RESEARCH PAPER

Issues in the development of open access to research data

Bridgette Wessels[a]*, Rachel L. Finn[b], Peter Linde[c], Paolo Mazzetti[d],
Stefano Nativi[d], Susan Riley[e], Rod Smallwood[f], Mark J. Taylor[g], Victoria Tsoukala[h],
Kush Wadhwa[b] and Sally Wyatt[i]

[a]*Department of Sociological Studies, University of Sheffield, Sheffield, UK;* [b]*Trilateral
Research & Consulting LLP, London, UK;* [c]*Library, Blekinge Institute of Technology,
Karlskrona, Sweden;* [d]*National Research Council of Italy, Area della Ricerca di Roma1,
Monterotondo, Rome, Italy;* [e]*Stichting LIBER Foundation, National Library of the
Netherlands, The Hague, The Netherlands;* [f]*Kroto Research Institute, University of Sheffield,
Sheffield, UK;* [g]*School of Law, University of Sheffield, Sheffield, UK;* [h]*National
Documentation Centre/NHRF, Athens, Greece;* [i]*eHumanities Group, Royal Netherlands
Academy of Arts and Sciences, Amsterdam, The Netherlands*

*This paper explores key issues in the development of open access to research
data. The use of digital means for developing, storing and manipulating data is
creating a focus on 'data-driven science'. One aspect of this focus is the develop-
ment of 'open access' to research data. Open access to research data refers to
the way in which various types of data are openly available to public and private
stakeholders, user communities and citizens. Open access to research data, how-
ever, involves more than simply providing easier and wider access to data for
potential user groups. The development of open access requires attention to the
ways data are considered in different areas of research. We identify how open
access is being unevenly developed across the research environment and the con-
sequences this has in terms of generating data gaps. Data gaps refer to the way
data becomes detached from published conclusions. To address these issues, we
examine four main areas in developing open access to research data: stakeholder
roles and values; technological requirements for managing and sharing data;
legal and ethical regulations and procedures; institutional roles and policy
frameworks. We conclude that problems of variability and consistency across the
open access ecosystem need to be addressed within and between these areas to
ensure that risks surrounding a data gap are managed in open access.*

Introduction

The development of digital means for producing, storing and manipulating data is
creating a focus on 'data-led science' and 'open access' to research data (Royal
Society, 2012, p.7). Developments in 'e-research' (Beaulieu and Wouters, 2009),
namely the use of digital technologies to support new and existing forms of research,
are fostering a reconsideration of the ways scientific and scholarly knowledge is pro-
duced and shared (Jankowski, 2009; Royal Society, 2012). Open access to research
data refers to making various types of data openly available to public and private
stakeholders, user communities and citizens. The development of open access
involves a reconsideration of the processes of the production and dissemination of
knowledge. In this paper, we address two related issues in the development of open

---

*Corresponding author. Email: b.wessels@sheffield.ac.uk

access to data, which are the uneven development of an open access ecosystem and how this might have consequences in the development of 'data gaps'.

The 'data gap' refers to the way data becomes detached from published conclusions through open access (Royal Society, 2012, p.26). What this means is that users can access data as a discrete entity and that they do not need to refer to publications related to the data. This is generating a different approach to the practice of open inquiry that underpins research in that traditionally research findings were verified through peer-reviewed publication in which primary data was not openly shared. In the traditional process, data and published findings are tightly integrated and remain connected. In open access, however, there is a gap between data and its published results. The consequences of this split need to be addressed to ensure the integrity of the relationship between research data and their interpretation in the drawing up of research conclusions.

The issue of a data gap is located in the broader move to open access, which in itself involves significant changes in the process and practice of research. Open access to data extends across the life cycle of the production of knowledge, including data collection, data analysis, data management and publication of findings, as well as the legal and ethical frameworks guiding research. Although some developments are shared across research practices, these are adapted within specific disciplines in the physical sciences, social sciences and humanities. This means that open access to data varies across research disciplines and in interdisciplinary research collaborations. The practices and norms of specific disciplinary research are embedded in wider disciplinary and interdisciplinary values of knowledge production. The range of research practices across scientific disciplines means that the development of open access varies across disciplines and is not fully understood in the same way across the research community.

To address both the data gap and the uneven development of open access to data across disciplines, we consider open access as an ecosystem of data generation, management, curation and access (Sveinsdottir *et al.*, 2013). This enables us to examine open access to data in relation to data collection and interpretation processes in terms of how open access may or may not ensure the integrity of the data and their interpretation. We argue that the process of enabling open access to research data involves addressing four main areas: stakeholder roles and values; technological requirements for managing, sharing, curating and using data; legal and ethical regulations and procedures; and institutional roles and policy frameworks. We further argue that it is important to address these areas because they provide the context in which to assess the issue of a data gap. It should be noted that this paper is addressing open access to research data, and that when we write 'open access' without qualification we are referring to the whole sector (i.e. open access publications and open access to research data).

Following on from this introduction, we delineate some important lexical caveats before discussing the context of the debate about open access. We then identify and discuss the following key areas of open access to data, which are: stakeholder roles and values, infrastructure and technology, legal and ethical complexities, and institutional issues and policy. In the conclusion, we argue that the unevenness of how open access is being developed has consequences in terms of creating data gaps, which need to be addressed to ensure that open access is beneficial to both the research community and wider stakeholders.

Defining research data and open access

The lack of attention paid to the specific characteristics of data is also evidenced in the way data and open access are being defined. The European Commission, for example, defines open access as 'free internet access to and use of publicly-funded scientific publications and data' (European Commission, 2012b, p.13). This definition is very broad and sees that research data are an integral part of the open access paradigm. Its broadness means it can cover a range of different research processes. However, as a result, the specificity of data is not especially highlighted. The Berlin Declaration's vision of open access is similarly broad in that it sees open access to data as having the potential to create 'a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community' (Max Planck Society, 2003, p.1). The broad approach is further seen in the way the Berlin Declaration states that open access contributions include all of the following: original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials, and scholarly multimedia material. The Declaration's criteria for open access contributions identify two specific points in terms of the process of open access. First, authors and rights holders must grant users free access to the materials, including a license to copy, use, distribute and display material subject to proper attribution of authorship and responsible use. Second, a complete version of the work should be in an appropriate standard format and submitted in an online repository with suitable technical standards that seek to enable open access, unrestricted distribution, interoperability and long-term archiving (Max Planck Society, 2003).[1]

