Research Data Sharing:

Developing a Stakeholder-Driven Model for Journal Policies

Professor Paul Sturges, Loughborough University, Leicestershire, UK, LE11 3TU, email R.P.Sturges@lboro.ac.uk

Corresponding author Dr Marianne Bamkin, Centre for Research Communications, Nottingham University, Nottingham, UK, NG7 2UH Tel +44 (0) 115 84 68602, email Marianne.bamkin@nottingham.ac.uk Fax +44 (0) 115 8467577

Jane HS Anders, Centre for Research Communications, Nottingham University, Nottingham, UK, NG7 2UH Tel +44 (0) 115 95 14341, email jane.anders@nottingham.ac.uk Fax +44 (0) 115 8467577

Bill Hubbard, Centre for Research Communications, Nottingham University, Nottingham, UK, NG7 2UH Tel +44 (0) 115 84 67657, email bill.hubbard@nottingham.ac.uk Fax +44 (0) 115 8467577

Azhar Hussain, Centre for Research Communications, Nottingham University, Nottingham, UK, NG7 2UH Tel +44 (0) 115 84 67235, email azhar.hussain@nottingham.ac.uk Fax +44 (0) 115 8467577

Dr Melanie Heeley, Nottingham University, Nottingham, UK, NG7 2UH email M.Heely@nottingham.ac.uk

Abstract

The conclusions of research articles generally depend on bodies of data that cannot be included in the articles themselves. The sharing of this data is important for reasons of both transparency and possible reuse. Science, Technology and Medicine journals have an obvious role in facilitating sharing, but how they might do that is not yet clear. The Journal Research Data (JoRD) Project was a JISC (Joint Information Systems Committee) funded feasibility study on the possible shape of a central service on journal research data policies. The objectives of the study included, amongst other considerations: to identify the current state of journal data sharing policies and to investigate the views and practices of stakeholders to data sharing. The project confirmed that a large percentage of journals do not have a policy on data sharing, and that there are inconsistencies between the traceable journal data sharing policies. Such a state leaves authors unsure of whether they should deposit data relating to articles and where and how to share that data. In the absence of a consolidated infrastructure for the easy sharing of data, a journal data sharing model policy was developed. The model policy was developed from comparing the quantitative information gathered from analysing existing journal data policies with qualitative data collected from the stakeholders concerned. This article summarises the information gathered, outlines the process by which the model was developed and presents the model journal data sharing policy in full.

Introduction

Research data is presently a publicly-funded resource that passes into private hands without explicit permission, or remuneration to the public purse. The overwhelming volume of research across the disciplines is funded by government via research councils and institutions of higher education and by non-profit-making institutions set up for the public good. Organisations wish to maximise value in their investment and there is growing opinion from funders that access to data is part of that value. The Organisation for Economic Co-operation and Development (OECD, 2007) has published guidelines on access to publicly funded data where it states that “Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research”. Yet after the creation of research outputs, the data on which these outputs depend has, in the first place, tended to be left in the possession of the researchers who may use or neglect it as they see fit. This is not easy to justify, but it seems even harder to rectify. More recently, publishers have identified the data as a resource and facilitating access is capable of producing further revenue streams, but this apparent solution to the problem promises to exacerbate the public/private dilemma. Therefore it is important that the strength of the case in principle for sharing research data, both for reasons of transparency and the potential for reusing it in new research, has gained formal recognition from international and national research bodies, research funders, learned societies and the researchers themselves. These are the key stakeholders in research and ultimately it is their interests that should drive the research data sharing process.

The data with which these stakeholders are concerned is in fact a more complex set of resources than it might at first seem. Our starting point was a Royal Society (2012) definition: “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”. We found that what tended to be discussed or listed in data-sharing policies ranged through software, video, geodata, geological maps, ontologies, web content, data models and a great deal more. Although we sought to confine our attention to research generated data as such, we found that there was impossible to totally ignore supplemental material deposited with data that was actually behind the research results reported by the articles. On supplemental Materials, the National Information Standards Organization together with National Federation of Abstracting and Information Services (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. These are intended to assist publishers and editors to guide authors and peer reviewers in dealing with supplemental materials. As such, the recommended practices feed directly into a journal policies of the kind we model later in this article.

