Access to Research Data: Addressing the Problem through Journal Data Sharing Policies

Paul Sturges
Nottingham University, r.p.sturges@lboro.ac.uk

Marianne Bamkin
Nottingham University

Jane Anders
Nottingham University

Azhar Hussain
Nottingham University

Access to Research Data:

Addressing the Problem through Journal Data Sharing Policies

Paul Sturges, Marianne Bamkin, Jane Anders, Azhar Hussain, Centre for Research Communication, Nottingham University, Nottingham NG7 2NR, UK. r.p.sturges@lboro.ac.uk

Abstract

There is a growing consensus in the broader research community, including libraries and other information repositories, that sharing of research data is vital both for transparency and possible reuse. Logically the sharing should be in the form of data held in suitable repositories which is linked to effective access points such as library catalogues. The journals in which the research appears have a central role in this process. The JoRD Project at Nottingham University investigated the current state of journal data sharing policies through a survey of sample titles, and explored the views and practices of stakeholders including the research community and its funders, publishers and editors. The project identified that although a percentage of journals did have a policy on data sharing, they were in a minority, and policies generally encouraged good practice rather than made it a firm requirement. Many of the policies examined had little to say on standardised formats for data, metadata, or the use of data repositories. If there is to be genuine data sharing, initiatives to encourage journals to set out policies that mandate sharing in well-specified and appropriate forms are essential.

Keywords: research data; data sharing; journals; policies.

Introduction

It is widely agreed that sharing of research data is important for both research transparency and because of its potential for re-use in further research. This overwhelming weight of positive comment in favour of sharing, partially conceals the fact that the mechanisms by which sharing might be effectively implemented remain topics for discussion rather than functioning aspects of the research world. The first major problem relating to sharing is the sense that at least two of the key groups of stakeholders support it with their heads rather than their hearts. The researchers themselves tend to have an attachment to 'their' data which no rational argument for sharing can really drive out. The position of publishers derives from the fact that their revenue streams depend on acquiring and defending copyright in research articles and could very well also extend to profitable commercial 'sharing' of the research data which attaches to the articles. The JoRD project at Nottingham University,* which we will report here, confirmed scepticism about the extent to which the principle of data sharing is operative in practice. The research concentrated on the role of research journals in mandating and enabling sharing. We looked at the data sharing policies of journals in the expectation that these would, for instance, provide good guidance on data structures and metadata, and direct authors to suitable web-linked repositories. In fact we found that the state of journal data sharing policies at the current time was what can be described kindly as patchy and inconsistent. We then undertook qualitative research to assess the views of stakeholders and this confirmed the perception, but offered perspectives on how effective sharing might be encouraged.

In the course of the research, a second problem emerged which complicates the implementation of sharing. This is the definition of data in this context. The discussion of data often concerns a more complex set of resources than it might at first seem. Our starting point was a Royal Society (2012) definition of data as: 'Qualitative or quantitative statements or numbers that are (or are assumed to be) factual. Data may be raw or primary data (eg direct from measurement), or derivative of primary data, but they are not yet the product of analysis or interpretation other than calculation.' In fact, we found that what tended to be discussed or listed in discussions of data-sharing ranged through software, video, geodata, geological maps, ontologies, web content, data models and a great deal more. Furthermore, we found that it was impossible to totally ignore what are described as supplemental materials which might be deposited along with data. On supplemental materials, the NISO/NFAIS (NISO, 2013) has recently issued a set of recommended practices to address the lack of guidance on selection, delivery, aids to discovery and preservation plans, which incidentally makes much clearer the range of materials that can be involved. Without ignoring such materials we sought to concentrate on the essential data that supports results.

