The International Journal of Digital Curation
Issue 1, Volume 4 | 2009

The Publication of Research Data: Researcher Attitudes and Behaviour

Aaron Griffiths,
Research Information Network

July 2008

Abstract

There is now widespread recognition that data are a valuable long-term resource and that making them publicly available is a way to realise their potential value – both as part of the scholarly record or for re-use by others. The Research Information Network (RIN) report, *To share or not to share: Publication and quality assurance of research data outputs* (June 2008), investigates whether or not researchers make their research data available to others and the issues they encounter when doing so. Importantly, it seeks to do this by seeking the perspectives of researchers themselves. This paper reflects on how this relates to the more top-down literature on the subject. The discussion of the significance of the RIN's main findings is correlated to the four themes of the RIN report. Firstly, it discusses some distinctions in the types of data that should be shared and preserved and what needs to done to do so effectively. Secondly, it reflects on the motivations for and constraints on researchers publishing their data, and how funders and publishers can address them. Thirdly, it reviews some issues around how data are discovered, accessed and re-used. Finally, it discusses the scholarly and technical quality of published data.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

It is increasingly widely recognised that research data are a valuable long-term resource and that making them publicly available is a way to realise their potential value – both as part of the scholarly record or for re-use by others. Funders, data managers and some researchers appear to be increasingly exercised by how to achieve more effective management and sharing of data in response to 'data deluge' (Hey & Trefethen, 2003) in the digital age.

Many international declarations and reports in recent years (e.g., Organisation for Economic Co-operation and Development [OECD], 2004; International Council for Science [ICSU], 2004) have emphasised the importance of research data and the need for infrastructure to manage it and make it accessible. National bodies have produced plans and frameworks. For example, in the US, the National Science Foundation (NSF) has committed itself to developing a national digital data framework comprising a coherent organisational framework, flexible technological architecture and coherent data policies (NSF, 2006). In Australia, the Australian National Data Service (ANDS) set out a vision for building the Australian Research Data Commons (ANDS, 2008). In the UK, there has been a feasibility study into a UK Research Data Service (UKRDS) as "a vehicle for achieving coherence in data management strategy and service provision across the UK" (Serco, 2008). Another significant report explored the roles, rights, responsibilities and relationships of institutions, data centres and other key stakeholders who work with data (Lyon, 2007).

What the Research Information Network (RIN) attempted to add to this in commissioning its report, *To Share or not to share: Publication and quality assurance of research data outputs*, was a bottom-up perspective: a focus on what researchers in the UK in a representative range of disciplines and subjects are actually doing, and what their motivations and constraints are. According to RIN Director Michael Jubb, we lacked "a clear picture of how researchers are responding to these challenges: whether they are in fact making their data available and accessible to others, and the issues that they are encountering when and if they do so" (RIN, 2008, p.5).

The RIN study, based on in-depth interviews with over 100 researchers, data managers and data experts, gathered information on researchers' attitudes and data-related practices in six discrete research areas – astronomy, chemical crystallography, classics, climate science, genomics, and social and public health sciences – and two interdisciplinary areas – systems biology and the UK's rural economy and land use (RELU) programme.

Studies of data management and sharing based on researchers' perspectives are not unique – for example see the survey-based research by Serco Ltd. for the UKRDS feasibility study (Serco, 2008) – but the RIN report is perhaps the most in-depth. The findings of the RIN report cannot be hailed as essentially new or surprising, but they do support the findings and recommendations of other reports with qualitative evidence from researchers themselves across a representative range of disciplines. This paper reflects on what the RIN's findings mean for the emerging policy frameworks.

Types of Data

Across the spectrum of subjects and disciplines, researchers create and collect many different kinds of data during the course of their research. Datasets are generated through different processes and methodologies, for different purposes and beneficiaries. Building on its previous work, and in line with research funders' efforts to classify different types of data (see a summary by Lyon, 2007, p.15), the RIN report encourages a nuanced view of the range of types of data and the distinctions which must be made in deciding which data should be shared and preserved. It recommends that "research funders and institutions need to take full account of the different kinds of data that researchers create and collect in the course of their research, and of the significant variations in researchers' attitudes, behaviours and needs, and to make clear the categories of data that they wish to see preserved and shared with others in each case" (RIN, 2008, p.17).

