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 D3.1:

Legal and ethical issues in open access and data dissemination and preservation

Author(s):	Rachel Finn and Kush Wadhwa (Trilateral Research & Consulting), Mark Taylor and Thordis Sveinsdottir (University of Sheffield), Merel Noorman (Royal Netherlands Academy of Arts and Sciences) and Jeroen Sondervan (Amsterdam University Press)

Dissemination level:	Public
Deliverable type:	Draft
Version:	2.0
Submission date:	30 April 2014

Table of Contents

Executive Summary 4
1 Introduction 7
1.1 RECODE case studies 8
1.2 Methodology 9
2 Legal issues 11
2.1 Intellectual property 11
2.1.1 Copyright 12
2.1.2 Trade secret 13
2.1.3 Database rights 14
2.1.4 Licensing 16
2.2 Data protection 17
2.2.1 Definition of personal data and data subject 18
2.2.2 Data minimisation 19
2.2.3 Data retention 20
2.2.4 Pseudonymisation 20
2.2.5 Research use exception (GDPR Article 83) 22
2.2.6 Consent or alternative legitimate basis 22
2.2.7 Fair processing 24
2.2.8 Right to erasure (GDPR Article 17(3)) 25
2.2.9 Summary and solutions 25
2.3 Open access legislation 26
2.3.1 Amended PSI Directive 26
2.3.2 Commission Decision 2011/833 27
2.3.3 National legislation 28
2.4 Summary 30
3 Ethical issues in open access to research data 31
3.1 The potential benefits of open access 31
3.2 Ethical concerns about open access 33
3.2.1 Unintended secondary uses and misappropriation 34
3.2.2 Dual use 37
3.2.3 Violations of privacy and confidentiality 38
3.2.4 Unequal distribution of research results 39
3.2.5 Commercialisation 41
3.2.6 Restriction of scientific freedom 44
3.3 Summary 45

4 Existing solutions and potential pitfalls 47
4.1 Licensing 47
4.2 Access management 52
4.3 Editorial review 54
4.4 Soft-law measures 56
5 Policy recommendations 58
6 Conclusion 61

EXECUTIVE SUMMARY

In this deliverable, we utilise a combination of literature review and case study interviews to identify legal and ethical issues relevant to open access to research data, to identify examples that illuminate these issues, and to identify potential solutions currently being used to address these issues. On the basis of our research, we indicate possible policy recommendations on legal and ethical issues raised by open access to research data, together with a set of good practice policy guidelines targeted at significant stakeholders and key policy-makers. This work was undertaken as part of the EU FP7-funded project "Policy Recommendations for open access to research data in Europe" (RECODE), within work Package 3 (WP3), Ethical and legal issues.

This document recognises and makes use of the European Commission's definition of "open occess" as "free internet access to and use of publicly-funded scientific publications and data". While this report examines open access to research data more broadly, the focus on "free internet access to and use of" research data is central to our definition. It specifically examines what legal and ethical issues arise in providing such open access, with specific reference to five disciplinary case studies. These include:

- Particle physics
- Health research
- Bioengineering
- Earth Sciences
- Archaeology

The report identifies the legal issues raised by open access to research data in these contexts. Specifically it examines intellectual property rights, including copyright, trade secrets and database rights, privacy and data protection as well as open access mandates. In addition, we identify ethical issues arising in relation to open access research data as including the unintended secondary use, misappropriation and commercialization of research data, unequal distribution of scientific results and disproportionate impacts on scientific freedom as well as other economic, social and scientific costs.

The consortium identified these issues through focussing their literature review on the impact of these issues for a range of different individuals on the knowledge production spectrum, including researchers, project managers, repository managers, policy-makers, and institutional representatives. Furthermore, we identify the level of government or policy that was impacted. e.g., institutional, local, national and supranational. This literature review is supplemented by 13 targeted interviews with key individuals from each of the five RECODE case studies, in order to elaborate on the legal and ethical issues they encounter in their research practice and in providing open access to research data. The combined results of this work were utilised to underpin a workshop on legal and ethical issues with stakeholders representing a number of different perspectives. The information gained from the workshop was added to the analysis, as research data, to further analyse the legal and ethical issues and solutions described in this report and to evaluate the efficacy of these different good practice solutions.

Our discussion in Chapter two reveals that many of the legal obligations to which stakeholders are subject are sometimes in conflict, specifically, intellectual property, privacy, data protection and open access mandates. For example, practitioners wishing to comply with open access mandates often have to navigate privacy, data protection and intellectual

property issues when doing so. In addition, multidisciplinary, multi-national research collaborations often have to navigate complex legal frameworks, including those outside of Europe, as the different organisations involved may be subject to a range of legal requirements. These requirements can be a drain on intellectual and physical resources, as researchers, in particular, struggle to gain the required expertise, and these complex and contradictory obligations prompt stakeholders to find practical solutions to navigate them in creative ways. These include using internal review boards with particular expertise (e.g., data protection) to review materials before they are released, the use of access controls to ensure those who gain access to the material have appropriate training or expertise and the use of licensing to control how other individuals re-use the research data. Ultimately, our analysis of legal issues related to open access to research data demonstrates the ways in which a range of legal instruments can impact the provision of open access to research data. For this reason it is important to establish the circumstances under which it is both lawful and appropriate to provide open access to personal data. In the absence of established frameworks, these legal regimes often create a complex landscape, with real consequences for researchers, organisations and institutions.

In Chapter three, the report examines the ethical issues relevant to open access to research data. We begin by examining open access as an ethical practice, and expand upon the first RECODE report that highlighted many benefits that support compelling moral arguments for open access to research data. Next, the report highlights some of the potentially negative impacts upon individuals, organisations and society as a whole which may arise in relation to ethical concerns associated with the provision of open access to research data. These ethical concerns include unintended secondary uses, dual use, violations of privacy and confidentiality, unequal distribution results, commercialisation and restricted scientific freedom. As above, evidence from the RECODE case studies demonstrates that researchers and institutions adopt particular strategies and measures to address these potential ethical issues, many of which are shared with the measures in evidence in the legal issues discussion. Specifically, ethical review boards, in universities especially, are useful for ensuring ethical treatment of research data and research subjects. Access control mechanisms, licensing (particularly in relation to commercialisation) and other "soft law" measures (e.g., established, agreed practices with now legally binding force) all assisted researchers and institutions in managing these potential ethical issues.

Chapter four of the report deals with a number of findings and tentative recommendations for addressing the legal and ethical issues identified. It details some of the solutions already emerging from the case study analysis, with particular reference to those that are cross-disciplinary or inter-disciplinary in nature. These themes include practices related to licensing, access management, ethical and legal editorial review mechanisms and other soft-law measures. We find that many of these solutions are reliant on existing practices and frameworks and argue that existing disciplinary organisations have a significant role to play in assisting open access stakeholders in addressing legal and ethical issues. However, this analysis also concludes that the existing solutions to legal and ethical issues conflict with the European Commission's definition of "open access" as free access over the Internet. Instead, many of these solutions (licensing, ethical reviews, access controls, etc.) result in some restriction on the data that can be accessed or the method in which it is accessed. Thus, these solutions are not adequate to fully meet the legal and ethical treatment of research data *and* open access requirements. Instead, researchers, institutions, industry and other stakeholders need additional advice about how to ensure that all three of these obligations are addressed simultaneously.

Chapter five makes the following policy recommendations to assist different categories of stakeholder in implementing open access to research data.:

1. Explore the use of licensing, especially Creative Commons or similar open licenses, to address legal and ethical issues;
2. Stakeholders associated with open access to research data should begin by trying to ask different questions to produce a relationship that is not viewed as trading off legal issues for open access;
3. Consider technical or institutional solutions to legal and ethical problems;
4. Establish and clarify circumstances where it is lawful and appropriate to provide open access to personal data
5. Make better use of internal review processes
6. Establish better institutional reward systems for high-quality data
7. Policy-makers, funders, institutions and researchers have to accept that some data cannot be made open.

As the field matures, new or more optimised solutions will become available to better provide open access to research data. In the interim, these solutions may represent a series of stepping-stones to support these early open access practices.

This report has revealed that despite the legal and ethical barriers to providing open access to research data, many solutions are already being utilised to meet ethical and legal obligations, while providing open access as far as possible. However, making data freely available to anyone and accessible over the Internet, in line with the European Commission's description of "open access", may leave some researchers, repositories, commercial organisation, research participants and members of the public vulnerable in important ways. While new solutions should be sought that are able to provide legal and ethical pathways to open access, the current policy push towards open access may need to accept some limits and caveats. These are likely to be in terms of intellectual property, data protection and ethical research practice. This will ensure both the public interest in opening research data, better informing citizens and assisting in innovation as well as the public interest in protecting knowledge production, maintaining privacy and data protection rights and ensuring ethical research practice.

1 INTRODUCTION

As a result of technological, social and policy changes there is growing interest in open access, data preservation and the dissemination of scientific materials. In Europe, a number of policies, initiatives, communities and projects have emerged in order to harness the potential benefits of open access to research data. Potential benefits include innovation, better-informed citizens, the creation of a knowledge society, networking and communities of practice for researchers, cost savings related to unnecessary duplicaiton of research and a reduction in scientific fraud. Many policy initiatives have sought to address the barriers, for example intellectual property issues, ethical considerations, conflicting stakeholder values, disciplinary differences,associated with making research data more accessible. However, many of these initiatives are fragmented by nation, region and/or discipline and are focused on different and particular aspects of the open access and data dissemination and preservation landscape.

The RECODE *Stakeholder values and relationships* report opens by outlining a range of definitions of open access. One such definition, utilised in this report is from the European Commission, which describes open access as "free internet access to and use of publicly-funded scientific publications and data".[1] While this report examines open access to research data more broadly, the focus on "free internet access to and use of" research data is central to our definition. Reasons for providing such open access to research data is described within the Berlin Declaration's vision of open access, which recognises the potential of open access to create "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community".[2] Against this positive backdrop, the report notes "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".[3] As the possibilities for good quality research improve, so do the possibilities for human understanding and improvements in the broadest range of scientific and commercial activities. At the same time, RECODE has identified a series of issues related to such a full commitment to open access.

Some of the key challenges raised by open access to research data are legal and ethical challenges. The legal issues surrounding open access to scientific data primarily include intellectual property considerations, data protection and open access mandates. In particular, data sets that contain personal data raise a number of distinct legal and ethical challenges. In addition to privacy, other ethical issues include the unintended secondary use, misappropriation and commercialisation of research data, unequal distribution of scientific results[4] and disproportionate impacts on scientific freedom as well as other economic, social and scientific costs. These ethical and legal challenges are currently being met with a range of solutions, such as open licensing regimes, including Creative Commons licenses, access management procedures, editorial review procedures and other soft law measures.

---

[1] European Commission, Commission Recommendation on access to and preservation of scientific information,
[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] Sveinsdottir, Thordis Bridgette Wessels, Rod Smallwood, Peter Linde, Vasso Kala, Victoria Tsoukala and Jeroen Sondervan, *Stakeholder values and relationships within open access and data dissemination and preservation ecosystems*, RECODE D1, September 2013, p. 13.
[4] This section refers to the fact that opening access to data may only enable use by a privileged minority and might exacerbate inequalities in access. This is described in more detail in section 3.2.4.

This report will describe the complexities raised by the various legal and ethical issues relevant to open access to research data. We describe each legal and ethical issue, examine its relationship with open access to research data and identify practical solutions being utilised by case study individuals or other research groups. The final section of the report evaluates the solutions to these legal and ethical issues already in evidence. We argue that although many of the solutions proposed do address a number of legal and ethical issues, they often limit the provision of open access to research data as defined by the European Commission. The report concludes by describing a series of policy recommendations for promoting open access to research data whilst meeting legal and ethical obligations.

1.1 RECODE CASE STUDIES

As mentioned above, disciplinary fragmentation is a significant barrier to realising the benefits of open access and data preservation and dissemination. The RECODE project seeks to understand and utilise this disciplinary fragmentation in order to address the grand challenges associated with open access and data preservation and dissemination, including technological and infrastructural, legal and ethical, institutional and policy issues. In order to do so, RECODE utilises five case studies from across scientific disciplines that bring a range of benefits to the project. We have selected five areas of scientific research and each area illustrates some of the key issues. We use these case studies to provide a comprehensive picture of open data ecosystems. The case studies range from open access to data that has already been generated to the collection processes of primary data.

1. 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 USFD[5]is a member of one of four regional Computing Grid Groups in the UK. We explore the legal and ethical issues involved in collecting, disseminating, storing and processing large quantities of numerical data from experiments related to particle physics where the expertise and resources necessary for storing and processing the data are only available to established experts in the field and/or very large consortia.

2. The collection and validation of personal data in clinical, health and biological contexts and its use in research poses problems of data protection, privacy, research ethics and commercialisation. We use the FP7 project EVA, and associated experts in clinical, health and biological data, to explore tools to ensure the ethical treatment of personal data and the provision of open access in this area.

3. 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. Furthermore, the use or processing of data from human subjects raises issues around privacy, data protection and consent. We will explore these issues with the Bioengineering Institute at the University of Auckland[6], and colleagues in the VPH community involved in ontology development, standards for model description, and curation of model repositories.

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

[6] The University of Aukland, "Auckland Bioengineering Institute", no date. http://www.abi.auckland.ac.nz

4. GEOSS[7] is an initiative that seeks to make existing systems and applications for geographic observation, including observations around drought, forestry and biodiversity, interoperable. In addition to providing interoperable access to data, GEOSS also seeks to develop an advanced operating capacity that provides access to analytical models that scientists from different disciples have used to make the data more understandable. In order to do so, GEOSS uses advanced modelling from a range of heterogeneous data sources to make data models usable by other communities, including through the use of natural language interfaces.

5. Open Context is a free, open access resource for the electronic publication of diverse types of research datasets from archaeology and related disciplines. It enforces editorial control through its editorial board, utilises open licensing frameworks and focuses on data portability. Open Context is maintained and administered by the Alexandria Archive Institute[8], a not-for-profit organisation[9], based in Berkeley, California, while IT development is carried out in collaboration with the Berkeley School of Information. Open Context furnishes useful information regarding attitudes, practices and policies within the ecosystem of archaeology, as well as significant information regarding the technical approach adopted for the deposition of, accessibility to and preservation of the data it contains.

Despite the specificity of these case study descriptions, as the RECODE project developed it became clear that we needed to take a broad perspective on the case studies and have extended our research to stakeholders within and related to these case studies. Therefore, for example, some of the quotes below from the Archaeology case study emanate not only from research participants from Open Context, but also from people involved in a range of other organisations related to the Open Context framework. Thus, they should be read, more correctly, as disciplinary case studies, rather than organisational case studies.

These case studies provide an inter-disciplinary grounding that will help the consortium to combat disciplinary fragmentation in the area of open access and data preservation and dissemination as well as maintain an awareness of discipline-specific issues and practices. The case studies will provide insights into the legal and ethical issues, including intellectual property rights, open access mandates, privacy and data protection, research ethics, commercialization as well as others. The case studies will also assist RECODE in identifying policy gaps and evaluating good practice solutions that will contribute to an inclusive and participatory development of policy recommendations.

1.2 METHODOLOGY

This report examines barriers and solutions associated with legal and ethical issues in open access to research data. It utilises a combination of literature review and case study interviews to gather descriptive information about each legal and ethical issue, and to examine solutions currently being used in different disciplinary case studies. This information is then consolidated to formulate policy recommendations for meeting legal and ethical obligations whilst providing open access to research data.

[7] Group on Earth Observations, "Group on Earth Observations", 2014. www.earthobservations.org
[8] The Alexandria Archive Institute, "The Alexandria Archive Institute", no date. http://www.alexandriaarchive.org/
[9] 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.

In order to meet these objectives, we conducted a literature review of legal and ethical issues, focusing on academic literature, reports from related research projects, disciplinary and industry materials, relevant websites and media materials. These resources were used to identify legal and ethical issues relevant to open access to research data, to identify examples that illuminated these difficulties and to identify potential solutions currently being used to address these issues. The consortium focused their literature review on the impact of these issues for a range of different individuals on the knowledge production spectrum, including primarily researchers and project managers, as well as repository managers, policy-makers and institutional representatives. Furthermore, for each legal and ethical issue examined, we identified the level of government or policy that was impacted, e.g., institutional, local, national (European Member State or third country) and supranational (EU, OECD or UN).

This literature review was supplemented with 13 interviews with different case study representatives, in order to provide a more practical and applied understanding of how these legal and ethical issues were experienced and managed in practice. For each case study, as far as possible we interviewed different stakeholders within the case study, i.e., project managers, repository managers, researchers and (institutional and governmental) policy-makers. While we endeavoured to interview individuals who made use of research data that was available via open access, we were unsuccessful in recruiting such individuals within the timeframe allowed by RECODE.

The information from the literature review and the interviews was utilised to underpin a workshop on legal and ethical issues with stakeholders representing a number of different perspectives. These included academics, policy-makers, library and repository representatives, representatives of civil society organisations, representatives from funding organisations and industry. The information gained from the workshop was added to the analysis, as research data, to further analyse the legal and ethical issues and solutions described in this report, and to evaluate the efficacy of these different good practice solutions.

2 LEGAL ISSUES

The first section of this report examines the legal issues associated with providing open access to research data. This includes intellectual property rights, data protection obligations and legislation that creates open access mandates in terms of either research data or scholarly publications. This discussion reveals that many of the legal obligations to which stakeholders are subject are sometimes in conflict, specifically, intellectual property, privacy, data protection and open access mandates create complex and contradictory obligations which stakeholders must navigate in creative ways.