Definitions of research data are similarly broad. Policy definitions suggest that any material used as a foundation for research can be classified as research data, whether published texts, artifacts or raw unprocessed data. The Organisation for Economic Co-operation and Development (OECD) definition, for example, includes any kind of resource that is useful to researchers (OECD, 2007). In the most recent survey on information in the digital age, the European Commission defines research data as data which 'may be numerical/quantitative, descriptive/qualitative or visual, raw or analyzed, experimental or observational. Examples are digitized primary research data, photographs and images, films, etc.' (European Commission, 2012b, p.45). Other definitions of data include datasets, which are collections of factual information, and linked data, where data is described by a unique identifier that enables the linking of data. The Royal Society does provide some criteria that provide a framework for defining open access to data in that it states that open data refers to data that is accessible, usable, assessable and able to be evaluated (Royal Society, 2012, p.12).

Despite the above-mentioned variability and difficulties of definition, there is a strong policy push to develop open access at the national level and within certain world regions (e.g. Europe, North America and Asia-Pacific). The European Union provides an exemplar of the issues involved in developing open access, and this region is the major focus of this paper. There are a number of policies, initiatives and projects in the European research community that seek to support the development of open access to research data, the linkages between research data and publications, and the preservation of scientific data (e.g. FP7 OA pilot (European Commission, 2008), APARSEN,[2] DRIVER and DRIVER-II,[3] DARIAH (http://www.dariah.eu/) and OpenAIREplus[4]). Many of these projects and initiatives address

the barriers associated with making data more accessible: for example, intellectual property issues, ethical considerations, conflicting stakeholder values, and disciplinary differences. However, each initiative focuses on a specific aspect of open access in general terms without necessarily addressing how to address the specificity of data and its management in an open access ecosystem. One notable exception is the newly issued European Commission Recommendation of July 2012, which integrates open access to research data (alongside open access to scientific publications, development of e-infrastructures and improved stakeholder collaboration) within a larger field of open access to scientific information (European Commission, 2012b). Given the lack of attention to some of the details of making data openly available, it is important to consider various disciplinary, stakeholder and research practices. This requires examining, in the first instance, stakeholder values and mechanisms for integration.

The impetus for developing open access and the challenges in making research data open

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. The OECD, for example, argues that, given that research is publicly funded, it should therefore be made available to a range of stakeholders (OECD, 2007). The perceived benefits of making data openly available are that researchers will be able to re-use data in subsequent work, preventing costly duplication. Open access to research data also enables the validation of research results by assisting reproducibility and ensuring quality control. It is further argued that policy makers could use the data to inform decision making and the private sector could use them in the development of new products and services, and thus data may also have economic value. Further, civil society organisations and citizens would have access to data to inform themselves about important scientific developments and to participate in public debates (Royal Society, 2012).

These perceived benefits rest on an approach that sees data in very broad terms. This approach does not fully consider the specificity of different types of data and research practice in relation to making data openly available. The case of particle physics, for example, illustrates the complexity of open access in the context of extremely large volumes of data. The large hadron collider (LHC) of the European Organisation for Nuclear Research (CERN) produces about 15 petabytes of data per annum, and analysing this vast quantity of data requires the world's largest computing grid, the LHC computing grid. This area of research involves collecting, disseminating, storing and processing large quantities of numerical data from experiments which have hundreds of academic partners around the world. Before recording the raw data, they are pre-processed to reduce the number of events from around 40 million per second to 200 per second. Even with this reduction, it is not clear whether it is actually possible to make the data publicly available. 'Big science' requires one-of-a-kind facilities, and the resources necessary for storing and processing the data are available only to very large consortia. In many ways, these data are already widely available as there are 111 countries involved, and 20,000 users of the LHC computing grid. Access is, however, controlled in that users are members of the field's scientific community (who may well be the only people who understand the data). Before making these data openly available, the following questions need to be

addressed: at what stage in the process can data be made open, are those who access the data knowledgeable enough to interpret them, and what tools will be needed to access these very large and complex data?

Another consequence of not addressing the specificity of different types of data is illustrated in the field of bioengineering. The data gap in this context relates to the scrutiny of findings in relation to research validation processes. For example, the virtual physiological human (VPH) project aims to develop computational models of the whole of human physiology as a route to improving human health and longevity. There is a perception within the VPH community that the data used for developing computational models of human physiology are, in a sense, fragile, and that the outputs of computational models of extremely complex systems may not be replicable in the manner that is expected for acceptance in the current scientific paradigm (Niederer *et al.*, 2009). There are many levels at which these issues can be raised: how is the initial reduction in complexity (which is essential in order to make the problem computationally tractable) validated; what is the effect of determining parameters in a variety of species; how is the lack of consensus on biological mechanisms dealt with in a robust manner (i.e. how does one handle missing information); how can a complex model be described in a manner which enables reproducibility; and, is reproducibility of results an impossible condition to meet when the results may be the end product of tens of person-years of work? This example highlights the importance of the link between data and interpretation in the cumulative production of scientific knowledge.

Given this kind of complexity around data collection, processing and interpretation, the Royal Society (2012) suggests that open access to data must ensure that the provenance and clarity of data and metadata are clearly understood by stakeholders and by those accessing the data. To ensure that open access to data meets these requirements, a coherent system for making data openly available needs to be developed. If the process that supports open access to research data lacks coherence, there is greater risk of data being misinterpreted, which will undermine the validity and robustness of open access more generally. Further, researchers will need to understand how to prepare data for open access and the system for generating open access to data. However, many researchers currently lack the tools, standards and information to make their data publicly available (Repositories Support Project, 2011). In broad terms, the lack of attention to the details in operationalizing open access to specific data and a lack of research skills in making data open are resulting in an uneven development of open access to data across research areas. This, combined with insufficient strategies, the need to expand repositories, and a lack of funding, is a major barrier to enhancing coherent open access to research data (Directorate-General for Research and Innovation, 2012, p.28). These concerns suggest that there is a need to understand the practices of scientific disciplines to ensure the rigour of scientific data generation and interpretation is sustained when making data openly available.