One distinction that the report considers in detail is "the raw data vs derived data issue" (RIN, 2008, p. 15). It notes that the convention regarding data produced in the normal course of research (i.e. as part of a process towards publication in journals, rather than for large datasets maintained for reference purposes) is that 'derived' data – data that have been processed or reduced in some way – are made available. There are practical reasons why data in their rawest form cannot be provided, such as sheer size and unwieldiness, as well as more cultural reasons: researchers may wish to keep the raw data to themselves to use in future work, or a community may have settled on a certain standard format and be content to work with that. Derived data are also generally easier to work with for those who wish to build on (but not reproduce) previous findings.

Posed against this is the idea that making raw data available means that checks and balances can operate at the most fundamental level. It can ensure that the research is reproducible – a cornerstone of the scientific method. Thus, the report claims, "it is not surprising that there is now considerable discussion in some communities about the lack of access to raw data" (RIN, 2008, p.15). Even in disciplines that have traditionally had a convention of sharing only derived data, such as chemical crystallography (where the convention is to share data in the CIF format), there is now discussion on about the merits of sharing raw data in the form of the diffraction patterns produced by the machines used to analyse the crystals.

Caring for Data

Who Cares for Data?

Data must to be cared for if they are to represent and remain a useful resource. Datasets can be made more accessible, re-usable and richer in content through annotation, aggregation, linking to other types of data, adding metadata, providing tools for manipulating and using the data, and curation.

The RIN findings supports the view that it is not uncommon for researchers to store data in a haphazard manner on their computers or on transportable storage media, with little or no idea of what will happen to them in the future, and with only rudimentary metadata. Relatively few researchers have the skills and resources necessary to care for data properly themselves, but established data centres and large

databanks do. However, some fields are less well catered for by good data services than others. Researchers working in the scientific disciplines that are catered for by good data facilities and services will often have received some training in how to use them, both for data retrieval and for data deposit. Researchers working on the larger and better funded projects in arts and humanities are likely to have sought advice from the Arts and Humanities Data Service (AHDS) or their own institution's computing centre about the best ways to manage the data that the project will produce. In other fields, even closely related ones, the story can be very different, with ad hoc, sometimes very temporary, arrangements in place for keeping and sharing data.

The RIN report pays close attention to two data caring issues in particular: metadata and long-term viability.

Metadata

Metadata should serve to provide information about an information resource, enabling efficient curation, management and re-use of the data. But where there is a lack of informative metadata or there are file format inconsistencies, datasets are, to all intents and purposes, lost to the community and consigned to obscurity. The RIN study found that the extent to which effective metadata schemes have been adopted varies considerably. In some fields, including astronomy, crystallography and areas of research that were covered by the AHDS (at least until it lost its funding as a national service in April 2008) there is a significant degree of standardisation deriving from datasets being stored and curated in professional or semi-professional databanks that require the provision of a structured set of metadata.

Long-term Viability

Data are prone to becoming unusable if they are not expertly curated. There are two main ways of storing and curating data reliably: using large, centralised national or international data centres; or using a distributed array of local data stores (based on or in research institutions, researchers' own resources, or formal publication outlets such as journals).