2.1 INTELLECTUAL PROPERTY

Intellectual property rights protect works by individuals that are the result of creativity, innovation, skill and specialist effort.[10] This may include music, design, inventions, processes or scientific discoveries, as well as others. Intellectual property rights are comprised of moral rights and exploitation rights. Moral rights include rights such as attribution, respecting the work or remaining anonymous, and they are often non-transferrable. Exploitation rights include the ability to reproduce, distribute, perform, broadcast or transform materials without permission.[11] Intellectual property rights are governed by intellectual property laws, and the US, Japan and all 28 European Member States are among the members of the World Intellectual Property Organisation (WIPO) and have signed up to the Berne Convention which seeks to protect the rights of authors in their literary or artistic works.

The governing of intellectual property rights in relation to open access to research data references both moral rights and exploitation rights for the researchers or institutions who created, collected or curated the data. In relation to moral rights, rights of attribution and respecting the integrity of the original work are implicated. With respect to rights of exploitation, these are related to open access to research data through copyright, database rights, trade secrets, patents, licenses as well as rights to reproduce, distribute and transform materials. (However, individuals may waive their exploitation rights or trade them through licensing, which will be discussed in more detail below in Section four.) Furthermore, individuals or organisations other than researchers or institutions themselves may claim "neighbouring" or "related" rights if they have curated the data in some way.[12] Many institutions and organisations are aware of the potential repercussions open access may have for the rights of intellectual property owners:

> *For us the IPR is probably the most burning issue in the sense of open access and open data and obstacles to having open data.* (Researcher, Earth science)

Such intellectual property rights include copyright, trade secrets, database rights and licensing, each of which will be discussed in more detail below.

---

[10] Korn, Naomi, and Charles Oppenheim, *Licensing Open Data: A Practical Guide*, June 2011 version 2.0. http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

[11] Rodríguez-Doncel, Víctor, Asunción Gómez-Pérez and Nandana Mihindukulasooriya, "Rights declaration in Linked Data", in Olaf Hartig, Juan Sequeda, Aidan Hogan, Takahide Matsutsuka (eds.), *Proceedings of the Fourth International Workshop on Consuming Linked Data (COLD2013), Sydney, Australia, 22 October 2013*, CEUR-WS, Vol. 1034, 2013, p.3. http://ceur-ws.org/Vol-1034/RodriguezDoncelEtAl_COLD2013.pdf

[12] Ibid.

2.1.1 Copyright

Copyright is a significant component of intellectual property law, and is a form of intellectual property right. Copyright is automatic. It is the right of an author or rights holder of a work (e.g., literature, science, arts), which determines where, when and how the work is made available to the public, including how it is used. Copyrights are exclusive rights that would ordinarily restrict usage of the works created as a result of research whether this refers to research data held by researchers, institutions or publishers.

Discussions about research undertaken within the RECODE disciplinary case studies highlight problems arising in identifying the true right holders. In the field of Archaeology a number of individuals, groups and institutions may have conflicting ideas about who the true copyright holders may be. As a repository manager in Archaeology states:

> [R]*esearchers are the ones that, […] create the data and in creating the content, they are the IP owners. They are the ones that have the legal rights to license content as they see necessary.*

However, the respondent notes that the rights of indigenous people, such as First Nations, Native American or Aboriginal peoples may in fact undermine researchers' or institutions' traditional copyright. Accordingly, such indigenous people "*might have very different kind of worldviews and traditions and perspectives and their own legal traditions around intellectual property issues."* Therefore, while the research material may enjoy copyright, these are not absolute, and ethical research practice demands that researchers not consider this right absolute, particularly when providing open access to research data. Identifying ownership of the research data in physics can also be problematic given the large, international consortia that often partner in physics projects. A data manager describes the experience of CERN:

> *The biggest problem is who actually owns the data. So the collaborations, so this consists of many institutes and people worldwide. They think that they own the data. The funding agencies who fund either CERN (the now 21 CERN Member States) as a whole or specific experiments (e.g. the US, not a Member State, but active in both ATLAS and CMS, as well as ALICE), they might think that they own the data. And then the lab might think it owns the data. So I would say that there has never been unambiguously resolved.* (Data manager, Physics)

Significantly, the members of these large, multi-national consortia are often not subject to the same intellectual property laws and may not have concurrent expectations around the intellectual property generated by the research. Some universities allow researchers to own their own data, while others insist that data are owned by the university. This can cause significant difficulties with collaborations between researchers in different institutions, and affect decisions about making data open. One remedy would be for universities to be more consistent and explicit in defining their data ownership policies, and to insist on written IP/data ownership agreements in collaborative projects.[13]

Other disciplinary contexts may be interested in using copyright to gain commercial value from their research. This will be discussed in more detail in Chapter three below; however, researchers, institutions and organisations have the right to maintain intellectual property in order to gain advantage from the material.

[13] Toby Burrows, University of Western Australia, personal communication, 24 April 2014.

> *We do retain proprietary ownership over particular data sets relating to a company's interest. And we build interface…giving interfaces based on that open source software framework where those guides can be tailored to the needs of the particular company and then the company will have ownership of that, that interface. So we do work in both the open source public domain area, as well as working with companies that need to preserve IP around particular areas.* (Professor, Bioengineering)

Such interests may conflict with legislated open access mandates or requirements from funders or institutions to provide open access to research data.

These examples illustrate the ways in which copyright may act as a barrier to providing open access to research data. This includes issues around identifying the "true" copyright holders as well as the retention of rights to gain benefit from their intellectual property though restricting access to the material and/or by trying to gain or preserve their proprietary rights to this material and the benefits such proprietary rights engender. Both of these issues run counter to the policy push to promote open access to research data and must be adequately addressed in order to provide such open access.

2.1.2 Trade secret

Another aspect of intellectual property rights relevant to open access to research data is that of a trade secret. Trade secrets are protected under article 39(2) of the TRIPS Agreement and they cover commercial information only. In order to be considered a trade secret under TRIPS, the following conditions must be met:

- The information must be secret (i.e. it is not generally known among, or readily accessible to, circles that normally deal with the kind of information in question).
- It must have commercial value because it is a secret.
- It must have been subject to reasonable steps by the rightful holder of the information to keep it secret (e.g., through confidentiality agreements).[14]

Trade secrets can be protected for an unlimited period of time and without registering the secret or any other procedural formalities. A trade secret can include "a formula, pattern, compilation, program, device, method, technique or process".[15] According to the US Patent and Trademark Office, trade secrets "must be used in business" and provide "an opportunity to obtain an economic advantage over a competitor".[16] This means it would be difficult for researchers at a publicly funded university to claim trade secret protection, as there is no profit or economic competitors in the traditional business sense. Thus, it is unclear whether private universities, non-profit organisations or research institutes could claim trade secret protection unless they were explicitly engaging in "business".

Trade secret protection presents a significant obstacle to achieving the benefits of open access to research data, including peer review, verification and replication of research results and re-use of data. Pharmaceutical researchers have had success in claiming research data as a trade secret. According to Payne, this barrier in data sharing has prevented the identification of

[14] World Intellectual Property Organization, "How are trade secrets protected?", no date. http://www.wipo.int/sme/en/ip_business/trade_secrets/protection.htm

[15] The United States Patent and Trademark Office, "Office of Policy and External Affairs: Patent Trade Secrets", 20 February 2013. http://www.uspto.gov/ip/global/patents/ir_pat_tradesecret.jsp

[16] Ibid.

efficacy and safety issues associated with particular drugs, since other scientists could not review or verify this data.[17] In the Earth Sciences case study, researchers often purchase data from private companies and must negotiate trade secret rules when using the data, which are often mandated by specific licensing arrangements.

Mandates to provide open access to research data may conflict with these intellectual property rights, particularly if research funding is contingent upon open access to data agreements. Regulators would need to decide whether the contractual obligation to release research data over-rides any trade secret protection. Furthermore, the use of trade secret protection and the abdication of public funding sources may eventually impact the viability of a private company's business model. This is particularly true for SMEs, who may rely on trade secret protection to gain advantage over larger competitors.

Box 2.1: Tamiflu and trade secret protection

In relation to Tamiflu, the manufacturer, Roche, was able to block access to research data about Tamiflu using trade secret protection. According to an article in the *British Medical Journal*, this included the release of data Roche presented to the Food and Drug Administration in order to verify the effectiveness of Tamiflu. The FDA has collaborated in this protection, where pharmaceutical companies and the regulator have largely agreed that data shared between them would be kept confidential as a trade secret.[18]

2.1.3 Database rights

Copyright can also protect collections of data that sufficiently original and creative, e.g., a telephone directory, or a collection of purely factual material, would not have protection.[19] Such database rights are often relevant to data sets. Copyright in a database is independent from copyright in content elements. Simple collections of data do not count as intellectual property; it is at the point of organisation and selection that intellectual property rights are recognised.

In Europe, a specific database right law, the 1996 Database Directive, protects the producer of a database, who has invested the necessary effort to constitute the database.[20] Database rights under the EU are created automatically, vested in the employers of creators (when the action of creation was part of employment), and do not have to be registered to have effect. A JRC representative explains this:

> *[B]ecause of the investment in terms of financial investments or in terms of work that has been put in, in order to put together or amalgamate the data, the European legislature has granted a specific special rights for databases. That will not exclude the application of copyright, it would run in parallel with it, as a duration there is a shorter than what it granted to a copyright holder. However,* [this] *is going to give also makers of databases which do not qualify for copyright protection an incentive to*

[17] Payne, David, "Tamiflu: the battle for secret drug data", *British Medical Journal*, Vol. 345, 2012. http://www.bmj.com/content/345/bmj.e7303
[18] Ibid.
[19] Tysver, Daniel. A., "Database legal protection", *Bitlaw*, 2013. http://www.bitlaw.com/copyright/database.html
[20] Marc de Vries, *Open Data and Liability,* European Public Sector Information Platform Topic Report No. 2012 / 13, December 2012.

> *make such databases and by granting this exclusive right for a short period of time. […] in many countries around the world, it gets a little bit in the way. If you are thinking about dissemination of data in Europe* […this] *special right is also to be taken into account*. (Legal expert, Earth sciences)

Many countries, for example the USA, do not have a counterpart protective mechanism.

In relation to open data, database rights prevent third parties from publishing, distributing and copying protected research data. Some of the re-use restrictions claimed by private companies in the Earth Sciences case study are based upon database rights, as organisations, such as the JRC, do not own the data they are using for their research. In his discussion on releasing public sector data, De Vries points out that public sector bodies collect information and data sets that they do not necessarily own; for example data produced by third parties as a result of research or other contracts.[21] This leaves public sector bodies vulnerable to legal action by the rights holders. Therefore, if a public body does not hold all of the intellectual property rights associated with the data, it may not be entitled to open up the data for re-use and may need to refrain from doing so.[22] The Revised PSI Directive recognises this barrier and advises public sector bodies not to release information which third parties hold intellectual property rights under the Directive.[23]

What this means in practical terms, as identified in the health case study, for example, is that the applicable database law and other rights related issues of the database may be dealt with under specific arrangements made between those responsible for the compilation of the database, and the researchers seeking to utilise the data contained within the database:

> *You usually have a project officer on a project who will help to set up all the material transfer agreements. And we usually decide which law will be in place and it's usually wherever the database is held.* (Legal expert, Health)

This interview participant also described how the advent of data storage via cloud computing presented additional difficulties associated with database rights, including how to enforce these rights in a cloud environment. Furthermore, the Safe to be Open report by the OpenAIRE consortium finds that in relation to database rights, "Applying the criteria developed by the ECJ to scientific databases, it is unclear whether the majority of research databases meet the formal requirements for the sui generis right".[24] Therefore, this instrument may not adequately protect the intellectual property rights of scientific database creators. This lack of clarity may represent a significant barrier for researchers and institutions both of which might be reluctant to provide open access to research data without significant safeguards in place.

[21] Ibid., p. 7.
[22] Ibid.
[23] European Commission, Directive 2013/37/EU Amending Directive 2003/98/EC on the re-use of public sector information, *Official Journal of the European Union*, L175, 26 June 2013, pp. 1-8.
[24] Dietrich, Nils, Lucie Guibault, Thomas Margoni, Krzysztof Siewicz and Andreas Wiebe, "Possible forms of legal protection: An EU legal perspective", In Guibault, Lucie and Andreas Wiebe (Eds.), *Safe to be Open: Study on the protection of research data and recommendation for access and usage*, University of Göttingen Press, Göttingen, 2003, p. 26. http://webdoc.sub.gwdg.de/univerlag/2013/legalstudy.pdf

2.1.4 Licensing

In relation to open access to research data, licensing is unique in that it represents both a potential barrier to open access and a potential solution to assist in providing open access to research data. In this section, we examine licensing as a "barrier", while in section 4, below, we examine it as a potential solution in relation to a series of legal and ethical issues. Licensing is a useful way to manage intellectual property when users other than the creators seek to work with research data. Licenses can vary from tight, contractual arrangements between two specific parties to "open" licenses, such as Creative Commons[25] or Government Open Licenses, which have minimal restrictions and are intended to assist in providing open access to materials. Together with open data management policies, licenses describe what a person or institute can do with data that has been made publicly available in relation to reuse, dissemination and preservation. Because of their integration in the scholarly communication chain, this review will focus on the relevance of licensing for data creators, disseminators and users.

Licensing emerges as a key barrier for organisations, such as the JRC, which purchase data from private data creators or data brokers. The JRC has experience purchasing data from private companies for the Earth Sciences research as well as other types of data for their varied research activities. In relation to satellite data for Earth Sciences, the JRC often has to negotiate with private companies regarding the use of the data, which is a significant barrier to providing open access to their research data. As noted above, there is a range of different licensing models available for databases and data. For example, the JRC agreements with data owners may vary from closed, restricted licenses to fully open licenses where it is possible to do anything with the created data and information. For example, some licenses agreed with the data owners include elements such as only being able to use the data on one computer, only one user being able to use the data and only being able to use the data for a very specific purpose.

Box 2.2: Licensing and data reuse

The recently published Biomed Central Open Data Policy[26] states that policies should aim for a clarification of the legal (copyright) status of data published in data repositories and (open access) journals and to maximise the potential for reuse of published science. This may include processes such as data and text mining, or visualisations of large datasets. For society to gain the full benefit from scientific data, it needs to be available ways such that it can be reused, scrutinised and built upon with the minimum of barriers.

A Researcher in the Earth Sciences case study describes this practice:

> *[E]ven though we manage to negotiate quite different licences than usual end-user licences agreements that private companies have, we still are not in the position in most of the cases to offer open access to those data. Because the licensing agreements that we have to sign with these companies limit further reuse, a number of applications not related to the purpose for which we initially were buying the data may not utilise the acquired data. […] because this data can be used for very different purposes, we saw it as an obstacle of buying the same data several times.*

[25] Creative Commons, "Creative Commons", no date. http://www.creativecommons.org
[26] Biomed Central, "Open data", 3 Sept 2013. http://www.biomedcentral.com/about/opendata

Thus, in addition to limitations on re-use of the data, these licensing terms also have an economic cost. For example, one organisation was forced to purchase the same material multiple times. Furthermore, the re-use limitations, in particular, had significant impacts in terms of its ability to meet its own open access principles.

> *The licensing agreements that we have to sign with these companies limit further reuse, […] So from this perspective, we are still struggling quite a lot to come up with schemes that allow wider access and more open use of the data that we acquire from private companies. These restrictions come mostly from the intellectual property rights attached to the data or database rights, or when the data are declared trade secrets have to be kept as such by us.* (Researcher, Earth Sciences)

Thus, licensing arrangements allow private companies to restrict the re-use of their data. However, it can interfere with some organisations' legal or funding obligations to make their data accessible to the public and available for re-use. Again, this requires organisations that rely on purchased data to navigate conflicting legal regimes in relation to open access to research data.

These examples introduce some of the issues central to intellectual property rights that may compromise open access research data. All of these issues present limitations on the preservation, dissemination, accessibility and re-use of research data. As such, they have spurred the development of practical solutions to navigate open access to research data. Furthermore, intellectual property rights may conflict with other legal obligations, especially open access mandates that require the provision of open access to research data.

2.2 DATA PROTECTION

Currently, personal data in the European Union are protected by domestic law implementing the Data Protection Directive (95/46/EC). This Directive was intended to protect *both* the fundamental right to data protection and also to guarantee the free flow of personal data between Members States of the European Union. Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) established the principle that "Everyone has the right to the protection of personal data concerning them." Article 16(2) established the process whereby the European Parliament and the Council could "lay down the rules" relating to the processing of personal data "and the rules relating to the free movement of such data".

In January 2012, the European Commission proposed that the Directive be replaced by a Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation, hereafter "the Regulation"). It is important to consider the title in full as emphasis is typically given to the intention expressed in the first part of the title: To protect individuals with regard to the processing of their personal data. The second intention of the Regulation – namely, *to promote and protect the free movement of personal data* – is sometimes overlooked or underplayed.

Since the proposal was first introduced by the Commission it has received a tremendous amount of critical attention and thousands of amendments have been proposed to the original text. In October 2013, the Civil Liberties, Justice and Home Affairs committee of the

European Parliament (known as the LIBE committee) voted to approve a compromise text on the proposed Regulation on behalf of Parliament. In March 2014, the European Parliament passed the proposed text by an overwhelming majority.[27] The Council working group is continuing to meet to discuss the Regulation but the Council has yet to agree its position. When the Council of the European Union has proposed its own amendments, the process of reconciliation can begin.[28] The Greek Presidency is aiming to reach a partial general approach on the Regulation by June 2014, to enable the Council to enter negotiations with Parliament after the summer. However, timescales are uncertain and expected to continue in the autumn. This process may not be completed until 2015.

The significance of the Regulation for EU Data Protection means that any implications for open access require consideration. If RECODE is to make recommendations in relation to EU data protection law, then those recommendations are most meaningfully to be made in relation to the anticipated Regulation. With a new Regulation on the horizon there is little that will be done to amend the interpretation or implementation of the current Directive. In what follows those elements of the Regulation that may be of particular significance to open access are identified and the possibilities of open access in terms consistent with the Regulation are described.