Firm statements on data sharing, calling for openness and freely available access to publicly-funded research data have been made by the International Council for Science (ICSU, 2004) and the UK Royal Society (Royal Society, 2012) in addition to that of the OECD (OECD, 2007). Similarly, funding bodies are requesting data management plans from researchers as part of their funding applications. This includes making the data openly accessible. For example, the AHRC (Arts and Humanities Research Council) funding guidelines “expects” digital outputs to be “freely available” to the research community. In the United States of America, the responsibility of authors to share data has been clearly set out by the National Academy of Sciences (2003), in a statement which also identifies the need for journals to specify data sharing policies for the benefit of authors. Furthermore, the Opportunities for Data Exchange project (ODE) underlines the need for publications and their supporting data to retain their essential integration (Reilly, Schallier, Schrimpf, Smit, Wilkinson, 2011). The Brussels Declaration (STM, 2007) is a statement from the publishing industry supporting the principle of free availability of access to research data although reflecting some of the unease about open deposit of accepted manuscripts in rights-protected archives. Yet despite all this weight of positive comment, the mechanisms by which sharing might be effectively implemented still remain topics for discussion rather than functioning aspects of the research world.

In the following article we explore what can be regarded as the pivotal aspect of any general mechanism for data sharing: the role of research journals and, in particular, the data sharing policies they present to their authors. This is an essentially pragmatic approach, recognising that the most effective policies are those which present themselves to researchers at a point in the research process where there is an immediate incentive for compliance and the opportunity to do so. The approach recognises that while both funders and employing research institutions may have policies which apply to the researcher, awareness of, and compliance with such polices can remain very low. Such policies are not typically presented to the researcher at the point where the data becomes available to be archived, nor do they offer an immediate incentive for compliance. However, a data policy that exists as part of the process of publication is presented after the research, and therefore data collection, is complete and has the incentive that compliance is needed for publication.

We believe that publishers and publisher policies have a key role to play in the wider adoption of data archiving and the development of model policies may assist in this. The article reports the findings of the Journal Research Data (JoRD) Project at the Centre for Research Communications (CRC) at the University of Nottingham, which was funded by JISC (www.jisc.ac.uk), and draws attention to the strong indications in these findings as to the shape of model data sharing policies for adoption by journals. It seems almost indisputable that the policies best capable of delivering transparency and reuse opportunities mandate deposit of data, provide guidance on structures and metadata, and direct authors to suitable web-linked repositories. Such policies not only benefit the researchers themselves and fellow researchers in the same and related fields, but also stimulate archiving and linked data activities that complement the basic act of deposit. Examining large numbers of existing policies, as we did, provides a view of what a model policy might say based on current practice. However, there is an alternative, that of a model policy that goes back to direct consideration of stakeholder concerns. We used both approaches in this study: analysis of existing policies and identifying stakeholder concerns through qualitative research.

Literature Review

The literature reveals that until quite recently, publications concerning what would now be framed as data-sharing issues frequently discussed them in terms of data-withholding. Campbell et al (2002) identified the pattern of data-withholding in genetics, based on the evidence of a substantial survey. Blumenthal et al (2006) and Vogeli et al (2006) also set out the issues in a context of data-withholding. Yet by the end of the 2000s Hodson (2009) could claim that the data culture had changed to one in which research collaboration, facilitated by the Internet, had led researchers generally to acknowledge the need to share data. It is, of course, open to question how deeply felt is the commitment of researchers and whether there is symmetry in attitudes towards others sharing data with a researcher and that researcher sharing data with others. What is more, it clearly varies across the spectrum of disciplines. Hrynaszkiewicz and Altman (2009) discuss the issue in terms of raw clinical data and Pianta et al (2010) show that there is a sharing culture in social sciences despite lack of structure in the available resources. Intellectual property issues are common to all disciplines, because by establishing the intellectual rights of synthesised ideas and the data from which the syntheses are derived, researchers can seek to consolidate their claims to research topics, innovations and conceptual direction. Reichman and Uhlir (2003) pursue the legal aspect of this intellectual property-based line of approach but the bulk of the current literature primarily concentrates on the value of sharing rather than defining obstacles. Neylon (2009) frames a positive treatment in terms of open data, and Fisher and Fortmann (2010) talk of data commons.

