IJDC | *General Article*

Disciplinary Data Publication Guides

Zosia Beckles
University of Bristol

Stephen Gray
University of Bristol

Debra Hiom
University of Bristol

Kirsty Merrett
University of Bristol

Kellie Snow
University of Bristol

Damian Steer
University of Bristol

Abstract

Many academic disciplines have very comprehensive standard for data publication and clear guidance from funding bodies and academic publishers. In other cases, whilst much good-quality general guidance exists, there is a lack of information available to researchers to help them decide which specific data elements should be shared. This is a particular issue for disciplines with very varied data types, such as engineering, and presents an unnecessary barrier to researchers wishing to meet funder expectations on data sharing.

This article outlines a project to provide simple, visual, discipline-specific guidance on data publication, undertaken at the University of Bristol at the request of the Faculty of Engineering.

Received 21 January 2018 ~ *Accepted* 20 February 2018

Correspondence should be addressed to Zosia Beckles, Library Services, University of Bristol, Augustine's Courtyard, Orchard Lane, Bristol, BS1 5DS, United Kingdom. Email: z.beckles@bristol.ac.uk

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

There is a lack of clarity around the definitions of 'supporting' or 'underpinning' data in certain disciplines, which can lead to confusion for researchers trying to determine what data they are required to share to meet funder and publisher data sharing policies. For example, the Engineering and Physical Sciences Research Council (EPSRC) clarification of expectations on research data[1] requires EPSRC-funded papers to include a statement explaining how supporting data may be accessed, but provides no examples, or detailed definition, of this term. In contrast, the Medical Research Council (MRC) policy for sharing data from population and patient studies[2] includes examples of the types of data that are covered by the policy, and their good practice principles for sharing individual patient data[3] (IPD) has an extensive definition of the elements that might be captured within this data type.

In fields where funder policies lack clarity but data publication standards are well established, such as astronomy (Borgman, 2015), or emerging, such as palaeobiology (Davies et al., 2017), a tacit definition of supporting data may be widely understood. However, this is often not the case in disciplines where data types depart from what might traditionally be thought of as data, for example fields using computational models and model outputs, or those where supporting data types vary significantly between related publications. The latter case occurs frequently in engineering, where a single study may produce papers using both traditional and non-traditional data, for example data logger capture from a physical experiment, and numerical models relating to these experiments.

Journal data publication polices often provide very detailed guidance on what is expected in terms of supporting data, including file formats, documentation standards, and recommended repositories for deposit.[4] As might be expected, these policies are typically in line with standards for the relevant discipline and therefore in cases where no standards have been established, journal data policies are very general.

Organizations such as the Digital Curation Centre (DCC) provide guidance on how to appraise and select data,[5] but this information is again general in scope. Similarly, the OpenAIRE guidelines on open access to scientific publications and research data in Horizon 2020[6] define 'underlying data' as "the data needed to validate the results presented in scientific publications," but does not provide any examples or more specific definitions. This is understandable given the wide range of research funded by Horizon 2020, but can leave researchers unsure whether they are meeting the requirements of their funders and publishers.

1 EPSRC Clarifications of Expectations on Research Data Management: https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/
2 MRC Policy and Guidance on Sharing of Research Data from Population and Patient Studies: https://www.mrc.ac.uk/publications/browse/mrc-policy-and-guidance-on-sharing-of-research-data-from-population-and-patient-studies/
3 Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials: http://www.methodologyhubs.mrc.ac.uk/files/7114/3682/3831/Datasharingguidance2015.pdf
4 For example, see Astronomy and Astrophysics Data Policy: https://www.aanda.org/author-information#Data
5 See: http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep
6 Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020: https://www.openaire.eu/guidelines-on-open-access-to-scientific-publications-and-research-data-in-horizon-2020

 doi:10.2218/ijdc.v13i1.603

Data management plan (DMP) guidance such as that provided by the DCC[7] is often a useful source for examples of the types of data intended to be collected for different projects. However, DMP examples are necessarily based on information available at the start of a project and thus do not go into the detail of what will be published at the end. Typically, the description of data to be published is limited to data which 'underpin or contribute to' patent applications and publications.[8] The recently-released Science Europe framework for domain data protocols (Science Europe, 2018) may generate more specific guidance around data publication expectations, but this depends on the levels of participation from research communities; the minimum requirements for domain data protocols intentionally do not go into detail with regards to data types and publication mechanisms to allow customisation within disciplines.