2.2.1 Definition of personal data and data subject

The definition of personal data provided by the Regulation is similar to that currently contained in the Data Protection Directive (in addition to the extended list of 'identifiers' and associated "factors"[29]) but Recital 23 points to one material difference. It is proposed that Recital 23 is amended so that:

> To ascertain whether means are reasonable likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, *taking into consideration both available technology at the time of the processing and technological development*. (emphasis added)

This revision may make it particularly challenging to determine whether data have been effectively anonymised and taken outside of the scope of the definition of personal data.[30] There are acknowledged challenges with establishing that data cannot be (re)identified with a particular individual (see section in ethics on privacy and confidentiality). It may be difficult for an individual to know what "technological development" will mean for the possibilities of identification in years to come. The appropriate response in some cases will be only to release data into a controlled environment. The controls may be a combination of technical, organisation, and contractual measure to prevent further linkage of data. This should not,

[27] European Commission, "Progress on EU data protection reform now irreversible following European Parliament vote", Press release, MEMO/14/186, 12 March 2014. http://europa.eu/rapid/press-release_MEMO-14-186_en.htm

[28] European Parliament, "Legislative powers: Ordinary legislative procedure", no date. http://www.europarl.europa.eu/aboutparliament/en/0081f4b3c7/Law-making-procedures-in-detail.html

[29] The proposed definition of personal data are "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, unique identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social or gender identity of that person"

[30] ibid.

however, be adopted without challenge as the default response where open access is the aspiration.

A solution in some cases will be *not* to rely upon a particular process of anonymisation to take data outside the scope of the Regulation: If one treats data as 'personal data', then one reduces the significance of marginal cases and the need to control the environment to reduce the risks of identifiability. Although, the proposed change to Recital 23 will make it more difficult to reliably anticipate when data disclosed have been sufficiently de-identified to no longer be classed personal data, the appropriate response in some cases may be to extend treatment of data as personal data. Where data can be robustly anonymised, such as through aggregation and the suppression of small number counts in cells, then data may be published without satisfying the requirements of data protection law. But, privacy protective techniques – such as cell suppression – can reduce the research utility of data.[31] The legal possibility, and ethical appropriateness, alongside questions of additional research utility, of publishing identifiers should always be considered before the decision is made to avoid the need to comply with the requirements of data protection law by anonymising data through techniques that may become increasingly destructive of the research value of the data as they become more robust to account for the growing possibilities of re-identification.

Certainly, it will not *always* be appropriate to anonymise data before providing open access. For this reason it is important to establish the circumstances under which it is both lawful and appropriate to provide open access to personal data.

2.2.2 Data minimisation

While it can be lawful to provide open access, and the response to marginal cases of identifiability will sometimes be to treat data as though it is capable of identifying people, that does not remove the pressure to remove identifiers that are not required for processing: The law of data protection adopts the principle of Data minimisation.[32] Identifiable data should not be processed if non-identifiable data are sufficient and the fewest identifiers necessary should be used. Notwithstanding this pressure to minimise the use of identifiable data, there are examples of activity where – assuming other legal requirements are met – personal data may be justifiably processed due to there being no practicable alternative to the use of identifiable data. For example, a video of somebody speaking might provide "information about their body movement, about their facial expression, [that] is very important for language research" (Legal expert, Earth Science) and which cannot be easily anonymised. As long as the requirements of data protection are met in relation to such data, then it can be lawful to provide open access to data in identifiable form[33] and it may be preferable in some cases (although it should be remembered that there may be applicable legal rules other than data protection[34]).

---

[31] Ohno-Machade, Lucila, Staal Vinterbo and Stephan Dreiseitl, "Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance", *Journal of the American Medical Information Association*, Vol. 9, No. 6, Supp. 1, 2002, pp. s115-119.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC419433/

[32] Article 6(1)(c): Personal data must be "kept in a form which permits identification … for no longer than is necessary for the purposes for which the data were collected or for which they are further processed".

[33] The principles relating to personal data processing are set out in Article 6 of 95/46/EC and are substantively preserved in Article 5 of the proposed Regulation. However, the lawfulness of research processing is, according to Article 6(2) of the LIBE proposal subject to safeguards set out in Article 83.

[34] The example provided in this interview required consent because of specific image rights in force within the jurisdiction that prevented use of a person's image without his or her explicit consent.

2.2.3 Data retention

An explanation of data retention is provided in Article 5(e) of the Regulation. As a general principle, data should be kept in identifiable form for no longer than necessary for the purposes for which data are collected or for which they are further processed. However, both the current Directive and the proposed Regulation allow an exception to this general principle in relation to the processing for research purposes. Researchers should continue to reflect upon the need to retain identifiable information and, in particular, the potential benefits of *not* retaining it in some circumstances.

> *I think that's another reason that people agreed to participate in my research is that there is that, ok, it's definitely anonymous but it's always going to be destroyed in five years. So five years from now, people will have moved on maybe to a different job or whatever and they won't have to worry about it ever coming back to haunt them. That's really interesting.* (Ethical editorial reviewer, Open Context)

As this interview participant points out, there will sometimes be potential tension between expectations or preferences of research participants and the data preservation elements of open access to research data, where obligations to protect researchers or to treat their data ethically may limit the extent to which open access to research data can be realised.

2.2.4 Pseudonymisation

A potentially significant change proposed in the draft Regulation is the introduction, for the first time, of the term "pseudonymised data":

> "personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution"

If data are only held in pseudonymised form, then particular responsibilities are removed.[35] This may have some positive implications for open access. If access is provided to only pseudonymised data, then although such data might continue to fall within the scope of the Regulation, the party accessing it will not have to discharge the same responsibilities as they would if the data were not pseudonymised. This will reduce the burdens of compliance for those processing only pseudonymised data.

However, it can be challenging to effectively pseudonymise data, particularly in the context of open access to research data. Examples from the earth science and archaeology case studies are illustrative:

> *I'll just outline the case. It was a database of land samples, so they were like collecting samples of land across Europe and we had a database of chemical analysis of the samples. Now, because the database contain also the GPS coordinates of where the sample was collected, here was an issue that was, like ok through the GPS*

[35] See Article 10 of LIBE proposals. Committee for Civil Liberties, Justice and Home Affairs (LIBE) of the European Parliament, Inofficial consolidated version after LIBE committee vote provided by the rapporteur, 22 October 2013.
http://www.janalbrecht.eu/fileadmin/material/Dokumente/DPR-Regulation-inofficial-consolidated-LIBE.pdf

coordinate you can identify the particular land and then if you go, you can verify also who is the owner of that land. So you can trace back the physical person behind the data. When the question reached the data protection coordinator, I think we agreed that we could only disclose the data and the area where it was, but not the GPS coordinate for that matter. So you keep the, for example, the maps that you can build with this data in a way that you can not really tell whether it is this land or the land that is nearby, that has a particular chemical component in it. So we could disclose data, but we just needed to adjust with how precise the GPS coordinates [are]. (Legal expert, Earth sciences)

Landowner details are a major issue of confidentiality. These records often times contain the names of one or several personal individuals or families, potentially their mailing address, phone number, which need to be scrubbed. And that can be a difficult, well not difficult, it can be a labour intensive process. […F]ortunately in most cases, there is a field related to some kind of, either it's called landowner name or contact person, so that's fairly easy to get at. But then there is also from the case of private information and fortunate American tendency to name sites after the person whose property it's on. But if you have Jones Farm, it might be called 'Jones Farm Site No. 2'! (Editorial reviewer, Archaeology)

The difficulties in pseudonymising data pose particular challenges given the proposal that research processing may *only* use pseudonymised data.[36] The difficulties that would be caused by this requirement extend beyond open access and there are efforts on-going to persuade the Council of Europe to resist this position in the final text of the proposed Data Protection Regulation.

The challenges to effective pseudonymisation are similar to those posed to effective anonymisation. Some of the techniques available to reduce the identifiability of data would be applicable to both and could be used to help satisfy the definition of pseudonymisation in some cases. Examples of techniques described in case studies include use by Open Context of Google or other finder software to identify given names and suggest these for "scrubbing". However, despite this seemingly effective method of anonymising data, a research manager in the VPH case study noted that modelling software and tools are becoming so detailed and sophisticated that it might be possible to identify someone based on images, for example of their heart, produced by these techniques.

The challenges to effective pseudonymisation are similar to those posed to effective anonymisation. Some of the techniques available to reduce the identifiability of data would be applicable to both and could be used to help satisfy the definition of pseudonymisation in some cases. However, the 'solution' to the challenge described in the case of marginal cases (where effectively anonymisation is doubted) – namely treating data as personal data – is not available if data can *only* be processed for research purposes if pseudonymised. One potential solution here is for the Council of Europe to remove this requirement entirely. As a position has yet to be adopted by the Council, lobbying on this point remains a viable option.

[36] Article 83. The *necessity* of satisfying the requirements of Article 83 depends in part upon the interpretation of Article 6(2). For explanation and discussion see Taylor, M. J., and B. Thompson, "Update on the Data Protection Regulation", *Public population project in genomics and society*, Briefing paper, 2013. http://www.p3g.org/sites/default/files/site/default/files/Taylor%20Thompson%20-%20Data%20Protection%20Regulation%20Update%2025%20February%202014%20V3.pdf

Alternative arguments in relation to the development of Article 83 are considered further below.

2.2.5 Research use exception (GDPR Article 83)

According to Article 83 personal data may be processed for historical, statistical or scientific research purposes only if:

> (a) these purposes cannot be otherwise fulfilled by processing data which does not permit or not any longer permit the identification of the data subject;
> (b) data enabling the attribution of information to an identified or identifiable data subject is kept separately from the other information under the highest technical standards, and all necessary measures are taken to prevent unwarranted re-identification of the data subjects.

The wording of Article 83(b) is unfortunately ambiguous. It could be interpreted to prohibit the processing of data for research purposes without pseudonymisation or anonymisation. This could be extremely problematic for any research that relies upon processing data that cannot be effectively pseudonymised.[37] As noted above, the challenges of effective pseudonymisation can be exacerbated in the context of open access. It may be particularly challenging for open access when the "highest technical standards" of pseudonymisation may be incompatible with open access.

There is a range of potential solutions to the problem posed by the current text of Article 83. They vary from lobbying the Council to remove the requirement of pseudonymisation entirely (i.e., Change the text, see above), to avoiding the requirement in some cases by an interpretation of the current text of Article 6, to mitigating the significance of the requirement by interpreting the text of Article 83 in a particular way.[38]

The suggestion at this stage is for each of these solutions to be pursued. This requires co-ordinated action by the research community. The effectiveness of an interpretation of Article 6 that recognised satisfaction of the requirements of Article 83 to be an *alternative* route to lawful processing depends, in part, upon the suitability of the other legal basis contained within Article 6. First amongst these is the consent of the data subject.

2.2.6 Consent or alternative legitimate basis

The current text changes the definition of consent to require "explicit" consent in all cases where a data subject's consent is relied upon to satisfy a condition of lawful processing. The requirement to provide explicit, or even unambiguous (as it is worded in the current Directive), can cause some challenges if it is interpreted to require *written* consent.

> *For instance I have heard about a linguist for instance working with speakers of obscure languages, like rare African languages, or with immigrants, so people who may have some bad thoughts about someone making them sign something, people who are not necessarily in this culture of written word and written agreements. People who sometimes can hardly read and write or who are illiterate. And definitely in such*

[37] See Taylor and Thompson, op. cit., 2013.
[38] For alternative interpretations of the requirement to pseudonymise, and a description of alternative understandings of Article 6, see See Taylor and Thompson, op. cit., 2013.

> *circumstances the privacy protection issues, are more problematic.* (Legal expert, Earth sciences)

The appropriate response in such cases may be to establish whether the relevant data protection authority will accept that consent might be lawfully recorded in an alternative way *or* establishing an alternative legal basis to consent for the processing in the circumstances. That alternative does not undermine the separate requirements to provide information to data subjects about processing, nor any responsibility to seek consent in an appropriate way to discharge any ethical obligations.

Relying upon consent to satisfy a legal condition of data protection can also be challenging due to the requirement that consent relate to "specific" purposes. This has been interpreted and implemented in different Member States in different ways. Some, e.g., the UK, have permitted broad consent but that is not a position universally adopted across the EU and is arguably inconsistent with the position articulated by the Article 29 Working Party.[39] If the Regulation is adopted, and "specific" consent is interpreted more narrowly than is currently the case in some countries, such as the UK, then this may make research processing more difficult generally. That said, there may be responses available:

> *[Under the proposed Regulation] it's unclear whether you can have a broad consent for a number of different purposes. … the thing about [this] is that it may have a real effect on epidemiology which is sort of longitudinal studies. And also on, well it will have an effect on sharing, because you will have to get an explicit…and we are developing, we have an IT interface for patients which would enable them to give that consent and to track that consent each time that it moves to different researchers. So, I personally don't think that is a problem.* (Legal expert, Health)

If broad consent is not permitted, then consent to open access can be challenging where arrangements for such dynamic consent[40] are not possible. It is difficult to reliably anticipate all of the research purposes to which data, if they are made open access, will be put. The solution in this case – to lawful processing under the Directive at least – is:

(i) to rely upon broad consent in those jurisdictions that permit it
(ii) to seek to adopt a sufficiently comprehensive approach to consent – as seen in personal genome project[41] – to specifically include all conceivable purposes *even if* data is made open access or
(iii) to rely upon an alternative to consent to satisfy the legal requirements for processing.

In relation to personal data, Article 7 of the Data Protection Directive recognises that processing which "is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed" may meet the requirement for legitimate processing "except where such interests are overridden by the

[39] Article 29 Working Party, Opinion 15/2011 on the definition of consent, WP187, Brussels, 13 July 2011 and Article 29 Working Party, Working Document on the processing of personal data relating to health in electronic records, WP131, 15 February 2007.

[40] Kaye, Jane, Edgar A Whitley, Nadja Kanellopoulou, Sadie Creese, Kay J. Hughes and David Lund, "Dynamic consent: a solution to a perennial problem?", *British Medical Journal*, No. 343, 2011 and Solum Steinsbekk, Kristin, Bjørn Kåre Myskja and Berge Solberg, "Broad consent *versus* dynamic consent in biobank research: Is passive participation an ethical problem?", *European Journal of Human Genetics*, Vol. 21, 2013, pp. 897–902.

[41] Personal Genome Project, "PGP-UK Participation Documents: Consent form", 20 Jan 2014. http://www.personalgenomes.org.uk/static/docs/uk/PGP-UK_FullConsent_06Jun13_with_amend.pdf

interests for fundamental rights and freedoms of the data subject". Research may be considered a legitimate interest by the relevant national supervisory authority and satisfy this criteria of legitimate processing, if "on balance" the nature of the data and the nature of the processing, by whom and for what, appears to represent no disproportionate interference with the interests of the data subject.[42] It will be particularly challenging, but potentially arguable in some cases, that this "balance" falls in favour of open access to research data. While the viability of option (i) under the Regulation will depend entirely upon the interpretation of "explicit" consent adopted at the EU level, options (ii) and (iii) may remain viable. The continued viability of (iii), even in exceptional cases, will depend in part upon the final wording of "the legitimate interests" provision adopted by the EU.

2.2.7 Fair processing

The text of the Regulation adopted by the European Parliament proposes a new standard information policy table. While aiming for simplicity, the "tick box" approach may actually create confusion and be potentially misleading: The distinctions that it seeks to make may be open to considerable differences in interpretation. For example, whether data are "disseminated to commercial third parties" may be difficult to establish. This is a general problem but can raise particular difficulties in the context of open access. This can be seen in the different approaches to interpreting creative commons licences; "after all creative commons are just simple contracts that can be interpreted in different ways" (Legal expert, Earth science). This may lead to a precautionary approach being taken with "boxes ticked" to guard against a particular eventuality. This may result in "defensive fair processing" with "standardised" information sheets designed to protect the data controller rather than inform the data subject. One perverse response might be to actually undermine understanding of what is done with data.

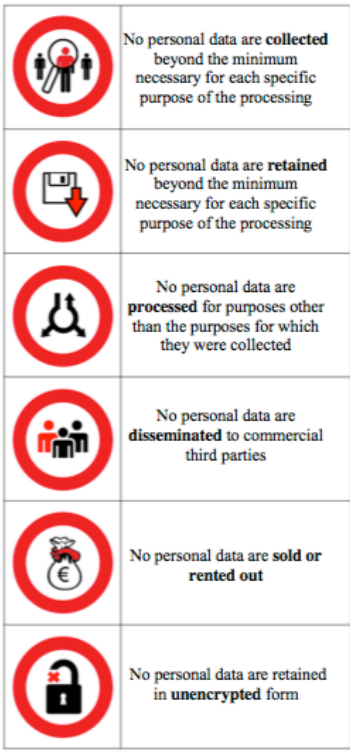

Figure 1: Data protection elements

The challenges associated with fair processing under the existing Directive are similar to those described above in relation to ensuring the purposes for which consent are obtained are sufficiently "specific": It can be difficult to describe clearly all the purposes for which open access data may be used. The response here is, however, rather simpler as the data controller (i.e., the party responsible for determining the purposes of processing) is only under a responsibility under Article 10 (or Article 11 if data are obtained via a third party) to provide information in relation to the purposes of the processing *by the data controller*. If it is intended to provide open access to data collected, then this should be made clear to a data subject together with any other information necessary "having regard to the specific circumstances" in which the data are collected or processed to ensure that processing are fair, consistent with the responsibilities set out in Articles 10 and 11 of the Data Protection Directive.

