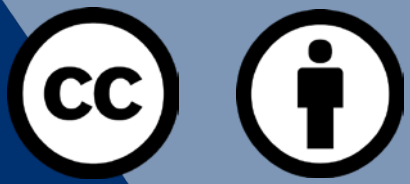
October 2016

Embedding cultures and incentives to support open research

Michael Jubb



Cite this as: Jubb, Michael (2016) Embedding cultures and incentives to support open research. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4055514

1. Introduction

This paper is one of five commissioned by the Wellcome Trust to draw together evidence on key challenges relating to open science, and to identify options to address those challenges. It focuses on cultures and incentives, and mechanisms to address them. Other papers focus on infrastructures, skills, interoperability, global equity; and there is inevitably some overlap between them.

Open science[1] and open research are increasingly- used terms, but it is not always clear precisely what they are intended to encompass. At their broadest, these terms are used to cover many different aspects of the opening up of research to allow others to participate at all stages, as well as to learn and benefit from it as freely as possible once it has been completed. In this sense, open research covers not only providing access to published findings and to underlying data (together with the code), but features such as open (and post-publication) peer review; open research notebooks; open source software; and 'citizen science'(Amsen, 2014). While some of those features will be covered in this paper, the emphasis is on what is produced by the end of a research project, especially in the form of data, related code and metadata.

The paper is based on desk research using sources from roughly the last five years, though it should be emphasised that it does not constitute a comprehensive literature review. The work of the Expert Advisory Group on Data Access established by Cancer Research UK, the Economic and Social Research Council (ESRC), the Medical Research Council (MRC) and the Wellcome Trust has been an important starting point; but many other sources have been used. A full list of the sources cited is given in Annex 1.

1.1 Open research data

The data produced by the new instruments and techniques that underpin data-intensive science come in many different forms: observational data which may be unique in time and place, and thus impossible to replicate; data arising from experiments of wildly differing scale and type, and which may or may not be replicable; computational data arising from large-scale simulations; and reference data sets and data banks (Protein Data Bank, International Cancer Research Data Bank), including those arising from major longitudinal studies. The processes involved in the creation and management of these different kinds of data vary widely, and have major impacts on researchers' attitudes towards them and how they can be made accessible to others: what may seem appropriate and acceptable in relation to data produced from a large-scale international project may not be viewed in the same way with data arising from a small-scale survey or experiment.

And just as there are different kinds of data, there are many different ways of making them available to others: the commonly-used term 'data-sharing' can mean very different things. At one end of a spectrum, it may refer simply to sharing data with members of a research team or consortium, or more broadly with trusted colleagues in informal networks. At the other end, sharing is also used to denote sharing with members of a research community, or making data openly available to anyone. And at all points in the spectrum, data may be made available under different conditions (on request, or by posting in a repository); at different stages in the research process (with attendant differences as to whether it is raw data or processed data that is shared); with different embargo periods; with varying degrees of usability; and with different licensing terms. Usage restrictions may range, for example, from a simple requirement to cite or acknowledge, to a joint authorship (Pasquetto et al 2016).

[1] 'Open science' was originally coined as a term by Paul David (2003) in an attempt to prevent the further extension of intellectual property rights over information resources produced by scientists, especially those funded from public sources

It is not helpful that ‘sharing’ is used to denote such different practices, with differing implications, and for this reason some commentators avoid the term, referring to data exchange or data publication rather than data sharing (Gallagher et al 2015). But it is perhaps not surprising that many researchers are confused as to the precise meaning of the various terms, and the varying requirements set by funders and others.

2. Cultural barriers and incentives

An international survey conducted by the DataONE project (Tenopir 2011) found that two-thirds of researchers agreed that lack of access to data generated by others is a major impediment to progress in science; and half said that their own research had been impeded as a result. But a third of respondents did not answer whether they made their data available to others; and of those that did respond, nearly half reported that they did not share their data. Only just over a third agreed that others could access their data easily, though two-thirds said that they would be more likely to make their data available if they could impose restrictions on access. These figures may be compared with the 58% of respondents to the ParseInsight survey in 2009 who said they made their data freely available to their team, and 25% who made it available to everyone publicly. But both surveys indicate a clear gap between researchers’ interest in using others’ data, and their willingness to provide access themselves.

A more recent survey of German researchers in 2014 (Fecher et al 2015) found that three-quarters of them (76%) agreed that other researchers should publish their data, while 83% stated that making data available to other researchers benefits scientific progress. But again significantly fewer had actually made their own data available to others: 58% said they had shared with researchers they know personally, 49% with researchers within their institute or organization, 40% with researchers working on similar topics and only 13% publicly.

A recent survey of UK researchers in universities (Wolff et al 2016) indicated that 80% of them often use their own computers to manage and organise their data during the course of their research, while under 40% often use institutional storage, and a slightly smaller proportion use cloud-based storage such as Google Drive or Dropbox. There was relatively little difference between researchers in broad subject areas, though rather more (over 50%) medical and veterinary researchers than the average often used institutional storage, with slightly less use of cloud-based storage. When it comes to preserving data at the end of a project, the survey indicated that as compared with a similar survey in 2012, there had been a slight fall (from over 70% to just over 60%) in the proportions of researchers relying on their own resources, and a rise (from under 30% to just over 40%) in those making use of an institutional or other repository. Although the increase in those using repositories is welcome, the figures nevertheless indicate that after a decade and more of advocacy and the development of policies and services for effective data management, more than half of all UK researchers rely on their own equipment to preserve the data they gather or create, with little chance of its being made accessible to others. Moreover, a stubborn 13% of researchers say that they do not preserve the data after the end of a project.

2. 1 Drivers and incentives.

Nevertheless, there are key forces driving researchers towards more openness, including the availability of new technologies and services, the growth of research collaborations and international partnerships, the growing emphasis on effective management of research at institutional level, and not least the policies of Government and funders, with greater pressure for accountability. Such external drivers are of crucial importance in influencing research practice, and some reports suggest (Knowledge Exchange 2014) that at least some researchers would welcome stronger leadership from

funders, with greater specificity of data management and sharing requirements, not least to create a level playing field for all researchers.

For there is a widespread view that while external drivers, along with the cultural norms in a few subjects and disciplines, can be powerful influences on behaviour, the direct incentives for individual researchers and teams are often weak (Walpor et al 2011;Tenopir et al 2011; Royal Society 2012; Noorman et al 2014; EAGDA 2014; OECD 2015; Brack et al 2015; EU 2016). As many commentators have pointed out, while researchers have a strong interest in maximising the dissemination and impact of the articles they publish, that is not generally true for research data. Evaluations of the quality and impact of their work exert a major influence over researchers' prospects of success in winning further research grants, and advancing their careers. They also influence the status and ranking of the institutions in which researchers work. Such evaluations currently focus largely if not exclusively on publications in scholarly journals: career rewards arise from publishing scholarly articles rather than the data on which the articles are based. And despite the campaigns of various individuals and organisations to persuade funders, institutions and publishers to take more account of the value and impact of all research outputs, including datasets and software (as in the San Francisco Declaration on Research Assessment), this remains the case. Hence in a reputation economy, data sharing and open data are perceived as not counting, This has a powerful influence in shaping researchers' attitudes, behaviour and decisions; and it can even underpin decisions to protect and withhold data rather than to publish it. As the Royal Society noted in 2012

> major barriers to widespread adoption of the principles of open data lie in the systems of reward, esteem and promotion in universities and institutes. It is crucial that the generation of important datasets, their curation and open and effective communication is recognised, cited and rewarded.