Stakeholder roles and values

The way stakeholders can be linked in developing a system for making data open is one of the challenges in developing open access to research data. Stakeholders include universities, publishers, public and private research organisations, software developers, libraries, funding bodies and repositories. Each of these stakeholders

tends to be connected to a specific area of open access, the open access process, and particular data dissemination and preservation initiatives. For example, OpenAIRE-plus focuses on researchers within the European Funding Programme, and DARIAH focuses on the arts and humanities. These connections occur at various levels via a wide range of different types of stakeholder organisations mentioned above. There is, however, a lack of clarity about stakeholder roles and responsibilities, including identifying which stakeholders are responsible for ensuring that open access to data is promoted and that data are maintained once they are made public (European Commission, 2009, pp.6–7).

Stakeholders have different values, drivers and interests (Bulger *et al.*, 2011). For example, in seeking to create profit, industry partners and funders may well restrict access to research data in order to protect their knowledge base and source of revenue. Academics often wish to restrict access in order to maintain their intellectual property rights, to develop future publications, to maintain their own careers or league table positions and to gain recognition among their peers. In the context of public-private research collaborations, partners have different motivations for producing data that pose unique challenges for data policy and practice (Wouters and Schröder, 2003). In this context, private-sector partners may wish to maintain commercial secrecy, making funding for research contingent upon such secrecy, whereas academic partners may require open access to data to publish their results in peer-reviewed journals. In the context of public policy, policy makers and funding bodies seek to increase access to research data to extract maximum (public) value from their investment. Furthermore, even within universities (a major stakeholder group), disciplinary differences affect open access and data sharing. Increasingly, research questions demand access to data from different disciplines, yet disciplines differ in their approach to data sharing and re-use. It can be difficult to use data sets produced by others without sufficiently descriptive and understandable metadata (Zuiderwijk *et al.*, 2012).

Research on the environment, for example, seeks to understand global environmental change and to mitigate its effects. This entails interdisciplinary research that cuts across many domains and needs to address interoperability in open access. The US National Science Foundation (NSF) recently launched the EarthCube initiative, which aims to transform the conduct of research through the development of community-guided cyberinfrastructure to integrate information and data across the geosciences. The Governance Working Group proposed that 'EarthCube governance shall strive for the free and open sharing of data, information, software and services' (Governance Working Group, 2012). These types of initiatives are facing a set of common challenges in developing flexible multidisciplinary systems of systems. To address this, EarthCube based itself on a Ning domain (see http://earthcube.ning.com/) and, given that Ning is a commercial service, it leaves EarthCube exposed to commercial agendas. Further, the commercial Ning branding is somewhat at odds with the EarthCube ethos. In this context, the European Commission-funded 'A European Approach to Global Earth Observation System of Systems' (EuroGEOSS) project developed an innovative operating capacity to make existing systems and applications for geoscience observation (including observations about drought, forestry and biodiversity) interoperable. In addition to providing interoperable access to data, this capacity provides access to analytical models that scientists from different disciples have used to make the data more understandable, which addresses the data gap and seeks to make open access intelligent, as suggested by the Royal Society

(2012). EuroGEOSS is an example of publicly funded development and support that to some degree counters some of the more commercially driven services.

Issues of differing stakeholder values are also evident in the humanities (involving disciplines such as archaeology, palaeoanthropology, zooarchaeology, palaeobotany and history). These disciplines are based on the collection and analysis of diverse types of data, ranging from collections of ancient primary texts to collections of animal bones, coins and various other artifacts found in the ground. The challenges regarding open access to research data in the humanities pertain to the particularities of scholarly communication in these disciplines. These include a very slow turnover of publications compared with the natural sciences, authors' unwillingness to share their data and the diversity of the data themselves, as well as issues of intellectual property, especially with regard to the public display of cultural artifacts. Nonetheless, advances in information and communication technologies have had a profound effect on these disciplines, giving rise to new approaches to the study of human societies and their past and present achievements. This is most evident in the recent explosion of digital humanities initiatives. The term is often used to describe use of computational methodologies to the humanities, usually involving research into large volumes of digitally born and/or stored data.[5] Research funding agencies around the world, especially in the US, the UK, the Netherlands and Canada, provide incentives to promote the use of current technologies for new, data-intensive approaches to the humanities, such as the Digging into Data Challenge (http://www.diggingintodata.org/) that funds computationally intensive humanities projects. The drive to fund digital humanities may disrupt established scholarly cultures in the humanities that can be resistant to an open approach to sharing data that individual scholars have to negotiate.

Despite some conflicting values demonstrated in our examples, stakeholders concerned with open access to research data, and the general open access process, are highly dependent upon one another. This dependency is a factor in finding ways to overcome conflicts of interest, such as the Digging into Data Challenge, and in developing a EuroGEOSS governance framework to support interoperability. The differences between disciplines suggest that a single model of open access will not be appropriate for all disciplines. For example, in medical research, privacy constraints may make the sharing of research data difficult, while sociologists studying risk scenarios may instigate undesirable economic impacts if data is shared. This suggests that a variety of institutional models will be needed to ensure that open access is workable given disciplinary constraints, but these models will also have to be compatible to overcome risks of uneven development and gaps in legal and ethical frameworks.

The picture becomes even more complicated in the international arena. As research becomes increasingly global, data-intensive and multifaceted (Nowotny *et al.*, 2001), it is imperative to address national and international data access and sharing issues systematically. Europe, North America and Australia (OECD, 2007) have a similar distribution of open access repositories as well as publishing and science infrastructures, yet policy and disciplinary differences remain between, and sometimes within, countries. Furthermore, the situation in other parts of the world significantly diverges from Europe, Australia and North America: stakeholder structures, motivations and interrelations can be influenced by political and cultural differences. A handful of countries (Japan, Taiwan, China and India) control two thirds of Asia's data repositories (OpenDOAR, 2013) and, in less industrialised countries, such as those in Africa, other challenges, including underdeveloped and often

unreliable Internet and electricity infrastructures (although mobile Internet access is improving significantly) and a lack of affordable access to expensive academic journals, persist. Funding bodies in developing countries may also face sustainability and reliability obstacles and, as in many highly industrialised countries, may be governed by commercial interests rather than user needs (Chan *et al.*, 2011). Finally, one of the key differentiators of stakeholders in open access is scientific discipline, where some disciplines have well-developed open access portals and collaboration mechanisms while others are significantly underdeveloped (Directorate-General for Research and Innovation, 2011).