[42] For a comparison of different approaches to interpreting 'legitimate interests' see Korff, Doewe, "Comparative Summary of National Laws", *EC Study on Implementation of Data Protection Directive*, Human Rights Centre, 2002, p. 80.
http://www.garanteprivacy.it/documents/10160/10704/Stato+di+attuazione+della+Direttiva+95-46-CE

2.2.8 Right to erasure (GDPR Article 17(3))

Article 17 introduces a right to erasure (this replaces the earlier proposed right to be forgotten). The general right does admit certain exceptions. One exception is where retention of data is necessary for historical, statistical and scientific research purposes (in accordance with Article 83). It may be that data provided in open access may be excepted from certain requirements of erasure where data is provided for research purposes and the right to erasure would fundamentally conflict with pursuit of that research purpose. The meaning and scope of when it is 'necessary' to except data from this requirement requires clarification.

2.2.9 Summary and solutions

There are circumstances in which it may be appropriate and lawful to provide open access to personal data. However, in *all* cases of processing personal data, including publication and dissemination, the requirements of data protection law must be met. There will be occasions when consideration should be given to treating even data of relatively low risk of identifiably as personal data so far as practicable. If marginal cases are treated consistently with the requirements of the Regulation, then this reduces the risk that – with technological development – changes to identifiability might take particular instances of open access outside the principles of data protection. However, the costs of treating data as personal data are currently heightened by the uncertainties around proper handling. Without clear guidance the temptation is to act in a precautionary manner and seek to take data outside of the data protection regime. If handling data as though it were personal data is not to represent a disproportionate impediment to research in general and open access in particular, then the conditions for lawful open access to personal data need to be more clearly established.

If the conditions for lawful open access to research data are clearer, then research access to personal data can be improved. Clarifying the conditions for open access will also clarify the circumstances in which open access may not be appropriate.[43] For example, if there are concerns about the potential identifiability of data *if linked with other data*, then the risks of identifiability might be effectively mitigated by only making data available in sufficiently controlled environments, e.g., closed 'data labs'. In this way a number of the aspirations of open data might be met without data meeting all the requirement of open access. Of course, anything short of open access can interfere with achievement of some of its ambitions but might be necessary and appropriate on occasion to protect other values. The point is that it should be considered whether it is appropriate and practicable to meet data protection requirements *and* provide open access to personal data. In some cases, it might be *more* appropriate to recognise that (even if steps are taken to minimise the risk) there may be a potential for identification and to treat data as personal data:

> *So I would say another solution to that issue, is we can never actually, never guarantee confidentiality of all data, because it would be hacked into and we can't anymore say that your data will be anonymous because that is a nonsense too, because we are able to bring in so many different kinds of data, ... that the potential*

[43] See, for example, the moves by funders to clarify when it is inappropriate to seek to identify data subjects. Callaway, Ewen, "UK Funders Get Tough on Privacy Breaches", *Nature News Blog*, 24 March 2014. http://blogs.nature.com/news/2014/03/uk-funders-get-tough-on-privacy-breaches.html

for people to be re-identified or distinguished in the data are quite high. […] So you could reframe that as how do we enable people to exercise their privacy rights. So do we do that through dynamic consent interface. […] We say that, we can no longer guarantee that the information will remain confidential, because things will be hacked into or there will be an unauthorised access but if, we find out about that, then we will let you know. And it is in order to maintain participation in research then in actual fact, we should do our utmost to make sure people know what's going on. But also we want really rich data sets. (Legal expert, Health)

Future work might not only consider how transparency and communication might be improved in an increasingly complex and networked environment but also the appropriateness of consent being explicitly delegated to a trusted third party (e.g., an access committee). This was one of the solutions to current challenges in relation to consent that was proposed during the WP3 workshop.

Certainly, better guidance on the expectations of control in different contexts, and the legitimacy of open access to data that may contain personal data, should be provided. This will be most usefully provided in context specific guidance with distinctions drawn between different categories of data due to the risks, by both data content and the context of use, in relation to the legal and ethical issues recognised by WP3. For example, a key distinction is likely to be between personal data and sensitive personal data, retrospective and prospective data collections, and between data collected for research purposes and data collected for other purposes but subsequently made available to researchers. It will be in those cases where data are not sensitive, they are collected prospectively, and processed specifically for research purposes that data protection requirements in relation to open access are likely to be most readily satisfied. Even in this case, however, it will be important to ensure that all principles, including for example, those that relate to data minimisation, fair processing, and respect for data subject rights, are met.

2.3 OPEN ACCESS LEGISLATION

There have been a number of legislative pushes for open access in the last decade, both from the European government as well as Member State governments. These include European legislation on access to and re-use of public sector information and national legislation mandating the deposit of publications resulting from publicly funded research in open access databases. Although only the European legal instruments directly address open access to research data, all of these legislative instruments provide lessons for such initiatives that can help to avoid pitfalls and capitalise on existing good practice.

2.3.1 Amended PSI Directive

At the European level, the most relevant legislation for open access to research data surrounds public sector information. Access to public sector information is governed by the 2013 Amended Directive on the re-use of public sector information 2013/37/EC. The purpose of the Amended PSI Directive is to provide harmonised rules for the sharing of public sector information because Commission documents "constitute a vast, diverse and valuable pool of

resources that can benefit the knowledge economy.".[44] The original 2003 Directive was amended by the current version in order to deal with the exponential increase in the amount of data available in the world and the "continuous evolution in technologies for analysis, exploitation and processing of data", including "the use, aggregation or combination of data."[45] This document, unlike the original Directive, applies to university libraries, archives and museums as well as more traditional sources of public information (e.g., local governments, national governments, agencies and others).

In relation to intellectual property rights, the Directive addresses many of the issues described above. It does not apply to situations in which a third party is the owner of the intellectual property rights. Public sector bodies are encouraged to exercise their copyright in ways that facilitates the re-use of the data, including through licenses that ensure that the data is attributed to the public sector body in question and states whether the re-user has modified or amended the document in any way. The Amended PSI Directive specifically encourages the use of open licenses for this purpose. Finally, documents that contain personal data are "incompatible" with the Directive due to protections included under the 1995 Data Protection Directive.[46]

As a public research organisation, the JRC, within which the earth science case study is primarily situated, is obligated to provide open access to its research data, in so far as it is compatible with other legal instruments like the Data Protection Directive. This is solved via an institutional process, where data sets are evaluated internally before they are released. This process is carried out both in relation to intellectual property considerations as well as data protection:

> *We have data protection coordinators in each Directorate General at the European Commission, so there is one also in the Joint Research Centre and we work close with them, whenever there is a data that contain data that can lead to identification of a physical person. So we contact the coordinator and then if it needs be, then he contacts the data protection officer of the European Commission and then he provides us with an opinion.* (Legal expert, Earth sciences)

As in other circumstances, the JRC uses an internal review process by dedicated legal experts to ensure that they are in compliance with both the Data Protection Directive and the PSI Directive; a process that has been working well to ensure that all of the legal layers they encounter are adequately addressed.

2.3.2 Commission Decision 2011/833

The layers of legislation and legal compliance the JRC must navigate are also demonstrated by the need to comply with Commission Decision on the reuse of Commission Documents.[47] The Commission Decision is contextualised by the older Regulation 1049/2001 on public access to European Parliament, Council and Commission documents[48], with which the

[44] European Commission, Directive 2013/37/EU Amending Directive 2003/98/EC on the re-use of public sector information, *Official Journal of the European Union*, L175, 26 June 2013, pp. 1-8, p. 1.
[45] Ibid.
[46] Ibid., p. 6.
[47] European Commission, Commission Decision of 12 December 2011 on the reuse of Commission Documents, 2011/833/EU, Brussels, 12 December 2011.
[48] European Commission, Regulation No. 1049/2001 regarding public access to European Parliament, Council

Decision is in compliance. The purpose of the Decision is to respond to the fact that the Commission feels that it holds a corpus of information, primarily documents, that could be used to benefit citizens, as well as companies seeking to provide new services. The Decision specifically mentions the JRC as an important resource of information, and states that a dedicated data portal should be set up. As above, the Decision states that internal departments must respect intellectual property rights and the Data Protection Directive when providing open access to documents.

Box 2.3: Public-private partnerships

Most of the data that are produced by the European Commission are in a scientific field, the result of a partnership with other research organisations in Europe or private public partnerships and so forth. So when this happens, the results of research that is done in partnership is not subject to the re-use policy of European Commission only. And has to be agreed with all the parties, before it can be released. (Legal expert, Earth science)

According to the experts we interviewed for the Earth Sciences case study, the JRC is committed to this policy. "[S]ince we are bound by the Commission's Decision […] we aspire to promote sharing and make these data available for further reuse by whatever users without any kind of restrictions." (Researcher, Earth sciences). This includes a lack of limitations on commercial reuse. However, as the Decision foresees, sometimes there is a conflict with other legal instruments in complying with the legislation, particularly with respect to intellectual property rights and the use of private sector data. This means that in practice, data is often not subject to the Decision. When conflicts occur, the JRC, again, uses a formal review process to determine whether the information can be released. A legal expert from the JRC describes the procedure as follows:

> *[W]e receive a request for support and we treat the matter [... we] study a particular file and we verify whether the Decision applies or not. And if not, then we verify what else we can do to make it available.* (Legal expert, Earth science)

This internal review process is working as a useful mechanism to enable the organisation to meet all of its legal requirements when releasing data under an open access regime.

2.3.3 National legislation

Some European countries have also introduced open access mandates via legislation, with particular respect to publications arising from public research funding. Although these decrees do not extend to research data, many national funding bodies have been mandating open access to research data and the open access to publications movement suggests that this may be a precursor to national, legislative mandates to provide such open access. Open access mandates by funding organisations will be discussed in more detail in later reports, specifically RECODE Deliverable 5.1 *Policy guidelines for open access and data dissemination and preservation*. However, the following countries serve as examples of legislative mandates to provide open access to publications arising out of publicly funded research.

---

and Commission documents, 30 May 2001. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:145:0043:0048:EN:PDF

Specifically, the Italian, Spanish and Belgian governments have all passed specific Decrees, Laws or Declarations relating to open access to publications. In Italy, Decree n. 91 of 8 August 2013[49] states that publications that result from public funding of at least 50 per cent must be deposited in an institutional or industry electronic archive no later than 6 months after publication.[50] The publication must be freely accessible both within and outside the European Union and must enable long-term storage in an electronic format.[51] Similarly, Spanish lawmakers have also demonstrated a commitment to open access through the 2011 Law on Science, Technology and Innovation. This instrument requires that an electronic version of any publications that has resulted from publicly funded research must be deposited in an institutional repository no later than 12 months after the original date of publication. [52] Finally, Belgian government has implemented a *Declaration on Open Access to Belgian publicly funded research*, which states that open access is the "default infrastructure for dissemination of Belgian scientific research results".[53] The Declaration requests signatories to "ask" researchers to deposit their publications in an open access repository within 6-12 months of publication (depending on discipline) or to publish using a gold open access route.[54]

Both the Spanish and Belgian instruments utilise the Liège model[55], whereby deposited publications are the only ones eligible for inclusion in official reviews of research. In Spain these open access publications are linked to formal reviews of the funding, which means that researchers who do not publish in open access could receive poor reviews for their activity, which could jeopardise their ability to secure public funding for their research in the future. Although the Belgian Declaration does not specifically mention the Liège model, it "requires immediate deposit in the institutional repository as the means of submitting work for research evaluation, and as a condition for research funding".[56] Therefore, researchers who do not conform to open access requirements will find themselves ineligible for funding, including especially by the Fund for Scientific Research, Belgium's largest public funder of scientific research.

Finally, in a slightly different vein, it is worth noting that multi-national research collaborations, especially those that span continents, often have to deal with a range of complex national legislative elements. For example, a representative of the Bioengineering case study described how research collaboration between their country and the USA meant

[49] Italian Government, Urgent measures for the protection, enhancement and recovery of property and cultural activities and tourism, *Official Gazette*, Vol. 186, 9 August 2013. http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto-legge:2013-08-08;91

[50] Abernard102@gmail.com, "Feed Item: And the decree of August brings open access to public research", *TagTeam Blog*, 28 August 2013. http://tagteam.harvard.edu/hub_feeds/928/feed_items/256078

[51] Ibid.

[52] Government of Italy, Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación, 2 June 2011. http://noticias.juridicas.com/base_datos/Admin/l14-2011.html

[53] Government of Belgium, Brussels Declaration on Open Access to Belgian publicly funded research, Brussels, 22 October 2012, p. 3. http://openaccessbelgium.files.wordpress.com/2012/11/signedbrussels-declaration-on-open-access.pdf

[54] Ibid.

[55] Rentier, Bernard, and Paul Thirion, "The Liège ORBi model: Mandatory policy without rights retention but linked to assessment processes", *Berlin 9 Pre-conference Workshop*, November 2011. http://orbi.ulg.ac.be/bitstream/2268/102031/1/Rentier-WashDC-2011.pdf

[56] Harnad, Stevan, "Open Access in Europe: The Blind Men and the Elephant", *EuroScientist*, 27 September 2013.
http://euroscientist.com/2013/09/open-access-in-europe-the-bear-and-the-tortoise/#sthash.nyMYxX6a.8iOI2dH6.dpuf

that the researchers had to navigate their own legislation as well as medical patient data protections in the US:

> [*W]e were collaborating, it was actually funded by the NIH, so US funded, And […] that meant that we had to comply with the…I think it's called Health Insurance Portability and Accountability Act (HIPAA).* (Laboratory manager, Bioengineering)

As this respondent explains, this ended up being a significant drain on project resources:

> *So quite a lot of effort was done by someone associated with that project to familiarise themselves and make sure that all the technologies were set up to protect the data in compliance with those regulations.* (Laboratory manager, Bioengineering)

Although each of these open access legislative mandates refer to open access to publications, many of their provisions could be useful for demanding open access to research data. Specifically, requiring open access as a default is a significant step forward in ensuring research data is accessible. The Liège model, in particular, provides a clear incentive for researchers to publish their research data in an open access repository, since institutional-level decisions will be based on the information that is publicly accessible. The threat of future inaccessibility of government funding also acts as a stick, with clear consequences for not making research data available.

2.4 SUMMARY

This analysis of legal issues related to open access to research data demonstrates the ways in which a range of legal instruments can impact the provision of open access to research data. Intellectual property rights, especially in relation to data that has been purchased from commercial organisations or cultural data, can act as a significant barrier to providing open access to research data, as sometimes the data creators may not hold the intellectual property rights to the material that they collect and to which they seek to provide access. Similarly, research participants, rather than researchers, institutions, repositories and other stakeholders, have primary control over the use of personal information for research purposes, which can limit the extent to which this data can be made open access.

Furthermore, these legal regimes often crate a complex landscape, with real consequences for researchers, organisations and institutions. Open access mandates from governments and funders may place researchers and institutions in a situation where they are pressured to provide open access to data, despite the fact that intellectual property rights or data protection rights specifically and explicitly limit their ability to do so. Open access incentives, such as the Liège model can incentivise researchers and institutions to participate in open access activities, but they should not have a detrimental effect on individuals and organisations that are prevented from participating in open access as a result of other legal obligations. The JRC, in particular, has to negotiate this interconnected and contradictory environment, given its position as part of the European Commission and as a research organisation. Furthermore, the navigation of such a complex environment is a significant drain on researchers' and institutions' time and budget, as expertise needs to be found or developed in order to respond effectively to these intersecting obligations.

3 ETHICAL ISSUES IN OPEN ACCESS TO RESEARCH DATA

This section discusses some of the ethical aspects of open access to research data. First, it revisits several benefits, previously discussed in first RECODE report, that provide support for the now increasingly common notion that making research data publicly available is a "good thing". Second, it addresses some of the ethical concerns that emerged during the literature review and in the case studies interviews.

3.1 THE POTENTIAL BENEFITS OF OPEN ACCESS

The first RECODE report about stakeholder values and ecosystems in relation to open access to research data looked at the different motivations for open access to research data among various stakeholders. It found that open access to research data is now commonly regarded as something to strive for that brings value to science and society. Two moral arguments seem to support this view: open access is good scientific practice and it serves the public interest.

Sveinsdottir, et al., identified a number of proposed benefits that provide support for the argument that open access is good scientific practice.[58] One such benefit is that open access will improve the quality of science, because it allows other researchers to verify and reproduce research results. Ultimately, this would be beneficial for the integrity of and public trust in science. Indeed recent fraud cases in various disciplines involving researchers that manipulated or made up data provide convincing argument to make research data more easily available. Other benefits for science that the report mentions are that it will allow for new forms of collaboration and data sharing; offer scientists a wider range of data to use, compare and re-analyse; minimize the duplication of effort and ultimately speed up the rate of scientific discovery. The successes of international collaborations that embrace the principle of open access such as the Human Genome Project provide compelling examples of these benefits. The immediate availability of DNA sequences has enabled researchers around the world to work on the data, without patent processes constraining or delaying them. As the RECODE report notes, such "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."[59] Anything that can improve the quality, and quantity, of research conducted is therefore, *prima facie*, valued positively.