Perceptions are of course critically important here, and it is a constant refrain of recent reports and surveys that data sharing is not rewarded by either institutions or funders. Moreover, although this is less commonly mentioned in the literature (EAGDA 2014), there appears to be little evidence across the broad scope of subject domains that researchers have suffered negative consequences from failing to meet the requirements of funders' and institutional policies on data sharing; there is, for instance considerable scepticism as to whether data management plans (DMPs) weigh significantly in the assessment of grant applications (EAGDA 2015b).

This is not to say that incentives to share data are entirely absent. For researchers in fields such as genomics or particle physics, sharing data is an essential part of the research process. This may be because many researchers with different sets of expertise across a large group are actively involved in analysing complex data that is shared as part of direct collaboration or peer exchange; or because data shared across a community sharing or in fully open fashion has become a cultural norm driven by the nature and needs of research (Knowledge Exchange 2014). In such cases, data sharing - in one or more of the senses outlined in Section 1.1 above – leads to better research and, crucially, to mutual benefits to all those engaged in the research., with publications typically co-authored. In other research areas, most researchers depend on the availability of data that cannot easily be replicated, either because of its scale and complexity, or because it derives from longitudinal and cohort studies (which may have been funded and created as community resources). The reward (and thus the incentive) for primary data creators in such cases may take the form simply of a citation in the secondary researchers' publications.

In yet other cases, researchers are motivated by an altruistic desire to promote and support the cause of open science, by publishing – via a community repository or other means - the raw data they create or gather, and thus seeking to ensure that their research and their findings are sound, and can stand up

to external critique. Such researchers see new analyses and interpretations of their data as a positive benefit to the community, as well as themselves.

But data sharing can also have more specific benefits. Some researchers see data sharing as part of a strategy to enhance recognition and the reputation of their (and their research group's) work, to extend their networks and potential for partnerships, and thus to promote their careers. In other cases, sharing data with trusted colleagues brings - simultaneously or later - a reciprocal exchange in the other direction; such exchange can engender feelings of gratitude and reassurance ("my data is valuable enough to be wanted by another researcher") as well as responsibility, particularly for early-career researchers. Reciprocity can be an important foundation for a research career. It is important to stress, however, that benefits of the kind outlined above can derive from one of the more restrictive kinds of sharing, such as informal exchange with trusted peer, noted in Section 1.1; they need not necessarily be the result of 'open science'.

Moreover, outside a few subject areas and sub-disciplines frequently mentioned in the literature, there is a relative lack of systematic evidence of benefits to individual researchers. In genomics a large-scale analysis of data sharing shows that studies that made data available in repositories received 9% more citations, when controlling for other variables; and that third party papers re-using the data continued to accumulate for at least six years (Piwowar et al, 2013). But in most disciplines and subject areas, the evidence of benefits to individual researchers or teams is anecdotal rather than systematic, and less persuasive as a result. In contrast to open access publishing, for data sharing there is no extensive literature on a citation advantage or other benefits.

As the EU RECODE project noted (2014) awareness about open research data and the incentives to make data open will increase only when research communities more broadly start valuing data produced as much as they value publications, and when research institutions, universities, funding bodies and learned societies start evaluating and rewarding researchers and research groups on the basis of their work in research data management and data sharing And the Expert Advisory Group on Data Access (EAGDA) noted in 2014 that despite an overwhelming consensus that the current lack of incentives militates against greater availability and use of data, very few concrete steps seem to have been initiated so far to address these issues. Unfortunately, despite some moves by publishers such as Springer Nature (Hrynaszkiewicz, 2016), and the adoption by a small number of journals of the badges promoted by the Center for Open Science to signal open science practices, that remains the case two years later.

2.2 Barriers

The lack of concrete measures is particularly regrettable because the barriers and disincentives to open data and the various forms of data sharing remain strong. As the DataONE project noted (Tenopir 2011), "barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves". The key barriers may be grouped under seven main heads: costs in time and money; lack of, or deficiencies in, infrastructure and standards; deficiencies in skills and training; concerns about data quality; issues relating to software and code; ethical or legal issues and concerns; and fears of competition and loss of control.

2.2.1 Costs

Preparing data for sharing, including the necessary documentation and contextual metadata, and then curating and disseminating them, are time-consuming and costly. The leading reason given by researchers in the DataONE survey in 2010 for mot making their data available to others was lack of time, followed by lack of funding; and this is confirmed by other studies (Tenopir 2011; ParseInsight 2009; DAMVAD 2014). In a context where the key incentives for researchers are to get their findings

published as quickly and effectively as possible, researchers prioritise achieving that result. Anything else that adds to their tasks and workflow is given much less priority: the next grant, the next project and the next publication are much more important. Thus support for data sharing in principle may not be turned into data sharing action in practice. Moreover, there is lack of understanding about the full costs of cleaning, preparing and formatting data, metadata and other documentation, and of handling and administering requests for data (where that is appropriate and needed); about the availability of funding; and concerns that project-based funding models militate against the sustainability of support for data management, curation and preservation services. In all these circumstances, perceptions of the costs in time and money of data sharing may outweigh the perceived benefits (EAGDA 2015b).

2.2.2 Infrastructure and standards

The research – and particularly the research data – ecosystem has become much more complex in the past decade: even the experts in the field find it difficult to comprehend the multi-layered relationships between funders, Government and international bodies, universities and other research institutions, regulatory bodies, research consortia and collaborative initiatives, repositories and other data services, journals and publishers, registries and discovery services, and standards bodies both formal and informal (Hook 2016;Enoch 2016). For individual researchers grappling with the complexities of the implications and relationships between their own activities on the one hand, and the policies, protocols, tools and services issued or provided by such bodies on the other has become an increasingly-difficult challenge; and there are complaints that funders' requirements, for example, are not harmonised or co-ordinated. It should also be noted that these complexities give rise to difficulties for data users as well as originators; lack of commonality in standards and services, and thus in interoperability, mean that they find relevant data less discoverable, accessible and usable than it would otherwise be (EAGDA 2015b; Gallagher et al 2015)..