Arguably it is effectiveness of deposit procedures that is the crucial issue. Data that is notionally open and sharable may be in practice nothing of the kind because it is insufficiently structured, lacks metadata or has not been deposited in a repository that offers the capacity to fully realise external access. Articles written from the standpoint of sharing as a given notion and concentrate on the most appropriate method by which the inherent value of data can be disclosed to the research community generally look towards the concept of linked data. Kauppinen and Espindola (2011) identify what they call the four silver bullets of linked data, but Bechhofer et al (2011) adopt a more nuanced view. Delivering data fit for linking from the cumulations of notes, measures, mentions, readings and statistics that arise in the course of research requires a substantial organisational input on the part of the researchers. This is a message that goes well beyond the requirement to simply agree that the data must be made available for sharing. It is a message that cannot easily be given the necessary detailed specificity in high level declarations of principle from governments, international bodies and learned societies. The policies of funding institutions need to set it out clearly and explicitly so that structured data gathering can be built into the research process and so make data capable of being structured readily available at the point of deposition, most likely at the time of contact between the researchers and the journals in which they hope to publish their findings.

Such policies from funders, and from the research institutions that employ the productive researchers, are of course "upstream" of publisher policies in the research process and so will produce data with deposit requirements already attached. Therefore, journal policies must be able to accommodate pre-existing conditions and choices for deposit that are already invested in the data, with some process for resolution of any potential conflict between different policies which may arise. In spite of the primacy of the funders' and institutional policies, for the pragmatic reasons noted above, it is journal policies that are thus central to the wider adoption of the whole data sharing enterprise and the recent literature is beginning to reflect this. In the mid-1990s McCain (1995) surveyed 850 journals, discovering that only 132 had identifiable policies. The important, though unremarkable, conclusion was drawn that the best policies set out strong compliance sanctions. A smaller survey of medical journals by Schriger, Aroa, and Altman (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 & Chapman, 2008b; Piwowar, 2010; Piwowar & Chapman 2010a; Piwowar & Chapman 2010b). Perhaps the most valuable to the JoRD Project (Piwowar & Chapman 2008a) 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. The authors conclude that there is a wide variation in policies; some evidence that where there is policy then instances of data sharing increase; no real suggestion that a strong policy discourages authors from submitting their articles to a journal; and they provide some evidence as to the factors that make data sharing difficult for authors. More recently, the Permanent Access to the Records of Science in Europe project (PARSE.insight) (Kuipers & 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 & Gruttemeier, 2011). Stodden et al (2013) 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

Survey of Journals

We chose four hundred international and national journals to represent the top 200 most cited journals (high impact journals), and the bottom 200 least cited (low impact journals), equally shared between science and social science, based on the 2011 Thomson Reuters Journal Citation Report. There was some duplication between the two available indices and in those cases one incidence of the journal was removed. This left a total of 371 journals. We did not top up the total so as to avoid disrupting the impact factor ranges analysed. Thirty six subject areas were covered over both the broad disciplinary areas. The selection of journals that we analysed originated from a mix of large commercial publishers, academic presses, and independent publishers.

We sought data policies on each journal's webpage. Typically we found policies in the notes for authors or statements of editorial policy. Once we had located a data policy we broke it 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 and Chapman (2008)'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.

Stakeholder Consultation

In order to complement the survey of journal policies we sought to establish the views of key stakeholders, using qualitative methodology based on the sampling and analysis techniques of grounded theory. This structured approach allowed us to focus on stakeholder perceptions within a short time frame and iterative data selection with comparative analysis ensured that gaps in knowledge were filled. Views of individuals working for the publishing industry in the UK were elicited on the principles underlying data sharing, the drivers for change and the challenges faced in effecting change. We selected the individual respondents by purposive sampling for their expertise. Twelve came from a range of publishing backgrounds, from large to small, subscription to open access enterprises, together with four representatives from funding agencies (two of which were interviewed jointly), one data service manager, one representative of research administrators and managers and two academics. Thirteen structured interviews were conducted for the project, each lasting one hour. Six written responses to the interview questions were also obtained. Later in the project interviews with four representatives of the academic library world were added.