Box 3.1: The benefits for science

The OECD report, *Principles and Guidelines for access to Research Data from Public Funding*, notes eight specific advantages for science:

1. Reinforces open scientific inquiry;
2. Encourages diversity of analysis and opinion;
3. Promotes new research;
4. Makes possible the testing of new or alternative hypotheses and methods of analysis;
5. Supports studies on data collection methods and measurement;
6. Facilitates the education of new researchers;
7. Enables the exploration of topics not envisioned by the initial investigators;
8. Permits the creation of new data sets when data from multiple sources are combined.[57]

[57] OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, 2007. http://www.oecd.org/sti/sci-tech/38500813.pdf
[58] Sveinsdottir, et al., op. cit., 2013, p. 8.
[59] Ibid., p8

Box 3.2: Sharing data to advance medical research

International collaboration in genomics has demonstrated the advantages of providing open access, for instance in the fight against infectious diseases like Malaria and Polio. In this field, international agreements, such as the Bermuda principles and the Fort Lauderdale agreements, ensure that genome sequences are made public within 24 hours of generation.[60] The aim of these agreements was to create the conditions under which genome sequence data could be released early before the data producers published their results. The principles were the results of an agreement between various stakeholders, including funding agencies, data producers and data users. Researchers and institutions around the world gained free access to large data sets that they could not have produced on their own. This proved to accelerate the scientific process and pace of discovery.[61]

The acquisition of knowledge, and the advancement of human understanding, however, is not a good valued only for the sake of understanding and the satisfaction of curiosity *per se*. Knowledge gained can have an instrumental value that serves the public interest. The OECD report, thus, considers the accessibility to research data to be an important condition in the "good stewardship of the public investment in factual information; the creation of strong value chains of innovation; [and] the enhancement of value from international co-operation."[62] The instrumental value can translate into a commercial as well as political value, for instance when scientific knowledge is used to effectively inform decision-making. It can level the playing field in terms of who has access to information and knowledge[63], which can serve as a counterweight against a government[64], a powerful company[65] or other influential stakeholders that have historically had significant control over the flow of information. In its report *Science as an Open Enterprise*, the Royal Society observed that the hope for open access is that it will "increase public trust and stimulate business activity".[66] Other reports have identified additional social benefits, such as an increased public understanding of science, inspiring the young, allowing the exploitation of the cognitive surplus in society, better quality decision making in government and commerce, and the re-use of data instead of new data collection reduces time and cost to new research results.[67] It

[60]Human Genome Organization, *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing*, Bermuda, 25-28 February 1996; Welcome Trust, *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility,* Fort Lauderdale Report, January 2003. http://www.genome.gov/pages/research/wellcomereport0303.pdf

[61] Knoppers, B. M., et al. (2011). From genomic databases to translation: a call to action. J. Med. Ethics 37, p. 515-516.

[62] Organisation for Economic Cooperation and Development, OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007, p. 9. http://www.oecd.org/sti/sci-tech/38500813.pdf

[63] Sveinsdottir, et al., op. cit., 2013, p. 36.

[64] An example exists in the UK where the government were accused of suppressing a report that would have provided useful information to critics of a government policy. See Wollaston, Sarah, "The government has failed to lead by example on open data", 8 January 2014. http://www.telegraph.co.uk/health/10558580/The-Government-has-failed-to-lead-by-example-on-open-data.html

[65] The pharmaceutical industry has been accused of "burying" the results of unfavourable trials. See Goldacre, Ben, *Bad Pharma*, Fourth Estate, London, 2013.

[66] Royal Society, *Science as an Open Enterprise*, June 2012, p. 7. http://royalsociety.org/policy/projects/science-public-enterprise/report/

[67] Dallmeier-Tiessen, Sunje, Robert Darby, Kathrin Gitmans, Simon Lambert, Jari Suhonen and Michael Wilson, Compilation of Results on Drivers and Barriers and New Opportunities, 9 July 2012, p. 16. Retrieved from http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/outputs/

also provides citizens with opportunities to more actively engage with scientific projects, for instance by helping to locate certain species or identifying the shape of galaxies.[68]

The many benefits provide compelling arguments for open access. However, this does not mean that the various infrastructural, cultural, and financial conditions necessary to enable open access and to contribute toward such outcomes are in place. Instead, the realisation of these proposed positive outcomes to follow open access will depend upon the contingent circumstances. Furthermore, as with any activity, there are also potentially negative impacts upon individuals, organisations and society as a whole. The following section discusses some ethical concerns in order to explore the conditions under which open access to research data can contribute to the identified advantages, while minimising the disadvantages.

3.2 ETHICAL CONCERNS ABOUT OPEN ACCESS

Open access to research data raises several ethical concerns. Many echo or exacerbate existing concerns about sharing research data in general. For instance, publishing data without restrictions may in some cases conflict with established principles of ethical research, including respect for the autonomy of individuals, justice and beneficence. Various disciplines have formalised such principles in codes of ethics, which urge researchers to properly inform participants about the nature of the research, to treat data confidentially and to ensure that benefits and burdens of research are equally distributed.[69] Failing to meet such ethical standards may not only cause harm to research participants, it can also be detrimental to the scientific enterprise or society. Open access to research data raises concerns about the ability of researchers to adhere to these standards and the disruptive effects it may have on existing infrastructures and practices.

The research community has developed a range of strategies to mitigate some of the risks of data sharing, for instance, by de-identifying data through statistical techniques or regulating and monitoring access to research data. The push for open access, however, can come into conflict with some of these strategies, because the aim is to lift restrictions as much as possible. New strategies and solutions may, thus, have to be developed and negotiated or existing ones may have to be adapted.

The sections below discuss some of the ethical concerns that sharing data can generate in more detail and consider how open access relates to these concerns. Note that they may not be an issue in every kind of research and may be evaluated differently in various disciplines. Indeed, some may consider several of the potential disruptive qualities of open access to be benefits, rather than drawbacks or risks. In some disciplines, ethical concerns about open access may not even arise, as some of the respondents in our case studies observed. Yet, in other disciplines some of the concerns can present significant obstacles to making data publicly available.

[68] See for instance Zoouniverse, http://www.zoouniverse.org and Galaxy Zoo, http://www.galaxyzoo.org

[69] Well-known examples include the Nuremberg Code, http://history.nih.gov/research/downloads/nuremberg.pdf; Declaration of Helsinki, http://www.wma.net/en/30publications/10policies/b3/; and the International Sociological Association's (ISA) Code of Ethics, http://www.isa-sociology.org/about/isa_code_of_ethics.htm

3.2.1 Unintended secondary uses and misappropriation

The secondary use of data to validate results, address new questions or apply new analytical methods may produce relevant new insights or scientific advances. It may also help to uncover errors or mistakes in research results, which contributes to sound scientific practices. Sometimes, though, data are misinterpreted, taken out of context or used for purposes that the original researchers or research participants did not intend or anticipate. In some cases this might be – from a researcher's perspective at least - an undesirable, but nevertheless an acceptable, drawback of publishing research results, be they data or publications

However, in some instances the intended secondary use or misappropriation of research data may cause unacceptable damage or distress to individuals and groups, as well as to research and the scientific enterprise. It can harm or wrong research participants or other stakeholders, particularly when results are perceived to be manipulated or distorted or when data are used for purposes that research participants themselves find objectionable. An example is the secondary use of culturally sensitive samples and data, such as human remains. In particular, misinterpretation or misappropriation can offend communities and individuals:

> *Well certainly there are a myriad of First Nations people who may feel offended or compromised if the raw materials related to religious locations, remains etc., are made publicly available and consumable in the wrong fashion. […] if you are putting native artefacts on display on line, information about them online, it really comes down to a whole hodgepodge of historic questions regarding how each particular tribal entity has been treated politically and also what their particular cultural feelings are about such matters. For groups that have less sensitivity about the remains of the deceased, you have to remember that these really represent scores of different cultural sensibilities*. (Editorial reviewer, Archaeology)

Unintended secondary use can damage identities, reputations and relationships between individuals, and may even endanger research subjects or sites. One concern is that the misinterpretation of publicly available medical health data by patients, for instance, can put these patients at risk.[70] It often requires considerable knowledge and expertise to evaluate and interpret research data properly and to use it to decide on medical diagnosis and treatment. Unintended use can be particularly problematic when it involves personal data about research participants' ethnic or racial origins, political opinions, sexuality, religious beliefs, criminal background, or physical or mental health. It may result in stigmatisation, discrimination or other kinds of harm. In addition, research participants may feel wronged or betrayed when their expectations about the use of their information do not match with intentions and practices of new studies.[71]

[70] See for instance: Rehman, Jalees, "Open Science and Access to Medical Research", Guest Blog, *Scientific American,* 24 April 2012. http://blogs.scientificamerican.com/guest-blog/2012/04/24/open-science-and-access-to-medical-research. Also, Janssens, A. Cecile J. W. "Raw Data: Access to Inaccuracy", *Science,* 343(6174), 28 February 2014, p. 968. http://www.sciencemag.org/content/343/6174/968.1.full

[71] Law, Margaret, "Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data", *IASSIST Quarterly* Spring, 2005.

Box 3.3: Ethically objectionable reuse of genome data

Fullerton and Lee have identified some ethically questionable secondary uses of data from the Human Genome Diversity Panel (HGDP)[72]. The HGDP is a collection that contains human tissue samples from 51 different human populations that were originally donated by multiple independent researchers over a period of years. The samples are archived together with geographic location and the sex of the individual from whom the sample was taken. Fuller and Lee reviewed the secondary uses of this collection and found that whereas the majority of studies were in line with the original intent of the collection, some published studies "could be regarded as controversial, objectionable or potentially stigmatizing in their interpretation".[73] One publication that they reviewed used samples from the HGDP to support the findings of a study that examined genetic signatures of Jewish Ancestry in European Americans, concluding that Jewish people are genetically distinct. Fullerton and Lee argue that such studies may cause indirect harm to participants, as they may support potentially unfavourable conclusions about populations from which participants were drawn. It may lead to discrimination or stigmatisation within populations or communities.

Another concern is that unanticipated or unintended uses may harm the reputation of researchers and the public trust in science or social institutions.[74] For example, a physician and policy-maker in the health case study described how the public release of data about hospital performance resulted in some journalists spreading information about the relative quality of different hospitals, despite the fact that the observed differences were not statistically significant.

When third parties use data in ways that participants might find objectionable, these participants, and perhaps future participants, may be more reluctant to cooperate with researchers. In human subject research it is good practice and often a legal obligation to properly inform research participants about the nature of the research and what will happen to the identifiable data provided. However, it is difficult to anticipate all future uses of data and researchers can only offer limited information. If expectations are not met, then the relationship of trust may be damaged, potentially resulting in fewer volunteers who are willing to participate. This may skew research results and the quality of the results.

The concerns about the misuse of data can lead researchers to shield their data, as one of the respondents in the Earth Science case study observed:

> *Some projects that deal with maritime security have to keep their data, or some of them, close. Or they cannot release the data with the same resolution, or only as quick looks, so that one cannot say, zoom in and identify a vessel in the sea. We also have some units that do medical or biodiversity research. […] So for example for*

[72] Fullerton, Stephanie M, and Sandra S-J Lee, "Secondary uses and the governance of de-identified data: Lessons From the Human Genome Diversity Panel", *BMC Medical Ethics,* Vol. 12, No. 16, 2011. http://www.biomedcentral.com/1472-6939/12/16

[73] Ibid., p. 3.

[74] It is salutary to note that such uses do not even necessarily have to have occurred but simply be associated with a particular research collection. See, for example, the frequently reported case of the Havasupai Indians and genetic research into schizophrenia. Lewis, Ricki, "Is the Havasupai Indian Case a Fairytale?", *DNA Science Blog,* 15 August 2013. http://blogs.plos.org/dnascience/2013/08/15/is-the-havasupai-indian-case-a-fairy-tale/

> *endangered species, the researchers are very concerned not to let information out about where these endangered species are concentrated to prevent negative effects from distributing such information.* (Researcher, Earth science)

This reluctance to share precise locations was also echoed in the archaeology case study in relation to discovery sites. Furthermore, writing about genomics, Kaye et al. note in this regard that, "The obligation to share genomic data may be perceived as an imposition on the relationships that have been built between researchers and participants".[75] Thus, researchers are often reluctant to share their data because of the moral obligation and responsibility they feel to protect from harm those who have cooperated with them, e.g., social groups, governments, informants or research participants.

The valid concerns described above are not necessarily reasons to avoid providing open access altogether. In some cases, the benefits of providing unrestricted access to data can offset the potential risks. Thus, although open research data could potentially harm research participants or relevant stakeholders, it may also offer them new opportunities to take control over their data or situation. In the Archaeology case study one respondent pointed out that putting sensitive cultural data online could also help the directly impacted community to gain more insight in the scientific process as well as in their own heritage.

> *If you have an anthropologist or a bio-archaeologist working on a collection of human remains that were excavated in the last century or the century before, then there may be descendants who do not want you to do anything with this material ... making the data available might alert them. ...it's an interesting ethical dilemma... It may be offensive to certain populations, but it's also interesting in that, if it was posted on open context, they actually would have greater access and know what material is being worked on.* (Ethical editorial reviewer, Archaeology)

Moreover, concerns about open access may cast a new light on some existing vulnerabilities of data curation. However, anonymity and confidentiality are increasingly difficult to guarantee in this technological age.[76] From this perspective, it is more ethical to inform participants that they should be aware that third parties could potentially have access to their data[77]. Similarly, open access can provide new opportunities to discuss and address the inherent risk of misinterpretation that data sharing carries. For instance, open access databases can also provide a platform to prompt community discussions and offer further information about the limits and constraints on the use of certain datasets.

Finally, some of these concerns may be mitigated through various kinds of measures, including technological tools, review boards, education, or agreements between multiple parties. In the next Chapter we will discuss some of these measures. Making research data openly available requires further reflection on evaluating the risks and possible ways of mitigating these concerns.

---

[75] Kaye, Jane, Catherine Heeney, Naomi Hawkins, Jantina de Vries and Paula Boddington, "Data Sharing in Genomics: Re-Shaping Scientific Practice", *Nature Reviews Genetics,* Vol. 10, May 2009, pp. 331–335.

[76] Cambon-Thompsen, A. E., Rial-Sebbag, and BM Knoppers, "Trends in Ethical and Legal Frameworks for the Use of Human Biobanks", *European Respiratory Journal*, Vol. 30, No. 2, August 2007, pp. 373–382. DOI: 10.1183/09031936.00165006

[77] Lunshof, Jeantine, Ruth Chadwick, Daniel B. Vorhaus and George M. Church, "From Genetic Privacy to Open Consent", *Nature*, Vol. 9, May 2008, pp. 406-411

3.2.2 Dual use

Some data can be used for research that could produce knowledge, products or technologies that benefit society, but could also pose a threat to public health, agriculture, plants, animals, the environment or material.[78] Such *dual-use* data present an ethical dilemma for data sharing and open access: do the benefits of providing access to research data outweigh the costs? Sharing data on a virus, for example, may facilitate research on an antidote, but people with ill intent may also use it to disrupt societies.

Box 3.4: Avian flu

One of the better-known examples of the dual use dilemma was the publication of two manuscripts that reported on the details of laboratory experiments with the H5N1 avian flu virus. The manuscripts concluded that the virus had a greater potential to be transmitted between mammals, including humans, than previously thought. After various reviews, the journals *Nature* and *Science* decided to publish the articles, because they believed that the benefits of publishing outweighed the risks.[79] After the publication scientists agreed to a one year moratorium "to provide time to explain the public-health benefits of this work, to describe the measures in place to minimise possible risks, and to enable organizations and governments around the world to review their policies (for example, on biosafety, biosecurity, oversight and communication) regarding these experiments".[80]

Concerns about dual use research have led the security and academic community to develop various safeguards and strategies, many of which involve limiting access to the data.[81] These include restricting access to research data and publications: Access may be denied completely, to random parts of data sets, to perturbed data[82] or to data at a particular level of granularity.

For some open access to research data might not be possible. However, that decision requires a thorough discussion on the advantages and disadvantages of releasing the data. Fortunately, mechanisms and measures for are enabling such discussions are, to a certain extent, already in place, including educating researchers about risk mitigation and the establishment of committees and boards that evaluate the risk of dual use for certain research. The Royal Society notes that there is a common sense of responsibility within academic communities regarding dual-use research.[83] Funders and publishers also screen research for potential dual uses. As the example in Box 3.4 illustrates, it is important to consider, as part of these

[78] Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology, *Biotechnology Research in an Age of Terrorism*, National Research Council, The National Academies Press, Washington, DC, 2004.
[79] Royal Society, op. cit., 2012, p. 58
[80] Fouchier, Ron A. M., Adolfo García-Sastre, Yoshihiro Kawaoka, and 37 co-authors, "H5N1 Virus: Transmission Studies Resume for Avian Flu", *Nature,* Vol. 493, No. 609, 31 January 2013. DOI:10.1038/nature11858
[81] Kelley, Maureen, "Infectious Disease Research and Dual-Risk", *Virtual Mentor*, Vol. 8, No. 4, 2006, pp. 230-234.
[82] Kargupta, Hillol and Souptik Datta, Qi Wang, Krishnamoorthy Sivakamur, "Random Data Perturbation Techniques and Privacy Preserving Data Mining", Working Paper. http://134.208.3.165/course/misc/class%20paper/3c_extend.pdf
[83] Ibid., p. 59.

evaluations and discussion, how open access could actually contribute to reducing the risks of dual use.

3.2.3 Violations of privacy and confidentiality

Providing open access to research data makes it increasingly difficult to maintain the confidentiality and privacy of research subjects. The standard practice in human subject research, as prescribed by current ethical codes and international laws, is to anonymise and de-identify the data that research participants provide and to properly inform them about how the data will be used and by whom. When researchers share data with each other or with the public, the risk of breaching confidentiality and privacy increases. Open access and privacy seem to be especially difficult to reconcile, as privacy is often discussed in terms that focus on the control of information.[84]

Anonymisation of data does not suffice to mitigate the risk for all data sets. As a result of technological advances and the availability of increasingly more digital data sets, anonymisation can be more easily undone, for instance by combining and integrating different data sets.[87] Furthermore, in some cases, measures or strategies to preserve confidentiality can be reverse-engineered. One of our respondents in our Archaeology case study noted:

> *[A] major point of concern was that we will indeed be creating a national map of, for sake of argument, everything known. […] if the actual coordinates were used to produce our maps and it was something that somebody could hack out of the XML, that would be pretty bad. And depending on the level of technological training of the different people we are working with, regarding web matting techniques, it's pretty scary and rightfully so. Because if we did it the wrong way, sooner or later some hacker, whether for a nefarious or simply the joy of hacking data, could recreate a CSV of XY data, and put it in a format that someone could download with their GPS unit and run around, so they could stand on our archaeological sites and then the wrong people would stick a shovel in them.* (Editorial reviewer, Archaeology)

Box 3.5: De-anonymising DNA data

In 2008, *PLoS Genetics* published a paper that demonstrated how a particular kind of widely shared genetic data could be used to identify individuals' DNA.[85] In response the US National Institute of Health and the Wellcome Trust, a UK scientific research organization, decided to remove various kinds of genetic data from their publicly accessible Web sites, due to concerns about the potential for re-identification.[86]

[84] Nissenbaum, Helen, *Privacy in Context: Technology, Policy and the Integrity of Social Life*, Stanford University Press, Stanford, CA, 2010.