The difficulties are exacerbated by the variations between what is provided (and/or taken for granted) in different subject areas and institutions. At a recent Jisc/CNI conference the research data infrastructure in both the US and the UK was characterised in terms such as 'radical scatter', 'piecemeal' and a 'patchwork of provision' (Wilkin 2016; Maguire 2016). And other reports (Noorman 2014, Tsoukala 2015) have pointed to the lack of clarity as to where responsibilities lie for the provision of infrastructure and services. Thus it is not surprising that in the DataONE survey (Tenopir 2011) only just over a quarter of researchers said they were satisfied with the tools for preparing metadata (many seemed unclear what the term 'metadata' meant); only two-fifths were satisfied with the tools for preparing documentation; and only a third were provided with the necessary tools and technical support for long-term data management. A biomedical symposium in 2015 (Academy of Medical Sciences 2015) noted that effective data sharing in neuroimaging requied the development of new infrastructure and software tools. Other reports (EAGDA 2014) have indicated that in many fields, a lack of widely-accepted data standards (or lack of awareness of those that do exist) remains as a significant barrier to data sharing: in a survey commissioned by EAGDA, 49% of respondents identified limitations on technical resources (including standards) as a key barrier, with the need for better and more user-friendly tools and infrastructure highlighted. In such circumstances, it is perhaps not surprising that a recent study found a bewildering range of research data software applications being used, and data file formats being deposited, in a single UK university (Mitcham 2016)

Standardisation thus remains a key challenge in a research landscape characterised increasingly by fragmentation and heterogeneity, as exemplified by the many different ways of formatting, storing, operating and standardizing data. In such a landscape there is clearly tension between the demands of a wide variety of disciplines and sub-disciplines with their own practices, cultures and standards on

the one hand (what works well in one filed might not be readily applied in another); and the widely-expressed need for common standards and interoperability on the other. This tension is as yet far from resolved.

2.2.3 Skills and training

A repeated refrain in reports over the past decade has been the need to address the skills gaps in data management and curation both within the research community, and among librarians and information sepecialists (OSTP 2013; OECD 2015; Tsoukala 2015; Brack et al 2015). Researchers in many fields require training in order to gain the knowledge and skills they need to work with varying formats and tools to make their data publishable and reusable as well as how to search for the data they need, to re-use it and incorporate it into their research process. And librarians and information specialists need to keep up both with technological changes and developments in different research domains so that they can support researchers in an effective way. There has been some progress in both these areas, not least in the UK through the efforts of the Digital Curation Centre; but many studies repeat the refrain that deficiencies in skills across many areas of the research community – in the UK and overseas - remain a significant problem.

Moreover, there are growing signs of what the High Level Expert Group on the European Open Science Cloud (EU 2016) has called a clash of cultures, between research domain specialists and data experts: an increasing divide between researchers and those who support research with data processing and software skills, with a consequential undermining of communication and collaboration between domain researchers and ICT experts. There are exceptions to this generalisation, of course; but there is a growing sense that researchers are "struggling with sub-optimal solutions, sometimes out of ignorance…but often because actual collaboration with computer scientists and engineers takes time and is not easy". There is also the risk that both the development and the use of tools and infrastructure become dominated by 'cyberinfrastructure experts, data managers, software engineers, data scientists, and others' rather than by domain scientists (Cutcher Gershenfeld 2016).

2.2.4 Data quality

The OECD Principles and Guidelines for Access to Research Data from Public Funding (OECD 2007) stress that "the value and utility of data depend to a large extent on the quality of the data themselves. Particular attention should be paid to ensuring compliance with explicit quality standards". But there are also concerns about data quality. Quite apart from the question whether the data – and the associated code - fulfils the scientific purposes for which it was created or collected, effective sharing depends on ensuring that data are discoverable, accessible, interpretable, assessable, and re-usable. But quality assessment or assurance are time-consuming and difficult tasks; and responsibilities for ensuring that data does indeed conform to such characteristics are not clearly defined. Funders, institutions, journals, publishers, repositories and data centres may all have a stake, and roles to play; but it is often unclear to both creators and users of research data what has been or will be done, and by whom.

Moreover, automated instruments and sensors can introduce unknown variations due to external noise, and when data from one discipline is combined with data from another, the uncertainties in the combined data may not be easily quantifiable. Such uncertainties may be addressed through the use of data quality indicators, but these are not as yet widely adopted, and many researchers are not in a position to specify the quality indicators needed for a specific analysis (Gallagher et al 2015).

All these problems are exacerbated by a general lack of availability of reviewers, who have few incentives to engage in the difficult and ill-defined task of data review. This in turn feeds into

scepticism on the part of data creators as to whether they will receive meaningful credit for making high-quality data available to others.

2.2.5 Software and code

Few of the policies issued by funders make more than a passing reference to software and code, though they can constitute a significant barrier to data sharing. For software and code are integral to the generation of research data; and access to them is thus essential if users are to be able to re-use data effectively or to validate research findings. A survey in 2014 of researchers in Russell Group universities in the UK (Chue Hong, 2016) showed that while 92% of respondents used research software, and 56% developed their own software, 71% had no formal software training. Hence many researchers often overlook the importance of producing and sharing well-written, accessible code; . Whether bespoke software generated for the specific research project is used, or software generated by a third party, researchers have to take steps to ensure its continued availability. Researchers may not be aware of the advantages of open data formats and open source software in this context, or of the complex licensing issues that may arise. This is in part because software development is often left to postdocs, who get little in the way of recognition for this work, so this it is difficult for them to develop a career in this area.

2.2.6 Ethical and legal issues

For data to be fully 'open' they must be available for use by anyone without technical or legal restrictions. But the OECD Principles and Guidelines on Access to Research Data also recognise that arrangements for access to data – and also code - must respect the legal rights and legitimate interests of all stakeholders and that restrictions on access may be required for a variety of reasons: national security, privacy and confidentiality, trade secrets and intellectual property rights, protection of endangered species, legal processes. In some cases both data and code may have a commercial potential which researchers and/or their funders or institutions may wish to seek to exploit. Similar points are made in the Concordat on Open Research Data recently published in the UK (Concordat 2016). But many reports and commentaries (Knowledge Exchange 2014; Tsoukala 2015; Gallagher et al 2015; Brack 2015) point to the complex issues that can arise in relation to intellectual property rights (IPR), especially with data that derives from third parties such as commercial organisations, and when researchers may not hold the IPR in the data to which they wish to provide access. Thus, for example, a quarter of respondents to the DataONE survey (Tenopir 2011) said that they had not made their data publicly available because they did not have the rights to do so, while 41% of respondents to the ParseInsight survey of 2009 said that legal issues were the key barrier to making their data freely available. On the other hand, many researchers are uncertain about the rights that they or their funders may have in relation to the data and/or code they have created, and of the need to consider carefully the appropriate licences before they make them available to others.

Researchers are also rightly wary of the need, when the data they have collected relates to individuals, to protect (through techniques including data reduction, cryptography, and de-identification methods ranging from anonymization to pseudonymisation or other means) the confidentiality and proper interests of research participants, to ensure that the terms of consent (which in the case of longitudinal and cohort studies may have changed over time) are fully met, and to ensure that no attempt is made to re-identify participants (EAGDA 2014; EAGDA 2015a and b; OECD 2015). These concerns imply careful processing of data before it is made accessible to others, but also close management (via a Data Access Committee or other mechanism) of the provision of access: the data cannot be made fully open to all.