Literature Review

There are numerous authoritative statements in favour of data sharing. The International Council for Science (ICSU, 2004) the Organisation for Economic Cooperation and Development (OECD, 2007) and the UK Royal Society (Royal Society, 2012) have made firm statements on the topic, calling for openness and freely available access to publicly-funded research data. Similarly, funding bodies such as the US National Academy of Sciences (2003) expect data to be made openly accessible. There is also the Brussels Declaration (STM, 2007), which nevertheless reflects the unease of the publishing industry about open deposit of accepted manuscripts in rights-protected archives. There is previous research on the potential of journal data sharing policies. In the mid-1990s McCain (1995) surveyed 850 journals, discovering that only 132 had identifiable policies. A smaller survey of medical journals by Shriger et al (2006) found contradictory approaches and little strong guidance. Since then there has been a series of important papers by Piwowar, usually with Chapman (including Piwowar and Chapman, 2008b; Piwowar, 2010; Piwowar and Chapman 2010a; Piwowar and Chapman 2010b). Perhaps the most significant is Piwowar and Chapman 2008a, which builds on McCain's work, using the data on gene expression microarrays to explore policies in depth. The article classifies policies according to their strength (strong, weak, non-existent); the relationship of policy strength to the journal's impact rating; and the number of instances of data submission that can be identified. More recently, the PARSE project (Kuipers and van der Hoeven, 2009) has produced helpful data on attitudes to data sharing, and a strong viewpoint on what needs to be done (Smit, 2011; Smit and Gruttemeier, 2011). There is also Stodden et al, (2013) which is based on research of a broadly similar type to ours conducted more or less contemporaneously, but concentrating on the sharing of code that will enable computational results to be replicated.

Methods

Data policies were sought on the webpages of a widely representative survey of nearly 400 journals. Once a data policy had been located, it was broken down into categories such as: what, when and where to deposit data; accessibility of data; types of data; monitoring data compliance, consequences of non-compliance and policy strength, based on Piwowar & Chapman (2008a)'s definition of strong and weak journal policies. These were then entered onto a matrix for comparison. Where no policy was found on a journal's website, this fact was indicated on the matrix. In the first stage of analysis we looked at a series of individual policies in considerable detail and continued adding to the number of policies looked at in this way until we ceased to discover fresh features. This exercise provided a set of criteria that could be used for the analysis of all the remaining policies. Our results were based on the use of these criteria.

To provide a different angle, a stakeholder consultation was used to supplement the lessons of the journal survey, using simple qualitative methods to establish the views of key stakeholders. Views were elicited from publishers, funding agencies, data services, librarians, research administrators and managers on the principles underlying data sharing, the drivers for change, and the challenges faced in effecting change. This was done in a series of stages, using what was broadly a grounded theory approach. First of all views of individuals working on the publishing industry in the UK were elicited on the principles underlying data sharing, the drivers for change and the challenges faced effecting change. We selected the individual respondents by purposive sampling for their expertise. They came from a range of publishing backgrounds, from large to small, subscription to open access enterprises, together with representatives from funding agencies, data services, and also research administrators and managers. Later in the project interviews with four representatives of the academic library world were added. Because the data collected from the interviews was biased towards the point-of-view of journal editors and publishers we needed to explore the opinions of researchers and authors. A focus group of UK researchers was organised. Participants were selected by snowball sampling, initially through a contact from a scientific debate forum. They represented a range of Arts and Science backgrounds. Then we used the results from the focus group discussions and indications from the literature review to formulate questions for an open survey of researchers which was posted online for one month via the project blog (convenience sampling). Seventy researchers world wide responded from every academic disciplinary area and their subjects ranged over a total of 36 different scientific areas. After each stage of data collection, we open coded the data and identified patterns in response that formed categories which allowed the comparison of views across the range of stakeholders.

Findings

The Survey of Journals

The state of journal data sharing policies at the current time emerged as what can be described kindly as patchy and inconsistent. To describe the situation as thoroughly inadequate in an environment in which the rhetoric and policy both point to data sharing would not be unjust. Approximately half the journals examined had no data sharing policy at all. Of the journal policies found, more than three quarters were by Piwowar and Chapman's definition weak, with the remaining quarter strong (76%: 24%). Significantly, the journals with high impact factors tended to have the strongest policies. Not only did fewer low impact journals actually have any data sharing policy, those policies these were less likely to mandate data sharing. In general they merely suggested that authors might wish to share their data. Our survey interrogated the policies we identified to discover whether they included any stipulation of which data might be linked to an article, where the data should be deposited and when in the publishing process it should be made available.