Finally, there are texts that provide support and guidance to researchers with regards to research data management and which provide a good level of detail with regards to data elements that should be shared: "share any data that supports the publication, unless applicable policies say otherwise. This means everything from data used in tables, data turned into figures, images that you performed analysis upon, etc. If someone will need a piece of data to reproduce your results, plan to share it" (Briney, 2015). The challenge then becomes presenting this information to the busy academic: a direction to consult a textbook is unlikely to meet with much success.

As a result of the lack of accessible guidance, the Research Data Service (RDS) at the University of Bristol received a number of requests from researchers in the Faculties of Engineering and Science for more specific information on which data they were required to publish in support of publications. There was some doubt as to whether we as support staff were best placed to provide such specific guidance – there is a strong argument that the 'what should be published' and 'what are data' questions are so fundamental and specific to each discipline that they should really be tackled by the academics, academic publishers and research organizations in question. However, the requests for guidance we received indicated that researchers did not feel sufficiently confident to answer them, and as has already been discussed, sufficiently detailed support was lacking from publishers and research organizations. We therefore felt that it was appropriate to make a first attempt at meeting this need, and decided to carry out a project to produce guidance that would address the needs of these researchers.

Aim

This project aimed to provide brief practical guides, preferably visual or with a strong visual element, on the minimum data elements that should be shared to support different types of paper, and how this sharing should be achieved. They should be produced in close consultation with academics to ensure that the information contained is accurate and detailed without being overly prescriptive. The guides were initially planned to address data and paper types found in the Faculty of Engineering, with a view to expanding this to the Faculty of Science at a later date if successful.

7 For example, see: http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

8 For example, see the University of Glasgow Synthetic Chemistry / EPSRC example DMP: https://www.gla.ac.uk/media/media_418166_en.pdf

Methods

Some information on the types of data used in the Faculty of Engineering was already available prior to the start of this project. The Research Data Service has had a data management review programme in place since 2016, which collects information on the data types, documentation and publication strategies in use in research groups throughout the University of Bristol. Participation by research groups is entirely voluntary, and has so far focussed largely on groups in science and engineering. Each review involves a one-hour structured interview with a member of the research group, usually the head or data manager, following a standard template, available for reference in the supporting data (Beckles, 2018). The programme is intended to identify and promote existing good data management practice within the university, to help match services to user need and to collect in a single location the information required to complete future data management plans from that research group.

In addition, two case studies[9] examining similar issues for large engineering projects had also recently been completed. These case studies had been a first attempt at meeting the need for advice on data publication, but feedback from academics within the Faculty of Engineering indicated that the case study format was too long, too difficult to quickly extract relevant information from, and therefore not suited to this type of guidance. Despite this, much of the information collected within the case studies was valuable and was able to be re-used for the present data publication guidance project.

Finally, a series of in-depth discussions were held with researchers and project and data managers from the Faculty of Engineering; these were free-flowing, unstructured conversations where we sought their feedback on the information already collected and thoughts on the nature and format of any new guidance. These sources formed the starting point for the construction of a simple ontology describing common data types and their relationships to types of paper in engineering.

Four types of paper, six data classes, 17 types of data and four publication mechanisms were initially identified:

Paper Type

1. Modelling
2. Experimental (physical)
3. Review (including systematic review)
4. Theory

Data Class

1. Model set-up information
2. Experiment set-up information
3. Review set-up information

9 Data publishing case studies: https://goo.gl/SyD1L5 and https://goo.gl/q1CnVw

 doi:10.2218/ijdc.v13i1.603

4. Software or code
5. Results
6. Physical samples

Data Type

1. Third party software
2. Code/software supporting workflow
3. Code/software integral to study
4. Model input or input conditions
5. Description of model behaviour
6. Complete model output
7. Representative sample of model output
8. Physical sample itself
9. Description of physical sample preparation/capture method
10. Experiment protocol
11. Complete raw experimental data
12. Representative sample of raw experimental data
13. Complete processed data
14. Representative sample of processed data
15. Review protocol
16. Summary statistics
17. Derived data

Publication Mechanism

1. In data repository
2. In supplementary information
3. In paper
4. Do not publish

The ontology shown in Figure 1 was established, describing the relationship between paper types, data classes, data types, and publication mechanisms. Initially, we intended that the guidance would address the preferred publication method solely for broad data classes within these paper types, for example the often-overlooked category of data relating to experiment or model set-up, data relating to software or code, and data relating to results or outputs. However, each data class contained several types of data with different recommended publication mechanisms, and providing several publication options for a single data class would not meet researchers' requirements for clear and specific guidance.