Researchers perceive that the centralised data centres are selective in what they will accept for curation and storage since they lack the capacity to take responsibility for everything that is produced in their disciplines or subject areas. Nor can data centres necessarily guarantee their own long-term existence, as the decision to stop funding the AHDS has shown. Distributed, local data storage is identified as a more "agile" approach, however the perception is that relatively few universities have the experience and expertise available for all that is involved in data curation and preservation. To succeed, such an approach requires further expertise and resources at the local level. The role of journals is highlighted as an interesting area, as there are two ways in which they make data available. The first is publishing datasets on their website (or insisting that the author deposits them in a databank); the second is eschewing the traditional journal article format and publishing datasets instead. A journal containing just a series of datasets (examples are *Acta Crystallographica E* from the International Union of Crystallography and, in project phase, the *Overlay Journal Infrastructure for Meteorological Sciences*), provide a formal way of citing them and ensuring that they are preserved at least in the medium term.

The RIN conclusion is that there is a critical role for research funders and institutions in seeking to ensure that long-term and sustainable arrangements are in place to preserve and make accessible the data that they deem to be of long-term value, and that such arrangements are not put at risk by short-term funding pressures. Yet importantly it notes that:

> Many research funders are putting policies in place to ensure that datasets judged to be potentially useful to others are curated in ways that allow discovery, access and re-use. But there is not a perfect match between cultural norms in some research disciplines and funder requirements. Some disciplines are well ahead of funding bodies in that they have had a culture of sharing data for a long time and have developed the infrastructures and methods for doing this. In other disciplines, data sharing is not commonplace and therefore funder policies may imply significant modifications to researchers' attitudes and behaviour. (RIN, 2008, p.12)

Much of the evidence the RIN offers on this matter is contained within the annexes to the report and is far too detailed to review here, but a summary is provided in the table reproduced below:

	Culture of sharing data	**Infrastructure-related barriers to publishing data**	**Effect of policy initiatives to encourage data publishing**	**Overall propensity to publish datasets (with appropriate metadata and contextual documentation)**
Astronomy	High	Low	Medium	High
Chemical crystallography	Medium	Low	Low	High
Genomics	High	Low	High	High
Systems biology	Medium	Medium	High	Medium
Classics	High	High	Medium	Medium
Social and Public Health Sciences	Low	Low	Low	Low
RELU	Medium	Low	Medium	Medium
Climate science	Low	Low	Medium	Low to Medium

Table 1. Summary of the position in each of the eight areas covered (c.f. RIN, 2008, p. 54).

Sharing and Publishing Data

There is widespread recognition that data are a valuable long-term resource and that sharing them and making them publicly-available is essential. Many UK research councils have introduced measures to encourage data publishing and sharing because they believe that datasets produced with public money should be available to other members of the research community – and indeed more widely. The RIN report provides some useful insights into the motivations and constraints faced by those whose data are at the centre of all this. (In doing so, it makes no practical distinction in terminology about "publishing" or "sharing" datasets, and, "publication of datasets" is used to mean "making datasets publicly available" in a general sense.)

In terms of the simplest mode of sharing data – responding to requests for access to it – the RIN research suggests that most researchers declare themselves happy to respond positively, especially if it may lead to co-authorship. However willing, though, many were often unable to meet such requests, most commonly by an inability to locate the data, or otherwise because of the time required to produce the necessary accompanying information (metadata or fuller explanations of the data and methodology).

Motivations for publishing datasets more generally were reported as altruism and acting for the good of scholarship, the expectation of reciprocation, positive feedback or esteem, greater visibility or opportunities for co-authorship or collaboration. Conversely, the report identified a number of factors that constrain researchers from publishing datasets:

- Lack of time and resources: there is a common perception among researchers that data management is time-consuming and costly. Even when funds have been provided to pay for it, they feel that those funds could be better used on the research itself.
- Lack of time to deal with requests for information: researchers may also worry that if they do publish datasets, they will have to spend valuable time dealing with requests and may also have to provide explanations, analytical tools, metadata, further data and so forth, all of which take time to gather and transmit.
- Lack of experience or expertise in data management (especially making it accessible and usable): there are many researchers for whom data management is an unfamiliar and daunting prospect. Even when expert support is available from research councils and data centres, researchers will not necessarily avail themselves of it.
- Legal or ethical constraints: it is not always clear to researchers who owns the datasets created during the course of their work and whether they have the rights to make the data publicly available. This rarely appears to prevent researchers sharing datasets directly with other individuals, but gives pause for thought when it comes to publishing the data more widely. In areas of research where personal data are collected, issues of confidentiality and data protection are an issue. Often interviewees' consent is sought only for the purposes of the original project, precluding re-use of those data for other projects.
- Uncertainity as to where to archive the data: if relevant data centres decide not to accept a dataset because it falls outside their selection criteria, researchers will often not have a fallback position. Although some researchers do publish and look after datasets themselves, many funders recognise that this is not an ideal use of researchers' time and resources.
- Competitive factors and fear of exploitation: the role of professional competition in limiting researchers' desire to publish datasets should not be underplayed. Many researchers wish to retain exclusive use of their data until they have extracted all the publication value they can. Many funders allow researchers a "reasonable" period of exclusive access to the datasets they have created. Nevertheless, some researchers still may not want to share their data or may wish to control who has access to them, fearing that their data may be misrepresented or that unwarranted conclusions may be drawn.

- Uncertainty as to demand for the data: although UK research councils increasingly convey the message that all data are unique and potentially valuable, researchers themselves do not necessarily take this view. Many find it difficult to believe anyone else will want access to their datasets – particularly with some data types such as model-run data or those deriving from small-scale projects. This supposition appears to be confirmed when requests for access are few in number, although this may partly be because data sharing is relatively uncommon in some disciplines, or because datasets are hard to discover.
- Limited or no specific reward: in an environment where researchers are assessed primarily according to their record of publishing in high-impact journals, there are few such career-related rewards for publishing datasets. Even where data management has become a mandatory part of research council grant application processes, researchers report that their dissemination behaviour is still primarily conditioned by the perceived strictures of research assessment exercises. Whereas the citation process is important for publishing papers in journals, it is limited with respect to datasets. Researchers will cite well known datasets within their subject area (though there is not always an accepted format for doing so), but for less recognised datasets citing articles based on them is more typical.

The report suggests that that researchers are more likely to publish data where they receive encouragement from peers or possess an interest in data-related issues, or where there is a data-sharing culture within their subject or niche. In areas such as astronomy, genomics and classics there is a tradition of sharing data and the infrastructure to do so. In other areas, such as climate modelling, data sharing and re-using other researchers' model-run data is not common practice, and hence researchers see little point in make data available for re-use. In social and public health sciences, there are several obstacles to data sharing, such as the right to confidentiality of those from whom primary data are gathered, or the expense of creating longitudinal datasets, and the RIN report found scant evidence of researchers wanting to publish datasets. It is perhaps this level of contemporary detail that distinguishes the report from other sources that tackle the incentives and disincentives for data sharing (see Borgman, 2007, pp. 192-222 and several sources cited therein). When asked what would encourage them to pay more attention to publishing or sharing their data, researchers interviewed for the RIN study typically pointed to one or more of the following incentives: evidence that there are benefits to be had from publishing datasets; standard, workable mechanisms for citing datasets; more explicit rewards in terms of career progression from funding bodies and research institutions; taking account of formal assessments of data sharing/publishing; closing the gap between reward for publishing papers and for publishing data; and taking account of past data-sharing/publishing track record when considering new grant applications.

Policies and Enablers

The challenge, then, is not further knowledge on incentives and constraints, but how to reflect these concerns in policy. Both funders and publishers have important roles. Many UK research councils require data management plans from researchers, but monitoring is difficult and is not routinely undertaken except in some long-term projects that are subject to interim review. Any monitoring inevitably places burdens on both researchers and funders at a time when there is pressure to reduce costs. Those

burdens have to be weighed, the RIN argues, against the benefits that can accrue from data sharing, and the desire to maximise the impact of funders' investment in the research process.