Some policies did specify types of data to be deposited. For example, data sets, multimedia or specimens, samples or material were the most commonly mentioned types of data. Structures, protein or DNA sequencing and program code or software were referred to but less frequently. Many policies were not at all specific, using the terms; supporting information, unspecified data and other data. Other policies made a distinction between data that was integral to the article and supplemental data. Supplemental data might enhance the article but was not essential to support its argument. Indeed 7% of policies asked for the quantity of supplemental data to be limited or to be included only after discussion.

What is even more important is that few of the policies specified where the data should be deposited. A few talked of deposit but were vague as to where (7%). Others (17%) referred to the use of a repository but were not explicit as to which repository. Only 15% named a specific repository. Statements on expectations as to access were notably lacking, with only 28 out of the 371 policies surveyed commenting on this. Four of these talked of free access, two of open access and eighteen of low cost access. Perhaps most damning of all, only one policy discussed the inclusion of metadata

with deposits. On the question as to when the data should be deposited (either before publication or when publication occurred) there was again a lack of consistency and direction. 51% of policies that were specific on this broadly mentioned depositing data along with the submission of the article, with another 23% of the policies indicating that the data should be available for the peer review process. The remaining 26% of policies basically remarked that deposit at some later stage, typically on publication, was acceptable. In summary, we found low numbers of policies (for barely half of the journals surveyed) with the overwhelming majority of them weak and confusing. The weakness can be illustrated by the fact that only 10% contained mention of sanctions in the event of non-compliance.

The Stakeholder Consultation

There were low levels of mutual understanding between the stakeholder groups that were sampled in the interviews, focus groups and online enquiries. Stakeholders made assumptions about each other's views and actions and had obviously made little attempt to investigate the broader landscape. Although all stakeholders purported to be in favour of shared data and were willing to list the benefits of data sharing, they all raised caveats and concerns and identified barriers to the sharing of data. For instance, it was clear from researchers' comments during the focus group and from the online survey that they understood the expectation that data will be shared. At the same time, the online survey demonstrated a less positive reality. Around 40% of the respondents admitted that they did not allow others access to their data, and the rest mainly shared only with collaborators and colleagues. Researchers are not yet sharers by instinct: this underlines the importance of policy clarity in changing behaviour and awareness and advocacy of policy from funders' institutions and publishers. As noted above, it is at the point of publication that policy needs to be set out in the most specific terms for it to be effective. The publishers who need to present policy to authors on their websites and in the pages of their journals, in fact reveal anxieties over the capacity of the current digital infrastructure to allow data to be reliably linked to articles, if the data was distributed amongst a variety of databases and other repositories. Some of them were also not confident that their own databases would be viable alternative places of deposit because of the increasing file size of research data deposits and requirement for greater storage capacity. This implies that research institutions and funders have the opportunity to take the archiving issue in hand and they need to do so through clear, enforceable policy and clear easy-to-use deposit venues and processes.

A series of other anxieties emerged from the consultation. Both researchers and publishers considered that it would be difficult to deposit and link data in the original state in which they were gathered. There was a need for data to undergo a certain basic level of refinement before it might be shared. Raw qualitative data, for instance, might well be recorded in ways only truly understood by the data gatherer. This difficulty in the sharing and interpretation of purely raw data has been corroborated by the findings of work package one of the RECODE project (http://recodeproject.eu/). Similarly, large collections of quantitative data would require the correction of statistical errors before being fit to share. The context of the data gathering was also a factor: it might have been gathered with a promise of confidentiality; or it might have been gathered in order to complete a study (report or PhD thesis) for which there is a commitment that it should remain undisclosed for a specified amount of time. The currency of data was also an issue, with the danger that some data might either be too out of date by the time of publication to be of value for subsequent research. This difficulty relates to a wider requirement, identified by the publishers, that linked data in a journal article should be "fit for use" and "replicable". Data has been saved unstructured, not supplied with sufficient metadata, and in formats which have subsequently become incapable of retrieval. These anxieties need to be addressed at the policy creation level. Whilst research data from published articles might sometimes be open data, it is seldom linked or linkable data. If there is to be genuine data sharing, initiatives to encourage journals to set out policies that mandate sharing in well-specified and appropriate forms are essential. Our work on policy design is to be reported elsewhere in an article at present in the reviewing process for JASIST. However, we will indicate our line of approach in the following.