2.2.7 Competition and loss of control

We have already noted that making data fully open – accessible and usable by all – is at one end of a spectrum of practices that may be termed data sharing. Sharing in a more restricted way is, as recent surveys have shown, much more common (Tenopir 2011; Fecher 2015). For as many reports have commented, there are widespread concerns about losing control over one's data: that making it openly available to others will reduce the options for producing publications based on it in the future; or that others may misuse it, or use and publish it without crediting the originator, or pre-empt intended publications by the data creator. In contrast to the DataONE survey in 2010 (where the option was not put to respondents), in the survey of German researchers in 2014 (Fecher 2015), fear that 'other researchers could publish before me' was the top reason, cited by 80% of respondents, for not making data openly available. It is worth noting, however, that a similar proportion said that 'to publish before sharing' was a major motivator towards data sharing. In this context, it is worth noting that many large-scale genomics projects (such as the Cancer Genome Atlas) adhere to the principles of rapid release of their data in advance of formal publication of their work, but subject to a publication moratorium lasting either until the data is published by a project member, or for a defined period (usually a year) after the release of the data.. But there appears to have been no analysis of the effectiveness of such moratoriums.

Data citation is widely seen as the key solution to many of the concerns about unfair competition, so that researchers can secure acknowledgement for the data they have gathered or created, and for the work of curation, through mechanisms similar to those used for citations of scholarly publications. But as we have seen, scholarly reputation, - and the rewards that come with it - are still seen as overwhelmingly based upon publications, and data citation is as yet far from widely adopted. Moreover, although there has been significant progress over the past five years, there remain a number of technical issues (for, example, granularity, version control, micro-attribution, identification of contributors) still to be fully resolved in different subject areas and for different kinds of data. Survey datasets, for example, are often recorded in a finely-grained way – differentiating between versions, samples, interview modes, location and so on – that is often ignored in a citation. Matching citation stings to datasets then poses a number of technical challenges (Mathiak et al 2015). There are then socio-cultural challenges in addition for data creators, research institutions, data centres, journal publishers and editors, and research funders that need to be addressed before good citation practices become widely adopted across all disciplines and subjects (Socha,2013; Herther 2013). As yet, most of the datasets included in Thomson Reuters Data Citation Index remain uncited.

3. Disciplinary differences

There is a lack of comparative studies, but it is noticeable that the provision of data-related services, and of community initiatives, varies widely across different subjects and disciplines. The wide arrays of services and initiatives in areas such as bioinformatics, crystallography or the environmental sciences is in contrast to the much smaller arrays in many areas of engineering or chemistry, or still more in key areas of the social sciences and humanities. As we have already noted, there are significant variations between the cultures and practices – and in the nature of the data created or gathered – in different disciplines and sub-disciplines; and successful practices in, say, astronomy or archaeology may not translate effectively even into other disciplines that might at first sight seem closely related.

It seems likely, however, that there is some relationship between the provision of services on the one hand, and the adoption of data sharing and open data on the other, in different disciplines. The evidence from surveys of researchers provides some support for such a hypothesis, though we must caution against the possibility of response bias in all the surveys; and it is unfortunate that the survey

reports aggregate subjects and/or disciplines into very broad categories, and even then there is a lack of consistency between them. Nevertheless, the ParseInsight survey in 2009, for example, showed an interestingly high proportion of humanities respondents – much higher than any other discipline - expressing a need for better infrastructure to guard against the risks to digital preservation. This perhaps suggests that humanities researchers who are actively engaged in digital research are uncomfortably aware of the relative lack of provision in their institutions and subject domain.

It is notable also that in the survey of German researchers in 2014 (Fecher 2015), the pattern of major impediments to sharing (risk of pre-emptive publication by others , efforts required to share, risk of misinterpretation etc) and of motivators (time to publish first, support from institution, data citation etc) shows relatively little variation between six subject areas: agriculture, engineering, humanities, human science, social science and natural science.

When it comes to actual practice, the DataONE survey (Tenopir 2011) found that respondents in medicine and the social sciences were less likely than their colleagues in computer science and engineering, or the physical, biological, atmospheric or environmental sciences (there was no category for the humanities) to make their data electronically available to others, or to be willing to place either some or all of their data into a central repository, presumably because they were more likely to be handling sensitive human data with ethical constraints on access. On the other hand, social scientists were most likely to say that lack of access to data generated by other researchers or institutions is a major impediment to progress in science, while medical researchers were the most likely to say that they received the necessary tools, funding and technical support for data management and sharing (but social scientists were the least likely to agree with such a view). Differences between other subject areas were much less evident: they were generally more satisfied with current situations and willing to share their data. It is also worth noting that the survey indicated that older researchers were more likely than their younger colleagues to share their data openly and without restrictions.

The recent Ithaka survey (Wolff et al 2016) categorised respondents in the arts and humanities, social sciences, sciences, and medical/veterinary. It highlights key differences in the kinds of data that are generated by researchers in those different disciplines. In the social sciences, arts and humanities, but also in medicine, a majority of researchers said that they generated qualitative data such as open-ended survey responses, interview transcripts, laboratory and field notes, text, documents, images, video, audio and so on. Only a minority – but a substantial one of 40% - of scientists indicated that they generated such data. A majority of researchers in social sciences, science and medicine (with the latter the most likely) generated quantitative data such as numeric or geospatial data files, or survey responses; not surprisingly, only around a quarter of arts and humanities researchers said they produced such data. Again, it is not surprising that only small minorities of social scientists and arts and humanities respondents produce experimental data, slides, physical artefacts or biological specimens and samples, while a majority of scientists and medical researchers did so; and a majority only of scientists produced computational data, algorithms, programmes and the like (about a quarter of medical researchers did so). Although the overall pattern here may not be surprising, it is worth noting that it would be over-simplifying to say, for example, that arts and humanities researchers do not produce computational data, or that scientists do not produce unstructured qualitative data. Minorities in each case do precisely that.

Moreover, when it comes to organising and managing their data, the disciplinary differences are much less visible. Thus the percentage of respondents who agreed that they often managed the data on their own computers varied from over 70% in medicine to just over 80% in social sciences and the arts and humanities. The differences in the use of institutional storage were greater, ranging from over half in medicine to under a quarter in the arts and humanities; but use of cloud services ranged between a

quarter and just over a third , while relying on the university library was an option used by only small minorities in all four disciplinary groups. There were similarly small differences in the value attached to various sources of support (ranging from 'freely-available software' through various university support services, to disciplinary repositories, publishers, or learned societies) in managing and preserving data beyond the end of a project. The major difference came in the actual use of institutional or other repositories: while over half of medical researchers used them, only a quarter of arts and humanities did so, making much more use of 'commercial or freely-available software or services'.

In sum, the evidence seems to suggest that although there are clear differences of culture as well as practice between disciplines and subject areas, these are strongly influenced by differences in the kinds of data produced in different areas of research (where fine-grained differences may be important but hidden by the evidence currently available); by constraints relating to the sensitivity of such data; by the availability and accessibility of infrastructures and services to support effective data management and sharing; and by the closeness or otherwise of relationships between domain specialists and data specialists.

4. Measures to support open research and data sharing

In an ecology characterised by the barriers militating against open science and by the disciplinary differences outlined in Sections 2 and 3, this Section looks at the measures currently being taken by funders, institutions, publishers, research support services, information professionals and others to support and promote open science and data sharing.