Infrastructure and technology

Open access to research data requires advanced infrastructures and technological solutions to assure cross-disciplinarity, sustainability and low entry barriers for both providers and users. The Internet provides the basis for such infrastructures because of the wide adoption of its protocols, technologies, machine-to-machine accessibility through web services, and user-friendly navigation paradigm. In particular, recent web-based innovations, such as the Web 2.0 approach and the semantic web (including linked data), contribute to lowering the entry barriers for data users.

However, a range of issues arises when considering the additional technological and infrastructural needs (in addition to Internet access) of an open-access system focused on scientific data. Issues include securing trustworthy data, utility, discoverability, access management, data selection, heterogeneous formats and structural complexity (Manzella *et al.*, 2009). The main challenges in developing full and effective open access to research data include interoperability to make different types of data or repositories work together (Habert and Huc, 2010; Bulger *et al.*, 2011), data curation and long-term preservation to address technological obsolescence (Muir, 2004), scalability to support storage and processing of large amounts of data, policy and security support, and data quality representation.

Different solutions have been proposed in the context of past and ongoing initiatives, programmes and projects. For example, in the Earth observation and geoscience domain, where datasets and systems heterogeneity are well-known issues, several architectural and technological solutions (e.g. protocols, metadata, data models) to address interoperability have been developed. Standards are widely adopted, but the heterogeneity of user requirements makes it difficult to identify a common set of accepted specifications. The system-of-systems engineering (SOSE) and, in particular, the brokering approach proved to be effective in addressing the remaining heterogeneity (Nativi *et al.*, 2011). Currently, two European and international programmes, INSPIRE (INfrastructure for SPatial InfoRmation in Europe) and GEO-GEOSS (Group on Earth Observation), have recommended a set of specific sharing principles pushing open data discovery, access and use:

- Keep the existing capacities as autonomous as possible by interconnecting and mediating standard and non-standard capacities.
- Supplement but do not supplant systems mandates and governance arrangements.
- Assure a low entry barrier for both resource users and producers.
- Be flexible enough to accommodate existing and future information systems.

- Build incrementally on existing infrastructures (information systems) and introduce distribution and mediation functionalities to interconnect heterogeneous resources (Craglia *et al.*, 2011, p.6).

These principles build on general information engineering principles, such as separation of concerns (Dijkstra, 1974), and on specific Internet and world wide web principles such as layered systems (Fielding, 2000) and extensibility (W3C, 2004). The SOSE approach and brokered architectures complement them with solutions addressing specific issues and requirements for connecting heterogeneous and autonomous systems. With approaches such as these in place, the benefits of relieving individual repositories from the complexities of implementing interoperability can be realized in supporting interdisciplinary scientific cooperation. These include increased information access and use, sustainability of cross-domain discovery and facilitating future technology insertion in a consistent manner (Craglia *et al.*, 2011).

Crowd sourcing is one approach being used within the widening frame of cooperation. Crowd sourcing enables specific groups of people to contribute data by undertaking some fairly straightforward research tasks. The results are published online by organisers of the research process. Popular examples from different disciplines include medicine (Laurenti, 2012), biodiversity (Sotiriou, 2013) and astronomy (American Scientist, 2013). However, some quality problems may arise during the process, such as non-serious submissions and people presenting vague solutions because they are trying to obtain a monetary reward. To make crowd sourcing successful, these problems need to be solved and programmes and initiatives which use these resources need to have some quality assurance for their products (Saengkhattiya *et al.*, 2012).

The Earth observation domain is addressing the problem of data quality and uncertainty in representation in crowd sourcing contexts. In particular, it is propelled by the need to include observations from unconventional and non-authoritative sources (such as crowd sourcing and citizen science applications). This experience in the Earth observation domain has resulted in several proposed solutions for the representation of data quality. These include UncertML (n.d.), which has defined uncertainty and quality data models, Bigagli and Nativi's (2013) related encodings definition in common metadata and data formats (e.g. netCDF-U), and the testing of technological solutions for adding quality information through metadata enrichment and users' annotation in the GeoViQua project (GeoViQua Project Consortium, n.d.). The concept of a quality label for GEO datasets is under investigation in several projects (e.g. GeoViQua, EGIDA [Coordinating Earth and Environmental Cross-Disciplinary Projects to Promote GEOSS]) (Parsons *et al.*, 2011). Although there is some development, further work in this area is ongoing.

In recent years, significant effort has focused on the long-term preservation of digital information. This has resulted in three approaches to digital preservation being proposed and adopted:

- Technology preservation – preserving the original software that was used to create and access the information.
- Emulation – making future powerful computer systems that can emulate older, obsolete computer platforms and operating systems as required.
- Migration – ensuring that the digital information is re-encoded in new formats before the old format becomes obsolete (Digital Preservation Coalition, n.d.).

Some standardised solutions have been defined for all three approaches, enabling long-term access to open archives. For example, the Internet Engineering Task Force (IETF) has described the requirements for long-term Internet archive services (Wallace *et al.*, 2007). With a specific focus on Earth observation (EO) data, the Long-Term Data Preservation Working Group (2012), which is composed of members of several space agencies, has also recently released guidelines for a European EO Long-Term Data Preservation Framework.

Legal and ethical complexities

Open access to research data raises considerable and fragmented legal and ethical challenges. The legal issues surrounding open access to scientific data primarily include intellectual property considerations, ownership, freedom of information (FoI) considerations, privacy laws, data protection laws and human rights considerations.

Intellectual property encompasses trademarks, design rights, patents and copyright. One key, unexplored intellectual property issue is trade secrets, which protect confidential business information. This could significantly disrupt the sharing of scientific information generated by private research, particularly health or pharmaceutical research (Payne, 2012). Copyright is also a relevant intellectual property issue, and provides exclusive control over the copying, distribution, performance and display of a piece of work (Smith and Hansen, n.d.). Some countries, for example the US, exclude information generated by governments from copyright as well as information contained within databases. However, in other cases, such as the UK, one finds that 'sweat of the brow' protections in common law protect databases and the associated scientific data content (Uhlir and Schröder, 2007). Another such example is the European directive on the legal protection of databases, which created a *sui generis* right for database producers that seeks to protect their investment of resources as well as to harmonise copyright laws applicable to the contents of the database (Europa, 1996). However, it is important to note that there is no legal definition of when a collection of data becomes a database, and none of these issues has been tested by courts of law in relation to scientific data. Furthermore, Rodrigues (2009) notes that providing open access to data may be precluded by publishing practices that sign copyright over to publishers, who themselves may have individual copyright policies. This heterogeneity has led Latvia, Austria and Greece to call for a harmonisation of European copyright law to assist scientists in disseminating their work, 'reflect the conditions of modern digital preservation', drive growth and innovation and preserve cultural and scientific heritage (Directorate-General for Research and Innovation, 2011, p.51).