The RIN report summarises various measures that have been suggested to facilitate and encourage data publishing. They include: promoting its benefits through the use of case studies; providing better access to sources of expert advice; promoting the control mechanisms available to data creators (e.g. embargoes, restricted access, licence conditions); ensuring that there is an adequate physical infrastructure of data centres and services; promoting better discovery tools and metadata standards; and promoting standard, workable mechanisms for citing datasets. It stresses that funding bodies and research institutions may need to consider how to offer career-related rewards to researchers who publish high-quality data, taking account of formal assessments of data publishing, closing the gap between reward for publishing papers as against publishing data, and taking account of past data publishing record when considering new grant applications.

Discovery, Access and Usability of Datasets

The third theme of the RIN report is how the data that are available are rendered discoverable, accessible and usable. Researchers looking for data within their own particular discipline or subject area tend to have little difficulty discovering the datasets most relevant to their work, even where, as in most cases, their discovery routines are the product of habit and far from comprehensive. These routines comprise approaches such as searching sources with which they are closely acquainted, turning to peers or colleagues for advice, using published articles as signposts to datasets, using specialised data discovery tools provided by data centres or funders, or using a generic search engine such as Google. Researchers from other disciplines or subject areas, or people working outside the research community (for example from the commercial sector or government) find it more difficult to discover research datasets because they do not normally have access to a discipline-specific peer network, nor are they familiar with the relevant specialist discovery tools.

That data are available and discoverable does not necessarily mean that they are accessible. As the annexe to the RIN report explores in great detail, obstacles to accessibility are rare in some subjects and disciplines, such as the classics, but more prevalent in others, such as the social and public health sciences. Common obstacles include fees and charges, the requirement for licences or specialist tools, confidentiality issues, or that datasets are too large to transfer electronically for local processing.

Assuming that a dataset can be found and accessed, perhaps the biggest challenge is usability, which is central to enabling effective data sharing and data publication. Data centres invest heavily in ensuring that the datasets they look after are readily usable, but usability is an issue often overlooked by researchers publishing data themselves. Commonly, datasets are insufficient in themselves to enable other researchers to use them effectively. Data files in pdf format are especially problematic, since it may be impossible to manipulate them: in some disciplines the practice of making files available only in pdf format is known as "protecting by pdf". Even when the file format is satisfactory, it is often necessary to provide contextual information about how the data were collected and what tools or syntax were used to derive new

variables or produce particular analyses. Doing so also provides a means to alleviate data creators' concerns about their data being misrepresented or used inappropriately.

Particular issues arise with dynamic datasets where original data may be amended, added to, or replaced by newer data at a later date. The "freeze-and-build" approach, where original datasets are preserved and made available alongside the newer datasets rather than being replaced by them, is recommended in the report, but it is noted that it is not always clear whether multiple validations have taken place, or whether earlier data have simply been incorporated and assumed to be correct without further validation.

As the social sciences in particular become more data intensive, data residing within the text of published articles are of growing interest to some researchers. Experimentation with text mining is becoming more common in many areas of research. Here researchers report some confusion as to publishers' policies with regard to allowing access for text-mining tools to their journal contents. The RIN concludes that current uncertainties need to be resolved if the potential of this technology is to be realised.

Quality Assurance

Research communities are to a large extent self-regulating in respect of data quality assurance. Most researchers interviewed by the authors of the RIN report replied that they generally take other researchers' outputs on trust in terms of data quality and integrity, and there is little evidence of dissatisfaction with this state of affairs.

The scholarly merit of data is assessed by the peer community by comment, re-use, and building upon data outputs. Peer review may involve checking supporting data in a more or less detailed way. In many cases, reviewers may not be able to judge the data satisfactorily, and especially in scientific disciplines the datasets may be too large or complex to review manually or in their entirety. Reviewers may check that the data are present and in the format and of the type that the work warrants, and leave it at that.