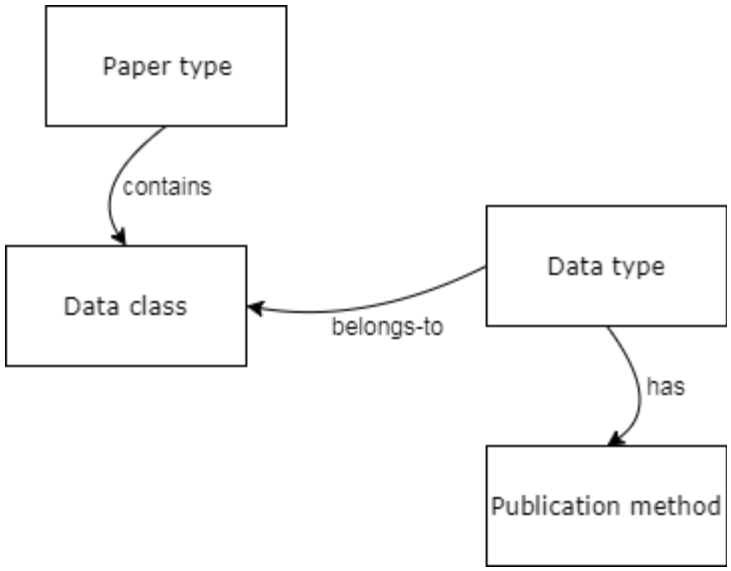

Figure 1. Ontology diagram for data in the Faculty of Engineering.

From this ontology a more granular decision tree was developed, incorporating recommended publication mechanisms for individual data types within each data class for each paper type. There was some discussion as to whether the purpose of data sharing should be included in the ontology – for example, data required for validation or verification of findings (repeating the same analysis on the same data) might be different to data required to replicate a study (repeating the same analysis on different data), which might be different again from a dataset intended for re-use for purposes unrelated to the original study. We decided that for the first instance of the guidance it would be sufficient to further divide the publication mechanisms into 'publish' or 'consider publishing' categories in order to distinguish those elements considered essential as supporting data, and those that would be useful but not critical, and to leave further consideration of purpose to a later iteration.

Initially all four paper types were captured in the same diagram, which is too large to reproduce here but is available in the supporting data (Beckles, 2018), and was circulated for comment to school research directors, heads of research groups, and project managers of key engineering projects. Following feedback from these stakeholders, the theoretical paper type was removed in order to simplify the guidance, as there was considerable overlap with other paper types with regards to data types and recommended publication mechanisms.

 doi:10.2218/ijdc.v13i1.603

For simplicity, certain restrictions to data sharing had been omitted from the diagram. Feedback indicated that restricting sharing of commercially sensitive data was a key concern to engineers due to the frequency of research partnerships with commercial organizations. Guidance on this was added in as a stop-notice alongside the decision tree; an attempt to capture data sensitivity in the ontology and decision tree led to hugely complex diagrams which, although highly specific, did not meet the requirements for clear guidance. At this point the decision trees for the three remaining paper types were also split into separate diagrams for simplicity as shown in Figures 2-4.

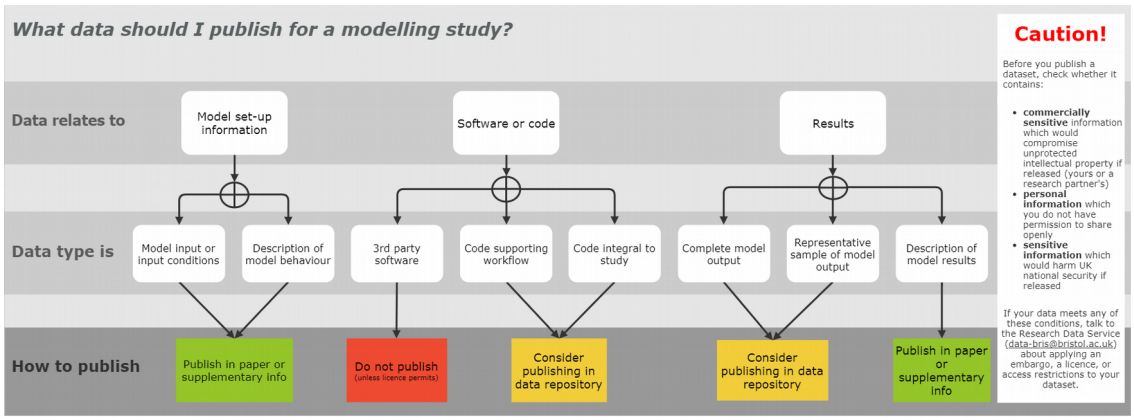

Figure 2. Modelling study data publication decision tree.

