Making sense of journal research data policies

This article gives an overview of the findings from the first phase of the Jisc Journal Research Data Policy Registry pilot (JRDPR), which is currently under way. The project continues from the initial study, 'Journal of Research Data policy bank' (JoRD), carried out by Nottingham University's Centre for Research Communication from 2012 to 2014. The project undertook an analysis of 250 journal research data policies to assess the feasibility of developing a policy registry to assist researchers and support staff to comply with research data publication requirements. The evidence shows that the current research data policy ecosystem is in critical need of standardization and harmonization if such services are to be built and implemented. To this end, the article proposes the next steps for the project with the objective of ultimately moving towards a modern research infrastructure based on machine-readable policies that support a more open scholarly communications environment.

Introduction

LINDA NAUGHTON

Head of Research, Jisc, UK

DAVID KERNOHAN

Senior Co-Design Manager, Jisc, UK

The research data landscape has changed considerably in recent years. In part, this has been driven by the momentum of the open agenda and an increasing number of funder policies which share a vision for accessible research outputs. In parallel, the development of the open publishing sector has accelerated the ease and speed with which publications are made openly available. There has also been a rise in the number of data journals which give researchers an alternative avenue to gain credit for data-based outputs. The debates around research data policy and research data management (RDM) have been drivers of good practice with views from diverse stakeholder groups coalescing on the importance of the accessibility of the 'data behind the paper.' Recent discussions, such as those at the UK Open Research Data Forum, have indicated that there is a need to encourage the development of journal policies for data – drawing on the growing evidence that mandatory data polices incentivize data sharing described in the literature review – and that recognition of data in the publishing process will help to incentivize data sharing and reuse.

In order to engage and comply with this new policy environment, researchers need access to clear guidelines on the journals' expectations when it comes to the deposit and accessibility of the supporting data. Currently, journal data policies can be difficult to discover and difficult to reliably understand after discovery. Our consultations reported that this presented an issue for researchers and supporting institutional staff. The initial aim of the Journal Research Data Policy Registry (JRDPR) pilot was to assess the feasibility of a service that enabled researchers, research support staff, publishers, journals and other interested stakeholders to create, search, view and update research data policies. The project was based on the recommendations of an earlier project, Journal Research Data policy bank (JoRD), funded by Jisc and carried out by Nottingham University's Centre for Research Communications.[1] The JoRD study found that a central service could assist data management in the following ways: by facilitating easy access to journal data policies; by providing clarity on when, where and what to deposit; by offering guidance on file and metadata formats; and in helping librarians to support researchers to deposit data. The JoRD study concluded that although the idea of making scientific data openly accessible for sharing is widely accepted in the scientific community, the practice is confronted with serious obstacles. The most immediate of these obstacles is the lack of a consolidated infrastructure for the easy sharing of data. In consequence, some researchers simply do not know how to share their data.

'some researchers simply do not know how to share their data'

Literature review

The high level of current interest in journal research data policies is reflected in the rapid growth in published literature and reports that deal with the issue. The JoRD project conducted an exhaustive review of the pre-2012 academic literature,[2] which is summarized in Sturges.[3] This paper makes a strong, evidence-led case for the standardization of research data formats and metadata descriptions via journal policy mandates, themes echoed in the presentation of a model policy for research data sharing.[4] Several other international projects have made recommendations concerning journal data policies as part of a wider examination of RDM issues, notably the EU-funded RECODE project,[5] an NSF-supported OceanObsNetwork collaboration,[6] and a 'publisher summit' led by PLOS.[7] These have generally taken the form of assertions that journals and publishers should have clear and visible data policies.

'the existence of a journal data policy is positively correlated with data-sharing practice'

Piwowar and Chapman[8] examine the strength of journal data policies in the field of gene expression microarray data, which serves as a useful case study of one of the more engaged sub-disciplines. They found that only 17 out of 70 journals in this field had a 'strong' (enforceable) data policy. They noted that the existence of a journal data policy is positively correlated with data-sharing practice amongst authors and that many journals had a microarray data policy that did not extend to other forms of data. This highlights the difficulties encountered in attempting to build on existing good data management practice in a sub-discipline, where data-sharing practice in one area is not easily adapted to cognate areas. Similarly, Vines[9] examined a range of journals within the population genetics subject area, measuring the level of data sharing against the existence or otherwise of a journal research data policy. A positive correlation was identified, with a more stringent data policy yielding significantly more data sharing.