Variability in the quality of peers' assessment of the content of the datasets that underpin publications is one of the key reasons why many researchers interviewed by the RIN "do not discount the idea of instituting a formal process for assessing the quality of datasets". However, the report continues, "no one can see it working effectively in practice" (RIN, 2008, p. 49). There are concerns that it would be difficult to find reviewers with sufficient expertise in highly specialised fields to understand the data, let alone appraise them. Reviewing datasets would also have costs and add time to the research process at a point in the project life cycle where researchers want to be writing papers for publication.

But, as the report notes, this lack of grassroots demand does not mean that research funders might not wish to see a more rigorous and consistent quality assurance process for datasets, particularly if they, along with other organisations, are investing heavily in the infrastructure required to support their publication. As Borgman notes, "As data reuse becomes more common, the pressure on reviewers to assess and certify data will only increase" (Borgman, 2007, p. 135).

The report notes that there is more to quality assessment than just the consideration of the scholarly merit of a dataset. If the process of data sharing is to become more effective and useful, much more consideration needs to be given to making datasets accessible (through the effective use of metadata) and usable (by providing the information and possibly software tools necessary for others to re-use the data). Whether the creators of datasets should be encouraged to gain such skills through education, persuasion, grant conditions or other means is an issue for research councils, other funders and the data centres to consider. An alternative approach would be to train and recognise the value of data scientists – whether from a research or information background – whose role would be to work alongside researchers, helping them devise and achieve the goals of effective data management plans. This suggests the need for further consideration by all stakeholders on what approaches to the formal assessment of datasets are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum.

Conclusions

The conclusion that the RIN report draws us towards is that however funders and other policy actors take forward the data management and data sharing agendas, they need to take into account the evidence on actual behaviours, motivations and constraints. A key policy imperative is to add to and reinforce the incentives and to reduce the constraints. Moreover, the report suggests that there are risks in doing this in ways that do not recognise disciplinary differences. Technology and policy evolve rapidly, but the RIN report shows that "… researchers' attitudes to data creation and dissemination are not keeping pace in all disciplines" (RIN, 2008, p. 11). One of the strongest messages to be drawn from the RIN study is the lack of uniformity across different research disciplines in terms of behaviour, policies or needs. Any solutions to the problems identified, therefore, will need to be tailored to the requirements, cultural norms and practices of each individual research discipline.

Acknowledgements

This paper reviews *To Share or not to share: Publication and quality assurance of research data outputs*, which reported on a study commissioned from Key Perspectives Ltd by the RIN in association with the Joint Information Systems Committee (JISC) and the Natural Environment Research Council (NERC).

References

Australian National Data Service. (2008). *Australian National Data Service (ANDS) interim business plan, 2008/9*, published September 2, 2008. Retrieved September 30, 2008, from http://ands.org.au/andsinterimbusinessplan-final.pdf

Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Hey, T., & Trefethen, A. (2003). 'The Data Deluge: An e-Science Perspective' in F. Berman, G. C. Fox, and T. Hey, (Eds.), *Grid Computing: Making the Global Infrastructure a Reality*. New York: Wiley and Sons.

International Council for Science. (2004). *Scientific data and information: Report of the CSPR Assessment Panel*. International Council for Science, December 2004. Retrieved September 30, 2008, from http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities and relationships*. UKOLN Consultancy Report, June 19, 2007. Retrieved September 30, 2008, from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

National Science Foundation. (2007). *NSF's cyberinfrastructure vision for 21st century discovery*. National Science Foundation Cyberinfrastructure Council, March 2007. Retrieved September 30, 2008, from http://www.nsf.gov/od/oci/ci_v5.pdf

Organisation for Economic Co-operation and Development. (2004). *Declaration on access to research data from public funding*, adopted January 30, 2004 in Paris. Retrieved September 30, 2008, from http://www.codataweb.org/UNESCOmtg/dryden-declaration.pdf

Research Information Network. (2008). *To Share or not to share: Publication and quality assurance of research data outputs*. Research Information Network, June 2008. Retrieved September 30, 2008, from http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf

Serco. (2008). *UKRDS interim report*. July 7, 2008. Retrieved September 30, 2008 from http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20(2).doc