Legal issues, such as ownership and FoI requirements, also have some applicability to open access to research data. In relation to the ownership of scientific data deposited in repositories, there is some confusion as to whether libraries or repositories have the ability to copy data, or possibly change it, into different formats in order to preserve it (Muir, 2004). FoI requirements give individuals the right to request recorded information held by a public authority, including public universities in some contexts (Wilson, 2011). Such requirements could lead to misunderstandings or information misuse, including suppression of particular pieces of work for political ends or enabling scientists to request competitors' research data (Corner and Bell, 2010). However, scientists may refuse FoI requests if there is a plan to make the data publicly accessible: for example, by depositing it in an open access database. Thus,

in the context of FoI, data preservation may protect scientists' intellectual property during the critical, pre-publication phase of the research.

Another set of distinct challenges is raised by access to, and preservation of, datasets that contain personal data. Many countries (including approximately 80% of all OECD member countries; OECD, 2006) have introduced privacy and data protection laws, designed to protect data subjects from violations of human rights – in particular, the infringement of the right to a private life that can follow the inappropriate access, use and retention of personal information. In Europe, since the treaty of Lisbon, the protection of fundamental rights and freedoms has been effectively incorporated into the European constitution through articles 7 and 8 of the EU charter of fundamental rights and freedoms and EC directive 95/46/EC on the protection of personal data. Inconsistencies in privacy protection laws, and in particular within Europe, in the implementation of directive 95/46/EC, have long caused problems for those seeking to share data across frontiers. The EC has recognised the need to harmonise data protection laws within Europe and has proposed a new general data protection regulation (European Commission, 2012a). This draft regulation includes stronger consideration of the issues of identifiability and consent. However, both commercial and scientific stakeholders regard this as an imperfect solution. Protections for personal data might hinder the development of new business opportunities, and inadequate data protection might undermine research ethics. Conversely, both scientists and commercial organisations need adequate and predictable legal frameworks within which to develop (Huuskonen, 2013).

These issues, together with others, such as the consideration of how subject access rights might operate where tentative research findings are being drawn which are only tangentially (but perhaps identifiably) linked with particular individuals, raises the question of whether anonymisation is itself a purpose requiring notification. This involves how the proposed right to be forgotten might intersect with other concerns (including identifiability). Furthermore, there is some growing support for the idea, expressed succinctly by Clark and Weale (2011, p.28), that

> consent as a mechanism of legitimation for use of data 'kicks in' too automatically – that is, wherever anonymisation is not feasible – and too quickly – that is, before consideration is given to the possible justifications offered by the public interest in research.

This tension between the public interest associated with releasing data that could be re-identified and data subjects' rights to anonymity must be simultaneously acknowledged by any policy framework concerned with providing access to and re-use of scientific information containing personal data.

In some disciplines, several initiatives and programmes have addressed some of these legal issues. For example, in the environmental sciences, GEO (promoted to GEOSS) is a global and flexible network of content providers. It allows decision makers access to an extraordinary range of information at their desks. One of the first accomplishments of the GEO was the acceptance of a set of high-level data-sharing principles as a foundation for GEOSS. Ensuring that these principles are implemented in an effective yet flexible manner remains a major challenge. The 10-year GEOSS implementation plan says: 'The societal benefits of earth observations cannot be achieved without data sharing' and sets out the GEOSS data-sharing principles:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognising relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products will be either free of charge or cost no more than the costs of reproduction (GEOSS, 2005).

Additionally, clinical trials research, both in academic and commercial settings, is governed by the application of good clinical practice (GCP) (ICH Harmonised Tripartite Guideline, 1996), which includes guidelines to address each of these issues, with a particular emphasis on the ethical treatment of human subjects and the anonymisation and quality control of clinical data. The complexity of these legal and ethical issues requires a sustained and informed set of actions by institutional and government policy makers. It should take account of discipline-specific issues, such as consent and privacy, as well as overarching intellectual property issues. However, institutional and policy actors are themselves burdened by challenges particular to their positions within the open access to research data ecosystem.

Institutional issues and policy

Institutions, such as libraries, universities and open access repositories, encounter specific problems regarding open access and data dissemination and preservation. Some areas experience institutional barriers, such as a lack of financial support for open access (Habert and Huc, 2010), and/or may struggle with how to evaluate the research data with which they are presented in order to ensure scientific quality and integrity (High-Level Expert Group on Scientific Data, 2010). Consequently, specific needs around financial support, staff training and strategies for evaluating the data that they hold are high on the agenda of many of these institutions. While the creation of open access and data preservation repositories is clearly advantageous for institutions and those they serve, a significant financial outlay is needed, as setting up and maintaining open access repositories can cost millions of euros annually (Habert and Huc, 2010). Although the July 2012 recommendation encourages all European governments to invest in the preservation and dissemination of scientific information (European Commission, 2012b), the European Commission Directorates-General (DG) for Research and Innovation recognises that 'research libraries often have to find creative solutions with a limited budget, and despite their increasing responsibilities in access and dissemination' (Directorate-General for Research and Innovation, 2011, p.8). Data-sharing agreements between institutions with an associated sharing of the costs or the re-use of existing Information and Communication Technology (ICT) infrastructures may be more cost effective than creating new systems from scratch (Habert and Huc, 2010). However, such existing infrastructures may have problems of their own, such as technological obsolescence, or may also require additional staff training.[6] Furthermore, libraries and universities are under cost pressures from scientific journals that continually increase subscription costs (Directorate-General for Research and Innovation, 2011).

