ Council of European Social Science Data Archives, *Manifesto of the New Global Data Generation*, 2012. <http://www.CESSDA.org/about/manifesto.html>

⁷⁶ European Commission, *Riding the wave: How Europe can gain from the rising tide of scientific data*, op. cit., 2010, p.4.

⁷⁷ Ibid.

⁷⁸ Swan, Alma and Sheridan Brown, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*, A report commissioned by the Research Information Framework, RIN, London, 2008.

⁷⁹ Council of European Social Science Data Archives, op. cit., 2012.

term and therefore data re-use for future science. If research data are well organised, documented, preserved and accessible, and their accuracy and validity is controlled at all times, the result is high quality data, efficient research, findings based on solid evidence and the saving of time and resources. Researchers themselves benefit greatly from good data management. The development of good data management has been strongly supported by universities with many universities providing data management guidelines that align with Funders' expectations and that support the needs of data archives and repositories. It is strongly advocated that data management should be planned before research starts and may not necessarily incur much additional time or costs if it is engrained in standard research practice⁸⁰. However, the responsibility for data management lies primarily with researchers, but institutions and organisations can provide a supporting framework of guidance, tools and infrastructure and support staff can help with many facets of data management. Establishing the roles and responsibilities of all parties involved is key to successful data management and sharing.

Another and more detailed aspect of data management can be seen in The Wellcome Trust Sanger Institute *Data Sharing Guidelines*.⁸¹ The context of these guidelines is located in the Institute's aim to provide rapid access to datasets of use to the research community and to place these in publicly accessible repositories when possible. The guidelines cover ethical considerations relevant to conducting genetic and genomic research, which means addressing the responsibilities to protect confidentiality and the privacy of research participants. Access to certain datasets is therefore carefully managed and granted in a transparent manner to all appropriately qualified researchers. They also cover the rights of data providers and the Institute recognises the need for researchers to be appropriately credited for their scientific contribution and investment in data generation. It is therefore expected that all researchers both honour agreements in line with "Fort Lauderdale's" data sharing principles, and appropriately acknowledge the contributions of others.⁸² A further aspect is that of optimising and translation because the Institute recognises that, in specific instances, the use of intellectual property protection and attendant potential delays to data sharing may be necessary to prevent inappropriately exclusive claims by others and to ensure health benefits occur.

What is important to note is that data management is a key requisite that underpins Open Access to data. There are guidelines for researchers, and universities, funders and archiving groups are strongly supporting the development of data management. However, as with other aspects of making data open, there are specific issues that relate to particular kinds of data that require informed and intelligent approaches to data management. The responsibility for data management, nonetheless, lies with individual researchers and research groups.

The FP7 funded project Permanent Access to the Records of Science in Europe (PARSE.Insight) (2008-2010) conducted major surveys in order to gain insight into the current state of affairs in digital preservation of digital research data (including publications), the outlook of data preservation, data sharing, and the roles & responsibilities of stakeholders in research and funding of research. The key aim was to define the needs for an e-Science infrastructure for long-term availability of research data. Surveys were sent out to the

⁸⁰UK Data Archive, *Managing and Sharing Data, Best Practice for Researchers*, University of Essex, May 2011. <http://data-archive.ac.uk/media/2894/managingsharing.pdf>

⁸¹The Wellcome Trust, *The Wellcome Trust Sanger Institute Data Sharing Guidelines*, July 2010. http://www.sanger.ac.uk/datasharing/assets/wtsi_datasharing_guidelines.pdf

⁸² Ibid. p. 10.

stakeholder groups: researchers, publishers and data managers. A brief summary of findings gives some indication to the perceived barriers to further implementing the digital preservation agenda, which gives some indication of barriers that the move toward Open Data access may include.

- ‘Researchers regard *the lack of sustainable hardware, software or support of computer environment may make the information inaccessible* as the most important threat to digital preservation. 80% believe this to be either important or very important,
- 58% of the research respondents believe that an *international infrastructure for data preservation and access* should be built to help guard against some of the above-mentioned threats.
- 25% of the researchers make their data openly available for everyone.
- Major barriers for sharing research data are *the fear of researchers regarding legal issues* and *the misuse of their data*.
- Data managers also regard *the lack of sustainable hardware, software or support of computer environment may make the information inaccessible* as the most important threat to digital preservation. 86% believe this to be either important or very important.
- 59% of the respondents to the data managers’ survey don’t think that the tools and infra- structure available to them will suffice for the digital preservation objectives they have to achieve.
- 71% of the respondents to the data managers’ survey believe that funding for preservation will be an issue now and in five years’ time.’⁸³

It is clear that different fields of science may need specific data processes put in place which take into account the specificities of the research in question. Examples of this are clinical and social research, where personal data is collected. Policy documents do acknowledge the need for robust data security measures, as well as processes and standards for anonymising data. In other research, e.g. pharmaceutical and research where commercial stakes are high, barriers of IPR and other legal issues regarding openness of data and discovery may complicate the process to making research data publically open. The lack of infrastructure may prove to be costly barrier for the implementation of digitization and Open Data access, and this will be further explored in WP 2 of the RECODE project.

4.3 STAKEHOLDER MOTIVATIONS FOR OPEN ACCESS

4.3.1 Libraries and Repositories

A common theme running through the library values and motivations, as presented in a review of library documents, on Open Access to publications and data, is that Open Access is seen as a solution to global scientific challenges by offering access without limits to all human kind.⁸⁴ Access to information and data is seen as vital in developing people’s

⁸³ Kuipers, Tom and Jeffrey van der Hoeven. *Insight into digital preservation of research output in Europe*. Final Report of the PARSE.Insight project, STFC, London, 2009, pp.4-5. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

⁸⁴ International Federation of Library Associations and Institutions, “Statement on Open Access to Scholarly Literature and Research Documentation”, 2013. <http://www.ifla.org/publications/ifla-statement-on-open-access-to-scholarly-literature-and-research-documentation>

understanding of the world and is seen to reduce information inequality, both within developed countries and between developed and developing countries. Open Access to knowledge is seen to further strengthen the public role of libraries and help them fulfil their responsibility as generators, distributors and guardians of knowledge derived from scientific research.

Libraries also see it as their role to defend the rights of authors, with regard to attribution and copyright protection, and at the same time oppose any attempts at censorship from government or commercial bodies. The International Federation of Library Associations and Institutions (IFLA) is to adopt peer review processes to assure the quality of all scholarly literature irrespective of mode of publication.

LIBER⁸⁵ (Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries) is the main network for research libraries in Europe and it also has a key role in spreading the value of Open Access. LIBER sees itself as a champion for Open Access, e.g. policies for publication, especially the Gold route, as well as Open Access to Data. It promotes some underpinning areas to Open Access such as efficient information services, access to research information in any form whatsoever; preservation of cultural heritage; efficient and effective management; high-quality services for all users of library and information services; intellectual freedom and access to scholarship; collaboration with campus/local/national/European and global partners; stewardship of collections and institutional resources, in the most appropriate format; leadership, innovation and a willingness to embrace opportunities for change and inclusivity, equality of opportunity and fulfilment of potential. LIBER's new strategic plan⁸⁶ outlines its key performance areas (Scholarly Communication and Research Infrastructures, Reshaping the Research Library, Advocacy and Communications), values and motivations for the period 2013-2015. It includes an implementation plan, a communication plan and the management of data. LIBER openly states that it supports the European Commission published Communication and the Recommendation to Member States on "access to and preservation of scientific information".⁸⁷ Both documents strongly support Open Access as a way of making the free circulation of knowledge in Europe a reality. LIBER formally acknowledges the Open Access policy presented by the European Commission. "Gold" Open Access will continue to be supported in Horizon 2020 by providing funds for Open Access publishing, and the Commission will consider the conditions under which Open Access publication fees can be reimbursed after the end of the grant agreement. "Green" Open Access is also stimulated by the Commission, accepting embargo periods of 6 months (and 12 for Social Sciences and Humanities) and will be built on the scientific information infrastructure delivered by OpenAIRE.

LIBER strongly believes in the key role of research libraries in Europe to support the European Commission and national governments in the implementation of Open Access in the new Scholarly Communication arena. European libraries are strategically situated

⁸⁵ Association of European Research Libraries, "Association of European Research Libraries", no date. <http://www.libereurope.eu/>

⁸⁶ Association of European Research Libraries, "LIBER Strategic Plan 2013-2015: Re-inventing the Library for the future", 2012. <http://libereurope.eu/strategy>

⁸⁷ Directorate-General Research and Innovation, *Online survey on scientific information in the digital age*, European Commission, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

between funders, researchers, publishers and institutions and have a vast experience providing researchers with access to information.

In the US and Canada, The Association of Research Libraries⁸⁸ is a non-profit organisation of 125 research libraries that share similar research missions, aspirations, and achievements. Its principal values comprise open and equitable access to information as a fundamental tenet to society, research libraries as active agents central to the process of the transmission and creation of knowledge and their responsibility to anticipate and prepare for the information needs of present and future users, as well as the collaboration among libraries that improves prospects for individual library success in fulfilling local needs. ARL's strategic plan⁸⁹ supports Open Access to publication as well as open science and Open Access to Data, which are embedded in its mission and principles, such as fair use, fair dealing, use of copyrighted materials, reducing economic, legal, and technical barriers to access and use of the research results from publicly funded research projects, new models for the management and access to government information, openness and transparency in contract agreements. ARL aims to analyse and advocate based on its values, to expand and strengthen alliances with organisations that share common goals to advance policy issues, to promote, facilitate and conduct research in relevant areas of public and information policy, to the openness and transparency in contract agreements, to sponsor, conduct, and promote research that will lead to the development and assessments of new models of scholarly communication, to develop and promote appropriate responses to unacceptable business practices, to accelerate and enhance outreach and communication efforts to inform and induce change among the educational and research communities, to promote and facilitate the development of a diverse group of library professionals.

4.3.2 Funders

Funders in the humanities are broadly supporting Open Access to Data, although that support is uneven. The values underpinning the varying levels of support relate to seeing the value of making data more open to a range of user communities. Funders on the whole however are adopting a step-by-step approach to promoting, fostering and developing Open Access to Data.

Strongly promoting Open Access to data

The Austrian Science Fund (FWF) has an Open Access policy for the projects that it funds⁹⁰. The policy concerns both publications and data, where Open Access is obligatory for all types of scientific output, except when it is legally not possible. Researchers can choose between Open Access repositories or to publish in an Open Access venue. The agency provides funds for Open Access publications for up to three years after the completion of projects. Self-archiving of research papers in the life sciences is obligatory in the repository of EuropePMC. Open Access activities are to be indicated in any reports to the FWF, and grant recipients are required to provide justification in cases where the FWF's Open Access policy could not be observed for legal reasons. With respect to Open Access in the Social

⁸⁸ Association of Research Libraries, "Homepage", 2013. <http://www.arl.org/>

⁸⁹ Association of Research Libraries, "ARL Strategic Plan 2010-2012", 2012. <http://www.arl.org/storage/documents/publications/strategic-plan-2010-2012.pdf>

⁹⁰ The Austrian Science Fund (FWF), "Open Access Policy bei FWF-Projekten", no date. http://www.fwf.ac.at/de/public_relations/oai/index.html (in German)
http://www.fwf.ac.at/en/public_relations/oai/index.html (in English)

Sciences & Humanities (SSH), the FWF established a fund for innovative Open Access journals in SSH from April 2013.⁹¹ Austria's main funder, FWF, does not appear to have specific policies on the funding of the SSH, but the values that they set as important are the excellence and competition, the independence, the international orientation, the equal treatment of all disciplines, the transparency and fairness, the gender mainstreaming, the equal opportunities and ethical standards. The data archiving policy⁹² registered in the SHERPA/JULIET requires deposition of research data in Open Access archives within two years after the project completion. The location of deposition is an appropriate institutional or disciplinary repositories (optional). The policy applies to all projects funded totally or partly by the FWF and the access provision must be free of charge.

Overall, the Research Councils UK (RCUK) strongly promotes data sharing within the field of each specific Research Council. RCUK has published a list of common principles⁹³ on data policy, which provides an overarching framework for the individual Research Council data policies. In summary these see publicly funded research data as a public good, which should be made openly available with as few barriers as possible, and also in timely manner which does not harm intellectual property. Data with long-term value should be preserved, and made accessible for future research, by use of institutional and project specific data management policies that shall be according to recognised standards and community best practice. Sufficient metadata should be recorded and made openly available to enable data to be effectively re-used. The RCUK states that users of research data should always acknowledge the sources of their data, and abide by the terms and conditions under which they are accessed.

The UK Natural and Environmental Research Council (NERC) currently makes a distinction between Environmental Data and Information Products. Environmental Data are individual items or records (both digital and analogue) usually obtained by measurement, observation or modelling of the natural world and impacts of humans upon it, including all the necessary calibration and quality control elements. Information Products are created by adding another level of intellectual input that refines or adds value to the original Environmental Data through interpretation and/or combination with other data or Information Products. This approach reflects current UK Government policy and guidance. In general, NERC makes Environmental Data openly available with an associated licence and charges for Information Products.

Taking some measures to Open Access to data

The Social Sciences and Humanities Research Council of Canada (SSHRC) in October 2004 took the position of supporting Open Access in principle and its Governing Council approved a policy on Open Access in 2006, deciding to take “an awareness-raising, educational and promotional approach to [the policy's] implementation, rather than imposing mandatory requirements”.⁹⁴ As part of this commitment, the three Federal research granting agencies

⁹¹ The Austrian Science Fund (FWF), “Aktuelle Information”, 2012. http://www.fwf.ac.at/de/aktuelles_detail.asp?N_ID=506

⁹² The Austrian Science Fund (FWF), “Open Access Policy for FWF-funded projects”, no date.

⁹³ Research Councils UK, “Excellence with Impact – RCUK Common Principles on Data Policy”, no date. <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>

⁹⁴ Social Sciences and Humanities Research Council, “Research Data Archiving Policy”, Government of Canada, 9 June 2013. http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx

have taken various measures to promote Open Access practices with a goal to ultimately adopt a joint policy on access to research results, starting with a Comprehensive Brief on Open Access to Publications and Research Data for the Federal Granting Agencies⁹⁵, which followed the Canadian policy on Open Access to research data⁹⁶. The SSHRC does not have a specific data archiving policy registered in SHERPA/JULIET.

The German Research Foundation (DFG) is the self-governing organisation for science and research in Germany. The DFG supports Open Access and encourages the deposition of publications in repositories and Open Access publishing, as well as the deposition of scientific data in appropriate repositories. It thus assumes a voluntary and non-policed Open Access policy, which concerns all disciplines with an embargo of 12 months. DFG Open Access policies are interestingly described in the DGF Magazin and not in the main section of the website of the funder, where overall policies are provided⁹⁷, indicating that they may not enjoy as wide or accepted support as processes of this funding organisation. According to the SHERPA/JULIET, the data archiving policy⁹⁸ encourages deposition of research data in Open Access archives within 12 months after the project completion, in appropriate institutional/disciplinary repositories. According to the policy primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin.

The UK Biotechnology and Biological Sciences Research Council (BBSRC) expects research data generated as a result of BBSRC support to be made available with as few restrictions as possible in a timely and responsible manner to the scientific community for subsequent research. Applicants should make use of existing standards for data collection and management and make data available through existing community resources or databases where possible. In line with the BBSRC Statement on Safeguarding Good Scientific Practice⁹⁹, data should also be retained for a period of ten years after completion of a research project. BBSRC recognises that different fields of study will require different approaches. What is sensible in one scientific or technological area may not work in others; therefore the policy aims to achieve the sharing of data in an appropriate manner and not to be overly prescriptive. Researchers are required to adhere to any relevant regulatory requirements, including those relating to the ethical use of data. BBSRC recognises the importance of data quality and provenance. Data should, wherever appropriate and possible, be accompanied by contextual information or documentation (metadata) to provide a secondary user with any necessary details on the origin or manipulation of the data in order to prevent any misuse, misinterpretation or confusion.