At this stage we suspected that the data collected from the interviews was biased towards the point-of-view of journal editors and publishers and did not sufficiently reveal the opinions of researchers and authors. Therefore 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. 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

We found at the time of analysis that the overall landscape of journal data sharing policies contained patchy and inconsistent coverage. Such a situation appeared inadequate in an environment in which the rhetoric and policy advise and encourage data sharing. For example, some journals had multiple policies (two or three) whereas 50% of the journals examined had no data sharing policy at all. Of the 230 journal policies found 76% were by Piwowar and Chapman's definition weak, with the remaining 24% being strong. 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.

Table 1 shows a summary of the main points that we discovered in the policies that we analysed. As can be seen, 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 and a small percentage (7%) asked for the quantity of supplemental data to be limited or to be included only after discussion.

Table 1: Summary of main points discovered from survey of journal data policies

What to deposit

Vague terms - Supporting information; Unspecified data; Other data; Supplemental data (after discussion)
Least commonly mentioned - Structures; Protein; DNA sequencing; Program code; Software
Most commonly mentioned - Data sets; Multimedia; Specimens; Samples; Material

Where to deposit

Vague, 7%
Un-named repository, 17%
Named repository, 15%

Expectations of access

Low cost access, 8%
Free access, 2%
Open Access, 1%

When to deposit

With submission, 51%
For peer review, 23%
On publication or later, 26%

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. Others 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 12% policies commenting on this. Accessibility options that were mentioned ranged from low cost to closed access, with only a low number of policies suggesting free or open access (see table 1). 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. Just over half of policies that were specific about this broadly mentioned depositing data along with the submission of the article, with roughly a quarter indicating that the data should be available for the peer review process and slightly more than a quarter of policies basically remarked that deposit at some later stage, typically on publication, was acceptable (table 1). 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 Policy RECommendations for Open Access to Research Data in Europe (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.

Developing a Model Policy

The 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. At this stage we assembled a draft policy based on relevant and useful aspects of existing policies. This took the following sixteen clause form.

- There should be a general statement outlining the benefits of data sharing
- The policy should clearly state whether it is the policy of the journal, the publisher, or that of a professional association
- The type of data to be included in the article or linked to the article
- The format of the data, covering any disciplinary guidelines
- Instructions related to the data, such as data citation, and other metadata
- Whether data is required or requested to be shared, and any limit to the quantity of data
- Where the data is the be held, according the data type
- Where to state what data is available and how to access it
- When during the publication process should data be made available
- Whether embargo periods are allowed and for what length of time
- Whether the data should be made openly accessible, free, low cost, or other levels of restrictions
- Any terms or conditions for the reuse of data should be stated by the author
- Whether exceptions to the data policy are allowable
- The method by which author compliance with the policy will be monitored
- A statement of the consequences to the author of non compliance with the policy
- A statement of the journal procedure for dealing with complaints from other researchers should their requests for data are not met

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 are generally in favour of sharing data, 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

To find a way through the difficulties this presented we brought the distinction made by Piwowar and Chapman (2008) between strong and weak policies to the centre of the process. They identified the following characteristics of a strong policy:

- A motivating statement for the benefits of data sharing to the scientific community
- A general statement implying support for data sharing
- Types of data which can be included in articles
- Whether the data should be available for peer review
- The wording of data sharing instruction, and whether data deposit is a condition of publication
- An instruction for the location of data archiving, for example, a webpage, or publicly accessible repository
- The format of data
- The completeness of data sets
- The timing of when data will be made openly available
- Possible consequences of non-compliance with the journal data policy

Consideration of these points assisted us in the process of identifying key findings from the qualitative research. A major finding of our study was that it would 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. 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. Certain publishers are indeed attempting to produce online journal articles that have the capacity to include many kinds of data, for example Elsevier's Article of the Future (http://www.elsevier.com/about/mission/innovative-tools/article-of-the-future). However, such a capacity is unlikely to be available for every journal. This creates a requirement that a journal policy should clearly state to what extent data can or cannot be included as an integral part of an article.