4.1 Open science policies: funders, institutions and publishers

The drivers for policies to promote open research are clear. Opening up access to the outputs of research can bring improvements in the effectiveness and productivity of research by exploiting the potential for additional research – addressing new research questions - using the vast amounts of data that researchers now collect and create, often at significant cost; reducing duplication in collecting and creating data; enhancing opportunities for involvement in the research process; facilitating interdisciplinarity and international collaboration in addressing large-scale global challenges; promoting wider engagement in science and research, including the possibilities of citizens' active participation in research projects; and enhancing the opportunities for applying research results so that they bring greater social and other benefits.. For all these reasons, open research is seen as bringing increased returns to the investments funders make in supporting research. Crucially, open access to research data also facilitates the testing and validation of research findings, and thus helps to address growing concerns about replicability and reproducibility of those results (Academy of Medical Sciences, 2015).

It is thus not surprising that funders in the UK and beyond have developed policies to promote and support open research, although the Funding Councils in the UK have so far eschewed any attempt to introduce any requirements relating to open data – as distinct from open access publications – into such a major influence on researchers' behaviour as the Research Excellence Framework (REF)[2]. On the other side of the dual support system, however, the Research Councils' Common Principles on Data Policy, published in 2011 and revised in 2015 (RCUK 2015), are built around the concept of data as a public good 'which should be made openly available with as few restrictions as possible'. Similar principles underlie policies introduced by funders such as the NSF and NIH in the US, and the

[2] Individual universities have similarly taken relatively few steps to require or to promote open research data, in contrast to the active steps many have taken to support open access to publications.

pilot policy on data sharing introduced by the European Commission for certain areas of the Horizon 2020 programme (European Commission, 2016). But there are significant variations between the different policies.

In the UK, the policies of the seven Research Councils vary not only in their precise requirements, but in the language in which they are expressed, and also in the ease with which they can be found on the respective Council's websites. Thus the MRC has established a set of 21 requirements relating to data sharing for population and patient studies, covering areas including data standards, governance of data access, data sharing agreements, and data preparation and transfer. BBSRC's policies are much less prescriptive, and focus on areas such as high-volume experimentation, long time series, and data from systems approaches, where there is a strong case for data sharing, and where the Council thus 'expects' data sharing. It merely 'encourages' sharing in other areas 'where there is strong scientific need and where it is cost-effective' though it reserves the right to introduce a more prescriptive approach in particular cases. At a more detailed level, while all the Councils – like the Wellcome Trust - require grant applicants to submit data management plans (DMPs), the specification of what is to be included in those plans differ very significantly, so that the Digital Curation Centre's DMPOnline tool has to provide bespoke templates for the different Councils as well as the Wellcome Trust and other major funders, including the NSF. The BBSRC states that where DMPs are assessed as unsatisfactory, specific feedback may be given, or a conditional award made; but some other Councils are silent on the matter. Like some other major research funders, the Research Councils make commitments to review and support the costs of implementing the plans, though the nature and terms of the funding commitments vary too. When it comes to monitoring of the implementation of the plans, there are again differences: MRC has detailed reporting requirements, while BBSRC relies on ResearchFish as its monitoring mechanism. ESRC explicitly mentions the possibility of sanctions if its requirements relating to the deposit of data are not adhered to. No information is available on the extent to which, if at all, sanctions have been applied by any of the Councils.

At an international level, under the open research data pilot in the EU's Horizon 2020 programme – which has now been extended to all thematic areas in the work programme for 2017 – open data is an option which researchers may decline to accept at any point: at application stage, or at any point after the grant agreement has been signed. Hence assessment of DMPs plays no part in the overall assessment of the grant application. The explicit aim is to monitor progress during the course of the programme, in order to develop policy further. But no information is yet available on the take-up of the pilot, or on its impact.

As is indicated by these kinds of developments, funders' policies are still in flux. At a generic level, the RCUK Common Principles and the guidance associated with them have been amended as new issues have emerged, and the same goes for individual Councils and other funders. Thus the most recent policy statements and guidance for BBSRC and NERC were revised in March 2016 (in other cases, it is not easy to tell whether or when policies and guidance were revised). From the researchers' perspective, however, keeping track of changes in policy and guidance represent a challenge: they cannot always be sure that they are up-to-date.

Two further issues should be noted. First, as stated earlier, the predominance of project-based funding for research data – even when some contribution towards indirect costs may be made - is widely seen as a barrier when it comes to building sustainable data management and preservation services. Reviews from the Wellcome Trust (Carr 2016) suggest that it is 'not clear' that the resource implications of DMPs are being adequately considered or provided for. It is notable that the Concordat on Open Research Data (Concordat 2016) is highly circumspect in its treatment of costs and funding. Some of the Research Councils – notably NERC and ESRC through their direct support

for their dedicated data centres – have provided funds for data archives and related services. But levels of support across the research sector as a whole remain patchy and inconsistent. Second, we also noted earlier some scepticism from researchers as to the weight given to DMPs in the assessment of their applications; and researchers are aware that it is a challenge for funders to check whether or not the sharing set out in the DMP actually takes place once a funding award has come to an end.

A recent OECD report (OECD 2015) concluded that policies relating to data sharing are "at a less mature stage" than those relating to open access publishing, mainly because of the more complex nature of data outputs as compared with publications. This 'lack of maturity' itself makes for difficulties for researchers, who are presented with a complex and uncertain set of requirements – or lack of them – especially if in the course of their work they seek support from more than one funder; and there are real uncertainties as to the real extent of the implementation, monitoring and enforcement of policy requirements. In this context, it is important that funders should not only seek to impose requirements on researchers, but take active measures to promote and support an enabling environment for open data, and to enhance incentives: carrots are needed as well as sticks.

Funders have as yet made little headway in promoting formal publication of datasets, mainly because of an understandable reluctance to direct researchers in where or how to publish disseminate their research. And it is notable that although researchers have been encouraged since 1988 to submit to the REF (formerly the RAE) outputs other than books and articles, including 'digital artefacts such as datasets,….. software…..[or] web content', only a tiny proportion (mostly in the arts and humanities) have done so, presumably because few researchers think that such outputs will be rated as highly as books or articles in scholarly journals. Nor is it clear that Research Councils and other major funders discriminate in favour of applicants for research grants who have a positive track record in making their data openly accessible.

Publishers themselves, however, may be beginning to have an impact here. The International Committee of Medical Journal Editors published early in 2016 a proposed set of requirements on the sharing of clinical trials data (though it should be noted that this has aroused some controversy) (Tachman et al 2016; Devereauux at al 2016). Beyond the area of clinical trials, a number of major publishers including Wiley Elsevier, and Springer Nature, along with smaller ones such as Pensoft, have recently established 'data journals': journals that promote the publication of papers that describe a dataset or group of datasets, accessible online, and publish according to standard scholarly journal practices. A recent survey (Candela 2015) has identified over a hundred such journals.