The value of data often depends on timeliness. Researchers have a legitimate interest in benefiting from their own time and effort in producing the data but not in prolonged exclusive use of these data. Timescales for data sharing will be influenced by the nature of

⁹⁵ Government of Canada, “Comprehensive Brief on Open Access to Publications and Research Data for the Federal Granting Agencies”, June 2011. <http://www.science.gc.ca/default.asp?lang=En&n=2360F10C-1>

⁹⁶ Government of Canada, “Canadian IPY 2007-2008 Data Policy”, 14 November 2011. http://www.api-ipy.gc.ca/pg_IPYAPI_055-eng.htm

⁹⁷ Deutsche Forschungsgemeinschaft (DFG), “Open Access und Forschungsförderung durch die Deutsche Forschungsgemeinschaft”, aktualisierungsdatum, 24 January 2012. http://www.dfg.de/dfg_magazin/forschungspolitik_standpunkte_perspektiven/open_access/index.htm.

⁹⁸ Securing a Hybrid Environment for Research Preservation and Access (SHERPA), “Research funders archiving mandates and guidelines (JULIET)” 2013. <http://www.sherpa.ac.uk/juliet/index.php>

⁹⁹ Biotechnology and Biological Sciences Research Council, “Safeguarding good scientific practice”, June 2006. <http://www.bbsrc.ac.uk/organisation/policies/position/policy/good-scientific-practice.aspx>

the data but it is expected that timely release would generally be no later than the release through publication of the main findings and should be in-line with established best practice in the field. BBSRC considers data sharing to be an important activity and whilst recognising the need to safeguard Intellectual Property and to protect opportunities for commercialisation of research outputs considers that this should not unduly delay or prevent data sharing. BBSRC supports the view that those enabling sharing should receive full and appropriate recognition by funders, their academic institutions and new users for promoting secondary research.

The UK Economic and Social Research Council (ESRC), in respect to Open Access to publications, broadly refers to the Research Councils UK (RCUK) policy, summarised above. The ESRC furthermore refers to the UK Government's initiative (gov.uk) which aims to make public data openly accessible. Within their Research Data Policy¹⁰⁰, from 2010, it is stated that the ESRC is committed to providing access to research data, in order to enable further re-use and thus strengthen the capacity for secondary data analysis. The funding body requires research data arising from ESRC-funded research to be made available to the scientific community in a timely and responsible manner. The ESRC supports a number of data service providers, which facilitate easy access, dissemination and promotion of existing data sources. These providers are responsible for providing clear guidance to ESRC grant holders in support of their data management and sharing plans, which is now required as part of all funding applications. The ESRC also states that they will provide adequate funding for the realisation of data management and sharing. The data service providers should also provide support to the researchers requiring access, and use of data held by these service providers.

Focusing primarily on Open Access to publications

The Arts and Humanities Research Council in the UK (AHRC) supports the belief that free and Open Access to publicly-funded research offers significant social and economic benefits. It focuses more heavily on Open Access to publication and follows the RCUK policy on access to research outputs, which was revised on 6 March 2013¹⁰¹. It supports both 'Gold' and 'Green' routes to Open Access, though RCUK has a preference for immediate Open Access with the maximum opportunity for reuse. The RCUK policy applies for peer-reviewed papers that are submitted for publication from 1 April 2013 and which acknowledge Research Council funding and are published in journals or conference proceedings. The AHRC Data archiving policy (2006), according to the SHERPA/JULIET, requires deposition of research data in Open Access archives at the earliest possible opportunity, in any appropriate repository or in named repositories (optional) or in the Archaeology Data Service (optional). The policy applies to all projects funded totally or partly by the AHRC.¹⁰² Data must be stored for at least three years after the end of the grant. Archaeology grants depositing in Archaeology Data Service (ADS) must consult ADS on formats to deposit. The ADS deposits after 31 March 2013 will incur a charge for deposit and must occur within three months of the end of the project. As such this is very much a first step in the preparation for Open Access to data.

¹⁰⁰ Economic & Social Research Council, *ESRC Research Data Policy*, September 2010. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

¹⁰¹ The Research Councils UK, *RCUK Policy on Open Access and Supporting Guidance*, 8 April 2013. <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf>

¹⁰² Arts & Humanities Research Council, "Deposits of resources or datasets", 2012. <http://www.ahrc.ac.uk/Funding-Opportunities/Research-funding/RFG/Annexes/Pages/Deposits.aspx>

The Danish National Research Foundation (DNFR) and all the Danish Funding Research Councils have a common Open Access policy in terms of access to publications.¹⁰³ With this policy, research councils and foundations want to establish Open Access as the standard in scientific publishing. The aim is to ensure that all scientific articles, the quality of which has been assured by peer review and which have been published in a scientific magazine, can be read and distributed without any financial, technical or legal restrictions. This policy means that published scientific articles which are the result of full or part financing by research council and foundations must be made freely available to everybody via Open Access with the permission of the magazine. This policy excludes research data, as well as monographs, anthologies, books, popular science articles. The DNFR does not have a specific data archiving policy registered in the SHERPA/JULIET.

The European Science Foundation's (ESF) membership organisation for all medical research councils in Europe, the European Medical Research Councils, released in late 2012 a call for the adoption of Open Access in biomedical sciences.¹⁰⁴ It is stated that there is a moral imperative for Open Access and that all research councils must work together to raise awareness of this and to support the establishment of a Europe-wide repository in biomedicine as a partner site to the US equivalent PubMed.

The Scientific Council for Humanities and Social Sciences is a part of the major Swedish research funding organisation Swedish Research Council. In 2010 the council mandated Open Access for all 'research results' produced via grants from the Council. The mandate is focused on publications, and Med Central. Open Data is not explicitly mentioned in this document¹⁰⁵. In 2012 the Council got an assignment from the Swedish Government to draft a national Open Access policy for both publications and data.¹⁰⁶

Low levels of guidance

In the USA, however, the National Endowment for the Humanities Grant Policy does not refer to Open Access directly, but some issues linked to copyright, reproduction and re-use of research results are mentioned under the chapter "Intangible Property".¹⁰⁷ In this chapter, in response to a Freedom of Information Act (FOIA) request for research data relating to published research findings produced under an award that were used by the federal government in developing, NEH is regulating the terms under which the agency will request the research data, so that they can be made available to the public, the time in which the data will be requested and possible fees that the NEH may charge the requester with. The NEH is not registered with the SHERPA/JULIET.

¹⁰³ Det Frie Forskningsråd; Danish National Research Foundation; Højteknologifonden; Det Strategiske Forskningsråd; Rådet for Teknologi og Innovation, "Open Access policy for public-sector research councils and foundations", 21 June 2012.

http://dg.dk/filer/fonden/open_access/Final%20Open%20Access%20policy%20English.pdf.

¹⁰⁴ European Science Foundation, "Open Access in Biomedical Research", *Science Policy Briefing 47*, September 2012. http://www.esf.org/fileadmin/Public_documents/Publications/spb47_OpenAccess.pdf

¹⁰⁵ Vetenskapsrådet (Swedish Research Council), "Fri tillgänglighet till forskningsresultat – Open Access", 15 February 2012. <http://www.vr.se/106.29b9c5ae1268d01cd5c80001275.html>

¹⁰⁶ Ibid.

¹⁰⁷ "The NEH reserves a royalty-free, nonexclusive, and irrevocable right to reproduce, publish or otherwise use these materials for federal purposes and to authorize others to do so" (see 2 CFR Part 215.36, Intangible Property, OMB Circular A-110).

The purpose of the Fonds de la Recherche Scientifique in Belgium (FNRS) is to promote free (fundamental) scientific research within the French-speaking Community of Belgium through the attribution of grants to researchers and institutions. There are guidelines regarding undertaking research in an ethical manner but reference to Open Access to Data is not made explicit. Their policy mainly refers to the dissemination of publications under Grants for scientific publications ISDT Wernaers funds for research and dissemination of knowledge¹⁰⁸. Other regulations apply to the Scientific Committees¹⁰⁹, the Support Committee¹¹⁰ and the Guidance Committee¹¹¹. The FNRS is not registered with the SHERPA/JULIET.

4.3.3 Publishers

From the view of publishing, there seems to be a gap between the publication of Open Access articles and books on the one hand and publishing Open Data on the other. Whilst Open Access to publications is greatly advancing, the publishing of Open Data is taking off somewhat slower. However, there are developments afoot, which can be seen e.g., in the establishment of Elsevier's new Research Data Group, whose principles focus on working towards completely Open Data and being transparent and collaborative. Furthermore, in the pilot work that Elsevier is currently working on with academic partners, two other principles have been realised:

'Data must be open and shared, with distribution controlled by the creator of the data (when possible).

The model must be derived in collaboration with the research community and funding agencies, not driven by Elsevier or any publisher¹¹²,

Barriers to fully realising Open Data from the publishers' side are detailed as a 'lack of the effort and the informatics expertise required to standardize and normalize research data, and to add sufficient provenance and the descriptive metadata required for domain-specific data repositories.'¹¹³ Elsevier is currently collaborating with academic partners in pilots, in order to understand more fully what contribution the publisher can make to get more data shared, exploring sustainable funding models and also to help set up a credit system for researchers.

Wiley, in partnership with the Royal Meteorological Society and the University of Leicester, is currently working on developing workflows for the Open Access Geoscience Data Journal¹¹⁴, within the remit of the JISC funded project *Peer REview for Publication &*

¹⁰⁸ Fonds de la Recherche Scientifique (FNRS), "Diffusion et publications", 31 March 2013. <http://www.frs-fnrs.be/fr/financer-les-chercheurs/ecoles-doctorales-congres-publications/diffusion-publications.html>

¹⁰⁹ Fonds de la Recherche Scientifique (FNRS), "Reglement Adopte par le conseil d'administration du F.R.S. – FNRS", 7 December 2012.

http://www.frs-fnrs.be/uploaddocs/docs/SOUTENIR/FRS-FNRS_Reglement_Commissions_Scientifiques_2013.pdf

¹⁰⁹ FNRS, "Diffusion et publications", op. cit., 2013.

¹¹⁰ Fonds de la Recherche Scientifique (FNRS), "Reglement du comite d'accompagnement", 9 December 2010. http://www.frs-fnrs.be/uploaddocs/docs/SOUTENIR/FRS-FNRS_Reglement_Comite_Accompagnement_2011.pdf

¹¹¹ Ibid.

¹¹² Marques, David, "Research Data Driving New Services", *Research Data Management*, Vol 1, No.1, 2013. <http://libraryconnect.elsevier.com/articles/best-practices/2013-02/research-data-driving-new-services>

¹¹³ Ibid.

¹¹⁴ Wiley, "Geoscience Data Journal: Vol. 2", 2013. <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html>

Accreditation of Research Data in the Earth sciences (PREPARDE).¹¹⁵ It is an international project that brings together a wide range of experts in research, academic publishing and data management to produce data publication guidelines applicable across a range of research disciplines and data types.

The Geoscience Data Journal works with recognised Data Centres around the world to develop a future strategy for data publication, the recognition of the value of data and the communication and exploitation of data to the wider science and stakeholder communities.

The content description for the Geoscience Data Journal reads as follows:

*'A data article describes a dataset, giving details of its collection, processing, file formats etc., but does not go into detail of any scientific analysis of the dataset or draw conclusions from that data. The data paper should allow the reader to understand the when, why and how the data was collected, and what the data is.'*¹¹⁶

SpringerOpen, Springer's portfolio of fully Open Access journals and books, currently collaborates with DataCite, the British Library, the Digital Curation Centre and the scientific community to develop and maintain a list of data repositories, in order to further Open Data access as part of Open Access publishing.

Some of SpringerOpen journals now additionally encourage or require authors, as a condition of publication, to include in some articles a section that provides a permanent link to the data supporting the article results, 'This section is then called 'Availability of supporting data' and is only included in an article if supporting data are available in an Open Access repository or included in the additional files published with the article. The aim is to provide links in a consistent place within an article to supporting data - regardless of the location or format of the data - and to make it clear to readers when they can also access the data as well as the article.'¹¹⁷

BioMed Central (part of Springer) has recently posted an Open Data policy, which came into effect on the 3rd of September 2013, following their 2012 public consultation on Open Data. Their Open Data policy aims to clarify the legal status of data published in their Open Access journals, as well as maximizing the potential for reuse of published science. They refer to the *Panton Principles for Open Data in Science*¹¹⁸, as they state that data needs to be available so that it can be reused, scrutinised and built upon with the minimum of barriers. To achieve this BioMed Open Data will be:

*'unless otherwise stated in an individual article's license, data included in BioMed Central's published Open Access articles are distributed under the Creative Commons CC0 1.0 Public Domain Dedication waiver'*¹¹⁹. *Anyone reusing data published in BioMed Central journals must, wherever possible, cite the source(s) of the data in a derivative work, although this is*

¹¹⁵ University of Leicester, "PREPARDE", no date. <http://www2.le.ac.uk/projects/preparde>

¹¹⁶ Wiley, op. cit., 2013.

¹¹⁷ Springer Open, "Availability of Supporting Data", 2013. <http://www.springeropen.com/about/supportingdata>

¹¹⁸ Murray-Rust, Peter, Cameron Neylon, Rufus Pollock and John Wilbanks, "Panton Principles, principles for open data in science", 2010. <http://pantonprinciples.org/>

¹¹⁹ Creative Commons, "CC0 1.0 Universal (CCO 1.0) Public Domain Dedication", no date. <http://creativecommons.org/publicdomain/zero/1.0/>

*not a legal requirement. The Creative Commons CC0 waiver applies to data included in the article, its reference list(s) and its additional files.*¹²⁰

This is an interesting push towards open research data publications coming from an important player in the field of publishing. However, information regarding costs and business models are not currently available. The Open Data Policy is currently only about changing the license for data published in BioMed Central journals. According to BioMed's policy, there are no plans to increase the maximum additional file size and number of files which can be published. Therefore data storage is unaffected by the policy.

When discussing Open Data access, with regard to publishers of data, it is important to make a distinction between publishers as enterprises (Elsevier, Springer, University Presses) and publishing initiatives with a non-profit base, like scholarly-led journals, library services and big European research societies and groups, which are becoming important information platforms with their own repository's and document infrastructure (e.g. CERN, Wellcome Trust, Max Planck Society). These stakeholders will have different values and motivations associated with Open Data due to the different funding/business models as well the broader context within in which they work.

For scholarly publishers, the importance of enriching publications lies primarily in the added value this can offer to researchers, both readers and authors. This immediately raises a number of questions. Which criteria must an enriched publication and its associated dataset(s) meet to actually be considered enrichment for an article or other type of publication? Are authors readily prepared to release their research data? Which applications can the reader use to read the data? What is the impact on current practice (e.g. reading and publishing)? These are all important questions that a scholarly publisher has to take into account when considering starting to offer this type of service to readers and authors. Close collaboration with research groups and university libraries is therefore an essential requirement, as the examples above illustrate.

University Presses seems to be slow adopters of Open Data publishing. Open Access publishing has been adopted quickly though. Amsterdam University Press, Manchester University Press, MIT Press amongst others, have moved towards Open Access publication policies. A good example of a collaboration of presses to develop an Open Access publication infrastructure is the EU FP7 project OAPEN library (Open Access Publishing in European Networks).¹²¹ But this project only focuses on Open Access to publications. In May 2008 Amsterdam University Press (AUP) began working with nine archaeological institutions in the Netherlands and Flanders on the development of the Open Access e-journal JALC¹²². Together with the University of Amsterdam, University of Leiden and the e-depot Netherlands Archaeology (EDNA), AUP has taken the first steps towards an infrastructure that will allow enriched publications to be integrated in the online publication environment of JALC. Five research articles have been 'enriched' so far; however this has not lead towards a clear publication policy of Open Data. Smaller presses may be likely to work together with national and international institutional repositories, as issues regarding technology, preservation, storage, etc. are too complicated to take up for small enterprises.