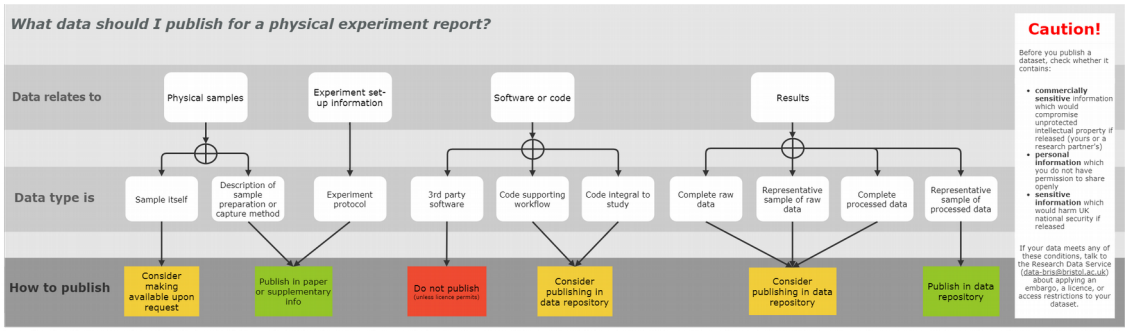

Figure 3. Physical experiment data publication decision tree.

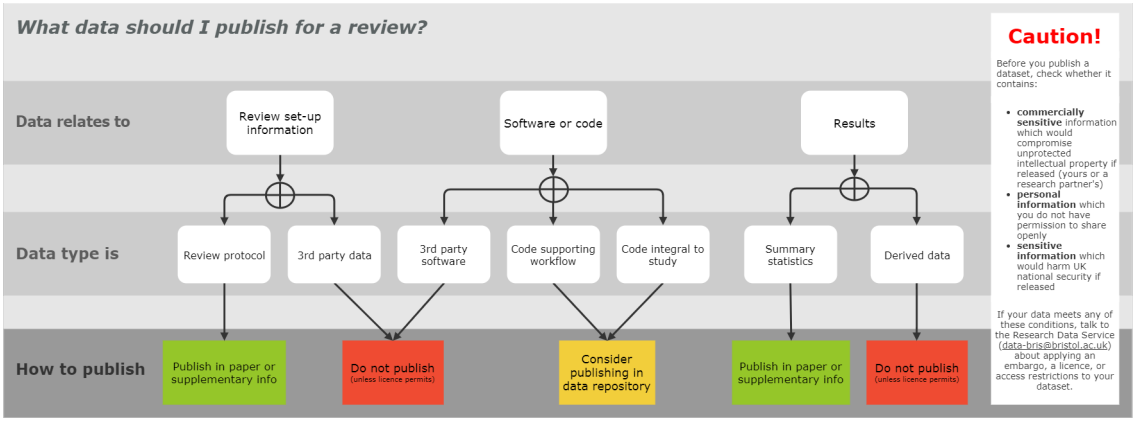

Figure 4. Review data publication decision tree.

A further round of feedback was sought from the previous stakeholders and from the broader research data management community via the Jiscmail Research Data Management discussion list. Feedback from the discussion list was particularly helpful and several members suggested refinements which, due to time constraints, could not be included in the current version of the guidance but have been planned for future iterations.

There was concern from researchers that certain nuances of data sharing, such as restrictions to protect IP, were still being overlooked. However, the prior attempt at integrating data sensitivity had illustrated the difficulty of combining brevity and completeness into a single document, so rather than publication as stand-alone items, the diagrams were embedded into a webpage. This page included an FAQ and contextual information that stressing the point that the information provided was intended as a guide rather than a set of hard-and-fast rules. The completed disciplinary data publication guides and contextual information were posted on Faculty of Engineering intranet on 31st July 2017. Copies of the text and diagrams are available in the supporting data (Beckles, 2018).