At the same time, some but not all publishers and journals have begun to establish and refine their policies relating to preservation and access to the data underlying the findings that they publish as journal articles; and Springer Nature, for example, has developed guidance, as well as a template, to help authors meet its requirements relating to the location and accessibility of custom code and software. Some of these policies, as at PLoSOne, have been controversial (Science Blogs 2014). Moreover, the landscape is currently complex, with many different approaches to data sharing – three different approaches have been identified just among the top ten general and internal medicine journals (Barbui, 2016); and attempts to create a register of such policies have so far foundered (Naughton, 2016). A number of commentators have recognised the need to make concerted efforts to harmonize standards and the language used in such policies, not least in order to help create greater awareness and action on data sharing, and so that a register *can* be developed to help researchers understand what they must do in order to comply with the policies (Hrynaszkiewicz, 2016). Such efforts are as yet only in their early stages. Journals are keenly aware, of course, of the cultures and practices of their different research communities; that those communities are at different stages in their preparedness to accept strong data sharing policies; and that they are at risk if they march too far

in advance of those different communities. But journals can perform a leadership role; and it seems likely that precisely because of the importance that researchers attach to successful publications in journals, the impact of acceptable and readily-understandable policies that were actually implemented and enforced could have a powerful impact on researchers' behaviour.

Finally, research institutions themselves have become more active in developing their own policies and guidance on data sharing and open data, as well as open access to publications. The Digital Curation Centre lists over 30 UK universities, from across all parts of the sector, with such policies, and more in the pipeline. Again, the nature and scope of the policies varies, but almost all of them cover such issues as a requirement for a DMP, and arrangements for preservation and access. In some cases, such as Manchester and Cambridge, the policies are accompanied by detailed guidance on issues such as file formats, metadata and documentation, as well as data storage and sharing data; and in the Cambridge case, guidance on sharing software and code.

4.2 Specialist services and community initiatives

We noted in Section 2 a number of perceived deficiencies in infrastructure and services, and in skills and training. A number of other papers in the series commissioned by the Wellcome Trust will address how some of these deficiencies might be addressed; but it is worth pointing to some of the current developments and initiatives that may help to improve the environment and incentives for the benefit of researchers. It is worth stressing at the outset, however, that the current complex landscape of top-down and bottom-up local, national and international initiatives and services is itself problematic for many researchers. Steps towards making it less complex and more readily-understandable, while respecting the needs of different research communities, would be most welcome.

4.2.1 International initiatives and services

The global nature of research means that some key parts of the infrastructure - and key efforts to improve it - perforce operate at international level. Thus data repositories such as GenBank, the Worldwide Protein Data Bank and the various World Data Centres operate as key services for preserving and providing access to data across a wide range of disciplines and subject areas; and given the large number of research data repositories and services across the world, the re3data service (since 2015 run under the auspices of DataCite) provides an essential searchable registry of such repositories, used by many publishers and funders to help them, their researchers and authors to identify the most appropriate data repository for their purposes. At a European level, OpenAire provides a similar service, and also aggregates content from a range of repositories. Some of the newer 'generic' repositories such as Figshare and Dryad seek to integrate with funders and publishers the workflows for deposit, the creation of metadata, links and so on, which can be helpful for researchers. Both have grown rapidly over the past three-four years: over 800,000 files have been uploaded into Figshare.

Other international initiatives seek to bring together specialists to develop solutions to facilitate data sharing, exchange, and interoperability; and to promote their adoption to improve the infrastructure at local and national as well as international levels. The biggest of these initiatives is the Research Data Alliance (RDA), established in 2013 with funding from the NSF, the European Commission and the Australian Government. It operates in the main through working groups and interest groups of experts with the aim of developing infrastructure that promotes data-sharing and data-driven research, in areas including metadata standards, interoperability, and security; and it has produced formal recommendations on topics including data type models and registries, workflows for data publishing, and machine-actionable templates for data policies. As yet, it is unclear how far these

recommendations will be taken up, or the impact they will have on policies and practices affecting researchers in any wide range of research domains. Similarly, it is not yet clear what impact will derive from the work of Force 11, which brings together experts from a variety of backgrounds with an interest in improving scholarly communications through the use of information technology. One of its most important outputs has been the FAIR principles, setting out four levels leading to an optimal state in which data – and other research objects – should be findable, accessible, interoperable and reusable by both humans and machines (Wilkinson et al 2016). The principles have been widely referenced, including, for example, in ESRC's latest version of its data policy, as well as in the Horizon 2020 data pilot. But any attempt to operationalise the principles brings the risk, unless relevant and easy-to-use tools are developed, that they will be seen by many data originators as simply another burden imposed on them for little reward in return; and in that sense as another example of the disconnect between data specialists – and the policy-makers they influence - and domain scientists.

In efforts to enhance the incentives for researchers to engage in open science, two international organisations are playing key roles. ORCID provides unique identifiers for researchers, distinguishing them from each other. It also seeks to build the identifiers into workflows for grant applications, manuscript submissions and deposits into repositories, with automated linkages between the different activities, so that it thus facilitates recognition of individual researchers and their work. Take-up has grown steadily, not least since the establishment of a UK national consortium, and there are now 2.5m live ORCID identifiers, linked to 6.5m persistent identifiers (DOIs) for research objects. DataCite provides DOIs for research data in order to help members of the research community locate, identify, and cite that data. It also provides a search facility, based on metadata collected (according to a prescribed schema) when DOIs are assigned, a citation formatter, and other services. There are close links with ORCID, and the aim, of course, is to promote data citation and thus to increase the incentives for data sharing. As we have seen (Section 2.2.6), there are still technical as well as cultural challenges to overcome before data citation becomes fully embedded and valued in practice; but if those aims are to be achieved, further development of DataCite and its services will be critical elements in the infrastructure. A number of publishers, repositories and other organisations have endorsed the Joint Declaration on Data Citation Principles issued by Force 11 in 2014, and further work is still under way on implementation, dissemination and adoption of those principles, including an expert group of publishers who are committed to the adoption of data citation. But citation remains in its early stages across most disciplines.

4.2.2 National and local services

We noted in Section 2.2.2 that the infrastructure of research data services both in the UK and the USA is fragmented. Jisc plays a key role in the UK in seeking to address that fragmentation, linking with international initiatives, providing services of its own and in supporting developments at institutional level. In so doing it works with groups of universities and with cross-sectoral organisations such as the Association of Research Managers and Administrators (ARMA), the Universities and Colleges Information Systems Association (UCISA), and Research Libraries UK. Services include a shared data centre, used by a number of universities and research institutes, particularly in London; pilot services for data discovery and usage metrics; and the safe share service to enable the secure exchange of sensitive data between different sites. It also provides guidance on good practice for institutions and individual researchers, most importantly through the Digital Curation Centre; and it supports and promotes developments in individual institutions and consortia, for instance in the development and adoption of services, software solutions, workflows, and training for researchers. And in specialist areas such as software and code, the Software Sustainability Institute has developed

guidance and training materials, as well as promoting recognition and rewards for the development of software as a research output in its own right, and career paths for research software engineers. Nevertheless, the challenges of developing a coherent and integrated set of services and infrastructure remain.

4.2.3 Approaches to infrastructure development

There are debates among information specialists about the most effective approaches to the development of infrastructure and services: distributed, federated, centralised, or some kind of hybrid. With the challenges of ever-increasing volumes of data, however, there seems to be a growing consensus that federated data services, combined with services that direct queries to all the repositories and databases in the federation, have the potential to provide significant benefits. From the perspective of the researcher, software and services that link to other services they wish to use can provide considerable advantages. Services with such linkages can be particularly attractive if they provide the flexibility to allow researchers to customise and adapt their own workflows, and if the documentation of those workflows is automated as far as possible. For this then allows for the automatic recording of key events that can bring transparency – whether managed or fully open – to the whole research process, as envisaged in the FAIR principles, and in development of services such as SHARE in the USA.