Institutions must also find effective ways of evaluating the quality, value and integrity of scientific data. Habert and Huc (2010, p.419) cite Hoog's 2009 warning

that data which are preserved become important, not because they are valuable, but because they are preserved.[7] Instead, the EU argues that institutions need better ways of measuring the quality and impact of the data they preserve (High-Level Expert Group on Scientific Data, 2010). Some institutions, for example the open context project described above, have established practices for providing peer review of data, but the standards of such evaluation have not yet been agreed by stakeholders and vary considerably. Academics, policy makers and other stakeholders have suggested a number of different potential strategies to evaluate such data, including establishing peer review practices for scientific data (Habert and Huc, 2010), citing data sets much as journal papers are currently cited in order to provide impact factors, and establishing peer review social media tools (Pöschl, 2010; Lin, 2012). In relation to these issues, the GEO (2012) has introduced the GEO label and the GEO data citation standard in order to provide: (1) valuable information to users of GEOSS (to help judge the quality and reliability of GEOSS components and services) and (2) an incentive for GEOSS providers to register their services and data. Secondly, Bulger *et al.* (2011, p.7) argue that the integrity of data needs to be preserved after they become accessible, and in particular that an institution must have a way to determine whether the data it holds in its repository remain exactly the same as the data that were originally deposited.

The online survey on scientific information in the digital age, conducted on behalf of the European Commission (Directorate-General Research and Innovation, 2012), makes clear that there is a significant amount of work to be carried out on institutional policies that regulate access to data, and specifically with respect to mandates imposing openness on research data, as well as with preserving scientific information (cf. European Commission, 2012b pp. 35–140). Institutions themselves require support from policy makers and other stakeholders in order to address their specific challenges. For example, academics could provide support by assisting in the evaluation of the quality of research data, and industry could assist in ensuring repositories are more interoperable. All of these initiatives require support from national, European and global governmental bodies in the form of effective policy making. Again, in the European Union context, the 2012 recommendation asks member states to define clear policies for the dissemination of, and open access to, research data resulting from publicly funded research, and at the same time to develop concrete objectives and indicators of progress, implementation plans and financial planning (European Commission, 2012b).

One concrete example of the way European policy has shaped developments in open access is the infrastructure for spatial information in the European Community directive (INSPIRE) (European Commission, 2007). The Directive establishes an infrastructure for spatial information in Europe to support Community environmental policies. Article 17(8) of the INSPIRE directive requires the development of implementing rules to regulate the provision of access to spatial data sets and services from member states to the institutions and bodies of the community. Thus, INSPIRE adopted a regulation on data and service sharing. The main points of the regulation are:

- Metadata must include the conditions applying to access and use for community institutions and bodies; this will facilitate their evaluation of the available specific conditions already at the discovery stage.

- Member states are requested to provide access to spatial data sets and services without delay and at the latest within 20 days after receipt of a written request; mutual agreements may allow an extension of this standard deadline.
- If charges are made for data or services, community institutions and bodies may request member states to provide information on how charges have been calculated.
- While fully safe-guarding the right of member states to limit sharing when this would compromise the course of justice, public security, national defence or international relations, member states are encouraged to find the means still to give access to sensitive data under restricted conditions (e.g. providing generalised datasets). Upon request, member states should give reasons for these limitations to sharing.

These institutional issues are significant because they can influence the development of open access. However, the complexity of developing open access means that it is proving difficult to generate an institutional framework. There are some examples of good practice emerging, such as INSPIRE. Nonetheless, greater understanding is needed of each aspect of open access and, crucially, of how developments in each need to be coordinated both at the phases of development and in sustainability models.

Conclusion

The development of open access to research data has consequences for the whole process of research. Although there are potential benefits in data being openly available, risks in the production and use of data emerge, and there are also risks in the development of open access itself. Two important factors are evident from the discussion in this paper. First, the development of open access is uneven across the stakeholder groups. Our discussion shows differences between the values of stakeholders across developments in open access to research data, in how best to develop integrated infrastructure and technology, in the legal and ethical complexities, and within and across institutions. These differences make it very difficult to bring together these key constituents to develop a coherent approach and development strategy for open access to data. Even if a general approach could be defined, it would need to be flexible enough to encompass the requirements of specific research areas and specific types of data. The unevenness of the development of open access to data and the lack of attention paid to the specific characteristics of different types of data generate risks to the data and also diminishes the potential benefits of open access. Overall, attention needs to be paid to the four areas we have discussed in order to ensure open access is robust and responsible in opening access to data and to ensure high quality access to stakeholders.

Second, the uneven development of open access involves considering how different types of data are collected and processed, and how they are subsequently interpreted, namely the data gap concern. Part of this concern is that different disciplines have different approaches to analysis and to the validation of research findings. Therefore, open access needs to be shaped in ways that respect the varying stages of the sensitivity and/or robustness of the status of data – and what can be claimed on the basis of those data. This aspect of the data gap is important not only for research

integrity and the accumulation of research results, but also for the social responsibility to enable access to data that requires informed interrogation.

Our main recommendation is that there needs to be a drive to motivate and integrate the stakeholders to reduce the unevenness of development in open access to research data. To ensure a coherent development of open access to data requires an international set of policy recommendations so that open access is equally developed for all at the global level. Although there have been national initiatives bringing stakeholders together to discuss joint use of metadata, protocols and common standards and so on, this is on a small scale (OpenAccess.se, 2012). There is a need for larger-scale initiatives to ensure a more even development of open access within an open access ecosystem, to ensure that open access to research data is one that is also sensitive to issues of the data gap.

Acknowledgements

Notes

1. For another widely used, similar definition of open access, see the Budapest Open Access Initiative (2002).
2. Funded under European Community's Seventh Framework Programme, FP7/2007-2013, Contract 269977. See http://www.alliancepermanentaccess.org/index.php/aparsen/
3. Funded under European Community's Seventh Framework Programme, FP7/2007-2013, Contract RI-212147. See http://www.driver-repository.eu/
4. Funded under European Community's Seventh Framework Programme, FP7/2007-2013, Contract RI-246686. See http://www.openaire.eu/
5. 'Digitally born data' refers to data that is digitally produced (not analogue material that has later been migrated to a digital environment).
6. Staff training in particular is necessary to increase awareness of the benefits and requirements of open access and data preservation. This includes training in how to use these new tools, and services to facilitate science as well as technical support (see Bulger *et al.*, 2011).
7. We do, however, acknowledge the adage that 'one person's trash is another's treasure', and recognise that future researchers may do creative things with data that are initially judged to be of little use.

References

American Scientist (2013) *Crowdsourcing Astronomy with Galaxy Zoo*. American Scientist, available from http://www.americanscientist.org/science/pub/crowdsourcing-astronomy-with-galaxy-zoo [accessed February 2014].

Beaulieu, A. and Wouters, P. (2009) 'E-research as intervention' in Jankowski, N. (ed.) *E-research: Transformation in Scholarly Practice*, Routledge, New York, pp.54–69.

Bigagli, L. and Nativi, S. (eds) (2013) *NetCDF Uncertainty Conventions (NetCDF-U)*. Open Geospatial Consortium, available from https://portal.opengeospatial.org/files/46702 [accessed February 2013].

Budapest Open Access Initiative (2002) *Budapest Open Access Initiative*, available at http://www.soros.org/openaccess [accessed February 2013].

Bulger, M., Meyer, E., de la Flor, G., Terras, M., Wyatt, S., Jirotka, M., Eccles, K. and Madsen, C. (2011) *Reinventing Research? Information Practices in the Humanities*, Research Information Network, London.

Chan, L., Kirsop, B. and Arunachalam, S. (2011) 'Towards open and equitable access to research and knowledge for development', *PLoS Medicine*, 8, 3, e1001016.

Clark, S. and Weale, A. (2011) *Information Governance in Health: Research Report*. Nuffield Trust, available from http://www.nuffieldtrust.org.uk/sites/files/nuffield/information_gover nance_in_health_-_research_report-_aug11.pdf [accessed February 2013].

Corner, A. and Bell, A. (2010) 'There's more to a story than facts and figures', *Times Higher Education*, 25 November, available from http://www.timeshighereducation.co.uk/story. asp?storycode=414350 [accessed February 2013].

Craglia, M., Nativi, S., Santoro, M., Vaccari, L. and Fugazza, G. (2011) 'Inter-disciplinary interoperability for global sustainability research' in Claramunt, C., Levashkin, S. and Bertolotto, M. (eds) *Proceedings of Geos 2011 Conference, Lecture Notes in Computer Science*, Springer, Berlin, pp.1–15.

Digital Preservation Coalition. (n.d.) *Digital Preservation Handbook*, available from http://www.dpconline.org/advice/preservationhandbook [accessed February 2013].

Dijkstra E. (1974) *On the Role of Scientific Thought*, available from https://www.cs.utexas. edu/users/EWD/transcriptions/EWD04xx/EWD447.html [accessed November 2013].

Directorate-General for Research and Innovation (2011) *National Open Access and Preservation Policies in Europe: Analysis of a Questionnaire to the European Research Area Committee*, European Commission, Brussels, available from http://ec.europa.eu/research/science-society/ document_library/pdf_06/open-access-report-2011_en.pdf [accessed February 2013].

Directorate-General for Research and Innovation (2012) *Online Survey on Scientific Informa-tion in the Digital Age*, European Commission, Brussels, available from http://ec.europa. eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digi tal-age_en.pdf [accessed January 2013].

Europa (1996) *Legal Protection of Databases Directive Adopted*. Europa press release, IP/96/ 171, 27 February, available from http://europa.eu/rapid/pressReleasesAction.do?reference= IP/96/171&format=HTML&aged=1&language=EN&guiLanguage=en [accessed February 2013].

European Commission (2007) *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*, 14 March, European Commission, Brussels, available from http://inspire.ec.europa.eu/ [accessed January 2013].

European Commission (2008) *Decision on the Adoption and a Modification of Special Clauses Applicable to the Model Grant Agreement Adopted on 10 April 2007 in the Context of the Implementation of the Seventh Framework Programmes of the European Community (2007–2013) and the European Atomic Energy Community (2007–2011), and to the Model Grant Agreement for 'Frontier' Research Actions and to the Marie Curie Model Grant Agreements Adopted on 16 April 2007 in the Context of the Implementation of the Seventh Framework Programme of the European Community (2007–2013)*. C(2008) 4408 final, 20 August, European Commission, Brussels.

European Commission (2009) *Communication on ICT Infrastructures for E-Science*. COM (2009), 108 final, 5 March, European Commission, Brussels, available from http://ec. europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf [accessed December 2013].

European Commission (2012a) *Proposal for a Regulation of the European Parliament and of the Council 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)*. COM(2012) 11 final, 25 January, European Commission, Brussels, available from http:// ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf [accessed January 2014].

European Commission (2012b) *A Reinforced European Research Area Partnership for Excel-lence and Growth*, COM(2012) 392 final, Brussels, 17 July, European Commission, Brussels.

Fielding R. (2000) *Architectural Styles and the Design of Network-Based Software Architec-tures*, available from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm [accessed November 2013].

GEO (Group on Earth Observations) (2012) *Task ID-03, C1, Activity 2.2 GEO Label. GEO user requirements registry*, available from http://www.scgcorp.com/urr/AppView.aspx?id=265 [accessed February 2013].

GEOSS (Global Earth Observation System of Systems) (2005) *The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan*, 16 February, available from http://www.earthobservations.org/documents/10-Year%20Implementation%20Plan.pdf [accessed February 2013].

GeoViQua Project Consortium. (n.d.) *QUAlity Aware VIsualisation for the Global Earth Observation System of Systems*, available from http://www.geoviqua.org [accessed February 2013].

Governance Working Group (2012) *Initial Governance Framework for EarthCube*, EarthCube, available from https://docs.google.com/spreadsheet/ccc?key=0AhE8IKrxOMWndDZfTXBUUFE1QmhOT01XT2JrZ3BQSWc#gid=0 [accessed February 2013].

Habert, B. and Huc, C. (2010) 'Building together digital archives for research in social sciences and humanities', *Social Science Information*, 49, 3, pp.415–43.

High-Level Expert Group on Scientific Data (2010) *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*, European Commission, available from http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204 [accessed Feb 2013].

Hoog, E. (2009) *Mémoire année zéro*, Le Seuil, Paris.

Huuskonen, M. (2013) *The Gap: ICT-revolution's Challenges to Legal Institutions*, European Commission, available from http://files.openinnovation-platform.eu/policydocs/the_gap_final.pdf

ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) Harmonised Tripartite Guideline (1996) *Guideline for Good Clinical Practice*. E6 (R1), 10 June, available from http://www.ich.org/ [accessed January 2013].

Jankowski, N. (ed.) (2009) *E-Research: Transformation in Scholarly Practice*, Routledge, New York, NY.

Laurenti C., (2012) *Crowdsourcing Medical Expertise*, available from http://www.technology-digital.com/web20/crowdsourcing-medical-expertise [accessed December 2013].

Lin, T. (2012) 'Cracking open the scientific process', *New York Times*, 16 January, p.D1.

Long-Term Data Preservation Working Group (2012) *Long-Term Preservation of Earth Observation Space Data European LTDP Common Guidelines*. Ground Segment Coordination Body, 30 June, available from http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommon Guidelines_Issue2.0.pdf [accessed February 2013].

Manzella G., Boer, C., Serra, M., Nativi, S., Mazzetti, P., Li Santi, E., Rossi, A. and Kaiser-Weiss, A. (2009) *Report on Research on Data Curation and Interoperability*, available from http://www.genesi-dr.eu/documents/GENESI-DR-JRA2-DEL-DJRA2.2%20v2.pdf [accessed February 2013].

Max Planck Society (2003) *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, available from http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf [accessed February 2013].

Muir, A. (2004) 'Digital preservation: awareness, responsibility and rights issues', *Journal of Information Science*, 30, 1, pp.73–92.

Nativi, S., Marsh, S., Mazzetti, P. and McCallum, I. (2011) 'Major networking – EGIDA: sustaining the system of systems', *International Innovation Magazine*, October, pp.45–7.

Niederer, S., Fink, M., Noble, D. and Smith, N. (2009) 'A meta-analysis of cardiac electrophysiology computational models', *Experimental Physiology*, 94, 5, pp.486–95.

Nowotny, H., Scott, P. and Gibbons, M. (2001) *Re-thinking Science: Knowledge and the Public in an Age of Uncertainty*, Polity Press, Cambridge.

OpenAccess.se. (2012) *OpenAccess.se – Scholarly Publishing*, Kungliga biblioteket, available from http://www.kb.se/openaccess/Hjalptexter/English/ [accessed February 2013].

OpenDOAR (2013) *Directory of Open Access Repositories*, available from http://www.opendoar.org/index.html [accessed February 2013].

OECD (Organisation for Economic Co-operation and Development) (2006) *Report on Cross-Border Enforcement of Privacy Laws*, OECD, Paris, available from www.oecd.org/dataoecd/17/43/37558845.pdf [accessed February 2013].

OECD (Organisation for Economic Co-operation and Development) (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD, Paris, available from http://www.oecd.org/dataoecd/9/61/38500813.pdf [accessed February 2013].

Parsons, M., Plag, H. and McCallum, I. (2011) *Proposal for a GEO Label*. Coordinating Earth and Environmental Cross-disciplinary Projects to Promote GEOSS (EGIDA), available from http://www.egida-project.eu/images/documents/proposalforageolabel.pdf [accessed February 2013].

Payne, D. (2012) 'Tamiflu: the battle for secret drug data', *British Medical Journal*, 345, e7303. doi:10.1136/bmj.e7303.

Pöschl, U. (2010) 'Interactive open access publishing and public peer review: the effectiveness of transparency and self-regulation in scientific quality assurance', *International Federation of Library Associations and Institutions Journal*, 36, 1, pp.40–6.

Repositories Support Project (2011) *Survey of Academic Attitudes to Open Access and Institutional Repositories – an RSP and UKCoRR Initiative*, November 2011, available from http://rspproject.files.wordpress.com/2011/12/attitudes-to-oa-basic-summary-report.doc [accessed February 2013].

Rodrigues, E. (2009) *Open Access Policies Toolkit*. Repositório Científico de Acesso Aberto de Portugal, 30 June, available from http://projecto.rcaap.pt/index.php/lang-pt/consultar-recursos-de-apoio/remsitory?func=startdown&id=336 [accessed February 2013].

Royal Society (2012) *Science as an Open Enterprise*, Royal Society, London.

Saengkhattiya, M., Sevandersson, M. and Vallejo, U. (2012) *Quality in Crowdsourcing: How Software Quality is Ensured in Software Crowdsourcing*, Masters thesis, School of Economics and Management, University of Lund, available from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3168789&fileOId=3168790 [accessed December 2013].

Smith, K. and Hansen, D. (n.d.) *Copyright and Authors' Rights: a Briefing Paper*. Open Access Scholarly Information Sourcebook, available from http://www.openoasis.org/images/stories/file/Copyright%20and%20authors'%20rights.pdf [accessed February 2013].

Sotiriou, M. (2013) *Crowdsourcing Biodiversity Watch*, available from http://www.youris.com/Environment/Biodiversity/Crowdsourcing_Biodiversity_Watch.kl [accessed January 2014].

Sveinsdottir, T., Wessels, B., Smallwood, R., Linde, P., Kallo, V., Sondervan, J. and Tsoukala, V. (2013) *Stakeholder values and Ecosystems, Workpackage 1*, Report of the Policy Recommendations for Open Access to Research Data (RECODE) project.

Uhlir, P. and PSchröder. (2007) 'Open data for global science', *Data Science Journal*, 6, 17 June, p.OD48.

UncertML. (n.d.) *UncertML: Describing and Exchanging Uncertainty*, available from http://www.uncertml.org/ [accessed February 2013].

W3C (2004) *Architecture of the World Wide Web, Volume One*, W3C recommendation, 15 December, available from http://www.w3.org/TR/webarch/ [accessed November 2013].

Wallace, C., Pordesch, U. and Brandner, R. (2007) *Long-Term Archive Service Requirements*. IETF Trust, RFC 4810, available from http://www.rfc-editor.org/rfc/rfc4810.txt [accessed February 2013].

Wilson, J. (2011) 'Freedom of information and research data', *Research Ethics*, 7, 3, pp.107–11.

Wouters, P. and Schröder, P. (eds) (2003) *Promise and Practice in Data Sharing*, Networked Research and Digital Information, NERDI-KNAW, Amsterdam.

Zuiderwijk, A., Jeffery, K. and Janssen, M. (2012) 'The necessity of metadata for linked open data and its contribution to policy analyses', in Parycek, R. and Edelman, N. (eds) *Proceedings CeDEM2012 Conference*, Donau-Universitat, Krems, pp.281–94.