Results

The Research Data Service monitors University of Bristol open access publications to ensure that they include funder-compliant data access statements (Beckles et al., 2017). We were able to interrogate these data together with information on use of the University's data repository, data.bris,[10] to investigate the impact of the new guidance, using rates of data publication and funder-compliant statements as a proxy for guidance uptake. The extent of any effect is likely to be muted due to the relatively short time since the release of the guidance and other ongoing research data management training initiatives, but Figure 5 shows that compared to the same period in 2016, the number of compliant data access statements in open access publications was higher in 2017 than 2016.

10 data.bris: https://data.bris.ac.uk/data/

 doi:10.2218/ijdc.v13i1.603

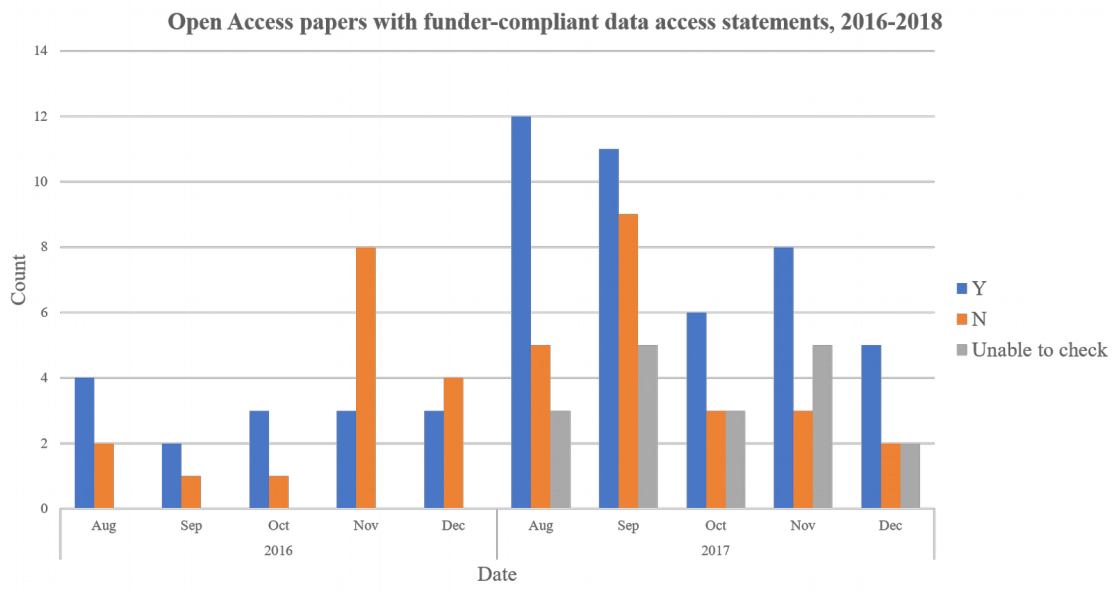

Figure 5. Faculty of Engineering open access papers with funder-compliant data access statements.

The period August-December has been used since it has not been a full year since the guidance was published, and there are seasonal variations in numbers of papers published meaning it would not be valid to compare the effect across different months. Similarly, Figure 6 shows that the number of datasets published by the Faculty of Engineering has also seen a steady increase from August-December 2016-2017. As noted previously, this might also be due to other data management training we offer to improve awareness of funder and publisher requirements around data sharing, but it is encouraging nevertheless.