5. Conclusions and recommendations

Any measures to help embed cultures and enhance incentives to support open research are of course means to an end. The overriding aims are to promote and support better research, and to increase the impact and the social and economic benefits that derive from that research. It is also important to emphasise once more that while the focus to date has tended to be on open access to publications and – as in this report – on data sharing and open data, open research covers the methodologies, tools and code used at all stages of the research process. But open science policies must take account also of the varying cultures and practices of different parts of the research community (even while in some cases seeking to change them); and of the interests of the different partners and participants in research projects, particularly when they come from outside the public or voluntary sectors. Moreover, policies and other measures must not undermine proper levels of competition, or the implicit social compact, between researchers. Clear definition of the boundaries of open research to take account of these crucial features of the research ecosystem, and of different circumstances, is essential.

The analysis set out earlier in this report also makes clear that moves to promote open research among the different parts of the research community involve many different players at local, national and international levels: researchers themselves, but also funders, regulators, universities and other research institutions, learned societies, repositories and data centres, publishers, and service providers in the public, commercial and voluntary non-profit sectors. Discussion, exchange of ideas and consultation between these various actors and stakeholders is essential if effective progress is to be made in developing and assessing different models, and in promoting good practice. Lead organisations need to be identified to catalyse these discussions and engagement from key players, and to foster the necessary collaborative work. The Wellcome Trust could play an important role here, as it has in promoting open access to publications.

Beyond these general considerations, there are opportunities for action on a number of fronts, and there are high levels of agreement in the conclusions and recommendations arising from the various studies and initiatives considered in this report. And again it seems clear that in those areas where action from funders is recommended, the Wellcome Trust could play a leading role.

First, key players including funders, universities and research institutions, and publishers should make sure that they have in place clear policies for data management and data sharing. The precise terms and the levels of specificity of those policies will vary according to the nature of the organisation, and the breadth of the research communities they cover. But it is important that all organisations should regularly review their policies, which should be accompanied by documents setting out clearly the reasons for them, and guidance for researchers on their implications and on the actions required to comply with them. So far as possible, the policies should employ a common structure, and be expressed using common terminology (matters on which the RDA has produced recommendations).

Second, funders should review their policies and practices relating to grant applications and their requirements relating to data management plans (DMPs), and the arrangements to be made – or not - for data sharing. The reviews should include consideration of the scope for using common templates for DMPs, and the guidance given to applicants in drafting them, and to reviewers in assessing them. It is crucial that researchers should understand that the assessment of DMPs is critically important in the evaluation of their applications for research grants; and that funders should commit to funding those plans appropriately. It is acknowledged that checking post-award on the implementation of DMPs is a challenge, but funders should consider what steps they can take to facilitate such checks. Funders may also wish to consider whether they should require researchers explicitly to review in their grant applications the data that has already been generated in the relevant area of research; and whether they should establish grant programmes for the secondary analysis of data that is already available.

Third, universities, research institutions, funders and learned societies should work together to support and promote measures to extend the scope and reach of data citation, including endorsement of the Joint Declaration on Citation Principles, and support for the continuing efforts to overcome the barriers to more widespread adoption of data citation. They should jointly make clear that objective evidence of data sharing and its impact will be given proper weight alongside other evidence in evaluations of research performance, including the contributions of early career researchers and of data specialists who provide support to projects; and that it will thus feature in the criteria adopted in assessing individuals in competitions for recruitment, promotions, and project funding. They should consider the steps they might most effectively take to ensure that they have access to good evidence on this point, including a requirement for grant applicants to address specific questions about their record in data sharing. And as part of these efforts, they should consider endorsing the San Francisco Declaration on Research Assessment.

Fourth, universities and research institutions, and funders on both sides of the dual support system, should work together as consortia to establish training programmes to raise awareness and to enable researchers – at all stages of their careers - to develop appropriate data skills; and to enhance (through targeted funding and fellowship schemes) the numbers and status of, and the career paths for, data scientists with specialist skills to support the data management needs of domain scientists. As part of these efforts, they should consider working with research communities in specific subject areas or disciplines to support them in developing their own practices, or to build evidence to demonstrate the value of data sharing. Stakeholders in the UK must also work with international organisations such as CODATA to support and promote measures to mitigate the risks of a growing divide between domain researchers and data specialists.

Fifth, funders should work together with existing data centres, platforms and service providers, along with international initiatives such as the RDA and DataCite to address the problems of (the lack of) interoperability, so that researchers can more readily query, re-use, analyse and compute on the

growing volumes of complex, heterogeneous data. This will involve co-operative work to harmonise the current plethora of data standards, protocols, models and formats.

Sixth, funders on both sides of the dual support system must work in partnership with institutions to ensure that the full costs of data management and sharing are accurately assessed, and that resources are provided to meet them. They should in particular consider how best to ensure that key data services and repositories have secure and sustainable long-term funding. This is essential if data is to be preserved and remain accessible for the long term; and also to ensure that current gaps are filled and that services continue to develop in user-friendly ways as new technologies emerge. Public-private partnerships and arrangements under which some of the costs are met by data users may be appropriate in some cases.

Seventh, given the crucial role that they currently play in the research ecology, publishers are in an especially powerful position to influence researchers' behaviour. Funders and learned societies (particularly those that publish journals) should therefore work with publishers to help develop and implement common frameworks of policies relating to data publishing, including micro-publishing; deposit of and access to the data relating to projects and findings reported in scholarly articles (not just the data supporting the published findings); standards for data and file formats, and for metadata and supporting documentation; the use of DOIs; and data citation (preferably in reference lists). Given the critical importance for publishers of peer review, key stakeholders should also work with them to try to develop guidelines and methodologies for the peer review of datasets, and possible solutions to the problems identified in assessing both the technical and the scholarly quality of research data.

Finally, since cultures, practices and services are developing rapidly, it is important that funders should join with other stakeholders in regularly monitoring key indicators of progress towards open data and data sharing, and disseminating the results of that work as widely as possible, as a means of further stimulating that progress, and identifying key barriers that need to be addressed..

Annex 1

Sources Cited

Academy of Medical Sciences (2015) Reproducibility and reliability of biomedical research: improving research practice Symposium Report October 2015 http://www.acmedsci.ac.uk/download.php?f=file&i=32558 Accessed 25 July 2016

Amsen, E (2014) "What is open science" http://blog.f1000research.com/2014/11/11/what-is-open-science/. Accessed 25 July 2016

Barbui, C (2016) "Sharing all types of clinical data and harmonizing journal standards" *BMC Medicine* 14.63

Brack, M et al (2015) *Data Sharing for Public Health: Key Lessons from Other Sectors*, Chatham House, London

Candela, L et al (2015) "Data Journals: a Survey" *Journal of the Association for Information Science and Technology*, 66(9)

Carr, D. (2016). Maximising the value of research data: Wellcome Trust perspective [Presentation file]. https://www.repository.cam.ac.uk/handle/1810/253402 Accessed 25 July 2016

Chue Hong, N (2016) "Doing Science in the Digital Age: skills, tools and practice" Presentation to Jisc/CNI conference, July 2016 https://www.slideshare.net/JISC/equipping-the-researcher-patterns-in-the-uk-and-us Accessed 25 July 2016

Concordat (2016) Concordat on Open Research Data http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/ Accessed 29 July 2016.

Cutcher Gershenfeld, J et al (2016 "Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis", *Data Science Journal* 15.8

DAMVAD (2014) Sharing and archiving of publicly funded research data: Report to the Research Council of Norway http://www.damvad.com/wp-content/uploads/2016/01/satellite.pdf Accessed 25 July 2015

David, P.A. (2003), "The economic logic of "open science" and the balance between private property rights and the public domain in scientific data and information: A primer", in P. Uhlir and J. Esanu (eds.), *National Research Council on the Role of the Public Domain in Science*, National Academy Press, Washington, DC.

Devereaux, PJ et al (2016) "Toward Fairness in Data Sharing", *NEJM*, 375 pp 405-407

Enoch, J (2016) "Cancer Research UK and Data Sharing", presentation at Gurdon Institute, Cambridge, 22 January 2016 https://www.repository.cam.ac.uk/handle/1810/253403 Accessed 25 July 2016

EU (2016) High Level Expert Group on the European Open Science Cloud first report and recommendations, *A Cloud on the 2020 Horizon, Realising the European Open Science Cloud*

European Commission (2016) *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf Accessed 26 July 2016

Expert Advisory Group on Data Access (EAGDA) (2014) *Establishing Incentives and Changing Cultures to Support Data Access* https://wellcome.ac.uk/sites/default/files/establishing-incentives-and-changing-cultures-to-support-data-access-eagda-may14.pdf. Accessed 25 July 2016

Expert Advisory Group on Data Access (EAGDA) (2015a) *Governance of Data Access* https://wellcome.ac.uk/sites/default/files/governance-of-data-access-eagda-jun15.pdf Accessed 25 July 2015

Expert Advisory Group on Data Access (EAGDA) (2015b) *Governance of Data Access :Annexes* https://wellcome.ac.uk/sites/default/files/governance-of-data-access-annexes-eagda-jun15.pdf Accessed 25 July 2016

Fecher, B et al (2015) *A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing* Deutsches Institut für Wirtschaftsforschung http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2568693_code428549.pdf?abstractid=2568693&mirid=1 Accessed 25 July 2016

Gallagher, J et al (2015) "Facilitating open exchange of data and information" *Earth Science Informatics* 8: 721-739

Herther, N (2013) "Data Citation, Open Access and Reusability: Key Issues for Research Institutions" *Against the Grain*, October 2014

Hook, D (2016) "Infrastructure and services to track research activity" Presentation to Jisc/CNI conference, July 2016 https://www.slideshare.net/JISC/tracking-research-and-research-systems Accessed 25 July 2016

Hrynaszkiewicz, I (2016) "Promoting research data sharing at Springer Nature", Springer Open Blog 5 July 2016, http://blogs.springeropen.com/springeropen/2016/07/05/promoting-research-data-sharing-springer-nature/?utm_source=SpringerOpen+blog&utm_campaign=260b6d2ce3-Blog-SpringerOpen&utm_medium=email&utm_term=0_0964436351-260b6d2ce3-129417833. Accessed 25 July 2016

Knowledge Exchange (2014) *Sowing the seed: Incentives and motivations for sharing research data, a researcher's perspective* http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf Accessed 25 July 2016

Maguire, D (2016) "Efficient Information Infrastructure for UK Research" Presentation to Jisc/CNI conference 6 July 2016. https://www.slideshare.net/JISC/closing-plenary-john-wilkin-and-david-maguire Accessed 25 July 2016

Mathiak, B et al (2015) "Challenges in Matching Dataset Citation Strings to Datasets in Social Science" *D-Lib Magazine* 21, 1/2

Mitcham, J (2016) "Addressing the preservation gap at the University of York" Presentation to Jisc/CNI conference 6 July 2016 https://www.slideshare.net/JISC/repository-and-preservation-systems Accessed 25 July 2016

Naughton, L et al (2016) "Making sense of journal research data policies" *Insights*, 29 (1)

Noorman M, et al (2014) *Institutional barriers and good practice solutions*. EU RECODE Project http://recodeproject.eu/wp-content/uploads/2014/09/RECODE-D4.1-Institutional-barriers-FINAL.pdf Accessed 25 July 2016

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD, Paris

OECD (2015) *Making open science a reality* OECD, Paris

OSTP (2013) *Increasing Access to the Results of Federally Funded Scientific Research* Office of Science and Technology Policy, Washington DC https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf Accessed 25 July 2016

ParseInsight (2009) Insight into digital preservation of research output in Europe: survey report. http://www.parse-insight.eu/wp-content/uploads/sites/9/downloads/2015/07/Deliverables/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf Accessed 25 July 2016

Pasquetto, I et al (2016) "Exploring Openness in Data and Science: What is "Open," to Whom, When, and Why?" *Proceedings of the Association for Information Science and Technology*, 52, 1

Piwowar H et al (2013) Data reuse and the open data citation advantage *PeerJ*

RCUK (2015) Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/ Accessed 25 July 2016

Royal Society (2012) *Science as an open enterprise* Royal Society, London

Science Blogs (2014) "Data sharing is always good, right? Well, not quite…" http://scienceblogs.com/insolence/2014/02/27/data-sharing-is-always-good-right-well-not-quite/ Accessed 26 July 2016

Socha, Y (2013) "Out of Cite, Out of Mind: the current state of practice, policy, and technology for the citation of data" *Data Science Journal* 12

Tachman, D et al (2016) Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors, *JAMA* 315, 5

Tenopir, C et al "Data Sharing by Scientists: Practices and Perceptions" *PLoS One* 6 (6)

Tsoukala V (2015) *Policy guidelines for open access and data dissemination and preservation* EU RECODE Project http://policy.recodeproject.eu/assets/recode-full-report.pdf Accessed 25 July 2016

Walport, M. et al (2011). "Sharing research data to improve public health". *Lancet* 377: 537-539

Wilkin, J (2016) "Infrastructure for research and collaboration in the United States" Presentation to Jisc/CNI conference 6 July 2016 https://www.slideshare.net/JISC/closing-plenary-john-wilkin-and-david-maguire Accessed 25 July 2016

Wilkinson, M et al "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, 3, 160018

Wolff, C et al (2016) *Ithaka S+R Jisc RLUK UK Survey of Academics 2015*, Ithaka S+R, New York

October 2016

Version 1

Wellcome exists to improve health for everyone by helping great ideas to thrive. We're a global charitable foundation, both politically and financially independent. We support scientists and researchers, take on big problems, fuel imaginations and spark debate.

**Wellcome Trust, 215 Euston Road,
London NW1 2BE, UK
T +44 (0)20 7611 8888, F +44 (0)20 7611 8545,
E contact@wellcome.ac.uk, wellcome.ac.uk**