[85] Homer, Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, David W. Craig, "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays", *Plos Genetics,* Vol. 4, No. 8., 2008. doi:10.1371/journal.pgen.1000167

[86] Couzin, Jennifer, "Whole-Genome Data Not Anonymous, Challenging Assumptions", *Science,* Vol. 321, No. 5894, 5 September 2008, p. 1278.

[87] Although the real world possibilities are subject to some challenge. See, for example, El Emam, Khaled. Elizabeth Jonker, Luk Arbuckle and Bradley Malin, "A Systematic Review of Re-Identification Attacks on Health Data", *Plos One*, Vol. 6, No. 12, 2 December 2011. doi:10.1371/journal.pone.0028071

In some research projects, anonymisation is not even possible because the data content enables identification and resists effective obfuscation. For instance, data from ethnographic studies of particular communities (e.g., transcribed interviews or field notes) may contain descriptions of practices and people that could be easily used to identify specific individuals. A concern is that the removal of identifying characteristics of research subjects may compromise the meaning, integrity and quality of the data.[88] Even if effective anonymisation *is* technically feasible, research participants may still feel uncomfortable:

> *The data that I collect is interviews with people. I have not engaged with open context because I deal with people who are doing illegal acts! Under my ethics review, I provide anonymity. To anonymise the data and to put it on an open context, adds another level. I doubt that many of my participants would agree if one of the stipulations was that I put my data into an open context.* (Ethical editorial reviewer, Archaeology)

As mentioned above, a response to the privacy risks can be to explain to research participants the extent to which subsequent use can be effectively anticipated and to ensure, so far as possible, that the principles that apply to the governance of the data are consistent with the prevailing privacy expectations.[89] False assurance, and associated drawbacks, may also be avoided by not "overpromising", i.e., being transparent and realistic about the possibilities of re-identification. This may affect recruitment strategies and have additional affects on how researchers design and conduct their investigations. It might deter individuals, but it may also provide new opportunities to empower research participants. For instance, identifiable data provides participants with more possibilities of keeping track of their data, and may enable important additional outcomes. For example, some health research initiatives may find additional markers, symptoms or characteristics that locate particular research participants in high-risk categories for specific diseases. In these cases, ethical research practice encourages that the individual in question be contacted with this information, which would not be possible unless identifiable information was stored and shared.

3.2.4 Unequal distribution of research results

Open access to research data can level the playing field, but it is no guarantee that all stakeholders will benefit equally. It may reinforce or even lead to an unequal distribution of those results. Those who lack the required scientific, technical or cultural capital and resources to make use of data are at a disadvantage, even when the data are formally open to all.[90] They may not be able to keep up with the increasing rate of scientific discovery or they are disadvantaged in terms of funding opportunities. Language, for instance, may pose a barrier for some: "how do you make sure that the researcher in Croatia looking for German data, using Croatian language is in a position to find the data that he or she is looking for" (Researcher, Earth science). Another example is that data producers in lower and middle-

---

[88] Parry, Odette and Natasha S. Mauthner, "Whose Data Are They Anyway? Practical, Legal and Ethical Issues in Archiving Qualitative Research Data", *Sociology*, Vol. 38, No. 1, 2004, pp. 139-152. http://dx.doi.org/10.1177/0038038504039366

[89] For discussion of privacy expectations and the relationship with control see Taylor, Mark. J., *Genetic Data and the Law: A Critical Perspective on Privacy Protection,* CUP, Cambridge, 2012, Ch.2.

[90] Mauthner, Natasha S and Odette Parry, "Open Access Digital Data Sharing: Principles, Policies and Practices", *Social Epistemology: A Journal of Knowledge, Culture and Policy*, Vol. 27, No. 1, 2013, pp. 47-67, DOI: 10.1080/02691728.2012.760663. Luo, Airong, and Judith S. Olson. "How Collaboratories Affect Scientists from Developing Countries", in Gary Olson, Ann Zimmerman and Nathan Bos, *Scientific Collaboration on the Internet*, MIT Press, Cambridge, MA, 2008, pp. 365–76.

income countries may not have the same opportunities to access and work with the data they produce as researchers in richer countries:

> *Because GEO is a widely represented organisation in terms of who the participants are, and there are quite a few developing countries, especially African countries who are members of GEO. They say that for them, due to capacity building process it is difficult to provide their users with amazing tools or amazing computers, or otherwise enabling technology or also knowledge of how to handle, how to process certain data to extract information from them. For them the necessity is to try to provide access to information. Because mere access to data cut off certain types of users with certain backgrounds from certain places, from actually benefiting from that access they technically have, but practically they cannot extract anything useful from the data they can access.* (Researcher, Earth science)

For some countries, differences in opportunity to benefit from the data have been a reason *not* to provide open access. A concern is that open access will negatively impact the leverage that these countries have in setting the research agenda in accordance with their interests and priorities or in making sure that the results of the research, such as new vaccines, are distributed equally. Indonesia, for example, refused to provide their biological samples of the Avian flu virus in 2007 to the WHO freely, because it felt that other foreign researchers, governments and companies were acting unethically in various ways, including not notifying Indonesia when samples were shared and only offering co-authorship on publications very late in the manuscript writing. A key objection was that companies in developed countries were profiting from the samples at Indonesia's cost:

> *Disease affected countries, which are usually developing countries, provide information and share biological specimens/virus with the WHO system; then pharmaceutical industries of developed countries obtain free access to this information and specimens, produce and patent the products (diagnostics, vaccines, therapeutics or other technologies), and sell them back to the developing countries at unaffordable prices. […] Moreover, in Indonesia's opinion, what has been emphasised by the current global system is merely the responsibilities of developing countries, leaving a big hole in the "rights" of these nations.*[91]

The unequal distribution of research results is also a concern for researchers. Other researchers may reap the rewards from the efforts that individual researchers have put in developing data sets. Those sacrificing valuable time to create and maintain a high quality data set may be at a disadvantage as compared to those researchers who can spend more time analysing the data and publishing the results. In the Archaeology case study one respondent puts it as follows:

> *So I can imagine people, there* [is] *this sort of class* [of] *researchers out there, alternative academics who are, they don't have normal mainstream faculty appointments. But they are heavily engaged with data and software development and […] that sort of thing. But they don't have any of the job security that normal mainstream faculty people have. So, they are typically funded grant by grant. And they are kind of hired and fired, disposable people. […] So, depending on how the*

[91] Sedyaningsih Endang R, Siti Isfandari, Triono Soendoro and Siti Fadilah Supari. "Towards Mutual Trust, Transparency and Equity in Virus Sharing Mechanism: the Avian Influenza Case of Indonesia", *Annals Academy Medical,* Vol. 37, No. 6, 2008, pp. 482-488.

> *sort of academic work is disturbed I can see, open data and open access mandates, […] have these kind of perverse effects. If it's not part of a sort of larger rethinking about the way that the profession is sort of, overall work place conditions basically, the conditions of labour do need to be considered in this whole work. So, in that sense, I can imagine an exploited class of data miners. Basically get the scut job of making lots of open data, while a couple of celebrity people get lots of the benefits of writing the article and gets published that uses all that data.* (Repository manager, Archaeology)

Another respondent also observed this concern in the health sciences:

> *[T]he real fear is that, people will be working on something and they will deposit the data which they have analysed to a certain extent and then someone will come along and pip them to the post. That's the big issue really.* (Legal expert, Health)

Researchers' concerns about others benefitting from their data is perhaps not so much a problem of open access, but more of the current incentive systems.[92] Evaluations of researchers tend to focus on publications, rather than the creation of high-quality data sets, models or code. Rewarding individual researchers and groups of researchers for producing and maintaining such data sets, may make them more inclined to provide open access to their data.

The potential for unequal distribution of research results is not a sufficient reason to avoid data sharing and making research data public. Rather, the concerns about unequal distribution highlight the need for balanced approaches to open access that take the interests of the various stakeholders into account. Such approaches could entail additional support for certain stakeholders, in terms of technological tools, training and funding. A level playing field may not be a given, but conditions to contribute to more equal opportunities for all relevant stakeholders can nevertheless be created. Effort must be made to ensure that those contributing toward open data, or funding it, are ideally in a position to realise the benefits in practice and, at least, are not placed in a worse position with no reason to accept open access.

3.2.5 Commercialisation

In this section we discuss the relationship between open access to research data and commercialisation, particularly the ability to obtain monetary value from public goods. There are two types of commercialisation relevant to open access debates. The first is an emerging trend within the academy to use research generated by universities for commercial patents, despite this being in obvious contradiction to the stated commitment of universities to generate open knowledge. Second is the use of publicly available research data by private companies to develop new products and services. For example, a respondent from the physics case study described the potential value of innovations generated by scientific research. The respondent describes a presentation by a colleague:

> *It looked at the direct cost of one machine, and compared it to an accepted financial value of the PhDs produced. And it showed such infrastructure was more or less cost*

[92] Kaye, Jane, Catherine Heeney, Naomi Hawkins, Jantina de Vries and Paula Boddington, "Data Sharing in Genomics: Re-Shaping Scientific Practice", *Nature Reviews Genetics*, Vol. 10, 2009, p. 331–5.

neutral. But if you included the technology advancements that building the machine created or drove, it's tens or hundred times more valuable. (Data manager, Physics)

With 'Big Data' or governmental data becoming more available and accessible over the Internet this is a significant area to investigate.

The pressure to 'translate' research into socially useful products more quickly and effectively has grown sharper since the addition of a 'third role' for universities in the 1990s; they became vehicles for economic development as well as institutions of knowledge production and knowledge transmission.[93] Universities can generate income by developing a patents-portfolio; however, requesting and managing patent portfolios can be a costly affair.

Box 3.6: Selling patient data

A 23 February 2014 article in The Telegraph newspaper describes how the records of hospital patients in the UK were were sold to insurance companies. Despite being anonymised, the records contained information like postcode, length of visit, disease and treatment information. This information was then combined with lifestyle data held by credit scoring company Experian in order to enable the insurance companies to revise their premium structure, and increase premiums for some groups.[94] This led to a change in the law to ensure that data could only be disclosed by the Health and Social Care Information Centre "(a) for the purposes of the purposes of the provision of health care or adult social care, or (b) the promotion of health"[95]

Universities have become increasingly prevalent in the commercial world by participating in start-ups and young companies. Several universities have guidelines for these activities, such as The Ohio State University with its Technology Commercialization and Knowledge Transfer Office[96] and and Imperial College in London, where the Imperial Innovations investment fund has been around for almost 30 years. Imperial Innovations seeks to exploit the gap between scientific research and successful commercialisation in the UK by investing private funds in companies and patents.[97] Genome and stem cell research projects are good examples of research that has a strong commercial value for research groups and universities.[98] Governments are also recognising potential benefits to the wider economy and, ultimately, taxpayers by making publicly funded data available openly to stimulate business innovation. For example, the UK government has provided funding for the relatively recently launched Open Data Institute.[99] However, such pushes towards open access sometimes undermine the sustainability of research infrastructure, such as software innovations. A Laboratory Manager

[93] D'Este, P and P. Patel, "University-Industry linkages in the UK: What are the Factors Underlying the Variety of Interactions with Industry?", *Research Policy*, No. 36, 2007, pp. 1295-1313.
[94] Donnelly, Laura, "Hospital records of all NHS patients sold to insurers", *The Telegraph*, 24 February 2014. http://www.telegraph.co.uk/health/healthnews/10656893/Hospital-records-of-all-NHS-patients-sold-to-insurers.html
[95] Care Bill [HL 93] 2013-14, Commons Amendments (13.03.2014), paragraph 45
[96] Technology Commercialization and Knowledge Transfer Office, "Technology Commercialization Office", Ohio State University, 2014. http://tco.osu.edu/
[97] Imperial Innovations, "What we do", 2012. http://www.imperialinnovations.co.uk/about/activities/#sthash.5cACJfl2.dpuf
[98] See for example Genome Canada, "Entrepreneurship Education in Genomics (EEG) Program", 2014. http://www.genomecanada.ca/en/portfolio/research/eeg-program.aspx and Stem Cell Network, "Commercialization", 2009. http://www.stemcellnetwork.ca/index.php?page=commercialization&hl=eng and Harmon, S. H. E., T. Caulfield and Y. Joly, "Open science versus commercialization: a modern research conflict?", *Genome Medicine*, Vol. 4, No. 17, 2012.
[99] See the Open Data Institute, "About the ODI", no date. http://theodi.org/about-us

in Bioengineering explains that when you develop a piece of open source software, it may be quite successful for the first few years, but sustainability becomes and issue.

> *So after a year or two, after the funded has ended and if we haven't actually managed to secure new funding to continue the project, which is a challenge because now there is sort of not much novelty left, you have created this whole sort of novel piece of software. Now you are trying to get funding to maintain it, not really add any novelty, your proposals just aren't competitive.* […] [You have] *got a whole lot of people actually using the software, they become dependent on it and you now leave them in the lurch because the software becomes unmaintained, all the effort they have put into adopting software is now wasted. And I think some people are actually discouraged from the software in the first place, because they actually see its unlikely to be maintained in future.*

Here, the push towards open access can cause some researchers to shy away from specific resources due to concerns about their sustainability.

Such issues are more likely to impact the science, technology, engineering and medicine (STEM) disciplines where patents for inventions and heavy investments are common. Within the humanities and social sciences, these challenges and issues are much less immediate. Most articles focusing on the commercial potential and value of (open) data deal with health studies and governmental data. Medical data, in particular, are highly sensitive and should be handled with great care, especially when commercial goals are at stake.

However, despite these potential issues, the idea of universities being more "commercial" themselves or investing in patents or other profitable secondary activities has spilled over into funding agency policies and has lead to mandates. Several research councils are very explicit in their commercialisation policies.[100] The JRC, in particular, is in a difficult location with respect to commercialisation, where they are prevented as an agency of the European government from interfering with commercial opportunities for companies, whilst also being subject to a mandate to open their data in order to enable European companies to create new products and services, as a Researcher explains:

> *[As an EU institution,] we aspire to promote sharing and make these data available for further reuse by whatever users without any kind of restrictions. So we don't ideally even restrict for commercial use. For us it doesn't matter whether the user will take the data and use it for research purposes, or for purposes of setting up an application that will be sold on the market. […] because we are part of the European government, we have to be very careful how we might affect the business in the market. So whenever there is a possibility of infringing some business interest, we have to be very careful because of our position within the European Union.*

Another interesting issue, which is not directly connected to open data and private companies using that data for new product developments, deals with the more passive possibilities that open access to data can have for companies (e.g., information about users). One of the Archaeology case studies respondents says the following:

[100] Harmon, S. H. E., T. Caulfield and Y. Joly, "Commercialization versus open science: Making sense of the message(s) in the bottle", *Medical Law International*, Vol. 12, No.1, 2012, pp. 3-10 [p. 6].

> *[W]e are publishing all this data, and most of the people that come to us are coming to us via something like Google, or even social media services, like Twitter or Facebook. And these commercial services are getting a huge amount of metrics on this traffic. They might not really care about the data that we have got so much. But they do care about our users. And collecting data about our users seems to be the sort of commodity that is really valuable in this space. In terms of commercially valuable, the information about the preferences and the allocation of attention that our users have is part of the ecosystem of "open data". And so it's that sort of thing that that's what they monetise, Twitter and everybody else.* (Repository manager, Archaeology)

Therefore, open access raises additional ethical issues for individuals at the fringes of the knowledge production ecosystem.

During our workshop an interesting example of commercialisation going the other way around was brought to the table. One of our case study representatives from the European Molecular Biology Laboratory (EMBL) knew of an example of a private company that was sitting on highly specialised genetic data, which was commercially non-usable. The company decided to give the data to the EMBL so others could use it for their own purposes. When the data became publicly available, it was used for different kinds of projects and commercial developments. There are several other examples where open access to data has lead to re-use, also commercially, and reinterpretation. It once again illustrates the complex network of interests and stakeholders, including private companies, researchers and research funders.

3.2.6 Restriction of scientific freedom

Open access requires that researchers take a particular approach toward the data they collect for a particular research project. In order for data to be locatable, assessable and usable by others, there are different kinds of restrictions upon the choices available to researchers in terms of what they can do and how they must do it. These include, but are not limited to, the attachment of standardised meta-data to the datasets produced or the use of specific technical formats and naming standards.

In some disciplines, such standards and conventions can represent significant methodological constraints, which raises concerns about the freedom researchers have to determine their research design and data sharing practices. Some researchers, for instance, argue that their data cannot be separated from context in which it was generated without limiting its meaning and scientific usefulness.[101] Qualitative researchers for example, stress the importance of having tacit or insider knowledge of the particular cultural and material contexts as well as of the relationships between people in order to interpret the data appropriately.[102] Adjusting the data to fit into homogenised and standardised schemes or formats that would make them better accessible and re-usable may affect the structure and quality of their research.