Linking crucial data to a journal article from a specific institutional repository is a reasonable alternative to overloading a publisher's server, although this transfers the associated long term cost to the host institution. Funders currently do not include such longer-term costs as part of a research grant and institutions may be reluctant to see these included within current overheads. Publishers also indicated a number of concerns about linking data from repositories. Firstly, hyperlinks should be permanent. A broken URL would not reflect well on the publisher or the author. Secondly, publishers queried whether there is a procedure for data citation because there are currently few standard data citation schemes. Both authors and publishers are concerned about intellectual property rights and at present the potentially divisive implications of this are not made fully obvious in existing policies. There is also the concern of continued data preservation should a repository close. It is also fair to say that similar concerns could be expressed in return by institutions should publishers host the material.

It is possible that the concerns expressed by the publishers can be allayed through the current development of data repositories that have the remit of securely storing data with reliable and easy linkages. For example, the Dryad Digital Repository collaborates with partner journals, data citation systems and uses permanent URLs (Queens University, 2013). Similarly, the Australian National Data Service (ANDS) is a national repository for research data generated by Australian Institutions (ANDS, 2014) that also incorporates data citation systems with Digital Object Identifiers (DOI, 2013). The concept of data citation is currently being explored by researchers, particularly with the rise of Data Journals, and the continuing development of DataCite which is a world wide organisation that works with data centres and publishers by providing persistent identifiers for datasets and other digital items (DataCite, 2013). Although digital repositories are a recent phenomena and their longevity has not been tested, responsible repository managers have policies that would come into play should a repository close. For example, the policy of Dryad, states that " In the event that Dryad can no longer maintain the Repository as an active service, all Dryad-registered DOIs will be updated to resolve to the copy at the CLOCKSS[1] archive, which will continue to provide free access to the Content under the same licensing terms." (Dryad, 2013)

We noted that a consistent message from the research was that a major barrier to the open sharing of data was not the reluctance of researchers, but 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 uniform resource locator (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.

While the JoRD project was looking at Social Science and Science journals in a global sense, the European Data Watch Extended (EDaWaX) project was examining the policies of Economics journals from the aspect of German economists. They started from a perception that Economics journals needed to mandate data sharing policies in order to ensure that economics research data would become available for replication and validation. The requirements for data availability policies that EDaWaX suggest in Vlaeminck (2013) are summarised as follows:

1. A journal data policy must stipulate that sharing data is mandatory
2. The original data with any necessary instruction for computation must be made available
3. The data files must be given to journal editors before an article is published
4. All the submitted files must be publicly available, unless they contain sensitive data
5. The journal data policy should contain a procedure of the method by which sensitive data sets could be used to replicate research
6. The journal should contain a replication section, which would include results of failed replications. This would encourage authors to provide good quality well documented data
7. Data should be submitted in open formats, preferably ASCII, to allow preservation and interoperability
8. The version of the operating system and software used for analysing the data should be supplied

The terseness of these recommendations is a merit, but they are not universal in their application. For instance, numbers 5 and 6 on replication are probably not relevant to a general research data policy. They are also quite categorical that sharing should be mandatory. A model journal research data policy, to cover many disciplines, might reasonably allow a journal to express whether the deposition of data is recommended or mandatory. More universal is the recommendation that data should be made openly accessible. EDaWaX considered the issue of the sensitivity of some data (for reasons including the personal, commercial and national). We also encountered these concerns. A model policy might respond by including exemptions, procedures for closed access, or embargo periods for sensitive data.

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, many stakeholder concerns about Intellectual Property Rights of data should be allayed by the inclusion of recommendations about metadata associated with authors, such as Digital Object Identifiers (DOIs) and Open Researcher and Contributor IDentity (ORCID) identifiers. ORCID identifiers are small pieces of unique code that can be used to identify academic authors entered on the ORCID registry, which can be found at: http://orcid.org/content/initiative. Other intellectual property rights (IPR) issues, particularly around funders' IPR, can be addressed by authors supplying clear statements as to the IPR status of the data and any re-use rights or restrictions. The quality issues of URLs and linked data also should be mentioned, with guidelines about choice of permanent URLs or universal resource indicators (URIs). Some researchers were under the impression that depositing data would automatically preserve or "future-proof" it. To respond to this misapprehension we felt that a policy should include a statement on the need for appropriate formatting and metadata as key contributions to the preservation process.

The following model framework for a journal research data policy was developed from the insights outlined above. We stress that it is not a policy in its own right, but that it is capable of being used as a kind of 'policy engine' from which journal policies could be developed. 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 is needed, which will as a prerequisite recognise the ultimate right of the funders to mandate the fate of the data which has been generated by research for which they - or rather the public - have paid.