Overall, recent literature reflects the consensus on the need for clear and credible journal data policies, both from a 'pure' policy perspective, and based on attempts to collect and classify policies. It is also relevant to note that these are themes that were also found in the literature on open access (OA) policies.[10]

'consensus on the need for clear and credible journal data policies'

The Project

As a part of our innovation process, RDM was identified by the sector as one of six priorities that would benefit from Jisc intervention.[11] Following a number of workshops and consultative events held in 2014, the Research at Risk workplan was created.[12] The aim of the plan is to realize a robust, sustainable RDM infrastructure and services that enrich UK research.

The discovery phase of the JRDPR pilot ran from April to September 2015. Four streams of parallel and interrelated activity were undertaken: engagement with the community; creating a data model and question set; the analysis of policies; and assessment of the technical feasibility of a prototype service. The following sections give an overview of the findings from this activity.

Engagement with the community

One of the first tasks in this strand was to establish an Expert Advisory Group to guide the project, with stakeholder representation from across institutions, data centres, publishers, journals, learned societies and funders from the UK and internationally. The project was promoted at Jisc events as we developed the Research at Risk portfolio. At Repofringe 2015 the session on JRDPR was well attended although many people acknowledged that this was an area in which they had very little experience or information. Despite a number of programme clashes at the Research Data Alliance 6th plenary meeting, the 'Birds of a Feather' session was well attended and again there was wide representation from across the stakeholder groups, demonstrating the widespread interest in the topic.

Despite the community support for and interest in the area of work, it became apparent that there was a lack of consensus on what constitutes good practice as the stakeholder groups have different objectives and incentives. For example, there were differences of opinion on where the responsibility lay for driving good practice. Some participants wanted to see more of a role for journals and publishers in driving data sharing because this levelled the playing field with regard to authors from different countries. For others, the onus was put on funders' mandates for data sharing at the level of the complete data set so that publishers would only be required to point to the subset of data which validated the article. For another group the norms were determined by practice at the domain level and not helpfully prescribed by policy. The engagement activity highlighted the need for a forum to engage all stakeholders so that solutions can be found that are a best fit with the varied interests.

'there was a lack of consensus on what constitutes good practice'

Data model and question set

A pragmatic approach was taken to develop a question set with which to query policy information that focused on the elements that could feasibly be collected from existing sources. This activity was run in parallel with the policy analysis activity to gauge the difficulty of collecting the information. Some elements were included to drive best practice e.g. data access statements and licensing information. A candidate data model and question set was developed for the prototype service.[13] The question set went through multiple iterations as data consistency issues were raised.

The majority of these issues related to a lack of standard definitions of terms. For example, there was no standard definition applied to 'the data' or 'the supplementary material.' The NISO definition for supplementary material makes the distinction between integral and additional content.[14] The first relates to material which is 'essential for full understanding of the work' and the second relates to that which 'provides additional, relevant and useful expansion of the work,' but these guidelines have not been widely adopted by journals. A similar problem exists with regard to terms such as 'data sharing,' 'the data set' and 'peer review of data.' These terms are commonly used in research data policies but are often defined by community practice or via domain norms with respect to the particular types of data. This creates a problem when trying to codify information at the generic policy level. While there are very few commonly applied definitions, there are moves towards the development of common principles (UK Draft Concordat on Open Research Data)[15] and standards (TOP Guidelines).[16]

The objective of the prototype was to assist authors and research support staff with the publication of their data. To this end we included questions on what, when, how and where to deposit data. As these questions are often answered at the domain level, this made codifying the answers an unscalable task. The level of granularity required to capture policy at the data-set level was too complex for the data model. In addition, the answers to these questions often related to the submission guidelines rather than the policy. In a registry of *policies* it would be necessary to delineate between the two, although this separation was not found in practice. If the registry was to include the more moveable feast of submission guidelines then it would be necessary to account for changes to web pages, which would be difficult given the current lack of version control.