Developing a Model Policy

Our initial assumption that many of the problems of data sharing could be addressed in the publication process through the presentation by journals of strong clear policies on the issue was not contradicted by the research. The goal of identifying a model policy that could be recommended to journals therefore became a consistent focus of our activities. As we began to cumulate information about a large number of journal policies, it seemed for a time that a model policy would emerge from analysis of this material. However, we gradually became convinced that was not an adequate basis for a model policy. The cumulated features of existing policies tended to reflect the confusion, amounting at times to contradiction, in what publishers and editorial committees had so far set out. It became clear that an effective process required us to focus our attention on the views of the various stakeholders in the data sharing process. The first lessons this emphasis offered were that the current digital infrastructure is in a state of flux with such variation between publishers, repositories and systems that no powerful encouragement to share data emerges. We were clear that:

- Publishers vary widely in their approach to sharing data on which articles are based;
- Guidelines to authors concerning what type of data is acceptable, where the data should be deposited and when it should be deposited in the publication process are mainly vague;
- Researchers of all disciplines claim to be in favour of sharing data in principle, but perceive barriers which they do not know how to overcome;
- Researchers considered that they would benefit from clear publisher and journal policies on data format and place of deposit;
- Publishers also perceive barriers to linking and embedding data.

We also came to believe that it might often be impractical to include all data which supported the results reported in a journal article. Data formats and file sizes vary across a wide spectrum, very often dependant on the overall methodology for the research. Qualitative research generates data in the forms of documents and text, for example excavation and field observation notes, or transcripts of interviews or reports. Quantitative methods produce numerical data which are held in spreadsheets or only accessible when relevant computer code is also accessible. Many types of data might be generated from one piece of research, so an article might have to include extra text, numerical data sets and digital images which would increase its file size. In particular, the publishers showed concern about the ultimate file size required should large data sets be integrated into each and every article.

We noted that a consistent message from the research was that a major barrier to the open sharing of data was their inadequate knowledge of where to upload the data. Many were not aware of data repositories and those who were showed concern about their general infrastructure. The obvious implication was that a journal data policy should state whether the data should be deposited in a named repository with a trusted content policy, whether a permanent URL should be used and if any data citation style is necessary. The timing of the release of data raises an interesting point, researchers were not concerned about what point in the publication process the data should be made openly accessible, but at which point in their research. Articles are not only written at the conclusion of some studies, but at intervals during the research process. It may or may not be appropriate to release the data at the same point of the article, depending on such things as the established PhD premise that the research must be unique, the possible sensitivity of some forms of data, and ethical constraints that should protect human subjects.

Our initial model policy draft of the JoRD project covered the three questions of where? what? and when? That is, where data should be deposited, what type of data should be deposited, in which format, and at what time during the publication process, with also the possibility of embargos for the release of data at the correct time during the research process. The handling of sensitive data was not specifically addressed. The initial policy briefly mentioned data referencing under other instructions regarding data, but a full and clear statement about data citation and metadata in general is required by stakeholders. Similarly, it became clear that many stakeholder concerns about

Intellectual Property Rights of data needed to be allayed by the inclusion of recommendations about metadata associated with authors, such as Digital Object Identifiers. A further model framework for a journal research data policy was developed from the insights outlined above. It is intended to be capable of being used as a kind of 'policy engine' from which journal could develop a policy appropriate to their needs. We envisage a process whereby such policies are developed cooperatively between funders and research institutions on the one hand and publishers on the other. In the event of difficulties a resolution process would be needed, which would recognise the ultimate right of the funders to mandate the fate of the data generated by research for which they (or rather, the public) have paid. As noted above, the model policy will be included in an article in review at the time of writing.