Journal Research Data Policy Model Framework:

1. Policy statement on the benefits of data sharing - for example:
    - XYZ Publishing believes that the data used to draw conclusions from articles should be made widely available to the research community in order to facilitate collaboration, prove validation and encourage replication and re-use of the data. XYZ Publishing considers that such transparency benefits the author by greater exposure to their work and increased citation and improves the quality of science.

2. Designation of the policy owners - for example either of the following statements:
    - This research data policy is the policy of the Society of XYZ
    - This research data policy is the policy of the editorial board of The Journal of XYZ
    - This research data policy is the policy of XYZ Publishing

3. The policy should request that authors provide a statement identifying the original funder/s of the research which produced the data, or different parts of the data - for example the following statement:
    - Authors are required to name the funder which sponsored the research and collection of data on which an article is based

4. The policy should clearly state whether depositing data is mandatory to publication or is a recommendation - for example either of the following statements:

- It is a mandatory requirement of the publication of the submitted article that all data on which the article conclusions are based will be deposited by the author or authors in a location that is freely and openly accessible
- It is recommended that all data on which the article conclusions are based should be deposited by the author or authors in a location that is freely and openly accessible
- It is not necessary to make data associated with this article openly accessible

5. A policy should clearly state whether the data can or cannot be included as an integral part of an article or that hyperlinks should be included in the article, or appendices which lead to the data saved on a server which is different to that on which the article is held – for example:
   - Data will be embedded into the published article or appendices
   - Data must not be embedded into the published article or appendices
   - Data will be accessible through hyperlinks in the article that lead to another server which is/is not controlled by XYZ Publishing
   - Arrangements should be made for interested researchers to have access to the data

6. The policy should state whether the data should be deposited in a specifically named repository or a location of the author's choice – for example:
   - Data must be deposited in the data depository, Dryad (http://datadryad.org/) where The Journal of XYZ is an integrated journal
   - Data must be deposited in a repository that is accredited by the Society of XYZ
   - Data may be deposited in repository that has the XYZ Data Seal of Approval
   - Data may be deposited in the lead author's institutional repository
   - Data may be deposited in a trusted repository on the discretion of the author or authors
   - Data may be obtained by arrangement with the author/s

7. Should the data be linked to the article from another server, the policy should be clear about the form of URL which should be used – for example:
   - URLs used to link to the data must be permalinks
   - URLs used to link to the data must be Digital Object Identifiers
   - Authors must/ may use Uniform Resource Indicators to link the data to the article
   - Authors must/ may use Persistent Uniform Resource Locators to link the data to the article

8. The policy should be clear about the type of data which would be accepted bearing in mind the distinction between essential and supplemental data – for example:
   - Acceptable forms of data that can be linked to or embedded in articles are Video images/ audio files/ software/ spreadsheets/ text based files/ DNA sequences
   - Unacceptable forms of data to be linked or embedded into articles are Video images/ audio files/ software/ spreadsheets/ text based files/ DNA sequences

9. Guidance should be given on the selection of data from larger data sets which would be the most relevant to the published article – for example:

- If the published article is based on a limited quantity of data that was taken from a larger data set, only the data necessary for the article need be deposited
- If the published article is based on a limited quantity of data that was taken from a larger data set, we require that the entire data set must be made publicly accessible
- If the published article is based on a limited quantity of data that was taken from a larger data set, the author may choose to deposit some or all of the data set

10. The format of data accepted should be clearly indicated with an explanation given about the expectations of data preservation– for example:
    - Data will be accepted in any format
    - Data will only be accepted in ASCII-format in order to aid data preservation and interoperability
    - Data will be accepted in open formats in order to aid data preservation and interoperability
    - Data that requires access to code so that findings can be replicated will be deposited with that code.

11. Guidance on data citation style should be given if data citation is required – for example:
    - It is not necessary to reference the data
    - Authors may choose to reference the data
    - Data should be referenced using the following method (example given from Dryad)

      In text

      Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN]

      Bibliography

      Heneghan C, Thompson M, Billingsley M, Cohen, D (2011) Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.585t4

12. It should be made clear whether data should be reviewed and by whom – for example:
    - Data should be submitted along with the article in order to be peer reviewed by our appointed review team
    - Data should be independently reviewed
    - Data will not be reviewed

13. The policy should state whether an embargo can allowed for the public release of data – for example:
    - Data must be made openly accessible at the time of publication of the article
    - Data must be made openly accessible before the article is published
    - Data must be made openly accessible at least XXX weeks after the article is published
    - Data must be deposited before the article is published
    - Data may be deposited when the article is published with an embargo

14. The policy should state that ethical concerns on the publication of data from human subjects can be reassured - for example:
    - Prior to deposit, identifiers should be removed from Human subject data, such as names, addresses, dates of birth, social security or national health numbers, telephone numbers, etc
    - Human subject and other sensitive data may be allowed an embargo before release
    - Special arrangements may be made by authors for individual researchers to obtain Human subject and other sensitive data
    - In special cases Human subject and other sensitive data may be allowed an exemption

15. The policy should supply guidelines to authors on procedures for enabling individual researchers access to sensitive data – for example:
    - In the case of sensitive data which should not be made public, authors should make arrangements with individual researchers to pass on data sets
    - In the case of sensitive data which should not be made public, authors should make arrangements with individual researchers as how to replicate the study
    - In the case of sensitive data the contact details of the author will be supplied to interested parties

16. In the event of the policy allowing exemptions for certain types of data, the criteria for exemption should be clearly stated – for example:
    - The editorial board of The Journal of XYZ will consider exemptions to the research data policy should the author/s be able to prove that publication of the data they gathered will:
        - Be seriously detrimental to the life or lives of persons or their families who were participants of the research
        - Provoke serious consequences for an established industry
        - Aggravate serious consequences for national security

17. The policy should require authors to provide a statement concerning the IPR status of the data, or different parts of the data. Where re-use will be allowed, there should be a clear statement as to the re-use rights allowed, for example, using the Creative Commons licences (http://creativecommons.org/licenses) as the clearest and most widely understood re-use rights specifications. The statement should accommodate pre-existing IPR and/or re-use requirements arising from applicable funder or institutional policies, including embargo periods and treatment of sensitive data. For example:
    - This data is the result of funding from XYZ Funders, with shared IPR between the authors, their institutions and the funders in line with relevant policies. The data is released on an Attribution Non-Commercial Share Alike (CC BY-NC-SA) License after 6 months embargo from the time of publication, in line with the funding policy

18. There should be guidance on whether the method of data analysis should be declared – for example:
    - The method of data analysis should be made clear in the related article
    - A detailed method of data analysis should be provided to allow replication of the study
    - The author/s may chose to outline the data analysis

19. The policy should provide information on metadata and author identifiers – for example:
    - Data sets must be given an overall Digital Object Identifier (DOI)
    - Each item of data must be given a DOI
    - Data should be submitted with a README file which describes; coding and software, abbreviations and terms used, units of measurement and details of any other associated data

20. The policy compliance expectations should be prominently and clearly stated including any reasonable time limits allowed between publication and data deposit – for example:
    - XYZ Publishing expect that all authors will comply with the research data policy
    - XYZ Publishing will not publish an article until a notification is received from repository X that it has been duly deposited
    - XYZ Publishing will allow authors one calendar month from the data of publication for the deposition of data

21. Finally, consequences of non compliance with the journal research data policy and monitoring methods of non compliance should be prominently listed – for example:
    - Should The Journal of XYZ receive complaints from other researchers who cannot access data associated with a published article, the authors will be approached and evidence of data deposit must be produced
    - Should an author not comply with the policy of the Society of XYZ, membership to the organisation will be revoked
    - Should data not be deposited within the given time limit, XYZ Publishing will no longer publish papers written by the author/s of the associated article

Conclusions

A model policy is no more than the term implies: a suggestion. The JoRD project was in a position to both cumulate the content of existing policies and to design a policy on the basis of qualitative research. The model outlined above is confidently offered to publishers, editors, editorial boards and organisations such as scholarly societies and research institutes. They will nevertheless need to examine it closely to assess its fit with their specific needs, and adapt it as necessary. What is utterly essential in the opinion of the authors of this article is that journals should offer a policy, and offer the best policy that they can devise. This model is intended to facilitate and strengthen that process.