'questions on what, when, how and where to deposit data'

At the end of the discovery phase, a basic question set and data model had been developed on the basis of the information that could be found and codified, albeit imperfectly. The 16 questions were at a generic level and could not capture all the information required by authors or support staff.

Policy analysis

The policy analysis exercise was conducted with the updated question set on the journals previously analysed as part of the JoRD project. The analysis covered 250 top ranked journals, split equally between sciences and social sciences. Due to the data

consistency issues outlined above, the analysis necessarily contains a degree of subjective interpretation. As the data model changed, it was necessary to return to the cleansed data repeatedly to apply new definitions and/or reclassify the answers as more options became known.

On average just over half of the journals (52%) had a research data policy (science 65%, social science 40%), which is a 7% improvement since the JoRD survey. On average 30% of journals mandated some data sharing when 'data sharing' is interpreted as deposit of data in a public repository, with a significant difference between science (45.8%) and social science (10.5%) journals. These results do not show a significant change of practice since the JoRD survey but this does not take into account the domain level where changes are more likely to occur. We can conclude that there has not been a large-scale movement towards mandating data deposit although it would seem there has been an upward trend in the science literature.

'there has not been a large-scale movement towards mandating data deposit'

Technical feasibility of building a prototype

During the discovery phase, a rapid prototype of the registry was built to test the data model using the MEAN stack (MongoDB, Express, AngularJS and Node.js). The administrative functions were tested by inputting data from a small sample of policies on the basis of the question set. The tests were successful and a prototype service was deemed technically feasible but such a service would not meet the primary use cases. As a registry of high-level policy information, it does not meet the needs of authors and support staff looking for detailed information on what, when, where and how to deposit data. A full technical specification is available.[17]

Conclusions and next steps

A number of conclusions were drawn from the discovery stage of the pilot which informed the next steps of the project. The prototype that could be built would not be an authoritative source of information for researchers or support staff as it would not contain the information required at the level of data type. Additionally, there would be issues of authority if the administration of the database was not adequately supported by the publishing community in terms of maintaining accurate records and implementing version control on web-based information sources. To answer the question set using the sources available, a high degree of subjectivity and interpretation had to be applied as there were very few standard terms or definitions. Interpretation of policy was often best undertaken at the domain level, which further compounded the problems of building a scalable, generic database to codify the information.

It was agreed that the development of a journal policy registry would be impacted by changes in other areas and that it was therefore necessary to look at journal policies within the wider ecosystem of research data publication. The pilot highlighted many of the trade-offs and tensions in this environment. The tension between what can be prescribed at a policy level and the detailed guidance authors need in order to fulfil their obligations was evident. Both funders and publishers justifiably avoid prescribing the mechanics of how data will be shared as this is likely to change rapidly as the infrastructure matures. This means the responsibility is often left to the journal to provide authors with all of the necessary information. For small under-resourced journals this may be a daunting task which may influence editors towards less prescription and stringency when setting policy.

'the responsibility is often left to the journal to provide authors with all of the necessary information'

Unlike OA policies which focus on the publication of a single article, a central service for research data policies has to factor in more elements when considering compliance: what data to publish, how to publish the data, where to publish, when to publish, how the data should be reviewed and how long the data has to be available. It is not hard to see why researchers and support staff trying to navigate this system are

looking for simple solutions. While these do not appear likely at this stage, lessons from the implementation of OA can be applied with particular reference to the use of standard schemas and vocabularies.[18]

At the end of the discovery phase, the project's Expert Advisory Group agreed that the prototype service should not be built at this stage. Instead, the project will continue with a re-focus of objectives towards policy standardization and enabling good practice. This activity will look at policy across the research data ecosystem, extending the scope to funders, data centres and institutions. The consensus from the Project Advisory Group confirmed the need for a stakeholder forum to agree the priorities, build consensus and look at how more data consistency could be achieved. The project team will work on exemplars, checklists and vocabularies which will ultimately feed into policy templates. The provision of case studies to better understand the barriers to data publication were also deemed important. The work will also extend to the domain level to show the differences between disciplines at the data level and demonstrate how data-sharing policy can drive good practice. The project will continue to monitor the possibility of a central service which could reduce the burden on researchers and support staff. As the environment and infrastructure matures, the goal will be to automate both data publication and compliance checking. With respect to the latter, the project can be seen as the first steps towards machine-readable policies.