Conclusions

The statements of principle on research data sharing have been made. The case is more or less unanswerable. However, the means to make sharing effective are currently lacking. The authors of this article are firmly convinced that a crucial intervention in the publication process should be made by the research journals in the form of data sharing policies. The JoRD project was in a position to both cumulate the content of existing policies and to develop ideas on the design of a policy on the basis of qualitative research. What we present here is evidence that needs to be fed into the process of policy creation so that researchers will have no doubts about what they are required to do to meet not merely the requirements of the journals, but those of the journal publishers and the whole research community that they serve.

References

ICSU (International Council for Science (2004) ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. Paris: ICSU.

Kuipers, T. and van der Hoeven, J. (2009) PARSE: Insight into issues of permanent access to the records of science in Europe. Survey report. Brussels: European Commission.

McCain, K. (1995) Mandating sharing: journal policies in the natural sciences. Science Communication 16, 403-431.

National Academy of Sciences (2003). Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. Retrieved Mar. 3 2014 from URL: http://www.nap.edu/catalog/10613.html

NISO (2013) NISO RP-15-3013, Recommended Practices for Online Supplemental Journal Article Materials. Retrieved 14 January, 2014 from www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf

OECD (Organisation for Economic Co-operation and Development) (2007) OECD Principles and Guidelines for Access to Research Data from Public funding. Paris: OECD.

Piwowar, H. and Chapman, W. (2008a) A review of journal policies for sharing research data In: Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing (ELPUB) June 25-27 2008, Toronto Canada. Retrieved Mar. 3 2014 from URL http://ocs.library.utronto.ca/index.php/Elpub/2008/paper/view/684

Piwowar, H. and Chapman, W. (2008b) Identifying data sharing in biomedical literature. AMIA Annual Symposium Proceedings, 596-600. Retreived Mar. 3 2014 from URL: http://www.ncbi.nih.gov/pmc/articles/PM2655927

Piwowar, H. and Chapman, W. (2010a) Public sharing of research datasets: a pilot study of associations. Journal of Informetrics 4(2) 148-156. Retrieved Mar. 3 2014 from URL http://www.sciencedirect.com/science/article/pii/S1751157709000881

Piwowar, H. and Chapman, W. (2010b) Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. Journal of Biomedical Discovery and Collaboration 5, 7-20. Retrieved Mar 3 2014 from URL: http://www.ncbi.nih.gov/pmc/articles/PMC2990274

Piwowar, H. (2010) Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLoS One 6:7 07. Retrieved Mar. 3 2014 from URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018657

Royal Society (2012) Science as an open enterprise: summary report, June 2012. London: Royal Society. Retrieved Mar. 3 2014 from URL: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf

Shriger, D. et al (2006) The content of medical journal instructions for authors. Annals of Emergency Medicine 48(6), 742-749.

Smit, E. and Gruttemeier, H. (2011) Are scholarly publications ready for the data era? Suggestions for best practice guidelines and common standards for the integration of data and publications. New Review of Information Networking 16(1) 54-70.

Smit, E. (2011) Abelard and Heloise: why data and publications belong together. D-Lib Magazine 17(1-2). Retrieved Mar. 3 2014 from URL: http://www.dlib.org/dlib/january11/smit/01smit

STM (International Association of Scientific, Technical and Medical Publishers) (2007) Brussels Declaration. Retrieved Mar.3 2014 from URL: http://www.stm-assoc.org/brussels-declaration/

Stodden, V. et al (2013) Towards reproducible computational research: an empirical analysis of data and code policy adoption by journals. PLOS One, June 21, 2013. Retrieved January 14, 2014 from doi;10.1371/journal.pone.0067111

NOTE * The JoRD project was funded by UK JISC (www.jisc.ac.uk).