[1] CLOCKSS: Controlled Lots of Copies Keep Stuff Safe

References

ANDS (2014) About ANDS. Retrieved March 20, 2014 from http://www.ands.org.au/about-ands.html

Bechhofer, S. et al (2011) Why linked data is not enough for scientists. Future Generation Computer Systems. 29 (2) 599–611

Blumenthal, D. et al (2006) Data withholding in genetics and the other life sciences: prevalences and predictors. Academic Medicine: Journal of the Association of American Medical Colleges 81(2) 137-145.

Campbell, E. et al (2002) Data withholding in academic genetics: Evidence from a National Survey. Journal of the American Medical Association 287(4):473-480

DataCite (2013) Why Cite Data? Retrieved January 14, 2014 from http://www.datacite.org/whycitedata

DOI (2013) The DOI System. Retrieved January 14, 2014 from http://www.doi.org/

Dryad (2013). Dryad terms of service. Retrieved January 14, 2014 from http://datadryad.org/themes/Mirage/docs/TermsOfService-A4-2013.08.22.pdf

Fisher, J. and Fortmann, L. (2010) Governing the data commons: policy, practice and the advancement of science. Information and Management 47(4) 237-245.

Hrynaszkiewicz, I. and Altman, D. (2009). Towards agreement on best practice for publishing raw clinical trial data. Trials 10(17) 1-5. Retrieved January 14, 2014 from http://www.biomedcentral.com/content/pdf/1745-6215-10-17.pdf

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

Kauppinen, T. and Espindola, G. (2011) Linked open science – communicating, sharing and evaluating data, methods and results for executable papers. Procedia Computer Science 4, 726-731.

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 January 14, 2014 from http://www.nap.edu/catalog/10613.html

Neylon, C. (2009) Scientists lead the push for open data sharing. Research Information 41, 22-23.

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.

Pianta, A. et al (2010) The enduring value of social science research: the use and reuse of primary research data. In: The Organisation, Economics and Policy of Scientific Research Workshop, Torino, Italy, April 2010. Retrieved January 14, 2014http://deepblue.lib.umich.edu/bitstream/handle/2027.42/78307/pienta_alter_lyle_100331.pdf?sequence=1

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 January 14, 2014 from http://elpub.architexturez.net/doc/oai-elpub.id-001_elpub2008

Piwowar, H. and Chapman, W. (2008b) Identifying data sharing in biomedical literature. AMIA Annual Symposium Proceedings, 596-600. Retrieved January 14, 2014 from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655927/

Piwowar, H. and Chapman, W. (2010a) Public sharing of research datasets: a pilot study of associations. Journal of Infometrics 4(2) 148-156. Retrieved January 14, 2014 from 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 January 14, 2014 from 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 January 14, 2014 from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018657

Queens University (2013) Permanent links Types for Popular Research Databases. Retrieved January 14, 2014 from http://library.queensu.ca/help/permanent-links-databases

Reichman, J. and Uhlir, P. (2003) A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. Law and Contemporary Problems 66(1/2) 315-462.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., and Wilkinson, M., (2011) Report on integration of data and publications. Retrieved January 14, 2014 from http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

Royal Society (2012) Science as an open enterprise: summary report, June 2012. London: Royal Society. Retrieved January 14, 2014 from http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf

Schriger, D., Aroa, S., Altman, D. (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 January 14, 2014 from http://www.dlib.org/dlib/january11/smit/01smit.html

STM (International Association of Scientific, Technical and Medical Publishers) (2007) Brussels Declaration. Retrieved January 14, 2014 from 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

Thomson Reuters (2011), Journal Citation Reports, 2011. Retrieved July 30, 2012 from http://adminapps.webofknowledge.com/JCR/JCR?SID=P13wBWytszGuKJWINir&locale=en_US

Vogeli, C. et al (2006) Data withholding and the next generation of scientists: results of a national survey. American Medicine: Journal of the Association of American Medical Colleges 81(2) 128-136. Retrieved January 14, 2014 from http://view.ncbi.nlm.nih.gov/pubmed/16436573

Vlaeminck, S., (2013) Data management in scholarly journals and possible roles for libraries – some insights from EDaWaX. LIBER Quarterly, 23(1).