'the project will continue with a re-focus of objectives towards policy standardization and enabling good practice'

'As the environment and infrastructure matures, the goal will be to automate'

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'Abbreviations and Acronyms' link at the top of the page it directs you to: http://www.uksg.org/publications#aa

Competing interests
The authors have declared no competing interests.

References

1. Journal of Research Data policy bank (JoRD): https://jordproject.wordpress.com/ (accessed 4 January 2016).
2. JoRD blog literature review: https://jordproject.wordpress.com/2012/09/27/literature-review-articles-relevant-to-the-field/ (accessed 4 January 2016).
3. Sturges, P, Bamkin, M, Anders, J and Hussain, A, Access to Research Data: Addressing the Problem through Journal Data Sharing Policies, 2014, *IATUL Conference Proceedings*.
4. Sturges, P, Bamkin, M, Anders, J, Hubbard, B, Hussain, A and Heeley, M, Research data sharing: Developing a stakeholder-driven model for journal policies. *Journal of the Association for Information Science and Technology,* 2015; DOI: http://dx.doi.org/10.1002/asi.23336 (accessed 4 January 2016).
5. Tsoukala, V, Angelaki, M, Kalaitzi, V and Wessels, B, Policy guidelines for open access and data dissemination and preservation, *RECODE Deliverable D5.1, Seventh Framework Programme,* 2015, p. 57 URN: urn:nbn:se:bth-6397 (accessed 26 January 2016).
6. Gallagher, J, Orcutt, J, Simpson, P, Wright, D, Pearlman, J and Raymond, L, Facilitating open exchange of data and information, *Earth Science Informatics,* 2015, 8(4), 721–739; DOI: http://dx.doi.org/10.1007/s12145-014-0202-2 (accessed 4 January 2016).
7. Lin, J and Strasser, C, Recommendations for the role of publishers in access to data. *PLOS Biology,* 2014, 12(10); DOI: http://dx.doi.org/10.1371/journal.pbio.1001975 (accessed 4 January 2016).
8. Piwowar, H and Chapman, W, 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) 2008. Toronto, Canada.
9. Vines, T H et al., Mandated data archiving greatly improves access to research data, *FASEB Journal*, 2013, fj.12–218164; DOI: http://dx.doi.org/10.1096/fj.12-218164 (accessed 4 January 2016).
10. Picarra M, Angelaki M, Dogan G, Guy M and Artusio C, Aligning European OA policies with the Horizon 2020 OA policy, *Insights*, 2015, 28(3), 32–43; DOI: http://dx.doi.org/10.1629/uksg.252 (accessed 4 January 2016).
11. Jisc Co-Design: https://www.jisc.ac.uk/rd/how-we-innovate/co-design (accessed 4 January 2016)
12. Jisc Research at Risk: https://www.jisc.ac.uk/rd/projects/research-at-risk (accessed 4 January 2016).
13. Spotlight Data project report: http://repository.jisc.ac.uk/6264/ (accessed 4 January 2016).
14. NISO, Recommended practices for online supplemental journal article materials, *NISO RP-15-2013,* Baltimore, MD, National Information Standards Organisation: http://www.niso.org/publications/rp/rp-15-2013 (accessed 4 January 2016).

15. Draft Concordat on Open Research Data:
http://www.rcuk.ac.uk/research/opendata/ (accessed 5 January 2016).

16. Centre for Open Science, Transparency and Openness Promotion (TOP) Guidelines:
https://cos.io/top/ (accessed 5 January 2016).

17. Spotlight Data project report, ref. 13.

18. Picarra, M et al., ref. 10.

Corresponding author: Dr Linda Naughton

Head of Research
Jisc, UK
E-mail: Linda.Naughton@jisc.ac.uk

ORCID iD: http://orcid.org/0000-0001-5458-3238

Co-author: David Kernohan
Senior Co-Design Manager
Jisc, UK

ORCID iD: http://orcid.org/0000-0003-1464-0714

To cite this article:
Naughton, L and Kernohan, D, Making sense of journal research data policies, *Insights*, 2016, 29(1), 84–89;
DOI: http://dx.doi.org/10.1629/uksg.284

Published by UKSG in association with Ubiquity Press on 07 March 2016