Strict standards can also lead to restrictions on who can conduct research under conditions of open access. In climate research, for instance, there are limits to who can contribute to the data collections. A professor quoted in a report on drivers and barriers in data sharing provides an illustration:

[101] Mauthner and Parry, op. cit., 2013.
[102] Broom, Alex, Lynda Cheshire and Michael Emmison, "Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing", *Sociology*, Vol. 43, No. 6, December 2009, p. 1163-1180.

> *In meteorological and climate research, metadata are very important, and generating them always implies high effort.* […] *The high effort in handling diverse calibration methods and standard verification procedures hampers data re-usability in meteorology.* […] *Therefore it is essential, that only adequate climate institutions are specialised to homogenise and archive climate data.*[103]

Another concern is that researchers will be increasingly driven to rely upon common software tools to exploit the potential of open access. These tools themselves can be proprietary, with analytics increasingly closely guarded even as data itself is released, and as commercial products available to privileged researchers. Such restrictions can negatively impact upon the academic freedom and autonomy of individual researchers in a number of ways.[104]

As researchers are pushed to adopt methodologies consistent with the demands of open access, and re-use data collected by others, the risk exists that the possibilities of scientific research are driven along a particular path dependency. The implications of this may be particularly significant if, over time, the levels of investment to change architecture prevent innovation. Researchers become "locked in" to particular ways of doing things.

Research funders should be sensitive to the risk of encouraging a particular path dependency within scientific inquiry. Instead, researchers should not be penalised for or prevented from using different technological systems if there is a sound scientific argument for an approach that is not aligned with current standards or technologies. This may need to be assessed independently with respect to the compatibility of the proposed, alternate system with policies regarding open access.

3.3 SUMMARY

In this section we have reviewed some of the potential benefits of open access to research data and discussed some of the ethical concerns it raises. These include unintended secondary uses, dual use, violations of privacy and confidentiality, unequal distribution results, commercialisation and restricted scientific freedom. These issues echo existing concerns about data sharing, in particular because open access seems to conflict with some of the strategies and measures developed to address these concerns.

In order to generate broad support for open access to research data, it is important to take ethical concerns into account in developing policies and approaches to making research data openly available. In some instances the potential harm that can result from open access may outweigh the benefits. Nevertheless, concerns about the possible harm that open access may cause should not always be taken at face value or considered to be a reason *per se* to dismiss the possibility of open access. Some concerns, upon further reflection, may turn out to be unjustified in particular instances. In other cases, making research data openly available may

[103] Schäfer, Angela, Heinz Pampel, Hans Pfeiffenberger, Suenje Dallmeier-Tiessen, Satu Tissari, Robert Darby, Krystina Giaretta, David Giaretta, Kathrin Gitmans, Heikki Helin, Simon Lambert, Salvatore Mele, Susan Reilly, Sergio Ruiz, Marie Sandberg, Wouter Schallier, Sabine Schrimpf, Eefke Smit, Max Wilkinson and Michael Wilson, "Baseline Report on Drivers and Barriers in Data Sharing", 28 October 2011, p. 42. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf

[104] Royal Society, op. cit., 2012, p. 7.

turn out to be the solution to some concerns about data sharing. Open access is not an end itself, and the decision to provide unrestricted access to data should take into account the various concerns as well as the potential ways of addressing those concerns. In the following Chapter we will discuss some of the strategies and solutions to mitigate concerns.

4 EXISTING SOLUTIONS AND POTENTIAL PITFALLS

This examination of legal and ethical issues associated with providing open access to research data has suggested a number of findings and tentative recommendations for addressing these issues, with particular reference to the public interest element of enabling open access to research data. Specifically, some solutions are already emerging from this brief analysis that are cross or inter-disciplinary in nature. These themes include practices related to licensing, access management, ethical and legal editorial review mechanisms and other soft-law measures. These solutions could be tested with other disciplines to evaluate if they have the potential to offer meaningful interventions in addressing legal and ethical issues.

However, solutions developed in these areas can also result in the introduction of new pitfalls that can, among other things, affect the degree of "openness" of the relevant research data. This is so when the level of access is considered in accordance with the European Commission's definition of open access. As mentioned above, that definition understands open access as "free internet access to and use of publicly-funded scientific publications and data".[105] A second emerging issue is that the public interest in providing open access to research data must take account of the legal obligations and potential ethical impacts associated with opening this data. As such, solutions need to be found which navigate through such legal and ethical obligations, rather than legal and ethical obligations acting as barriers to open access to research data.

4.1 LICENSING

As mentioned in Section two, licensing provides a useful way to address intellectual property issues as well as ethical issues such as commercialisation and misappropriation and misuse of scientific information. These licensing models include Creative Commons licenses (as the most commonly employed forms of licensing), and other Licenses such as Government Open Licenses. The creators of research data and/or the repositories in which they are stored may make use of licenses to establish clear conditions related to how the research data should be used, including, for example, attributing content to original researchers and restrictions on modifying data. However, whilst licensing presents solutions such as important protections for stakeholders, it also introduces pitfalls such as the availability of licensed material that does not fully comply with the European Commission's definition of open access. Irrespective of this, licensing continues to be a commonly employed practical solution in the move towards open access to research data.

In particular, licenses have been a useful solution to the practical need for multiple-use and re-use of research data within the area of earth sciences, with specific reference to the example of the JRC using research data purchased from private vendors:

> [We have] *a project relating to running the Common Agricultural Policy (CAP) of the European Union* […] *So over time we were able to renegotiate the conditions according to which data were licensed to us and we managed to significantly widen the scope of reuse of the data that we were buying. And for example now when we are*

[105] 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

buying the data for the purposes of maintenance and enforcement of the CAP, we have in our licensing agreement clauses that enable use of this data for all the nine societal benefit areas of GEOSS: health, disasters, etc.: the data can be reused the same data for the purposes of different societal benefit areas. And this was quite a step because if you see the licences that the same companies have for the ordinary users that allow only use of the data on one computer and most often your for internal purposes. So we manage to negotiate with private providers of data for more room for the potential uses of this data. And now, we went as far as even displaying the data on our website which is the website of the Community Image Data (CID) portal. (Researcher, Earth Sciences)

Here, licensing has enabled the JRC to use the purchased material for a range of purposes and to display some of it in a data portal.

The implementation of licenses also provides practical solutions by assisting in achieving varying degrees of access to research data. This may be specifically related to material that includes personal data or other types of protected materials. Particular licensing conditions also allow researchers to remain aware of who is using their work and to have some control over how it is being used. For example, in bioengineering one of the licensing criteria is that the original researchers must be approached about their material and be listed as co-authors on any publication resulting from their data:

You would have to describe your intended use of the data. And then the people who originally were the researchers who gathered that data, would all have to agree to consent to each application. And so they still retain the control of the data. And I think one of the conditions usually if they granted you access to the data, was that you would make them a co-author on your publication. (Laboratory manager, Bioengineering)

Such conditions ensure that researchers are credited for their data, and that the use of their data conforms to the ethical principles against which it was collected. Some requirements only insist on informing researchers about the use of their data, largely because "you want to know what's going on with your work, you want to know who's interested in this, what it can be used for" (Legal expert, Earth science). Some organisations have also found creative ways to communicate licensing conditions, including the use of laundry symbols.

Box 4.1: Laundry symbols

We use the system of what's called "laundry symbols". The terms of each license are represented graphically and graphic symbols are put on the files, on the resources in our infrastructure, so that everyone can see what he can do with the content, without reading the license. […] it consists of three symbols, three colours, green is "public" which basically means open access, it's free for all use, yellow is "academic", so only academic research and red is "restricted", which in principle means that, if you want to use the resource, you have to contact the copyright owner, we'll provide you with his contact information and you'll have to ask him for a specific permission. (Legal expert, Earth science)

A significant number of the projects and initiatives discussed in this report are using the Creative Commons framework for licensing. As for data publications, or more specifically,

database publications, the CC0-licence[106](waiver) has been developed, which means that there are no rights reserved whatsoever. The information stored in these databases is in the public domain without any legal restrictions, including commercial use. Two important aspects stand out regarding the licensing of open data and the issues being faced. First of all there is the difference between *ported* and *unported*. Previous versions (or at least a few) of the CC-licenses are ported licences, which mean that the specific CC-license being used is only applicable to national law and legislation. With the CC 4.0 version it has changed to a solely *unported* license, which means that the license meets international requirements concerning copyright legislation. With online data, and more specifically research data, travelling to potentially every country, this new approach can offer better solutions. Another important change deals with the rights beyond copyright. For example, *sui generis* database rights were not explicitly covered in the previous versions of the Creative Commons licences, which has led to misunderstandings. In version 4.0, applicable s*ui generis* rights have been aligned within the scope of the specific license unless explicitly excluded by the licensor.[107]

However, some CC-licenses are not completely compatible with open access to publications and/or research data. The CC-BY 4.0 license requires only that the user attribute the original creator or source and is compatible with open data. CC0, where no rights are reserved, is also compatible. The other types of Creative Commons license are included in Table 1, below:

License	Comments
Creative Commons - Attribution - Share-Alike (CC – BY - SA)	A user is obligated to credit the original creator and to apply the same licence model to the remixed work.
Creative Commons - Attribution - No Derivatives (CC – BY - ND)	A user must credit the original creator and may only use verbatim copies of the work.
Creative Commons - Attribution - Non-Commercial (CC – BY - NC)	A user must credit the original creator and may not use the material for commercial purposes.

Table 1: Creative Commons licensing

Here, requirements to "share alike" as well as prohibitions against commercial use or derivatives do not allow the "free" use of the material, as preferred by the European Commission.

Nevertheless, Creative Commons licenses, in particular, have proven useful in the field of open access to research data. Open Context promotes the use of Creative Commons licences (CC0 or CC-BY) for their data portal:

> *[I]n the absence of not just international law, but even international disciplinary sensibilities, about how data should be controlled and shared, Creative Commons provides an extremely robust and definitive statement about how we expect things to be used […] obviously these are public data, we didn't build them, we have provided a bridging ontology but we also make that available through an attribution license, we consider that a published public data set. But we prefer that we be given credit for it. But that's it, it could be reused for any purpose.* (Editorial reviewer, Archaeology)

[106] Creative Commons, "About CC0 – 'No rights reserved'", no date. http://creativecommons.org/about/cc0
[107] Creative Commons, "What's new in 4.0", no date. http://creativecommons.org/version4

However, their preference for these least restrictive licences may conflict with national regulatory issues or additional complexities related to the context in which archaeologists may be working. The same sentiment was expressed by the JCR representative in the Earth Sciences data case study in provision of research data that is as close to open access as possible:

> *And we can first of all encourage of the use of open licences but we also help drafting licenses that would be as open as possible. And of course what we do actually quite a lot is we give presentations in which we simply inform the community about what open access is and what they need to achieve it. For instance we are dealing with some common myths like the one that Creative Commons equals open access. And so we are explaining basically what open access really is and how to achieve it.* (Legal expert, Earth science).

Therefore, although Creative Commons Licenses are employed to facilitate open access, they do not always enable that outcome, as they are also used to uphold creator rights. Furthermore, the Creative Commons Licenses commonly employed in relation to research data are not supported by hard law but are instead, customarily used. These barriers to open access exist despite the creator's(s') employing licenses with the intention to facilitate open access by transferring a number of their traditional rights non-exclusively.

Besides the Creative Commons licenses, there are other examples of non-reusable licenses, such as government open licenses. Specifically, both the UK and Canadian government have constructed open licenses [UK Open Government Licence 2.0 (OGL-UK-2.0) and Open Government License – Canada (OGL-Canada-2.0) respectively]. These open licenses allow a user to utilise the material if it is attributed to the original creator. The UK government open license also allows for the resulting material to be released under CC-BY or ODC-BY licenses. Finally, Science Commons has released a *Protocol for Implementing Open Access Data*, which encourages scientists to release data in the public domain, and to rely on "norms" rather than legal requirements. They argue that such a practice "allows for different scientific disciplines to develop different norms for citation." However, expectations should be expressed in "clear, lay-readable" language.[108] Practical examples of this approach include the interdisciplinary Polar Information Commons, the Personal Genome Project, and the Sage Bionetworks Commons for disease research.[109]

Unlike ordinary copyright licenses, a Creative Commons license is not exclusive and thus, on a superficial examination, friendly to the concept of open access. However, such open licenses have particular pitfalls in relation to open access. Although, the aforementioned types of Creative Commons licenses enable creators to make their works conditional, this results in the provision of research data in a form contrary to that envisaged by the European Commission's definition of open access. The conditions and/ or restrictions placed upon the use, dissemination and preservation of the research material by its creators means that not all publicly funded data is truly "freely" available on the Internet. This represents an identifiable pitfall. The only type of Creative Commons license that meets the breadth of the European Commission's definition, are those that do not place any restrictions on the works. This is explained by a legal expert in the earth sciences case study:

[108] Science Commons, *Protocol for Implementing Open Access Data*, Creative Commons, no date.
[109] Parsons, Mark, *Expert report on data policy – open access*, GRDI 2020: An vision for global research data infrastructures, 24 January 2011.

> *It's only Creative Commons BY (attribution) and Creative Commons BY-SA (attribution share alike) that are open licences, the others do not enter into the scope of the definition of open access. For instance, Creative Commons licences that haven the non-commercial (NC) requirement can not be qualified as open licences. The same for No derivatives (ND). […] And we have noticed in our committee that scientists in general know relatively little about Creative Commons; […] especially in the current version, creative commons 4.0. So even trained lawyers are not necessarily aware of all the details of this requirement, […]And there is relatively little case law related to creative commons.* (Legal expert, Earth Science).

This issue is compounded by many researchers making the assumption that Creative Commons is the same thing as open access:

> *[W]hat we do actually quite a lot is we give presentations in which we simply inform the community about what open access is and what they need to achieve it. For instance we are dealing with some common myths like the one that Creative Commons equals open access […] And there is the common confusion I believe that if there is this double C sign on something, it means that it is available in open access and you can do whatever you want with this.* (Legal expert, Earth Science)

More generally, another pitfall introduced by the licensing as a solution is the their potentially limited enforceability at law. Creative Commons Licenses lack legislative and judicial backing. Creative Commons Licenses, in their most open form, are not supported by hard law such as international conventions or legislation. Creative Commons Licenses are instead reliant upon mutual recognition and respect by members of the copyright community. Creative Commons Licensing is a voluntary scheme. Further, in the absence of judicial precedent, the outcome of an open access dispute before the Courts involving a Creative Commons License is unpredictable. Creative Commons Licenses "are just simple contracts" that are subject to interpretation, and as such, represent "a bit of a 'wild west' of intellectual property, because no one really knows what some requirements may mean exactly" (Legal expert, Earth science).

Further, a legal expert from the Earth science case study highlighted another pitfall introduced by licensing for open access. Although the licenses appear to enable open access, the more commonly employed licenses with restrictions can also prohibit the data integration:

> *One of the other things that we struggle with is the, interoperability of the data […] in the legal sense. Where if you deal with data integration especially and you have data from different sources, then you are bound by the conditions restrictions on use on your data and then you might be in a situation where you are locked. Because you cannot, say, combine a certain dataset with another dataset, but that's what you need to make your research or to achieve your research goals, etc. And for this, we were also working on using common licensing conditions, so that there is not much difference in terminology, the restrictions and the conditions of use are understandable in different jurisdictions in the same way. […] So we are looking now at some of these conditions, especially within the framework of working with GEOSS: what kind of licences from the licences of Creative Commons or the ODC licences are compatible with each other. So that the user knows that if it's CC-BY licence, then it's going to be ODC licence, that is equivalent to it, even if their terms might not be a*

> *100% the same. And I think this can be quite an important issue for open access. Because it might be open access on paper, but then once you start actually accessing and trying to use the data you might, face problems that come from incompatible licensing conditions, or terms that are used in different licenses."* (Researcher, Earth science).

With respect to Government Open Licenses described above, they too introduce the same pitfall typical to Creative Commons licenses in that they are not always unrestricted. This is because Government Open Licences may merely provide access that is conditional upon fulfilling one or more requirements. Thus, these licenses will not always enable unconditional, open access. The Repository manager from Archaeology stated that issues related to cultural heritage, national regulations and other complexities might make the use of such licenses impossible or inappropriate.

Despite these pitfalls, licensing is a realistic and practicable solution to the provision of open access to research data. Licenses assist in resolving some of the legal implications of accessing and using works that would otherwise be prohibited under traditional intellectual property laws, such as copyright law. Although some of the Creative Commons licenses are less restrictive than is widely perceived, they still assist in the process of moving towards open access. In any event, it is likely that "authors" of the relevant "works" will continue to push for some recognition for their work.

4.2 ACCESS MANAGEMENT

Archaeology, physics and clinical data all require some form of professional accreditation or other access management review in order to enable researchers to access data. This professional gate-keeping solution allows these disciplines to manage legal and ethical compliance in relation to open access to research data. Specifically, they serve to identify true "professionals" who will have expertise in research methods or legal requirements such as confidentiality, privacy, data protection and research ethics. This solution ensures that the data is used responsibly and any potential issues associated with misuse are identified and mitigated. It also serves as a mechanism for enforcement, whereby individuals who do not use data responsibly may not be "approved" a second time.

The case study participant representatives refer to examples of access management processes implemented by their organisations and institutions. These processes seek to uphold a form of professional "gate-keeping". However, the varying processes implemented to produce solutions to access management are not universally adopted. Access management solutions reflect the needs of the field of research, and the organisations, in which they have been adopted. This arguably introduces a new pitfall in the form of fragmented approaches to access management, which could be resolved by the implementation of a more universal approach to access management in the field of research per se.

Almost all of the case study disciplines made use of some kind of access management strategy. For example, in order to access research data held by government organisations, Archaeologists often have to apply to those governments to gain access to material. In the USA:

> *Each state is within its legal rights to craft their own guidelines as to who can access their materials and in what manner they may access them. […]The usual, the most common guideline is that, somebody is that a person accessing the records, should be on a list of professionals who are authorised to do archaeological work with in the boundaries of that state, […] we have in the United States a group called the 'Register of Professional Archaeologists' which serves as a kind of national level accreditation for somebody who should be capable of managing and completing as a principle investigator, a high level cultural resource management project. If you want to make an analogy to other professions it would be like passing your medical board or bar exam. So you are extensively qualified to do the job anywhere, although in any particular questions of particular local culture histories can come into play. […] So these requirements come into play in different ways as to whether or not people can gain access to archaeological resources, data resources. Each state basically within their rights to deny access to somebody who the feel doesn't meet the requirements, I have been denied by several*. (Editorial reviewer, Archaeology)

Researchers in the Physics case study implement slightly more informal processes:

> *For our own access it's currently based on certificates, which are issued by recognised grid certificate authorities. But that's not something scalable to the general public.* […] *It's simply a mechanism whereby if you are running in a fully distributed environment, you can get authentication, so I run a job in Australia, because I present a certificate that comes from a trusted authority CERN, there is an agreement that they would accept that. So I don't know, maybe you could think of it, as something like a passport, […] it means that someone has, somewhere has checked your credentials.* (Data manager, Physics).

Similarly, the health data case study provides another example of a formalised process of access management:

> *And those who participate in the quality registry or the bio-banks registry all of them have the same, nowadays, have the same possibility to use the data for research purposes but it has to go through a board to approve that the researcher can actually handle the data and has the skills to use the data. But, so in that way, it is open, but it is still, every research question has to be passed through the board.* (Physician and policy-maker, Health)

Finally, in relation to Bioengineering, a large, multi-national data bank of biological material uses the following strategy:

> *They would identify from our website, which data sets they want, because the data sets are listed, […] They would write the access request email, which is on our website […] And they would say, 'I'm interested in these data, can you pass me onto the relevant data owners.' And then there is a form to fill out, which might not be the same for each dataset. And you basically have to explain who you are and why you would to use the data and what you want to use it for, and that's just passed onto the data owners. And if the committee says yes, then we give them access. So the data are encrypted, so we would send a creater user password.* (Scientific services manager, Bioengineering).

These processes are particularly effective in meeting requirements around intellectual

property rights, data protection, secondary or dual use of research material and commercialisation. Therefore, it prevents unethical usage of the research data and aims to achieve and maintain legal compliance.

However, in the absence of a universally adopted and inter-disciplinary access management process, each discipline, institution, organisation or even each project is required to develop its own tailored access management process. Furthermore, as organisations and disciplines employ their own access management processes, issues with respect to cost and efficient access to research data may subsequently arise:

> *I'm very much in favour of different stakeholder communities maintaining their own information systems and repositories where the communities have more direct control over governance issues. It's more expensive, less efficient to be sure, since it means, many more systems that need to be maintained and standards will be harder to implement. But the key point is [...] trying to maximise autonomy and ethical approaches toward data management, and that's not necessarily going to be the cheapest or most efficient approach.* (Repository manager, Archaeology)

Although this pitfall is not overly obstructive for "true professionals" seeking access to research data, those wishing to access research data nevertheless likely face numerous processes during which they may be required to conform with a myriad of differing access requirements subject to the organisation granting access. Furthermore, the reliance on professional accreditation undermines the spirit in which the Commission is seeking to implement access to data. For example, interested members of the public may wish to access data, and would not "pass" these barriers.

A homogenisation of access management practices is taking place in the implementation of Data Management Plans (DMP) as requirements for researchers and the development of integrated data management policies by research insttutions and funders, which include access management as one of the policy items to be dealt with. Such integrated solutions are gradually gaining ground and, in collaboration with various disciplinary instruments, are expected to bring a certain homogenisation in data management and access practices that will enable smoother and wider access to research data.[110] These issues will be discussed in more detail in later RECODE Deliverables (specifically D4.1 on institutional and policy issues in open access to research data). However, more robust support for data management plans and advice about how to construct them would assist researchers in developing high-quality access management procedures that would enable access to research data as far as possible.

4.3 EDITORIAL REVIEW

Third, the use of editorial review mechanisms emerges as a useful tool in ensuring ethical data practice and legal compliance. Internal processes have been adopted amongst our case study participants as a solution to the publication of research data that may have resulted from unethical practices and/ or in a manner that may be contrary to applicable laws. However, the editorial review solution may also introduce new pitfalls not so dissimilar to those associated with access management as described above. By way of specific example,

[110] Significant work in this direction is being carried out by the Digital Curation Centre in the UK, where important resources on research data management. Digital Curation Centre, "Policy resources", 2014. http://www.dcc.ac.uk/resources/policy-and-legal

Open Context adhere to an editorial review process that involves participation from local governments:

> *So what we do is, before it even goes to open context, our data go through a cleaning process, where the sites are allocated to a grid in the grid system and then we scrub the coordinate data and any data that are considered sensitive by our state partners, which can potentially differ state to state, and then we put it up on open context. So the only location information relates to our grid.* (Editorial reviewer, Archaeology)

Another mechanism of editorial review is mandating attribution to the published version of the collaboration:

> *[W]ithin a collaboration people can interpret the data incorrectly and try and publish something that is wrong. […] So to get around that, some collaborations agree somewhat ad hoc that if you publish a paper there has to be at least one member of the original collaboration on the author list.* (Data manager, Physics).

Ensuring ethical practice and legal compliance is an ongoing process and editorial review mechanisms are also useful after the material has appeared in the public domain. This process might also rely on professional bodies or review procedures:

> *[W]e will essentially take down something, we won't delete it until we try to resolve a dispute. And the dispute resolves in the direction of well, the complaint against something is something that our editorial board and what we would probably do is, if it is something that we feel like is that we can't figure this out, we would take it to an ethics committee of a society for American Archaeology and say, yes, help! And that's one of the other reasons, that being part of IPINCH is important for us, because that has built up a very good network of researches […] having this sort of social ties and availability of expertise to figure out those dispute and come to fair resolutions. […] But essentially the idea would be, we would take stuff down and if the dispute resolves in one way that says that basically the data are ok to be up, we will flag it as under dispute. And it would be extra metadata that we would add with some documentation about what the nature of the dispute is.* (Repository manager, Archaeology).

In some instances, the editorial review process is necessary to ensure compliance with specific laws such as the Data Protection Directive (95/46/EC). Researchers at the JRC must submit their data to Data Protection Officers or to legal experts in intellectual property in order to evaluate whether data can be made openly accessible. The additional benefit of adopting this solution is that some staff members gain legal and ethical expertise that frees researchers from needing to devote resources to developing this expertise themselves.

> *It was a database of land samples, so they were like collecting samples of land across Europe and we had a database of chemical analysis of the samples. Now because the database contain also the GPS coordinates of where the sample was collected. There was an issue that was, like ok through the GPS coordinate you can identify the particular land and then if you go, you can verify also who is the owner of that land. So you can trace back the physical person behind the data. When the question reached the Data Protection Coordinator, I think we agreed that we could only disclose the data and the area where it was, but not the GPS coordinate for that matter. So you keep the, for example, the maps that you can build with this data in a*

> *way that you can not really tell whether it is this land or the land that is nearby, that has a particular chemical component in it. So we could disclose data, but we just needed to adjust with how precise the GPS coordinates.* (Legal expert, Earth science)

However, whilst editorial review mechanisms act as a solution facilitating the ethical and legally compliant provision of open access research data, they can introduce a new pitfall. As editorial review mechanisms are tailored to meet publication practices of specific organisations, rather than the world of research data per se. Perhaps each organisation or discipline may favour areas of legal compliance and ethical data practices over others. This may result in less streamlined review processes. More robust data management practices and editorial review mechanisms could be drafted and promoted. Again, the Digital Curation Centre, in particular, is doing important work in this area. They have drafted a *Checklist for a data management plan* that introduces key elements that will assist in establishing good practice in this area.[111]

4.4 SOFT-LAW MEASURES

Finally, the use of existing ethical and legal guidance instruments, such as checklists or professional codes of conduct are also employed by our case study participants as a solution to assist stakeholders in effectively evaluating their responsibilities. However, soft-law measures carry the potential for pitfalls to the extent that although they encourage ethical practices and legal compliance, they do not mandate them.

An ethical editorial reviewer in the Archaeology case study explains the adoption of soft-law measures by their organisation:

> *[W]e take a lot of our clues on the ethical front from various journals and other kinds of venues where people publish this kind of material routinely and most journals and publishing houses have ethical guidelines that they follow. And we look to them sometimes for clues, because it's quite similar in many ways.*

This suggests that existing disciplinary organisations have a significant role to play in assisting open access stakeholders in addressing legal and ethical issues. With this assistance, organisations may implement internal soft-law measures such as the example provided by a representative of Open Context:

> *So the main thing that we could do is to provide that sort of the very general sort of guidance about what our own ethical expectations are. And these expectations go well beyond what's legally required. So it's really about professional ethics that…it's sort of "soft law" I guess, […] and it's not necessarily coded in any sort of statutes.* (Repository manager, Archaeology).

In addition to ensuring that the research data is accessed and utilised in an ethically sound way, and in compliance with the applicable laws, soft-law measures may also seek to ensure that the data are accessed by adequately skilled professionals.

---

[111] Digital Curation Centre, *Checklist for a data management plan,* v.4.0, Edinburgh, 2013. http://www.dcc.ac.uk/resources/data-management-plans

Organisations may implement additional measures, such as internal policies regarding research data management, as a condition of access to the research data. This works to prevent unintended secondary use, commercial uses and intellectual property or data protection violations. The following example was provided in the health case study:

> *And especially for the new developments going on so that you can have patents and things like that, it makes it complicated and so it is of course, the institutions where you are advocating it depends on a lot of things including how we then use the value that can be generated* […] *it's a three party discussion – those who pay for the research, those who do the research and those who might benefit from the research.* […] *We had a very heated discussion in* [my country] *where health records were being used in a prospective way and the person who has to sign this thing he thought that he owned the data where the county council who paid for the people who agreed to having it done because they saw that they could use it for policy prediction in the future, they thought they owned it.* […] *So, we try to have as much as possible written in the policy beforehand.* (Physician and policy-maker, Health case study).

However, soft-law measures do not carry the weight of enforceable guidelines. This issue is highlighted by an Archaeologist who highlights that soft-law measures retain the status of guidelines that carry only an *expectation* that they will be adhered to:

> *So this is one of the harder things about this from a policy making perspective because essentially you can't really set up, really clear guidance except for saying to researchers, make sure you understand in your local context and build trust relationships.* (Repository manager, Archaeology).

Nevertheless, the flexibility of such soft-law measures, precisely because they are not strictly interpreted and do not carry the force of law, make them useful mechanisms to guide data collection, storage and preservation practice. In other cases, more robust, enforcable mechanisms are appropriate. In these casese, soft-law measures can and should be followed by more strict instruments.

5 POLICY RECOMMENDATIONS

This initial analysis has suggested some preliminary recommendations that may assist with providing open access to research data, whilst addressing legal and ethical obligations. While these preliminary recommendations are not exhaustive, they may assist stakeholders in managing legal and ethical issues as the open access sector matures.

1. Explore the use of licensing, especially Creative Commons or similar open licenses, to address legal and ethical issues

As described in the section above, open licenses, and especially Creative Commons licenses, provide a useful way for open access stakeholders to outline how information should be accessed, shared, and re-used. While there are some pitfalls associated with licensing, i.e., they are difficult to enforce, they are often confused with open access, they may restrict re-use, their popularity among open access stakeholders is a testament to their utility and efficacy. Further options for open licensing should be examined, mechanisms to enforce these licenses should be identified and new, interoperable licenses should be developed. Finally, to truly allow data reuse both the database and the datasets should have an open licence wherever possible.

2. Ask different questions

This analysis reveals that responses to legal and ethical obstacles often entail a restriction on access to research data, either through access management systems, editorial review procedures or soft law measures, as well as some licensing frameworks. However, the relationship between open access and (other) legal obligations should not be viewed using a "trade-off" model. Adherence to one should not be considered as necessarily requiring distance from the other. Instead, stakeholders associated with open access to research data should begin by trying to ask different questions. For example, where consent is an issue in relation to personal data, researchers and/or institutions might consider how they might give research participants maximum control over how their data are used, rather than simply closing data sets which might contain personal data, or where research participants may be uncomfortable with the outcomes. One response to this concern can be found in the idea of either "dynamic consent" or a modified "open consent" where communication technologies are utilised to allow research participants to track the use of the data they contributed.[112] While open access poses particular challenges in this regard a requirement to consistently publish a particular identifier, searchable by research participants, might allow subsequent use to be traced and preferences expressed to the nominated contact point. Finally, moving towards open data and open science must occur whilst maintaining rigorous ethical research practices. Researchers within disciplines should explore how their data collection methods can be modified to ensure that the standards for ethical research can be preserved alongside providing open access to their research data.

3. Consider technical or institutional solutions to legal and ethical problems

Not all legal and ethical solutions must be met with legal and ethical instruments. Instead, one of the most interesting elements of the Legal and ethical issues workshop was the ways in

[112] Steinsbekk, K. S., B. K. Myskja and B. Solberg, "Broad Consent *versus* dynamic consent in biobank research: Is passive participation an ethical problem?", *European Journal of Human Genetics*, Vol. 21, No. 9, 2013, pp. 897-902.

which some stakeholders recommended using technical or institutional solutions to meet legal and ethical obligations. One example was of the use of virtual machines or other proxies, whereby the researcher wishing to re-use the data submits their query to the software, and the software runs the query without ever giving the researcher access to the raw data. Other solutions such as "scrubbing" data fields, reducing the detail of the data or access management procedures within institutions also assisted in ensuring that legal and ethical obligations were met. These alternative solutions should be further examined and new technological solutions (especially) should be proposed to ensure legal and ethical practice in open access to research data.

4. Establish and clarify circumstances where it is lawful and appropriate to provide open access to personal data

The discussion of data protection in this report demonstrates that the proposed General Data Protection Regulation may have unintended conflicts with the provision of open access to research data. The Commission should use this opportunity to align their commitment to the strengthening of personal data protection with their commitment to providing open access to research data. This could occur in the form of a working group, or further research to establish areas where specific clarification is necessary and potential solutions for linking the two commitments. This is an especially opportune moment to undertake such an examination, as the GDPR has not yet been finalised.

5. Make better use of internal review processes

Internal review processes are functioning particularly well in some disciplines and institutions to meet requirements realted to particular legal issues (i.e., intellectual property rights and data protection obligations) and ethical research practice. The primary benefit of these review mechanisms is the removal of pressures on researchers to allocate resources to familiarising themselves with legal instruments, in particular, as the navigation of such complex frameworks is a significant drain on researchers' and institutions' time and budget, becuase expertise needs to be found or developed in order to respond effectively to often intersecting obligations. Review committees that specialise in data protection, intellectual property or research ethics can assist researchers and institutions in meetin obligations whilst conducting high-quality research.

6. Establish better institutional reward systems for high-quality data

In order to combat concerns about data mis-use and mis-apprpriation as well as concerns around potential inequalities resulting from open access provision, funding bodies, institutions and other organisations should establish clear rewards for individuals contributing to knowledge by producing high-quality data. As is already recognised in relation to intellectual contributions in the form of research publications, high-quality research data can further scientific knowledge and produce societal benefits, and the production and maintenance of high-quality data sets should be similarly rewarded and supported. Such support would include access to technologies, skills and infrastructure related to data storage and management. This particularly includes providing more robust support for data management plans and advice about how to construct them, which would assist researchers in developing high-quality access management procedures and would enable access to research data as far as possible. Better, standardised data citation practices should also be established to assist in the visibility of high-quality data sets.

7. Accept that some data cannot be "open"

While it is important to consider ways in which open access to research data and legal obligations are met, policy-makers, funders, institutions and researchers have to accept that some data cannot be made open. This may be because of privacy or data protection rights, intellectual property rights, potential for dual-usage or other issues. While all stakeholders should strive to support open access to research data, it may be wise to follow the example of DANS – "open if possible, restricted if necessary".[113]

Furthermore, while institutions, funders and other organisations should encourage researchers to provide open access to data, particularly through incentivising them through practices such as the Liège model, researchers to do not participate in open access to research data for reasons such as data protection requirements, intellectual property rights and/or ethical concerns should not suffer negative consequences as a result of this inability. Researchers, particularly early career researchers, may feel restricted in the types of research they can or should carry out, if the reward mechanisms for providing open access outweigh the benefits of carrying out research that makes use of materials protected by intellectual property or data protection rights, or which may raise ethical issues if they are released. Thus, open access itself should not restrict scientific freedom.

These recommendations are not a magic bullet, however they may assist many categories of stakeholder in implementing open access to research data. As the field matures, new or more optimised solutions will become available to better provide open access to research data. In the interim, these solutions may represent a series of stepping-stones to support these early open access practices.

[113] As part of its mission, the Dutch national repository DANS supports the Open Access principle "open if possible, restricted if necessary", and is aware of the fact that not all data can be freely available and without limitations at all times. Data Archiving and Networked Services, "About DANS", 2009. http://www.dans.knaw.nl/en/book/export/html/178

6 CONCLUSION

This analysis has revealed that despite the legal and ethical barriers to providing open access to research data, many solutions are already being utilised to meet ethical and legal obligations while providing open access as far as possible. Although these solutions all engender some pitfalls, all are useful in making open access a reality. However, making data freely available to anyone and accessible over the Internet may sometimes leave researchers, repositories, commercial organisations, research participants and members of the public vulnerable in important ways. While new solutions should be sought to provide legal and ethical pathways to open access, the current policy push towards open access must accept some limits and caveats. These may be related to intellectual property or data protection and ethical research practice. This will ensure both the public interest in opening research data, better informing citizens and assisting in innovation as well as the public interest in protecting knowledge production, maintaining privacy and data protection rights and ensuring ethical research practices.