¹²⁰ BioMed Central, "Open Data", 3 September 2013. <http://www.biomedcentral.com/about/opendata>

¹²¹ Open Access Publishing in European Networks, "Welcome to OAPEN", no date. www.oapen.org

¹²² Amsterdam University Press, "Journal of Archaeology in the Low Countries", 2013. www.jalc.nl

4.3.4 *Advocacy Groups, Professional Organisations and CSOs*

Working across individual level stakeholders are organisations, which promote and support Open Access. Organisations of European research libraries are key in promoting Open Access as well as providing practical steps to achieving Open Access. These include LERU¹²³, the League of European Research Universities, which is an advocacy body for the promotion of basic research at European universities and has an essential role in the innovation process. It has published the *LERU Roadmap towards Open Access* in June 2011 with a view to “investigate new models for scholarly communication and the dissemination of research outputs emanating from LERU universities”.¹²⁴ On December 2012 it also published the *LERU statement on Open Access to Research Publications*¹²⁵ and the *LERU statement on Open Research Data*¹²⁶. These documents provide a clear steer and guidance for associated bodies, for the development of Open Access to research data.

Other stakeholders working across organisations include Science Europe¹²⁷, which is an association of European Research Funding Organisations (RFOs) and Research Performing Organisations (RPOs) whose policy¹²⁸ aims to strengthen the dialogue between science and society, and one of its areas to facilitate cross border co-operation of RPOs, to ensure shared funding and efficient exploitation of medium-sized research infrastructure to develop a common policy on Open Access to scientific publications and "Permanent Access" to research data and to connect European research to the world. Its values and motivations are based on excellence, cross-border collaboration, joint strategies, expertise, co-operation and Open Access, in order to fund and perform excellence in research on a common basis with a view to develop common positions.

Another cross-organisational stakeholder is Securing a Hybrid Environment for Research Preservation and Access (SHERPA)¹²⁹, which investigates issues in the future of scholarly communication. It actively supports Open Access by developing Open Access institutional repositories in universities to facilitate the rapid and efficient worldwide dissemination of research. Its services include gathering publishers' copyright and archiving policies (RoMEO), research funders archiving mandates and guidelines (JULIET), a worldwide directory of Open Access repositories (OpenDOAR) and simple full-text search of UK repositories. SHERPA's main values are Open Access, copyright and IPR, advocacy issues, preservation, licenses and policies, legal aspects and comprehensiveness. Along with its partners' experience and expertise, SHERPA aims to “offer the ideal environment for exploring and testing ideas for repository development, which can be evaluated and disseminated to the wider community”.

¹²³ League of European Research Universities, “League of European Research Universities”, 2010. <http://www.leru.org/index.php/public/home/>

¹²⁴ League of European Research Universities, *The LERU Roadmap towards Open Access*, Advice paper No. 8, June 2011. http://www.leru.org/files/publications/LERU_AP8_Open_Access.pdf

¹²⁵ League of European Research Universities, *Open Access to Research Publications*, no date. http://www.leru.org/files/publications/Open_Access_to_Research_Publications-FINAL.pdf

¹²⁶ League of European Research Universities, *Open Research Data*, no date. http://www.leru.org/files/publications/Open_Access_to_Research_Data-FINALdocx.pdf

¹²⁷ Science Europe, “About us”, 2013. <http://www.scienceeurope.org/>

¹²⁸ Science Europe, “Policy at Science Europe”, 2013. <http://www.scienceeurope.org/policy/policy-2/>

¹²⁹ SHERPA, “Securing a Hybrid Environment for Research Preservation and Access”, 2006. <http://www.sherpa.ac.uk/>

According to the SHERPA/JULIET the data archiving policy¹³⁰ requires deposition of unpublished data, research data and program code in Open Access archives within 6 months after the project completion in any appropriate repository (e.g. GenBank¹³¹, PDB¹³²). Examples of data include: nucleotide/protein sequences, macromolecular atomic coordinates and anonymised epidemiological data. Preferably the deposit should be immediately after the publication of results.

The Data Seal of Approval¹³³ was established by a number of institutions committed to the long-term archiving of research data. By assigning the seal, the DSA community seeks to guarantee the durability of the data concerned, but also to promote the goal of durable archiving in general. The Data Seal of Approval is granted to repositories that are committed to archiving and providing access to scholarly research data in a sustainable way.

The main goal of the European Research Council¹³⁴ is to encourage high quality research in Europe through competitive funding. ERC's values are focused on scientific excellence, high quality research, bottom up approach, visibility to the best brains in Europe, high quality peer-review, international benchmarks of success, up-to-date information of who is succeeding and why and the establishment of better strategies. The ERC considers that providing *free online access* to research outputs is the most effective way of ensuring that research findings can be accessed, and re-used for further research. It requests that electronic copies of any research papers and monographs that are funded in whole, or in part, by ERC, to be made publicly available as soon as possible, and no later than six months after the official publication date of the original article. The ERC strongly encourages their funded researchers to make their publications available in Open Access, using discipline-specific repositories. If none is available, researchers should make effort to make their publications available in institutional repositories or on their own webpage. The ERC considers it essential that primary data, as well as data-related products such as computer codes, is deposited in the relevant databases as soon as possible, preferably immediately after publication and in any case not later than six months after the date of publication. The ERC encourages Institutions to cover Open Access fees of any research papers and monographs that are supported in whole, or in part, by ERC funding which arise in the period up to 24 months after the end of a grant. The ERC reminds ERC funded researchers that Open Access fees are eligible costs that can be charged against ERC grants.¹³⁵

4.3.5 A Study on Subject-Specific Requirements for Open Access Infrastructure – the Case of OpenAIRE

This study¹³⁶ addressed subject-specific requirements for research infrastructure with a focus

¹³⁰ European Research Council, “ERC Scientific Council guidelines for open access”, 2007. <http://erc.europa.eu/documents/erc-scientific-council-guidelines-open-access>

¹³¹ National Center for Biotechnology Information, “Genbank Overview”, 1 April 2013. <http://www.ncbi.nlm.nih.gov/genbank/>

¹³² Research Collaboratory for Structural Bioinformatics, “Protein Data Bank”, no date. <http://www.rcsb.org/pdb/home/home.do>

¹³³ Data Seal of Approval, “About the Data Seal of Approval”, no date. <http://www.datasealofapproval.org/>

¹³⁴ European Research Council, “European Research Council”, no date. <http://erc.europa.eu/>

¹³⁵ European Research Council, *Open Access Guidelines for researchers funded by the ERC*, 2012. http://erc.europa.eu/sites/default/files/document/file/open_access_policy_researchers_funded_ERC.pdf

¹³⁶ Open Access Infrastructure for Research in Europe, “Studies on Subject Specific Requirements for Open Access Infrastructure”, 12 January 2012. <http://www.openaire.eu/en/component/content/article/11/335-studies-on-subject-specific-requirements-for-open-access-infrastructure>

on the influences of Open Access. Open Access is treated in a broad sense covering Open Access to literature, Open Data and open science. The study took a case-based approach and six partners (institutions and organisations representing Health and Life Sciences, Information and Communication Technology, Environment data, Socio-economic sciences and Humanities, and e-infrastructures) were chosen to provide their subjective view on Open Access infrastructure.

The study, whilst focused on research infrastructures, raises very important points regarding diversity within scientific practices and how to accommodate varied practices within the Open Access agenda, which seems to aim for standardisation. The comparative case analysis highlighted how the characteristics of the research life-cycles (data collection, processing, enriching, archiving and re-using), along with literature and data management shows a large variety in both tooling and practice. The comparative analysis shows that Open Access to literature is a growing or established practice in the subject areas but not yet fully developed. Open Access to Data is considered an important future activity. This indicates that an Open Access infrastructure can be built immediately and in a rather generic sense for literature and has to be built with more patience and consideration for subject-specific requirements in the future. The key challenge is seen as the development of research infrastructure that operates in an open mode and, thereby, supports the diversity of research practices through increased information flows between subjects.

Digital literature and data resources are an essential precondition of research. The provision of digital literature and data resources through infrastructural services are perceived as a matter of course (or implicitness) and are not questioned unless they are obviously missing. The study raises the following points regarding Open Access to data:

- Open Access is described as a *modus operandi* for working with digital literature and data resources rather than as an end in itself or an ethical principle.
- Open Access to literature and Open Access to Data refer to very different parts of the research process. While literature shows universally generic characteristics, data is much more related to subject-specific methodologies and facilities. Even though the benefits are the same for literature and data, the obstacles vary broadly and require that Open Access to literature and Open Access to Data are differentiated in policy and infrastructure development.
- Open Access to Data has (yet) to be reflected in a fully subject-specific way in policy and research infrastructure development. The emerging practice of mandatory project-specific *data management plans* that address the question of Open Access to data could be sharpened by asking the question: “Are data open and if not, why not?” Open Access in data management plans could be supported by providing a generic Open Data policy with subject-specific *addendi* to such a generic policy. Such a subject-specific addendum to a generic Open Data policy may well be mandatory in a given subject area.
- The difference between Open Access to literature and Open Access to Data may be transient as more and more systematic connections between literature and data can be observed. Explorations towards infrastructural linkage between literature and data (e.g. enhanced publications) should be intensified.¹³⁷

¹³⁷ Meier zu Verl, C., and W. Horstmann, “Studies on Subject Specific Requirements for Open Access Infrastructure – Attempts at a synthesis”, in C. Meier zu Verl, and W. Horstmann (Eds.) *Studies on Subject-Specific Requirements for Open Access Infrastructure*, Universitätsbibliothek, Bielefeld, Chapter H, 2011, p.5.

4.4 CONCLUSIONS AND DISCUSSION

This review clearly found an overall drive for Open Data access within the policy documents, which is part of a wider drive for open science in general. The values underpinning this move are the view of science as an open enterprise, where knowledge is sought and where discovery rests on scientists working together to solve specific challenges, which increasingly are becoming interdisciplinary in nature. The argument for publically funded science to be open to the public is also strong, although it is not entirely clear often how this openness should be operationalised.

When discussing Open Data there is a clear tendency to refer to science as a whole sector, thus there is a danger that differences between disciplines are ignored in further policy making. Each discipline has different methods for gathering data, and analysing it. Data may be images, numerical, narrative, statistical and presented in small, medium or large datasets, interlinked – or not. Some disciplines deal with sensitive data, others with data that may have IPR or legal issues attached. It is important that these differences be acknowledged in further policy for Open Data as it will inform the debate about whether we require subject specific requirements, or common infrastructure for open data access.

One of the key distinctions one sees in reviewing the literature about Open Access is that it is addressed differently by stakeholders in the research ecosystem. High level policy makers focus at the very general level and argue for Open Access in terms of very broad social and economic benefits as well as seeing it as a development that will improve science. In this case little attention is paid to the research infrastructure, governance and practice. Further, some of the claims made lack strong evidence and some claims have yet to be fully proven. For example, the genomic exemplar that is cited as evidence for the benefits are open to question given that genomic research is not hypothesis based and is largely foundational at the moment. There is as yet little evidence that other science will benefit in the same way as genomic science.¹³⁸ The approach to Open Access to Data amongst funders is not as broad as seen by the high policy makers. However, funders increasingly are motivated to ensure that the allocation of publically funded research yields good value for money. To this end, funders have encouraged, or made it mandatory, that research data is deposited in a national data service so that it can be accessed openly. Funders are encouraging the research community to make data open in this way as well as moving towards Open Access in publishing whether at the Green or Gold level.

Stakeholders from within the infrastructural, libraries, repositories, and associated services focus largely on their particular role in enabling Open Access to Data. These stakeholders see value in Open Access to Data as a way of improving the means by which data is made more accessible, and they are motivated to meet the needs of Open Access within their business cases and service provision. For some stakeholders, such as LIBER, there is a clear value base in their organisational view that all information should be made freely available to everyone. There is however debate as to how these changes and demands will be funded, with the costs and benefits still unclear.

<http://www.openaire.eu/en/component/content/article/11/335-studies-on-subject-specific-requirements-for-open-access-infrastructure>

¹³⁸ Berman, Francine and Vint Cerf, “Who Will Pay for Public Access to Research Data?”, *Science*, Vol. 341, No. 6146, 9 August 2013, pp. 616-617.

Publishers are adapting to the open publishing environment and are developing new types of business models to facilitate that. Here the question of where the cost for Open Access publishing will rest is still undecided. There is resource in grant provision, but further exploration is needed in this area. When one moves towards the area of ethics and governance one starts to see that despite a belief in Open Access there are detailed concerns about how this can be carried out ethically. This is especially the case when dealing with humans and human subject matter. Here one starts to move closer to the details of research practice and the specificity of data, which starts to open up the meaning of Open Access in terms of scientific practice as well as in terms of the economic value of science. We explore some of the details of Open Access from a practice based approach below but a key point to register here is that values, motivations and barriers to Open Access are interwoven with these practices. This is because at this level the consequences of making data open are often recognised through deeply held knowledge of data, and its ethical and governance frameworks. This is not to say that scientists do not support Open Access to Data in principle, rather they can identify the issues involved in ensuring that making data open is done so in a responsible way, both at the level of the integrity of the data and at the level of scientific practice (which includes reputation and reward processes). What one sees therefore from these reviews is that Open Access to Data means different things to different stakeholders and there is some fragmentation. The need for a co-ordination at a high policy and implementation level is to be recommended. For instance in Australia the Australian National Data Service has been very well funded to support and develop Open Access to Data across the country.

The reviews therefore provide an indication of the values and motivations for Open Access across the science ecosystem. It is however important to note that whilst the Open Access agenda is being driven by values and perceived benefits, as yet there is a lack of robust evidence that making data open will improve science and yield economic and social benefit. The reviews when aligned with the stakeholders in Open Access also provide an overview of fragmentation in the development of Open Access, which can then be better addressed.

5 CASE STUDY RESEARCH: VALUES, MOTIVATIONS AND BARRIERS TO OPEN DATA – THE VIEW FROM SCIENTISTS WITHIN FIVE SCIENTIFIC DISCIPLINES

In addition to stakeholder mapping and document review, this WP draws on five case studies, which aim to further understanding of the values, motivations and barriers to Open Data from the viewpoints of practicing scientists from five different disciplines: particle physics and astrophysics, health and clinical research, bioengineering, environmental research and archaeology. The aim of the case studies is to ascertain if and how values, motivations and barriers, identified in the document review, map on to these scientific fields. The objective is to bring an added depth to the on-going discussion of Open Data by interviewing practicing scientists about what Open Data means to their work and their respective fields. Firstly, we present findings from each case study, in order to present the specificities of each field, before discussing both generic and specific findings of values, motivations and barriers to Open Data implementation as presented to us in the case study research.

5.1 PARTICLE PHYSICS AND PARTICLE ASTROPHYSICS: THE PPPA GROUP, THE UNIVERSITY OF SHEFFIELD

The case study on particle physics took place within the Particle Physics and Particle Astrophysics Group of the Department of Physics and Astronomy at the University of Sheffield, UK. Five physicists were interviewed who work on range of experiments and projects such as the Large Hadron Collider project ATLAS, Neutrino Physics, Dark Matter Research, Gravitational waves and Astrophysics.

5.1.1 Research practices within Physics

Much of the research is large scale collaborative research with tens or even hundreds of international partners. The data is numerical, computer generated and analysed using complex and custom made equipment, hardware and software, which is built and programmed by the scientists. Much of the research data is generated by particle physics detectors in which the tracking and measuring of particles is undertaken, in some instances over a long period of time. These detectors produce large quantities of data, which is stored, processed and analysed. Experiments like ATLAS are kept running 24/7 over a period of a few years, and researchers work in shifts monitoring the detector and any events. Analysis of the vast quantities of data produced each year by ATLAS cannot be undertaken with a single desktop computer or a single large supercomputer. Therefore, thousands of interconnected machines around the world are used to form a virtual computer called the Grid. The data from ATLAS is hundreds of petabytes and is stored in double copy.

5.1.2 Values and motivations

The disciplines of Particle Physics and Particle Astrophysics (PPPA) are both recognised as areas that are on the cusp of making great discoveries. These discoveries are seen to be at the heart of the quest to understand the Universe both at very small scales, the building blocks of matter, and at the very largest, and the structure of the cosmos itself. Consequently, the stakes are perceived as high, for the individual scientist, the research collaboration, as well as society itself. The ATLAS experiment for example is designed to study the fundamental

constituents of the Universe with goals including understanding the origin of mass and the nature of dark matter.¹³⁹

PPPA in many instances relies on large scale research equipment e.g. the Large Hadron Collider at CERN, Telescopes and Detectors. Consequently, many of the projects carried out within the PPPA group are carried out by large and multinational consortium of partners. Scientific collaboration is thus commonplace but due to the competitive element within some subfields, data sharing is limited.

‘One of my research areas, the main one is gravitational waves so there is some worldwide effort to detect gravitational waves directly using optical pyrometers and they are very expensive and cost in the order of \$100 million each. You have to collaborate in this kind of project obviously, so I am involved in the Ligo Scientific Collaboration involving some 60 institutions involving 3 of these detectors and it also has links to the other detectors. It is a worldwide group of scientists working together.’ (Senior Researcher, Physics)

Motivations for moving towards open data access within the fields of PPPA were described by respondents as a way of easing access to data, for the purpose of comparison, error testing and avoiding duplicate efforts.

‘If you combine the 4 experiments you have 4 times as much data and means you naively have halved the errors, which is a worthwhile gain in terms of its usefulness for theoreticians. Similarly if you are trying to exclude the production of the Higgs Boson at LEP it may be that different experiments have better sensitivity to different channels, one better at electrons and one at neutrinos so by combining all of them you get a better limit than you would with any individual one...’ (Senior Researcher, PPPA)

As experiments are carried out using different calibrations and parameters, there is an apparent benefit of having access to data from other researchers, and one respondent described to us the value of using Open Access data, for one of her PhD student’s work, to generate new knowledge.

‘The other thing is people trying to do an experiment who would like data with a different base line. What my student was doing was trying to use data from 2 experiments to look at the possibility of additional types of neutrino that don’t interact, so called sterile neutrinos. Our original plan was that we were going to combine these data with data from T2K but then decided T2K did not have enough data’ (Senior Researcher, PPPA)

The PhD Student had access to a variety of data, but in order to be able to re-use it for the purposes of his research needed first-hand knowledge from the data producers, due to lack of metadata and context.

5.1.3 Barriers to implementing open data

Although most of our respondents were positive about data sharing, with exception of the competitive element discussed earlier, within the scientific community, they doubted whether

¹³⁹ University of Sheffield, “Research in Particle Physics and Particle Astrophysics”, no date. <http://www.hep.shef.ac.uk/research/>

this data was suitable for Open Access, due to a lack of interest and relevance to the general public, and due to the enormous size of many of the datasets that were yielded through many of the experiments.

Making physics data, e.g. from the ATLAS experiment, would require the hiring of a number of full time staff, as well as high costs in order to prepare the data and make it meaningful for Open Access. One researcher claimed that doing all that work would mean a waste of already dwindling funds for scientific research:

“You might end up wasting millions of pounds and then only 10 people are interested, that is a waste of money and work. A small number of interesting events could be interesting for someone. For example if you talking about climate change data, that is interesting for many people but in ATLAS it is really complicated so if someone gets the processed data and carefully chosen it can be done but to access to all the data, it is not doable or interesting.” (Senior Researcher, PPPA)

The physicist in question told us that a selection of ATLAS data was made available for outreach purposes and to introduce particle physics and the experiments in schools. A small proportion of the data was then prepared, processed and annotated for students to play with. A full time member of staff undertook that work, and if Open Data access would be implemented, it would mean a great increase in staff numbers.

With the types of physics data, as well as the quantity there are specific challenges when considering Open Access. The following quote from a physicist working on Neutrino research outlines well the complications of sharing raw physics data.

“If you were to try to make the raw data available to outsiders you would have to make available the raw data, the reconstruction programs, the simulation and its database, the programs that handle the simulation and you would need to ensure there was access to the physics generators which are usually written by theorists and not the experimenters who wrote the experimental simulation and it just gets too complicated.” (Senior Researcher, PPPA)

In the few instances that Open Data was reportedly used, it was difficult to make sense of the data without speaking to the physicist who had carried out the experiment at the time, in order to clarify the context within which the experiment was conducted. Although access was straightforward, the data was neither useable nor intelligible without an explanation.

“The data was public but you needed an insider’s viewpoint to use it so it was not very public. It was not that they were being deliberately obstructive. It was just the problem of knowing the data well, then some things are obvious to you, they are not necessarily obvious to someone who does not know the experiment. It is like trying to write an instruction manual for something you do all the time.”

PPPA was portrayed as a competitive field, where a single event or discovery could bring enormous prestige to a research group. This is especially true of the search for the Higgs Boson, Dark Matter Particles and Gravitational Wave research. This undoubtedly will affect the willingness to share data, or make it Open Access, at least for the few years after each experiment.

‘gravitational waves are undiscovered and if we discover them for some people in the collaboration, not everyone, that will be the high point of their career and the main thing they get done in their career. So those people are wary of outsiders getting hold of the data in case they get scooped.’ (Senior Researcher, PPPA)

In the interviews with physicists it became clear that firstly to establish Open Access to the entire raw or processed data collection from large scale experiments, such as ATLAS, represents a substantial challenge from a technical and financial point of view; the data is too large.

5.2 HEALTH AND CLINICAL RESEARCH: THE FP7 PROJECT EVA

The case study on health and clinical research was conducted within the FP7 funded project EVA.¹⁴⁰ The EVA consortium aims to do ground-breaking research of chronic obstructive pulmonary disease (COPD) by bringing together clinical medicine, radiology, image analysis and genetics (including gene expression analysis), laboratory diagnostics and bioinformatics. The consortium consists of thirteen partners, including ten clinical partners, who provide cases and samples and two partners, to analyse the samples. We interviewed five scientists associated with the EVA project, whose research focus was in clinical studies, cell culture studies and gene expression.

5.2.1 Research practices in Health and Clinical Research

Health research is increasingly interdisciplinary, as can be seen by the mix of disciplines involved in the EVA project. Each discipline brings to the table their specific research practices and traditions, and different view of health, disease and patients. Health and clinical research benefits most from being able to link patient data from many sources, but complications arise when sharing genome and patient data due to legal and privacy issues. 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.

5.2.2 Values and Motivations

In the case of health research our respondents stated clear benefits of both sharing and opening access to data, in the sense that by having access to more research would lead to faster advancement within the field, and results would become more reliable, as they can be drawn from a bigger pool of data.

The benefits of Open Data Access were described as particularly relevant to clinical studies, where one respondent stated that there is a tendency to publish those studies only where a significant finding could be reported, rather than those where none was present. Therefore, openly sharing the data from those studies could prevent duplication of effort. This was also reported to be of great significance for patient groups, as patients would be presented with

¹⁴⁰ Emphysema vs Airways Disease, “Welcome to EvA”, 2008. <http://www.eva-copd.eu>

fewer tests and research participation requests. Health data, especially genome data, was also seen as particularly ripe for re-use and re-analysis due to its inherent qualities.

‘Genome information has a lot of content, some of it we know what it means, some we don’t - in the future we might learn what it means.’ (Director, Health)

Health and clinical project collaborations work with data in different formats, such as CT scans, gene expression data and measurement data. Data sharing and combining is common within collaborative projects but in many instances it takes specific forms, according to what type of data it is and what agreements there are between partners within the consortium. There is a tendency for each partner to firstly publish from their data before sharing it with others, and this is seen to stem from competition within the field of health and clinical science. One respondent described the research community as “open in a closed sense”:

‘This is something like semi open when you first say you want to publish your own aspects and then you go ahead and combine. That is something that is happening quite a bit in the genetic field because of power you need large numbers in many fields in genetics’ (Director, Health)

5.2.3 Barriers to implementing open data

Open Access was put into question by one of our respondents, who claimed that in order to fully understand and being able to work with data the user would need adequate metadata and information on how the data was generated. He claimed to have downloaded data via Open Access which included a spread sheet with “cryptic labels” which made little sense to him as a user, thus rendering the data unusable.

‘The important thing in my view is that just data is not everything you need. You need to know more about the data than just the numbers that have been generated in order to interpret that in a proper way. ... The methods and the context that has been used to collect the data are an integral part of the validity of the data.’ (Senior Researcher, Health)

The respondents agreed that, this would require extra work from the hands of scientists and that the benefit of doing so was unclear at this moment in time, as there are no reward systems in place for publishing data.

‘The disadvantage, especially with experimental work to put it in such a format that it can be directly and easily used by others, that involves quite a bit of work. That is in my view quite a substantial hurdle. It is one thing to put data in an Open Access data base but it is another thing to put it in such a way that you do not need an extensive explanation to be able to use it.’ (Senior Researcher, Health)

Ethical issues around data in the health and clinical sciences are complex and all respondents agreed that openness to data should be implemented with great care. Two main concerns were expressed in which respondents were concerned that although data was anonymised, there would be a way of linking it back to certain individuals. Firstly, that this could happen once researchers start integrating datasets from a number of studies, and secondly, if Open Access to genome information would be implemented a higher risk of identification would occur.

‘The concern is that by combining a number of data, which are very extensive when it comes to genetic traits combining it with another database, can lead to the identification of an individual. ... When you publish the entire genome of an individual and you make it publicly accessible then this whole genome is actually more than a fingerprint.’ (Director, Health)

Concerns over the sensitivity of health data led one of the respondents to reflect on what the meaning of ‘open’ would be when applied to data within the health sciences. Currently, the field was described as fairly open; however, access to ‘open’ data might require an application process or a more informal email request.

‘So it sounds as though it is open but in reality it is not open at all because I have to make an application to have a look at this data and then you move from there. You do the research and you do it on open data that you have to sign that you will keep confidential and sign all sorts of bits and pieces, which means it is not open at all. It is closed to a selected group of people who have a bona fide institution behind them’ (Senior Researcher, Health)

5.3 BIOENGINEERING: AUCKLAND BIOENGINEERING INSTITUTE AND THE VPH COMMUNITY

The case study on bioengineering was carried out with members of the Bioengineering Institute in Auckland, NZ¹⁴¹ and Virtual Physiological Human (VPH)¹⁴² Community. We interviewed six bioengineering scientists on the development towards open data within the field. The respondents all work on both modelling and experimental work based on physiological data. The institute’s work also encompasses computational physiology, instrumental medical devices and medical informatics. Some of the researchers work with patient data, but do not work actively with patients during the data collection phase.

5.3.1 Research practices in Bioengineering

Over the past decade, clinical medicine and biomedical science have been transformed by the emergence of bioengineering. Developments in imaging have enable the development of a characterisation of structure and function in cells, tissues, organs and the living body in exquisite 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. Bioengineering was described by respondents as a fundamentally open field, in terms of sharing models and code, but less so in the sharing of experimental data. There is some push from funders such as the NIH, EPSRC and the Wellcome Trust, to make data available for Open Access, and a recent trend is that once a model is submitted for publication, some publishers now ask that it is also made available via Open Access.

5.3.2 Values and Motivations

Openness and sharing of data and models are very important to the field of bioengineering, as it is in essence an international and collaborative effort toward understanding physiological processes and the diagnosis and treatment of injury or disease.

¹⁴¹ Auckland Bioengineering Institute, “Our Research”, no date. <http://www.abi.auckland.ac.nz/en/about/our-research.html>

¹⁴² VPH Institute, “Welcome to the VPH Institute”, 2012. <http://www.vph-institute.org/>

'I think one of the drivers is the fact that because you want to combine the work of many different groups throughout the world who are working on different aspects of understanding physiological mechanisms ...There is no way that you would be able to create this larger collaborative effort without people buying into an Open Access framework'(Director, Bioengineering)

Our respondents agreed that the field of bioengineering is very open and models, code and data are quite freely shared. It seems that the access to experimental data, however, is less developed and one respondent attributed it to the level of competition that exists within bioengineering, and the amount of work it takes to annotate and prepare data so that it makes sense to users.

All our respondents mentioned the model sharing platform CellML¹⁴³, which allows for access to computer-based mathematical models. Like with data sharing, there are however specific issues that need to be taken into account before a model is submitted.

"if you want to use this model in combination with another one and import it into a larger model that uses models from different places and with different layouts you need to have the models annotated so that people are referring to using the same terminology when they are talking about a particular biological process. You cannot combine things if people are using different terms to talk about the same thing."

The scientists were generally positive about sharing and the potential of sharing data via Open Access. They however stated that in order to share data, they would like have published from it first. Furthermore, in some cases ethical issues may arise if the experimental data is derived from patients. Data derived from animals is however easier to share as there are no ethical constraints to contend with.

5.3.3 Barriers to implementing open data

Many of our respondents agreed that the move to Open Data made sense for bioengineering, as an already open field. However, some of them expressed concerns regarding how Open Data would fit into an already established culture of scientific conduct.

'I would say that the impediments there are mainly to do with trying to engender a culture where scientists get used to annotating their data according to well specified ontologies and metadata constructs that can then be deposited into these general, freely available data repositories so they are then searchable. It is going to take a change in culture and a change in funding and attitudes to encourage the extra effort to do the annotation and make data available via Open Access' (Director, Bioengineering)

As the quote above illustrates, the reward for the individual scientist, who puts in extra work to make data meaningful for Open Access, is unclear. One of our respondents went so far as to say that this could prove the most difficult barrier to tackle when a move towards Open Data was considered within the field of Bioengineering. Although the data comes in digital format and is thus easily shareable, annotation, assigning of metadata and clarification of research design and context, would require a lot of work. Respondents also raised concern

¹⁴³ CellML, "The CellML Project", 2013. <http://www.cellml.org/>

that this data would not be meaningful to anyone outside of the field, as a potential user would need in-depth knowledge, expertise and specialised software in order to understand and work with the data at hand.

As some of our respondents were not directly involved in gathering patient data, they were careful when discussing the potential sharing or Open Access to research data, which involved use of data from human subjects.

‘I think, certainly from my perspective, that makes me very careful about sharing data, particularly because I am not being directly involved in the collection of the data and the ethical situation surrounding the recordings themselves.’ (Senior Researcher, Bioengineering)

When it comes to depositing experimental data, there is no one clear choice of repository, as CellML is for the sharing of models. Available Open Access Data is thus made available on individual or institutional websites, which may render search and use of data difficult. Furthermore, the longevity of the data is not ensured. One respondent told us that the data that is currently available for Open Access tends to be data that is more suitable for education rather than research and analysis. It seems to suggest that for furthering Open Access Data within the field there is a lack of common infrastructure to host and search for available data.

5.4 ENVIRONMENTAL RESEARCH: JRC, EUROGEOSS AND INSPIRE

This case study addresses environmental research and in particular multidisciplinary interoperability models for heterogeneous datasets. Interviews were carried out with six researchers within the Joint Research Centre, Institute for Environment and Sustainability, Digital Earth and Reference Data Unit.¹⁴⁴ Within the group of respondents were mostly technical researchers working on data access and interoperability, and legal issues regarding licencing of satellite data.

5.4.1 Research practices in Environmental Research

The mission of Centre 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. The JRC coordinates the scientific and technical development of the European Directive INSPIRE (Infrastructure for Spatial Information in the European Community)¹⁴⁵ establishing an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment. The centre also contributes to GEOSS, the Global Earth Observation System of Systems¹⁴⁶, which is a global initiative, with around 80 nations and international organisations co-ordinating and sharing information on the earth. JRC led a

¹⁴⁴ Joint Research Centre, “Digital Earth and Reference Data Unit”, European Commission, 15 February 2012. <http://ies.jrc.ec.europa.eu/the-institute/units/digital-earth-and-reference-data.html>

¹⁴⁵ Joint Research Centre, Institute for Environment and Sustainability, “INSPIRE - Europe's infrastructure for spatial information”, European Commission, 22 February 2012. <http://ies.jrc.ec.europa.eu/our-activities/support-for-eu-policies/inspire.html>

¹⁴⁶ Joint Research Centre, Digital Earth and Reference Data Unit, “GEO”, European Commission, 23 February 2010. <http://ies.jrc.ec.europa.eu/DE/de-our-work/policy/policy-geoss.html>

research project contributing to the development of GEOSS by building multidisciplinary interoperability across the thematic areas of Drought, Forestry and Biodiversity.

None of the respondents within the case study were working with data in the ‘traditional’ sense, i.e. that they themselves had gathered through the research process. Some of the respondents deal with environmental data (spatial and measurement data) coming from research projects, and public authorities within the EU Member States, with the aim of providing a single access point to discover data through metadata information. Some of the technical staff furthermore, also work with satellite data provided by the Member States, or space agencies world-wide. Furthermore they act as support services for searching and accessing metadata of datasets and services created and handled by other JRC scientists for research activities and Impact assessment analysis. As data are derived from many different sources and have different purposes, different limitations of access and use are attached with each data set.

5.4.2 Values and Motivations

As a field, environmental research is multidisciplinary, and addresses complex challenges, and thus requires diverse data derived from different scientific fields and sub-fields, e.g. biology, agricultural research, weather research, oceanography. Data is thus of great value to the on-going development within the field, and the key motivation for making data available through Open Access is the prospect of integrating datasets from different disciplines to produce new knowledge.

As much of environmental research is focused on studying the development of trends (e.g. weather) over time, access to data ranging from different time periods and different geographical locations is imperative. Access to data, integration and re-use are thus key values that underpin environmental research.

‘(Opening access to data is important for) fostering multidisciplinary research. (...) Also, the more you can make the data not only accessible in physical terms but also when documented, when explained, when linked to publications that explain how the data was collected and used and what was the context of data. I think all of these are important aspects. That obviously has a broader value for research and society in making much more transparent and clear what are the basis the research and findings are based.’ (Director, Environmental Case Study)

The key motivation for advancing Open Access to environmental data was identified to be first and foremost, speeding up the progress of environmental research to better grapple with the global challenge of environmental change. This was seen to require a concerted effort from wider society; hence public access to environmental data was seen important, as well as motivating industry to use this data for commercialisation, to help with the above mission, as well as producing growth in times of austerity.

“there is a belief that by providing full and Open Access to data there is also an opportunity to foster private sector developments, services that will create more jobs. So there is an economic dimension to open data, which I think at the moment is really very strong given the financial situation in the western world” (Director, Environmental Case Study)

5.4.3 *Barriers to implementing open data*

Integrating large datasets from different sources, as our respondents' work involves, with the aim of providing access to them is a huge technical and legal undertaking. In many instances, access is somewhat controlled in that a user will need to request access to the data and will be granted a data licence, which state the limitations of use. The group wants to move further towards Open Access Data, but due to sensitive nature of much satellite and environmental data, they are aware that this may not be entirely possible, and that access to these datasets will always be under some sort of control.

“some of the JRC-IES Units work with data or produce data quite sensitive that need expertise to be interpreted: release geographical data on flood risk to the general public could be misinterpreting, the same for data on chemical concentration/exposure measure in specific location.” (Senior Researcher, Environment Studies)

The respondents agreed that it is not only access that they are providing but also data services, e.g., data management, cleaning, providing a permanent reference in the form of an URI, and making sure that metadata and models for understanding the data are also made available. It makes little sense to Open Access to a data set that is unusable due to lack of metadata and context. Also, each data set must have a permanent location for the purposes of correct citation. One of the respondents reported that some researchers are concerned about the opening of access and integration of datasets due to fear that information and data may be misreported or misunderstood. It is therefore the role of the unit to communicate with researchers and institutions in order to present their data accurately.

“If you talk about these huge terabyte sets which satellite data is, there are technical issues such as the datasets must be interoperable, we must have accuracy standards so you can make relevant overlay analysis or GIS analysis or modelling and it has to be correct geometrically so you know what you are looking at ... so a lot of this preparation of the data and making the data map correct is an important issue” (Environment Case Study, Technical Staff)

Data management and long term data preservation was an issue that one respondent claimed to be very important when thinking about opening access to and storing any data. There are, as of yet, not specific standards for storing data that state how long data should be stored and in which format.

“this brings us to a huge chapter which is challenging for our project and that is what we call long term data preservation. How should all this be built up, I am talking here about satellite data but you can project this on to all types of data. You need to have rules for this long term archiving and there are rules set up but they are not completely finalised” (Environment Case Study, Technical Staff)

The respondent mentioned that once data reaches 5 years in age, it is sometimes stored on a cheaper storage device to save on costs. Access to that data may then still be open but not immediate.

Integrating, and opening access to multidisciplinary research data, derived from different producers and owners, does invite complexities regarding licencing, ownership, IPR and

copyright issues. With regard to the work that the case study representatives carry out these are apparent on a daily basis:

‘The data are so different that even from the legal perspective if we are not just talking about the issue of licensing itself or other legal ways of making Open Access each data type in the first is regulated by some other field of law e.g. the privacy law or security law or other fields of law that in the first place can restrict access to the type of data that they regulate and only based on these types of restrictions can we talk about opening the data up to other users’ (Data and Licencing Legal Expert, Environment Studies)

5.5 ARCHAEOLOGY: OPEN CONTEXT

Open Context¹⁴⁷ is a free, Open Access resource for the web based publication of diverse types of research datasets from archaeology and related disciplines. Open Context is developed by the Alexandria Archive Institute¹⁴⁸ and backed by the California Digital Library¹⁴⁹, USA. We interviewed seven archaeologists who work with Open Context, alongside working on archaeology in the traditional sense, i.e. on excavations, research and documentation. Open context focuses on data publishing and adding value to existing datasets by use of computer and web technologies, which allows for integrating datasets and linking of different data formats to present a holistic view of archaeological data.

5.5.1 Research practices in archaeology

Traditionally, archaeologists work on site excavations, which yield data in many different formats, e.g., diaries, artefacts, spatial data, images and statistical data. According to our respondents, excavations can take years and are done in phases, which render the journey from data collection to possible publication of results very long. Archaeology was described as a very personal undertaking, which drives specific and individual research practices, rather than standardisation in language, terminology and measurement.

5.5.2 Values and Motivations

Data has a particularly strong value for archaeology - as one respondent stated:

“most fundamentally when you excavate a site you destroy it, so the documentation record in your excavations is really the only record that anybody will be able to use again in the future, because they can’t go back to the original site because it has been destroyed in the excavation process.” (Senior Researcher, Archaeology)

Consequently, the ability to re-use, re-analyse and integrate datasets was described as being very valuable for the advancement of archaeology. Archaeological knowledge from any one site is cumulative, as excavations are carried out in phases and often over the period of many years or even decades. Hence, in relation to any one site there may be a vast amount of data available, which can be held by different archaeologists, within different institutions.

¹⁴⁷ Open Context: Web-based research data publication, “Welcome to Open Context”, no date. <http://opencontext.org/>

¹⁴⁸ The Alexandria Archive Institute, “The Alexandria Archive Institute”, no date. <http://alexandriarchive.org/>

¹⁴⁹ California Digital Library, “California Digital Library”, 2013. <http://www.cdlib.org/>

The data that archaeologists yield and work with comes in different formats, and interlinking is necessary to make it meaningful, through creating a narrative and establishing context:

“you have structured data, amenable to statistical analysis, for example spread sheets, relational databases and that sort of thing. You have a lot of images, both drawings and photos. You also get a lot of narrative text, which are excavation diaries that describe the thinking of the excavators as they are going down and their rationale around that is supposedly useful to have an idea of what the excavators are thinking as they are working.” (Senior Researcher, Archaeology)

According to one of our respondents, data sharing is currently not common practice within the field of archaeology, which may result in slower process and possibly duplication in research efforts.

‘Most of the ways people get what they call legacy datasets within the world of archaeology happen when someone has passed away.’ (Junior Researcher, Archaeology)

The length of the research process, results in a delay in publication of results, which is seen to slow down the process of archaeology as a scientific field. Having access to data and results more quickly was described to have immediate benefit for the individual scientist, especially junior researchers, and for the discipline as a whole as access to diverse data is seen to fuel new discoveries.

The key motivation for Open Data Access in Archaeology is linked to the inherent researcher practices, which are grounded in the method of excavation. Excavation of any one site can take years, even decades, where researchers re-visit the same site more than once, or visit sites which have been excavated by other teams of researchers. Hence, the knowledge and data about any one site is cumulative, and there would be great value of archaeologists having access to data of earlier excavation teams, e.g. for comparison purposes. Furthermore, as different subsets archaeology may work on the same site (e.g. zoo archaeology or archaeobotany) there is a definite value in integrating data gathered by different sub-disciplines to gain a holistic knowledge of any one site, and make new links between knowledge sets.

5.5.3 Barriers to implementing open data

Archaeology is faced with the challenge of organising vast amounts of data gathered by excavations, some of which may date back a few decades (i.e. legacy data), which in some instances may require digitization before it can be opened for access. Different data formats from any one excavation will need integrating and linking, and metadata will need to be ascribed. The meaning of archaeological data is seen to arise from its context, therefore a dataset, where metadata including parameters, terminology, research questions, coding systems and methodologies are not clear, or missing, is likely to be un-usable, un-intelligible and un-assessable by others.

“people can do limited work with a dataset that is not well documented. I do worry that people will just think “Oh I need to archive my data and then it’s done” but it goes beyond archiving. The key question here is, “is the data set going to be re-useable 40 years down the line, when you are not around anymore?” Just because it is archived, does not mean it is re-usable.” (Director, Archaeology)

Thus preparing data for Open Access does require substantial work, especially when some time has passed from the data collection phase to the data publishing phase. This is, in part, due to how personal and idiosyncratic archaeology datasets can be, which is built on a long history and tradition of archaeology practice. For example, excavation diaries may come in handwritten format, spread sheets may have missing codes, or contain abbreviations that may need further explanation from the authors. Also, research parameters may be unclear.

“the interesting issue was that one datasets would have molluscs in it and another dataset from a nearby site would have none and the question would be, did this site not have any molluscs in it, or was it that the molluscs were being analysed by another researcher and were not present in the dataset. So these issues can be very challenging when you get into data integration” (Senior Researcher, Archaeology)

Preparing a relational and linked database from excavation data, which in some instances may be derived from several excavations carried out over a long period of time, requires a specific skillset (e.g. programming, coding, archiving, data maintenance), which, according to our respondents, is currently not common within archaeology.

“There is no programme that teaches archaeologists how to interact with these tools or code or programme. If someone wants to do it you have to pick it up on your own, it is not really taught unless it conforms to the existing tools that are adopted.” (Junior Researcher, Archaeology)

There is thus a perceived need for the training within archaeology, and also services focused on preparing, archiving, publishing and curating archaeological data for Open Access.

Issues of research funding were raised throughout the interviews, and fell within two broad themes: firstly, the respondents pointed out that the cost and work required to make data amenable for Open Access was not realised by funding bodies.

“If Open Access is supposed to be sustainable something has to change in terms of that – it is crazy that a grant will pay hundreds of thousands of dollars for excavations over many years and then when it comes time for dissemination there is no budget for data or publications.” (Director, Archaeology)

Secondly, respondents raised the issue of sustainability of data storage and hosting, pointing out that in many instances the future funding of repositories and data platforms was insecure, as it was run on short term funding. Consequently, it was therefore not clear to scientists that their data would be secure and managed long term.

Within archaeology, there are also complex ethical and political issues to contend with when Open Access is considered. Archaeology deals with human remains, religious and/or valuable artefacts and sacred/historically significant sites. Displaying data about these via Open Access may not be possible without posing risk to the site and/or archaeologist in question.

“A country like Peru they have a really large problem with looting and a lot of the looting happens with local armed groups. It is not a safe place.” (Junior Researcher, Archaeology)

“Another sensitive issue (is) cultural heritage collections from native American or indigenous communities in South America who would not appreciate you having online pictures of ritual objects for example or objects only certain members of the tribe are allowed to see or skeleton material which for many are very sensitive” (Senior Researcher, Archaeology)

Hence, the move toward Open Data will need careful consideration, and decisions will need to be made on which data will be released for Open Access, and which may require access control, or even complete confidentiality.

5.6 OVERALL VALUES, MOTIVATIONS AND BARRIERS EMERGING FROM THE CASE STUDIES

Throughout the case studies there was a clear echo of the values and motivations expressed in the document review. Overall, researchers within all fields were positive about openness to data within their respective fields, but remained sceptical about the practicalities, as will be detailed below.

As became evident, fields vary greatly in terms how open they are, in terms of sharing within the research community and by Open Access. While archaeology was presented as a field with relatively limited sharing of data, bioengineering was very open, in terms of sharing models and methods. The scientists we interviewed recognise the benefit of having access to increased amounts of data as it has the potential of driving faster advancement of science and discovery within their fields, as well as reduce duplication of effort. Results were described as more reliable, the more data you could draw from (health and clinical studies) and the capability for error testing in physics was described to increase with data from other experiments.

The majority of the scientists we interviewed were partners in larger research collaborations, and recognised the importance of collaboration and sharing of data and methods when it comes to solving increasingly complex problems (environment studies, health and clinical research and bioengineering) and to fund large scale equipment for experiments (PPPA). In some instances these collaborations were seen to foster sharing (bioengineering) and others, due to competition within the field, the collaborations in some instances had built memorandums around the sharing of data.

5.6.1 Potential Barriers to Implementing Open Data within Science

The scientists, although positive about Open Data, expressed a variety of concerns, deriving from their own personal experiences and knowledge of scientific practices within their respective field.

Competition within the Field of Science

Competition within science was mentioned frequently throughout all of the case studies, and it was seen to drive specific practices, some of which could remain barriers to successfully implement Open Data Access.

Competition for prestige and funding was seen to make scientists reluctant to publish data openly due to the fear of being scooped. This was also seen to hinder successful sharing of data within research communities and collaborations. Therefore many respondents suggested that a certain time limit on openness would need to be set up, to allow the scientist to publish their findings first. However, within archaeology this would remain an issue, as excavations may take place in phases over a few years, and the journey from data collection to publication can be very long.

Competition within science was also seen to make scientists reluctant to make data available for Open Access, due to the work it would take to make data meaningful, by annotation, application of metadata and necessary context. This extra work would take up time from other research activities such as data collection, analysis, publications and applications for funding, all of which bring clear and demonstrable rewards and benefits to scientists and their careers. Publishing data still remains largely unrecognised as a valuable scientific activity, which will help with career progression and merit.

Lack of Funding for Data-Related Activities

As data publishing is not seen as a worthwhile scientific endeavour, funding bodies, all the while increasingly requesting data management plans and that data be made openly available, are seen to lag behind in funding the actual activities that are needed for making data available through Open Access. This is especially seen as a detriment to the move towards Open Data in archaeology, as the field is both grappling with vast amounts of un-digitised data and dwindling funds overall. Other fields, such as PPPA also reported that to fully comply with the move towards Open Data, there would be increasing need for funds for storage, and staff costs.

Increasing Interdisciplinarity within Science

Although the scientists recognised the benefits of interdisciplinary collaboration within science, this was seen to bring a set of potential barriers to the move towards Open Data. Those scientists who worked within interdisciplinary collaborations (health, bioengineering, environment research) recognised that different disciplines may have different attitudes to data and their subject of research. This became especially apparent in the health and clinical research and bioengineering case studies where ethical considerations come with the use of data derived from human subjects. Furthermore, environmental research grapples with the reality of often combining research data with data from public authorities, which come with a set of legality issues surrounding the use and integration of data.

Sensitive Data

Working with sensitive data caused many of our respondents to hesitate when it came to implementing Open Access to Data. The types of data that cause concern fell in to three broad strands: data derived from human subjects, culturally sensitive data and location data. Human data, as used in health and clinical research and bioengineering is currently subject to strict ethical guidelines and laws regarding privacy and confidentiality. Furthering access to this data might prove problematic, as it may take a lot of work to fully anonymise it and make sure that it cannot be traced back to individual research participants.

Location data can be sensitive as was presented to us by the case studies of archaeology and environment research, where in the latter, complex legal frameworks and licencing issues were presented as linked to sensitive location data. Archaeology can deal with sensitive cultural issues such as religious or cultural sites, as well as human remains and providing Open Access to this data would be ill possible without possible repercussions for the archaeologist in question, as well as the site itself, due to concerns over looting.

The Context of Research Data

What became apparent in all case studies is that Open Access to research data does involve opening access to the context of which the data is collected and analysed. It will not suffice to simply make access to data open by putting raw datasets in a repository or posting it online. What the case studies drew our attention to is that research data does not stand alone and in order for it to be meaningful for any potential user, standardisation of language, clarity of annotation, models, methods, metadata and context will need to be supplied. This may in many instances require vast amount of work for which there is no funding and for which benefits and rewards are currently lacking.

6 FINDINGS FROM THE VALIDATION WORKSHOP

In the validation workshop the findings of the document review and the case studies were presented by the RECODE team and by representatives from three of the case studies (Open Context, EVA, and PPPA). The findings were then discussed in with the 38 participants in group sessions and then in open whole group discussion.

In general terms the participants validated the presented findings and each could identify with the different dimensions of Open Access to Data. The key areas that emerged in the discussion were that the development of Open Access to Data is beginning to generate some level of consensus at policy level and at the level of supporting services such as developments in repositories, data archives and data management plans. The participants knew about the EU agenda on Open Access as well as national level agendas. The overall consensus amongst the workshop participants that there were benefits in making data openly available although the benefits as yet were not fully known across the whole range of scientific and scholarly disciplines. There was also concern that the costs of making data open were as yet still not fully known, and that attention needed to be paid to the development of business cases for Open Access to research data. Participants did express surprise at the fragmentation within disciplines and not just between disciplines. They had previously assumed that there would be disciplinary differences in terms of Open Access and they were surprised that there were also sectarian divisions within disciplines. This highlights one of the key conclusions of the workshop that echoes the findings of the case studies, which is that openness is context-specific. For example, some areas (e.g., astronomy) are completely open, whereas others (e.g., genomics) could be seen to be very asymmetric, with the data user gaining much more than the data provider. In part this relates to 'fair use' – that the generator of the data should be able to exploit their work before making the data openly available – and in part it is a consequence of the reward system being heavily biased towards the publication of research outputs.

The workshop participants raised a series of points about some of the details of making data open. Their attention on these details reasserts our broad finding that stakeholders need to pay close attention to the details of making data open throughout the research ecosystem. This is significant for two reasons: a) for researchers to support Open Access they need to trust that the Open Access system will protect their data and will ensure that the data is used responsibly; b) Open Data needs to be made meaningful for new users for the full value of Open Data to be realised, which requires the context of specific data to be made transparent.

Some of the more specific points that the participants raised can be categorised in the following way: policy level co-ordination; open data quality, management, security and storage; Open Data in the research process and research reward systems.

The main policy points made were that there should be a policy for all research institutes to link and store data and that data should be stored in open formats so that it is accessible to anyone, no matter what infrastructure they have access to. Another point made was that storing data might act as a first policy step, and then there could be an intermediate period before mandatory data sharing that will enable institutions to build an infrastructure. Policy could thereby mandate a step by step process. Incentivising data re-use was also seen as an important aspect of policy.

Many of the comments related to Open Data quality issues. These include: that plagiarism needed to be prevented and the institution of some sort of mechanism to identify plagiarism is important. Further, for data to be useful, the process which generated the data must be known, for example the Standard Operating Procedures that are used in genomics.

In terms of Open Data and the research process one suggestion was that there should be a simplified way for researchers to understand what will happen to their data, this for example might be something like a creative commons. The data use policy should be “human readable”, not overly technical or full of jargon so that is quickly and easily understood by data users. To develop Open Access to Data a reward structure is essential, through for example licensing and data citation license (most Workshop attendees appeared to be unaware that mechanisms for this were available through DataCite). There is also an educational aspect to Open Access, which is that researchers and other stakeholders need to be trained in how to design intelligence into data collection, analysis and storage methods to facilitate sharing and preservation. Finally, there is a tension between the discourse of “data as public property” and “data as a financial resource”. In this context, researchers will not be incentivised to share if it puts them at a financial or recognition based disadvantage while funders want to maximise their investment in scientific research. In conclusion, the workshop validated our findings in terms of the stakeholder policy and also in terms of the perspectives of researchers. There is a consensus that Open Access to research data may have benefits there is however concern about some of the details that need to be considered in making data open. There is also concern about how best to incentivise the use of Open Access and how to reward researchers who make their data open.

7 INTERNATIONAL ADVISORY BOARD COMMENTS

The members of the International Advisory Board were sent a draft of this WP1 Deliverable prior to e-meetings to discuss the Deliverable. 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) and Boyong Wang as representative for Dr Cao Jing (Handan City-EU Affairs representative; Dr Jing's representative for EU related projects in China).

Each Board member thought that WP1 had captured the key points from the review of Open Access to research. The Board also said that it had identified key points in developing Open Access to Data from the point of researchers.

At the level of policy, Toby Burrows pointed out that in Australia there had been a concerted effort to co-ordinate the development of Open Access to research data. In Australia it is also recognised that there is an overlap between government data and research data. The development of Open Access to both research data and government data is being done by the Australian National Data Service. This organisation is organising the development of Open Access to research data and is also co-ordinating that with developments in open government data. There has been a significant amount of funding put into Open Access to data through the Australian National Data Service to develop services, create support for Open Access and to seek solutions for sustaining Open Access to Data. In terms of ensuring that Open Data is useful the Australian National Data Service is just in the process of rolling out a system that will provide context about the data set, this will include profiles of the researchers who produced the data, details of the grant that supported the collection of the data and a list of publications based on the data. This will ensure that the data is meaningful for re-use and it goes some way to providing the context of the data.

Max Craglia raised similar issues to Toby Burrows. He stressed the need at the European level to co-ordinate Open Access to research data with Open Access to government data. He felt that the European Commission needs to co-ordinate across both these areas of Open Access and that it needs to bring both sets of stakeholders together to facilitate a more co-ordinated approach. He argued that there needs to be closer co-ordination also at the level of Open Access to research data, and that Open Access needs to be seen as a 'bundle of activities' that involve data management, curation, data quality and data citation to ensure Open Access to data generates a public good.

Jerome Reichman raised some pertinent points in terms of cost and in terms of incentives. He raised some of the difficulties that are emerging in the development of science based on large data and large collaborations. In particular he raised the issue of the cost of preserving data, which he felt had not been fully worked through. He also noted that the example of genomic data as a case for making data open needed to be treated with care. This is because genomic research is still in a foundational phase and was a non-hypothesis science. How this might translate to science based on hypothesis was still to be understood. He also noted that in terms of making data open, that 'open-ness' needed better definition. His research indicates that open might be defined by a 'whole commons', which is totally open and by a 'semi-commons', where data is restricted to particular users. In terms of the take-up of Open Access to data, he and his co-writers argue that this will need to be based on what he calls

‘reciprocal benefits’ between researchers¹⁵⁰. This will evolve through a research culture where groups, departments, and/or research areas find that they mutually benefit from making data open to each other.

Boyong Wang reiterated the main points made above and he said the main barrier was to create a culture of data sharing and openness. In his experience in making government data open the main barrier was encouraging people to share what they perceive to be ‘their data’. Other issues such as technology, storage and curation need to be addressed but the main point is that data producers – and indeed users – of data need to trust that any system of Open Access to research data (and government data) will ensure that the provenance, sustainability, context and responsible re-use of data.

¹⁵⁰ Reichman, Jerome H., Paul F. Uhler, and Tom Dedeurwaerdere, “Governing Digitally Integrated Genetic Resources, Data, and Literature”, in *Global Intellectual Property Strategies for the Microbial Research Commons*, Cambridge University Press, 2014 (forthcoming).

8 DISCUSSION

In our case study research we found that the values, motivations and barriers in the development of Open Access to data were tightly related to the practice of research. In this discussion we first give an overview of the scientific process and then discuss the values, motivations and barriers to Open Access. This underpins one of the main points that emerged from our case studies and their relationship to the document reviews, which is that to understand how the values, motivations and barriers relate to Open Access to data is through the way they are embedded in practice.

The case studies show clearly that each area of science has a distinctive approach that shapes the way research is undertaken and the way in which data is generated, managed, interpreted, stored and shared. There is however an overarching theme across the case studies. This is that whatever kind of science is being practiced the focus on a data pathway or data journey is at the heart of what it means to be undertaking research within the scientific frame of reference. The focus of different research areas even within disciplines as well as in the emerging areas of multi- and interdisciplinary research means that the production of data follows different paths but each path follows a typical set of scientific stages: identification of research issue; decisions about which ontological and epistemological approach and aligned methodology to take; research design; methods to be used; data to be generated; data management, storage and processing; data interpretation including validation and reliability process; identification of findings; peer review process; publication and other dissemination processes; depositing of data. This process helps to ensure that both the production of the data and its interpretation are rigorous and that scientists can make certain claims with confidence in the strength of their data. By following this process, findings have a certain authority, which may be refined or even refuted through further research.

What this process demonstrates is that data is created and used within certain knowledge frameworks that involve technical aspects of managing specific types of data; knowledge that enables interpretation of the data, and can assess the significance of findings; ethical and governance frameworks in creating and using data; and the value of data. This process operates within broader interoperability and technical frameworks, institutional frameworks and ethical frameworks, and publication and reward processes. The specificity of particular areas of research clearly indicates that data does have unique characteristics, which require particular approaches to data pathways. The distinctiveness of each area of study, however, has implications for the way data is managed, is shared and then can be made open.

Another slight variation is the relationship between overarching scientific practice and the way research scientists are rewarded. The overall model is that scientists gain prestige, reputation and promotion through publication. Good data forms the basis for publication and data provides the evidence for making original contributions to the progression of science. This mechanism for promotion and reward means that data are an important for career progression and such a competitive environment does not lend itself to fostering sharing data and making it open to others who may then gain competitive advantage.

The specificity of the data is also a significant factor in how data might be made openly available. Data varies in terms of size and type. For example, huge data, such as physics data requires very expensive processing and storage resources; environmental data is often drawn from a variety of datasets and types and requires a way to make those data sources interoperable; bioengineering requires complex models to work with the data; and

archaeology involves a bespoke mix of data in relation to each specific site. These types of distinctions means that the way in which data can be made open will require different processes, tools and governance frameworks. To summarise this point data is not a 'one size fits all' concept. A significant consequence of the last point is the impact different data characteristics have on the cost of making data open. There are other related points that need clarification, which include having some consensus regarding definitions of what we mean by 'open' and whether we have different levels of openness to data.

Comments from our validation workshop illustrate some of the issues surrounding data pathways, which are both specific to particular data but they also relate to overarching issues in making data open. One area concerns the ethical, political and security issues surrounding Open Access to Data. For example, the issues of ensuring patient confidentiality in medical and related health research and the sensitivity of ritual sites in archaeology and the risks of looting of archaeological sites. These types of issues raise the broader concern regarding ethical, political and security issues in various types of data. These concerns indicate that there are areas that may require controlled level of access to data and the removal of sensitive information. Another area to address comes under the remit of 'ownership and sharing' and here there are still issues of ownership (copyright) and confidentiality of data that limit ability to make the data openly available.

Related to this issue is that the use of (existing) licensing schemes is essential for open sharing and re-use of data. Further, workshop participants felt that there is much to learn from areas such as the Creative Commons¹⁵¹ and Reproducible Research¹⁵² communities in developing Open Access to Data. A further area of concern relates to the size, complexity and distribution of some data and the consequences of these issues in developing accessibility to broader communities and indeed to make them openly available. The cost of being able to access datasets that are extremely large, and some are distributed across extensive computing networks, is very high in that individuals would need access to high powered computing to access the data, thus these data may well only be of value to very well resourced groups with great expertise.

There are questions surrounding the curation, validation and reliability of data when making it open. There are significant overheads for curation and validation of datasets, and currently there is ambiguity about how these costs will be funded. There is an added dimension, which what processes need to be in place to ensure that open data is reliable. The context of data and the implicit knowledge about data and its interpretation needs to be conveyed in some way, and the key issue here is how to share knowledge that researchers accumulate often over a lifetime to remote users of the data. In summary, data does not exist in isolation - understanding and re-use of the data requires implicit knowledge to be made available as well, the process by which the data is generated must be fully described, and for derived data (e.g. climate modelling outputs), the raw data, computational models, and software frameworks and so on are all necessary for understanding the data.

Thus, although there is a high-level push for Open Access to Data, and broad support for Open Access to Data by the research community provided it is done intelligently with the appropriate checks and balances, the nature and context of data must be addressed in all its detail.

¹⁵¹ Creative Commons, "Creative Commons", no date. <http://creativecommons.org/>

¹⁵² Reproducible Research, "How to", 13 January 2009. http://reproducibleresearch.net/index.php/How_to

8.1 OVERVIEW OF RESEARCH PRACTICES IN MAKING DATA OPEN

The above description of practice provides a context with which to understand not only the values and motivations but also the barriers to making data open. The case studies found that the researchers in all the case studies saw value in making data open if it was done in an appropriate way. The value of Open Access to Data varies between areas of research in terms of the way a particular research field relates to its stakeholders and broader community and society. Thus for instance those working in environmental sciences are often more proactive in developing Open Access to Data because they draw on a wide range of data themselves and research tends to have a tighter relationship with action research and policy agendas within the broader society-wide environmental agenda. Physics is embedded in a different set of research relations and it does already share data amongst its large research community network. Although it produces data products for various types of outreach work, it does not directly make access to its data open to all. The largeness of physics data and the cost of producing and storing it – as well as the highly specialised knowledge required to interpret it means that make the data open will be a very expensive process. Further, it will require extensive explanations about how to interpret it as well as access to extensive computational resources to access the data. Bio-engineering exists within another set of relations in that it works across both the human biological sciences and the physical engineering sciences, and each has a different set of ethical and governance relations. The way data can be managed within the research teams means that complex ethical issues can be discussed and dealt with. However, making this type of data open raises real issues in terms of ethics. The case of archaeology highlights the way in which not all data is highly organised and designed for computational analysis. The variety of artefacts, different evidence bases, records and so on that are also site-specific shows that data comes in many different forms. This is the case not only within a discipline but that data can be highly varied within specific projects. In order to make this type of data open will require various platforms and links so that the different data can be assessed but that also they can be linked in the appropriate way to ensure that different data objects are located and understood within the overall framework of any project.

The practice-based findings raise several questions in the development of Open Access to data. These include:

- What are the requirements in developing Open Access to Data in terms of making access to data meaningful?
- Do we need clearer definitions of ‘data’ (raw, derived, processed etc.) and if so, what would be the implications for developing Open Access to Data?
- What are the issues in accessing Open Data in terms of the responsibilities of the accessing user? Should we develop licensing process to ensure we know what the data is being accessed for?
- How do we ensure an ethical framework for Open Access to Data that goes beyond data management and Open Access publishing guidelines?
- Is there a relationship between the Royal Society four principles - accessible; intelligible; assessable; and usable - in developing Open Access to Data? If so, what are the key aspects of each area that link with other aspects in the other areas? Do we need to find a way of linking developments between each of these areas to ensure that Open Access to Data is taken forward in a coherent way, in other words, do we understand the dependencies between these principles?
- How can we address researcher career paths in terms of Open Access to Data?

- Should we develop frameworks/guidelines of what data can be released when? Should there be bodies that regulate this?
- What are the 'real' costs of Open Access to Data?
- How important are knowledge communities in ensuring that data is understood properly and is used properly?
- What kind of peer review system is needed for Open Access to Data? (c.f. Open Source software and the peer review process for Open Access publishing)

9 CONCLUSION

The aim of WP1 is to provide a broad overview of the values and motivations in the development of Open Access to Data across a wide range of stakeholders as well as gaining an understanding of Open Access from the point of view of researchers. To address the broad overview of values and motivations we created a functional taxonomy of the stakeholders in the Open Data ecosystem that mapped out the functions of the Open Data ecosystem and the performers who conduct those functions, the activities they conduct and the types of records they generate. WP1 addresses the ecosystem of stakeholders at the level of document review. To address the perspectives of researchers, who produce data, we used a case study approach that focused on the micro-level of research practice. Researchers are positioned in various ways to stakeholder groups in the Functional Taxonomy, they are part of stakeholder groups such as universities, research institutes, professional bodies, and so on, and they also draw on repositories, libraries and publishers in the research process. What became clear in the case study work is that researchers also sit in a unique position in the Open Data ecosystem because of their in-depth and detailed knowledge of the production of data. The concerns the scientists raised in WP1 highlight the importance of understanding the specific needs of different types of data if it is to be made open in a responsible way.

In terms of values and motivations across the broader ecosystem of stakeholder values in Open Access and data dissemination and preservation there is an overarching consensus of the value of making data open. Amongst the groups working across the functional taxonomy there is a general value consensus that Open Access to Data will improve the productivity and quality of scientific work, Open Access to Data also may have the potential to yield economic benefits, and there is a clear sense that Open Access to Data is a general public good. Institutions and organisations across the funding and initiating, creating, disseminating, curating and using categories are addressing Open Access to Data from their respective perspectives. The stakeholders from within each functional area are, however, also aware of the practical issues in developing Open Access such as issues of interoperability, ethical and research governance issues, institutional issues and so on. What is evident is that the stakeholders across the taxonomy are taking steps to prepare the research ecosystem for Open Access. The development of data management protocols and ethical protocols, for example, are creating an awareness of best practice in managing data that underpins future Open Access procedure.

What has become evident from our case studies is that the values and motivations of the stakeholders in the Open Access ecosystem are understood differently by researchers, and understood differently from within specific disciplinary and interdisciplinary perspectives. The research community at the level of the researchers engages with each function in the taxonomy from the point of view of undertaking research, producing data and sharing and disseminating findings. Each operates within a specific framework of disciplinary ethics and governance and all operate within a broad remit of research integrity. Even though in principle, researchers can see the value of Open Access there is a range of attitudes to Open Access to Data that are based on the practical issues, ethical sensibilities and reward systems that form the framework of the practice of 'doing science'. From a researcher perspective there are five concerns that are potential barriers to developing Open Access to Data, which apply across the research ecosystem. One area of concern is competition for prestige and funding, which makes scientists reluctant to publish data openly due to the fear of being scooped, and they recommend that certain time limits on openness would need to be set up so that they can publish their findings first. Another area scientists point to is the amount of

work that involved in making data meaningful in Open Access such as annotation, application of metadata and necessary context, which currently is not recognised or funded in the way it will need to be in an Open Access context. Further, publishing data is largely unrecognised as a valuable scientific activity and will need to be recognised and rewarded in the way publication is. The case study researchers also point out that funding bodies are not funding the work required in data management plans that will enable data to be openly available. There are also concerns about making sensitive data open, which makes researchers hesitant about making such data open. The types of data that cause concern fell in to three broad strands: data derived from human subjects, culturally sensitive data and location data. Fifth, researchers feel that it is important that the stakeholders across the broader Open Data infrastructure need to be aware that Open Access to research data also involves opening access to the context of which the data is collected and analysed. It will not suffice to simply make access to data open by putting raw datasets in a repository or posting it online. What the case studies draw our attention to is that research data does not stand alone and in order for it to be meaningful for any potential user, standardisation of language, clarity of annotation, models, methods, metadata and context need to be supplied. This may in many instances require vast amount of work for which there is no funding and for which benefits and rewards are currently lacking.

WP1 has identified that in terms of possible conflicting value chains and stakeholder fragmentation that there are two parallel processes emerging which need to be interlinked in order to align the practice of science with the boarder ecosystem of Open Data and its stakeholders. The conflict of values between these processes is not in terms of the general principle of Open Access to Data, and some of the potential benefits of Open Access. Rather, there is concern about ensuring that the current quality processes are maintained and improved in science, in the care of data, interpretation of data, and researchers and knowledge communities. The knowledge of particular research communities of their data and how to manage and interpret that data is detailed, and in the development of Open Access to Data, the stakeholder groups in the ecosystem need to be aware of these details and need to be able to address them. Certainly, in terms of the broader ecosystem stakeholders are addressing the requirements for Open Access and each stakeholder is beginning to gauge a sense of the new interdependencies that are currently emerging. In particular, funders, institutions and various repository-based groups are beginning to create an overall framework of data management, curation, and storage of data. There are still many ambiguities about what types of relations will emerge between stakeholders, and in particular there is a lack of clarity about where the cost of Open Access will be leveraged. The key issue is ensuring the participation of the research community in further implementation of Open Access so that the detail of making data open in ways that make it accessible; intelligible; assessable; and usable. Without these four Royal Society points, the full benefits of Open Access may not be realised.

10 REFERENCES

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

Alliance Permanent Access, “About APARSEN”, no date. <http://www.alliancepermanentaccess.org/index.php/aparsen/>

Amsterdam University Press, “Journal of Archaeology in the Low Countries”, 2013. www.jalc.nl

Angrist, Misha, “Genetic privacy needs a more nuanced approach”, *Nature*, Vol. 294, February 2013.

Arts & Humanities Research Council, “Deposits of resources or datasets”, 2012. <http://www.ahrc.ac.uk/Funding-Opportunities/Research-funding/RFG/Annexes/Pages/Deposits.aspx>

Arts and Humanities Research Council, *Digital Transformations in the Arts and Humanities: Big Data Research Call for Proposals*, July 2013. <http://www.ahrc.ac.uk/Funding-Opportunities/Documents/Big-Data-Projects-call-document.pdf>

Association of European Research Libraries, “Association of European Research Libraries”, no date. <http://www.libereurope.eu/>

Association of European Research Libraries, “LIBER Strategic Plan 2013-2015: Re-inventing the Library for the future”, 2012. <http://libereurope.eu/strategy>

Association of Research Libraries, “ARL Strategic Plan 2010-2012”, 2012. <http://www.arl.org/storage/documents/publications/strategic-plan-2010-2012.pdf>

Association of Research Libraries, “Homepage”, 2013. <http://www.arl.org/>

Auckland Bioengineering Institute, “Our Research”, no date. <http://www.abi.auckland.ac.nz/en/about/our-research.html>

The Austrian Science Fund (FWF), “Open Access Policy bei FWF-Projekten”, no date. http://www.fwf.ac.at/de/public_relations/oai/index.html (in German)
http://www.fwf.ac.at/en/public_relations/oai/index.html (in English)

The Austrian Science Fund (FWF), “Aktuelle Information”, 2012. http://www.fwf.ac.at/de/aktuelles_detail.asp?N_ID=506

Beagrie, Neil, Julia Chruszcz, and Brian Lovoie, *Keeping research data safe: a cost model and guidance for UK Universities*, Final Report to JISC, JISC, London, 2008. <http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>

Beaulieu, Anne and Paul Wouters, “E-research as intervention”, in Jankowski, N. (ed.) *E-research: Transformations in Scholarly Practice*, Routledge, New York, 2009, pp. 54-69.

Berman, Francine and Vint Cerf, “Who Will Pay for Public Access to Research Data?”, *Science*, Vol. 341, No. 6146, 9 August 2013, pp. 616-617.

BioMed Central, “Open Data”, 3 September 2013.
<http://www.biomedcentral.com/about/opendata>

Biotechnology and Biological Sciences Research Council, “Safeguarding good scientific practice”, June 2006. <http://www.bbsrc.ac.uk/organisation/policies/position/policy/good-scientific-practice.aspx>

Bromley, Dennis B., *The case-study method in psychology and related disciplines*, John Wiley & Sons, Chichester, 1986.

Budapest Open Access Initiative, “Read the Budapest Open Access Initiative”, 2002.
<http://www.budapestopenaccessinitiative.org/read>.

California Digital Library, “California Digital Library”, 2013. <http://www.cdlib.org/>

CellML, “The CellML Project”, 2013. <http://www.cellml.org/>

The Charity Commission, “The regulator for charities in England and Wales”, no date.
<http://www.charitycommission.gov.uk/>

Council of European Social Science Data Archives, “Manifesto of the New Global Data Generation”, 2012. <http://www.cessda.org/about/manifesto.html>

Council of European Social Science Data Archives, “Depositing Data – Benefits”, 2012.
<http://www.cessda.org/sharing/depositing/1/>

Creative Commons, “Creative Commons”, no date. <http://creativecommons.org/>

Creative Commons, “CC0 1.0 Universal (CCO 1.0) Public Domain Dedication”, no date.
<http://creativecommons.org/publicdomain/zero/1.0/>

Data Seal of Approval, “About the Data Seal of Approval”, no date.
<http://www.datasealofapproval.org/>

DataCite. “Helping you find, access and reuse data”, 2009.
<http://www.datacite.org/whycitedata>

Det Frie Forskningsråd; Danish National Research Foundation; Højteknologifonden; Det Strategiske Forskningsråd; Rådet for Teknologi og Innovation, “Open Access policy for public-sector research councils and foundations”, 21 June 2012.
http://dg.dk/filer/fonden/open_access/Final%20Open%20Access%20policy%20English.pdf.

Deutsche Forschungsgemeinschaft (DFG), “Open Access und Forschungsförderung durch die Deutsche Forschungsgemeinschaft”, aktualisierungsdatum, 24 January 2012.
http://www.dfg.de/dfg_magazin/forschungspolitik_standpunkte_perspektiven/open_access/index.htm.

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

Directorate General for Research and Innovation, *Survey on Open Access in FP7*, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-open-access-in-fp7_en.pdf

Digital Object Identifier System. “The DOI System”, 2013. <http://www.doi.org>

Digital Repository Infrastructure Vision for European Research, “DRIVER”, 2013. <http://www.driver-repository.eu/>

Digital Research Infrastructure for the Arts and Humanities, “DARIAH-EU”, 2013. <http://www.dariah.eu/>

Economic & Social Research Council, *ESRC Research Data Policy*, September 2010. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

Emphysema vs Airways Disease: The EvA Project, “Welcome to EvA”, 2008. <http://www.eva-copd.eu>

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

European Commission, Open Data, an engine for innovation, growth and transparent governance, COM(2011) 882 final, Brussels, 12 December 2011.

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF>

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

European Commission, *Biobanks for Europe: A Challenge for Governance*, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/biobanks-for-europe_en.pdf

European Commission, Towards better access to scientific information: Boosting the benefits of public investments in research, Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions, COM(2012) 401 final, Brussels, 17 July 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

European Commission, *Scientific data: open access to research results will boost Europe's innovation capacity*, IP/12/790, 17 July 2012. http://europa.eu/rapid/press-release_IP-12-790_en.htm

European Commission, “Digital Agenda for Europe: Open Data”, 2013. <https://ec.europa.eu/digital-agenda/node/70>

The European Organisation for International Research Information, *Rome Declaration on CRIS and OAR*, 2011. <http://www.eurocris.org/Documents/RomeDeclaration.pdf>

European Research Council, “European Research Council”, no date. <http://erc.europa.eu/>

European Research Council, “ERC Scientific Council guidelines for open access”, 2007. <http://erc.europa.eu/documents/erc-scientific-council-guidelines-open-access>

European Research Council, *Open Access Guidelines for researchers funded by the ERC*, 2012. http://erc.europa.eu/sites/default/files/document/file/open_access_policy_researchers_funded_ERC.pdf

European Science Foundation, “Open Access in Biomedical Research”, *Science Policy Briefing 47*, September 2012. http://www.esf.org/fileadmin/Public_documents/Publications/spb47_OpenAccess.pdf

Fonds de la Recherche Scientifique (FNRS), “Reglement du comite dáccompagnement”, 9 December 2010. http://www.frs-fnrs.be/uploaddocs/docs/SOUTENIR/FRS-FNRS_Reglement_Comite_Accompagnement_2011.pdf

Fonds de la Recherche Scientifique (FNRS), “Reglement Adopte par le conseil dáministration du F.R.S. – FNRS”, 7 December 2012. http://www.frs-fnrs.be/uploaddocs/docs/SOUTENIR/FRS-FNRS_Reglement_Commissions_Scientifiques_2013.pdf

Fonds de la Recherche Scientifique (FNRS), “Diffusion et publications”, 31 March 2013. <http://www.frs-fnrs.be/fr/financer-les-chercheurs/ecoles-doctorales-congres-publications/diffusion-publications.html>

Fry, Jenny, Suzanne Lockyer, Charles Oppenheim, John Houghton and Bruce Rasmussen. *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes*, JISC, London, 2008. http://repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf

Government of Canada, “Comprehensive Brief on Open Access to Publications and Research Data for the Federal Granting Agencies”, June 2011. <http://www.science.gc.ca/default.asp?lang=En&n=2360F10C-1>

Government of Canada, “Canadian IPY 2007-2008 Data Policy”, 14 November 2011. http://www.api-ipy.gc.ca/pg_IPYAPI_055-eng.htm

HM Government, *Open Data White Paper: Unleashing the Potential*, Cabinet Office, London, 2012. <https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential>

International Federation of Library Associations and Institutions, “Statement on Open Access to Scholarly Literature and Research Documentation”, 2013. <http://www.ifla.org/publications/ifla-statement-on-open-access-to-scholarly-literature-and-research-documentation>

Jankowski, Nicholas W. (ed.) *E-Research: Transformations in Scholarly Practice*, Routledge, New York, 2009.

JISC, “Open Access for UK Research - JISC’s contributions”, 2 March 2012. <http://www.jisc.ac.uk/publications/programmerelated/2010/openaccessmainbrochure.aspx>

Joint Research Centre, Digital Earth and Reference Data Unit, “GEO”, European Commission, 23 February 2010. <http://ies.jrc.ec.europa.eu/DE/de-our-work/policy/policy-geoss.html>

Joint Research Centre, “Digital Earth and Reference Data Unit”, European Commission, 15 February 2012. <http://ies.jrc.ec.europa.eu/the-institute/units/digital-earth-and-reference-data.html>

Joint Research Centre, Institute for Environment and Sustainability, “INSPIRE - Europe's infrastructure for spatial information”, European Commission, 22 February 2012. <http://ies.jrc.ec.europa.eu/our-activities/support-for-eu-policies/inspire.html>

Kuipers, Tom and Jeffrey van der Hoeven, *Insight into digital preservation of research output in Europe*. Final Report of the PARSE.Insight project, STFC, London, 2009, pp.4-5. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

League of European Research Universities, *Open Access to Research Publications*, no date. http://www.leru.org/files/publications/Open_Access_to_Research_Publications-FINAL.pdf

League of European Research Universities, *Open Research Data*, no date. http://www.leru.org/files/publications/Open_Access_to_Research_Data-FINALdocx.pdf

League of European Research Universities, “League of European Research Universities”, 2010. <http://www.leru.org/index.php/public/home/>

League of European Research Universities, *The LERU Roadmap towards Open Access*, Advice paper No. 8, June 2011. http://www.leru.org/files/publications/LERU_AP8_Open_Access.pdf

Lin, Thomas, “Cracking Open the Scientific Process”, *The New York Times*, 16 January 2012, p. D1.

Marques, David. “Research Data Driving New Services”. *Research Data Management*, Vol 1, No. 1, 2013. <http://libraryconnect.elsevier.com/articles/best-practices/2013-02/research-data-driving-new-services>

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

Meier zu Verl, C., and W. Horstmann, “Studies on Subject Specific Requirements for Open Access Infrastructure – Attempts at a synthesis”, in C. Meier zu Verl, and W. Horstmann (Eds.) *Studies on Subject-Specific Requirements for Open Access Infrastructure*, Universitätsbibliothek, Bielefeld, Chapter H, 2011. <http://www.openaire.eu/en/component/content/article/11/335-studies-on-subject-specific-requirements-for-open-access-infrastructure>

Murray-Rust, Peter, Cameron Neylon, Rufus Pollock and John Wilbanks, “Panton Principles, principles for open data in science”, 2010. <http://pantonprinciples.org/>

National Center for Biotechnology Information, “Genbank Overview”, 1 April 2013. <http://www.ncbi.nlm.nih.gov/genbank/>

Open Access Infrastructure for Research in Europe, OpenAIRE, no date. <http://www.openaire.eu/>

Open Access Infrastructure for Research in Europe, “Studies on Subject Specific Requirements for Open Access Infrastructure”, 12 January 2012. <http://www.openaire.eu/en/component/content/article/11/335-studies-on-subject-specific-requirements-for-open-access-infrastructure>

Open Access Publishing in European Networks, “Welcome to OAPEN”, no date. www.oapen.org

Open Context: Web-based research data publication, “Welcome to Open Context”, no date. <http://opencontext.org/>

Organization for Economic Cooperation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD, Paris, 2007. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

Organization for Economic Cooperation and Development, *Report on Cross-Border Enforcement of Privacy Laws*, OECD, Paris, 2007. www.oecd.org/dataoecd/17/43/37558845.pdf

Parcher, Jean, “Benefits of open availability of Landsat data”, Presentation, U.S, Geological Survey, 2012. www.oosa.unvienna.org/pdf/pres/stsc2012/2012ind-05E.pdf

Reichman, Jerome H., Paul F. Uhler, and Tom Dedeurwaerdere, “Governing Digitally Integrated Genetic Resources, Data, and Literature”, in *Global Intellectual Property Strategies for the Microbial Research Commons*, Cambridge University Press, 2014 (forthcoming).

Reproducible Research, “How to”, 13 January 2009. http://reproducibleresearch.net/index.php/How_to

Repositories support project, *Survey of academic attitudes to Open Access and institutional repositories – an RSP and UKCoRR initiative*, 2011. <http://rspproject.files.wordpress.com/2011/12/attitudes-to-oa-basic-summary-report.doc>

Research Collaboratory for Structural Bioinformatics, “Protein Data Bank”, no date. <http://www.rcsb.org/pdb/home/home.do>

Research Councils UK, “Excellence with Impact – RCUK Common Principles on Data Policy”, no date. <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>

Research Councils UK, “RCUK Policy on Open Access and Supporting Guidance”, 8 April 2013. <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf>

Research Information Network, The Finch Group Report, *Accessibility, sustainability, excellence: how to expand access to research publications: Report of the Working Group on Expanding Access to Published Research Findings*, London, 2012. <http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf>

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

Science Europe, “About us”, 2013. <http://www.scienceeurope.org/>

Science Europe, “Policy at Science Europe”, 2013. <http://www.scienceeurope.org/policy/policy-2/>

Securing a Hybrid Environment for Research Preservation and Access (SHERPA), “Research funders archiving mandates and guidelines (JULIET)”, 2013. <http://www.sherpa.ac.uk/juliet/index.php>

SHERPA, “Securing a Hybrid Environment for Research Preservation and Access”, 2006. <http://www.sherpa.ac.uk/>

Springer Open, “Availability of Supporting Data”, 2013. <http://www.springeropen.com/about/supportingdata>

Social Sciences and Humanities Research Council, “Research Data Archiving Policy”, Government of Canada, 9 June 2013. http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx

Swan, Alma and Sheridan Brown, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs, A report commissioned by the Research Information Framework*, RIN, London, 2008.

UK Data Archive, *Managing and Sharing Data, Best Practice for Researchers*, University of Essex, May 2011. <http://data-archive.ac.uk/media/2894/managingsharing.pdf>

UK Data Archive, “Create & Manage Data: Consent and Ethics”, 2013. <http://www.data-archive.ac.uk/create-manage/consent-ethics/legal?index=4>

University of Edinburgh. “Research data management guidance”. 20 September 2013. <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/documenting-data>

University of Leicester, “PREPARDE”, no date. <http://www2.le.ac.uk/projects/preparde>

University of Sheffield, “Research in Particle Physics and Particle Astrophysics”, no date. <http://www.hep.shef.ac.uk/research/>

United Nations Educational, Scientific and Cultural Organisation, *Policy Guidelines for the Development and Promotion of Open Access*, Paris, 2012. <http://unesdoc.unesco.org/images/0021/002158/215863e.pdf>

Vetenskapsrådet (Swedish Research Council), “Fri tillgänglighet till forskningsresultat – Open Access”, 15 February 2012. <http://www.vr.se/106.29b9c5ae1268d01cd5c80001275.html>

VPH Institute, “Welcome to the VPH Institute”, 2012. <http://www.vph-institute.org/>

The Wellcome Trust, “Expert Advisory Group on Data Access”, no date. <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/index.htm>

The Wellcome Trust, *The Wellcome Trust Sanger Institute Data Sharing Guidelines*, July 2010. http://www.sanger.ac.uk/datasharing/assets/wtsi_datasharing_guidelines.pdf
The Wellcome Trust, *Minutes of the Second Meeting of the Expert Advisory Group on Data Access (EAGDA)*, 19 October 2012. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp041115.pdf

Wiley, “Geoscience Data Journal: Vol. 2”, 2013. <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html>

APPENDIX 1- LIST OF WORKSHOP ATTENDEES' INSTITUTIONS

Representative from the following institutions attended the RECODE WP1 workshop on 4th September at the University of Sheffield:

Amsterdam University Press
Blekinge Institute of Technology
Centre for Research Communications
Consiglio Nazionale delle Ricerche, Istituto sull'Inquinamento Atmosferico (CNR-IIA)
Digital Humanities Research Institute, University of Sheffield
EDINA, University of Edinburgh (JISC-designated centre of expertise and centre for online services)
JRC data centre
Max Planck Digital Library
MyScienceWork
National Documentation Centre (EKT/NHRF)
Open Context
Open Knowledge Foundation
Oxford Internet Institute
Research Information Network
Sheffield University Library
Social Science Data Archive
TERENO Earth Observation network
The British Library
Trilateral Research & Consulting
UK Data Archive and Economic and Social Data Service
UK Economic and Social Research Council
University College London
Wellcome Trust Sanger Institute

APPENDIX 2 – RECODE WP1 WORKSHOP AGENDA

Policy Recommendations for Open Access to Research Data in Europe (RECODE) Perspectives in understanding open access to research data: stakeholder values and motivations in research ecosystems

4 September 2013 – University of Sheffield

Background: The *Policy Recommendations for Open Access to Research Data in Europe* (RECODE) project (funded by the European Union) is addressing the drivers and barriers in developing Open Data Access (ODA) in Europe to draft recommendations for policy. The first work-package explores stakeholder values and motivations by undertaking a review of policy and related documents, and by conducting 5 case studies that seek to understand ODA in research practice. The case studies are: particle physics; health research; bioengineering; environmental research; and archaeology.

RECODE partners: Trilateral Research & Consulting, Royal Netherlands Academy of Arts and Sciences (KNAW), The University of Sheffield, Stichting LIBER Foundation, National Documentation Centre, National Research Council of Italy, Blekinge Institute of Technology and Amsterdam University Press.

Aim of the workshop:

- Present key findings from a review of select European policy documents and related documents about Open Data Access and findings from the five case studies of research practice.
- Solicit participants' and case study partners' feedback on the effectiveness of these policies as well as any gaps to assess their significance for policy development and for research practice.
- Better understand how to match policies with stakeholder drivers and motivations to increase their effectiveness in promoting open access to research data.

10.00am Participant registration and coffee

10.30am Introductions to RECODE project (Kush Wadhwa, Trilateral Research & Consulting)

10.50am Panel discussion: exploring the values of scientific practice, along with key motivations for, and barriers to scientists partaking in Open Data Access. (Professor Rod Smallwood, Dr Bridgette Wessels, Dr Thordis Sveinsdottir (UoS) and case study representatives)

11.20 Group Discussion (Facilitated by a member of the RECODE team)

- What are the implications of different disciplinary value systems for open data access?
- What are disciplinary motivations for furthering open data access?
- What are the key disciplinary barriers to furthering the development of open data access?

12.30pm: Lunch

1.30pm: Presentation of key findings from the five case studies: Current research practices and their implications for advancing open data access (Thordis Sveinsdottir, University of Sheffield)

2.00pm Group Discussion (Facilitated by a member of the RECODE team)

- What lessons can we take from current research practices for advancing open data access?
- How can we take different disciplinary practices into account when considering open data access?
- Given the distinctiveness between disciplinary practices, how can we build generic guidelines around open data access?

3.15pm Coffee

3.30pm Group Discussion: How can we better design policy to address the issues identified so far? (Facilitated by a member of the RECODE team)

4.00pm Next Steps: Infrastructure and Technology in Open Data Access. Introduction to RECODE WP2 (Lorenzo Biagagli, National Research Council of Italy)

4.20pm Q/A Session

4.45pm Finish

7.00pm Workshop Dinner

APPENDIX 3 – INTERVIEW PROTOCOLS

The interviews aim to identify and examine researchers' views on open data access how it fits within the values of scientific practice and what the key motivations/barriers might exist in partaking in Open Data Access. The interview will begin by exploring participants' views on open data access, and the culture of sharing within your immediate research community. We will then explore how moving to open data access, and consequently opening up your data to a wider user group might raise different issues, e.g., regarding trust, interoperability etc. As open data access rests on the how amenable to sharing data is, we will also ask questions relating to the types of data you work with, how shareable it is and what preparation it may need before it is released for open access. The interview will conclude by exploring your experiences and thoughts on using open data access in your work.

Directors

- Where does your topic/research centre or group sit within the broader field/discipline?
- What are the key drivers for advancing open data access within your field?
 - (Open discussion of open data access – where you, and your research community they see the motivations and drivers coming from – how does that feed into your and your discipline's values)
- How does open data access fit in within your discipline's values? (values underpinning disciplines)
- What do you see as the value of open data access:
 - What value does open data access have for the individual scientist?
 - What is the value of open data access for your Research Community?
 - What is the value of open data access for society? (societal value –social and economic impact)
- Open data access within your field:
 - How advanced is it? What is the general understanding of open data access in your field
 - What is the value of open data access in your field, or is there value of open data access?
 - Are there any developments in open data access in your field? If so, what?
 - What is driving these developments? What are the motivations for developing open data access?
 - What do you see as the main barriers to advancing open data access further?

- How important is open data access for the sustainability of your centre/group? (E.g. for securing on-going funding etc)
- What are the practicalities associated for your group/centre to making your data accessible to your immediate research community?
 - Enablers? – (research community culture and norms, interoperability, technology, ethical and legal framework, and any other enablers?)
 - Barriers? (Research community culture and norms, interoperability, technology, ethical and legal framework, and any other barriers?)
 - If any, how have you overcome these barriers?
- What are the practicalities of making data publically accessible in terms of open access?
 - Enablers (interoperability, funders, discipline norms, copyright and IPR (e.g. creative commons), legal issues, licencing, changing research practice, governance structures and any other enablers you have found)
 - Barriers? (funders, copyright and IPR, legal issues, licencing, discipline norms, changes in research practice, governance framework and any other barriers you have found)
 - If any, how will you overcome these barriers?
- Does the prospect of open data access to your research data change your process of collecting, archiving, organising and cleaning it?

Senior Academics

- Are you aware of open data access within your area, if so what do you think about it? Are there any key debates/discussions on-going with regard to open data access?
 - What do you see as the advantages/disadvantages of making data openly accessible? How does it relate to the value of the work you are doing?
- What is the value of data for your discipline/project?
- How prepared are you to make your data open access?
 - Are you happy for all your data to be available via open access? (why/why not)
 - Is there any data you would rather keep to yourself/for your research group? If 'yes' why is it not appropriate to make it openly available?

- Are you comfortable with the idea that data users will be able to re-use and even integrate your data with other datasets for analysis? (why/why not)
- Does the prospect of open access to your research data change your process of collecting, archiving, organising and cleaning it?
 - Are you aware of data standards that you must adhere to? If so, what are your thoughts on these?
 - Are you aware of any interoperability issues that your data might raise?
 - Any technical issues that you need to overcome to be able to make your data open access?
- What is the value of data for your professional accountability and reputation?
 - Do you receive any recognition for sharing data?
 - Do you receive any recognition for making data available for open access? (would you like to see a citation for any use of your data in published material?)
- Which funding bodies do you mainly work with?
- Are your funders' criteria pushing you toward open data access?
- What are the drivers towards open data access associated with funding?
 - Are researchers embracing them, and if so why?
 - Are researchers resisting these drivers, if so why?
 - Are there any particular issues with any of the funders' criteria that you find problematic in your research process?
- Is it clear to you where to deposit your data to for facilitating open access? Are there any options more feasible or attractive to you? If so, for what reason?
 - Is this choice based on tradition, ease of depositing, technological advances, interoperability issues kept to a minimum, safety of data?
- Are you willing to deposit any of your methods or research process/models or metadata for open access?
- Are you willing to provide “work in progress” data for open access?
 - What would be the value of doing so? (critique, feedback)
 - What would be the (opposite of value) of doing so?

Data sharing:

- What are the characteristics of the data you are working with?
- Does your work yield many different types of data?
- How shareable is the data you are working with?
 - (Large quantities, derived from models, ethics, data security, high cost etc.)
 - Do you share any of the process?
- If you share data, is there a peer review process in place? If so, how is it organised?
 - How are validity, reliability and quality ensured?
- How practical is it for you to share data? (Do you have the staff needed to prepare the data e.g. anonymising, formatting etc., do you have data repository/data centre to submit it to?)
- How costly is it for you to share data? (Are costs associated with open data access factored in at the proposal stage?)
- How, and with whom do you share data/are you comfortable with sharing data?
- Are you concerned about IPR and confidentiality issues?
 - Ask about issues of trust within Science
 - Ask about issues of trust in cases where there are industrial partners and commercial stakes

Using Open Access Data

- Do you use open access data in your work?
- If so, do you have any concerns when it comes to using data that you yourself have not gathered?
 - Reliability ,validity, robustness, completeness of datasets

Additional question for Junior Academics

- Do you receive any research training on the issues of sharing data or making data available for open data access?

Technical Staff

- Are you aware of a drive towards Open Data Access within your field? What are your thoughts on this?
- How important do you see open access to data as being within your field?

- In your view, is it practical to share all data?
- How shareable is the data you work with? (Large quantities, derived from models, ethics, data security, high cost etc.)
 - Does your work involve making it more shareable? If so, what does that involve? (e.g. Cleaning up the data, categorisations of data, interface)
 - Does your work involve advising academics on sharing and managing data? If so, what are the key issues that you offer assistance with?
 - How important is data management in developing open data access?
 - What are your guidelines for data management?
- Are there any technical barriers to sharing (making accessible) this data? How are these being overcome?
- How important is data set interoperability in your work? If important, is it straightforward to achieve?
- What is the cost of maintaining data for open access? (Please give examples)
- Are there technical challenges associated with managing, storing and preserving data?
- For how long is it viable to store and preserve data?
- Are there established timeframes for storing and preserving data? If so, what are they?