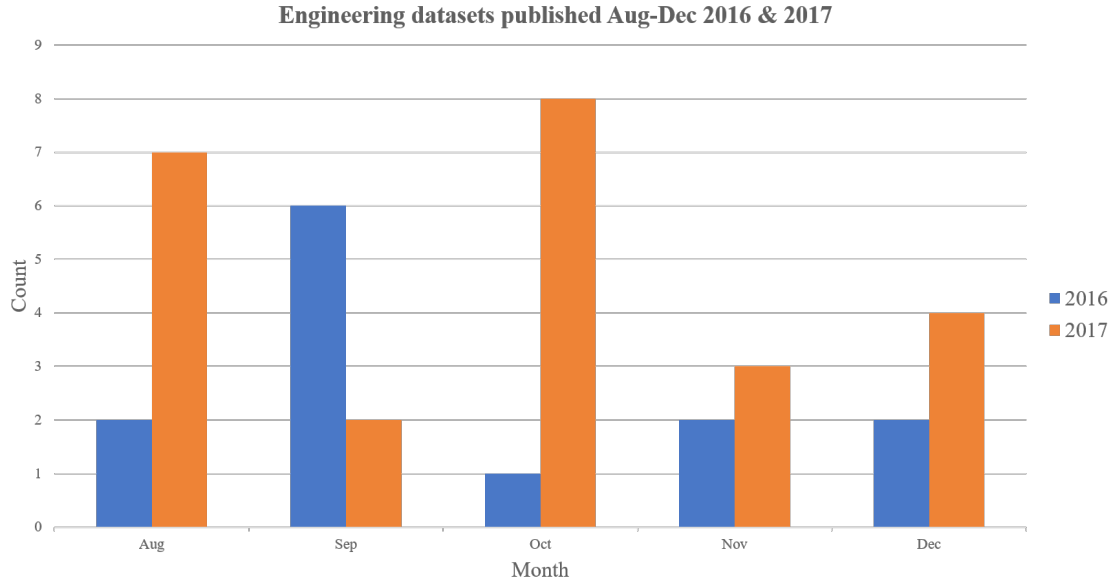

Figure 6. Faculty of Engineering data publications.

Feedback was sought from researchers in the Faculty of Engineering, including principal investigators and senior academics; their comments were generally very positive and noted that the guidance was very useful as a reference. The Research Data Service has not received further requests for this type of information from the Faculty of Engineering since the publication of this guidance. The diagrams have also been adapted by other institutions, for example the University of Utrecht, where they have been welcomed as a 'practical overview of what is shared where' (Pronk, 2017).

Conclusions

Disciplinary data and metadata standards are key for embedding understanding of what constitutes data in general, and 'supporting' data in particular. The approach described above worked well for the Faculty of Engineering, and seems likely to be easily extended to other scientific disciplines where data types are generally clearly defined and common across sub-disciplines; it remains to be seen whether it can also work for disciplines in the arts and humanities where there is even greater variation in data types, and indeed greater uncertainty over what even constitutes data.

In the absence of robust standards developed and accepted by the research community in question, it is important to engage with researchers to provide interim guidance they understand and will subscribe to. It should also be noted that the process of creating the guidance for a single faculty took around seven months, and involved input from multiple stakeholders both internal and external to the University of Bristol. In addition, we drew upon several pieces of pre-existing work (case studies and the information collected within the data management review programme), without which it might have taken much longer. It is therefore important to identify the audience for such guidance clearly in advance, and to ensure that there is an accepted need for it from academics; it would have been much more difficult to get researchers to engage with the process of creating the guidance had the prompt to create it not come from them in the first place.

Next Steps

The Research Data Service will work with disciplines within the Faculty of Science lacking established data standards to provide detailed guidance on data publication. Further iterations of the guidance may include information on additional publication mechanisms ,such as data papers, and may address the data elements needed to support different end uses of shared data (e.g. study validation versus broader re-use). We are also investigating ways to produce versions of the diagrams that are accessible to screen readers.

Acknowledgements

The authors would like to thank the members of the Jiscmail Research Data Management discussion list for their feedback. We are also grateful to the reviewers for their constructive input.

References

Beckles, Z., Gray, S., Hiom, D., Merrett, K., Snow, K., & Steer, D. (2017). Making a statement: Funder-compliant data access statements and support offered at the University of Bristol. *SCONUL Focus, 69*. Retrieved from https://www.sconul.ac.uk/sites/default/files/documents/30.%20MAKING%20A%20STATEMENT_0.pdf

Borgman, C.L. (2015). *Big data, little data, no data: Scholarship in the networked world*. London: The MIT Press.

Briney, K. (2015). *Data management for researchers: organize, maintain and share your data for research success*. Exeter: Pelagic Publishing.

Davies, T.G., Rahman, I.A., Lautenschlager, S., Cunningham, J.A., Asher, R.J., Barrett, P.M., . . . Donoghue, P.C.J. (2017). Open data and digital morphology. *Proceedings of the Royal Society B-Biological Sciences, 284*(1852). doi:10.1098/rspb.2017.0194

Pronk, T. (2017, 04/10/2017). [Personal communication].

Science Europe. (2018). *Science Europe guidance document presenting a framework for discipline-specific research data management*. Retrieved from http://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf