Springer Nature response to NIH RFI on Strategies for NIH Data Management, Sharing, and Citation

This is a reformatted version of Springer Nature's response to the NIH RFI on Strategies for NIH Data Management, Sharing, and Citation submitted online at http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation on 18th January 2017. The order of the text reflects the order of the response form fields.

Submitted by

Iain Hrynaszkiewicz, Head of Data Publishing, Open Research Group, Springer Nature

Responses to 'SECTION I. Data Sharing Strategy Development'

Highest-priority types of data to be shared

As a publisher of research, primarily via scholarly journals and books, data that support peer-reviewed publications are important to share, to enable reuse of research by the research community and independent replication and verification of results. For more specific types of research data and disciplines, priority might be given to research data that are difficult to generate and hard to recreate, such as data from human research subjects and rare or vulnerable species. Data important to public health, such as in response to a global epidemic, are also important to share rapidly. Similarly, data relevant to public policy and/or with high social impact might be viewed as high priority. Publishers have responded to public health emergencies by making research articles freely available. However publishers - which provide services for the research community - are generally not well placed to set the priorities for the types of data that should be shared but should support the needs of the research community in their policies and services.

Length of time these data should be made available; means for maintaining and sustaining such data; and long-term resource implications

The length of time research data remains useful to the research community is best determined by the research community and its funding agencies, and these expectations should be supported by publisher policies and services. Several UK Research Councils' data policies stipulate that "research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party" (e.g. https://www.epsrc.ac.uk/about/standards/researchdata/expectations).

While we can publish small datasets (up to 10-20Mb) as electronic supplementary materials in journals, a requirement of all Springer Nature journal data policies is the preference for archiving of research data in repositories. Public datasets supporting publications should be preserved indefinitely (whether in repositories or journals) to maintain the integrity of the

published record. Community specific data repositories are preferred and general/institutional repositories can also be used. Springer Nature manages a list of more than 80 recommended repositories across all research domains (http://www.springernature.com/gp/group/data-policy/repositories) including health sciences (http://www.springernature.com/gp/group/data-policy/repositories#c10106444). Our criteria for approving repositories (http://www.nature.com/uploads/ckeditor/attachments/3243/SciData_respository_evaluation_Aug2016.docx) include ensuring long-term preservation of datasets. In practice this means sustainability plans and data preservation for a minimum of 10 years.

Barriers to data sharing and mechanisms to overcome them

Barriers to data sharing reported by researchers include:

I. Copyright and licensing
II. Data standards
III. Uncertainty about compliance with funder policy
IV. Lack of time
V. Perceived lack of credit
VI. Lack of a data repository
VII. Uncertainty about covering costs
VIII. Concerns over inappropriate data reuse
IX. Protecting human research participant privacy

Mechanisms to overcome them (taking each of the above stated barriers in turn):

I. Data repositories and publishers with clear terms of use for data.
II. Promote community standards (e.g. https://biosharing.org/standards/) and provide for researcher training
III. Funder policy compliance advisory services (e.g. Springer Nature's Research Data Support helpdesk (http://www.springernature.com/gp/group/data-policy/helpdesk)
IV. Appoint or promote involvement of research data management experts in research, and recommend services/products for researchers to provide data management/sharing services, including those which might already be available commercially
V. Encourage formal data citation and quality/time stamping e.g. badges for open practices
VI. Promote data repositories including generalist repositories (http://blogs.nature.com/scientificdata/2016/11/14/expanding-our-generalist-data-repository-options/)
VII. Provide a proportion of grant funding for research data management and sharing (e.g. 5%)
VIII. Promote scholarly norms of citation and provenance/evidence when research data are reused (https://www.biomedcentral.com/about/policies/open-data)
IX. For sensitive data, establish data use agreements (DUAs) in partnership with specialist repositories (see http://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-016-0015-6 for a list of these repositories) and provide resources for anonymisation of datasets for sharing; and modify participant consent

procedures accordingly (https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-11-9).

Other topics for NIH to consider

The provision of explicit funding for research data sharing and management and the recognition of data curation as a vital skill and resource for research projects and researchers are important. These issues are recognised in the UK Concordat on Open Research Data (July 2016 http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/) and European Commission Horizon 2020 funding programme's data management guidelines (http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf). Researchers need support (financial and education) to meet the expectations of data management policies and where appropriate guidance on tools, services and publication venues and formats to aid compliance.

Researchers should, also, be encouraged to share and describe datasets in a way that facilitates reuse and reproducibility. Perhaps NIH could support creation of discipline specific guidance/standards on this.

Finally, harmonisation and standardisation of research data policy between publishers and funders and other stakeholders (repositories and institutions) is important. Springer Nature has defined 4 standard research data policy types which share principles of data citation, use of repositories and providing support to researchers (http://www.springernature.com/gp/group/data-policy). More than 600 journals (http://www.springernature.com/gb/group/media/press-releases/over-600-springer-nature-journals-commit-to-new-data-sharing-policies/11111248) have implemented one of the policies to date, including all Nature journals and BioMed Central journals. Springer Nature is leading an initiative via the Research Data Alliance (RDA) to engage funders and others on policy standardisation (https://rd-alliance.org/groups/data-policy-standardisation-and-implementation).

Responses to ‘SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications’

Potential impacts of increased reporting of data and software sharing in RPPRs and grant applications

Firstly, a likely impact will be creating cultural change amongst researchers and research assessment procedures to increase the recognition of data and software as legitimate - shareable and citable - scholarly outputs. The NSF in 2014 began asking for research “products” rather than papers in principal investigators’ bibliographic sketches (http://www.nature.com/nature/journal/v493/n7431/full/493159a.html). Secondly, reporting of data and software - particularly if done via persistent identification mechanisms - will help improve reproducibility and provenance tracking of research outputs. It will, also, help in the assigning of credit for researchers’ outputs that are not papers/publications.
Thirdly, reuse and assessment of data and software might be encouraged, improving scholarly discourse and return on investment in research, which might help also increase the reliability of scientific research funded by the NIH.

Finally, funding agencies encouraging or requiring the citation of data and software in RPPRs (and grant applications) might help make this practice more widespread in other scholarly literature.

Important features of technical guidance for data and software citation

For many researchers citation of research data and, even more so, research software is a new concept. It is important, therefore, that data citation with DOIs is communicated as a simple and easily achievable practice that is supported by publishers. Citing datasets and software can be done in much the same way as citing papers, provided the necessary information (metadata) about the data and software are available. Supporting these basic principles of data citation are mandatory parts of all Springer Nature’s standard research data policies: “Datasets that are assigned digital object identifiers (DOIs) by a data repository may be cited in the reference list. Data citations should include the minimum information recommended by DataCite: authors, title, publisher (repository name), identifier.” (http://www.springernature.com/gp/group/data-policy/policy-types#c10305772)

There are other considerations in implementing effective data citation, from the more technical perspective of publishers and repositories and persistent identifier issuing bodies. A working group to define a roadmap for implementing data citation (co-chaired by Springer Nature https://www.force11.org/group/dcip/eg3publisherearlyadopters) is working to enable consistent implementation of data citation by publishers.

For researchers who are more advanced with the practice of data citation there are standards emerging for the citation of dynamic objects and citation of data with additional metadata that infers more meaning to citations (e.g. data generated by a research study or

data referenced/analysed by a study).

Inclusion of links to data/software with citations in RRPRs

Availability of research data and software are increasingly reported in research papers in a designated section, often called a “Data availability statement”, “Code availability” statement or “Availability of data and materials” statement. This approach has several benefits:

- Makes links to research data and software easier to find in publications
- Supports the requirements of some funder policies, such as the UK Research Councils, which often require data availability statements in publications
- They complement data citation in reference lists by providing a narrative to describe availability of data and software

Springer Nature’s research data policies support and provide detailed guidance - and examples - for writing data availability statements. For more information see http://www.nature.com/authors/policies/data/data-availability-statements-FAQs.pdf and http://www.springernature.com/gp/group/data-policy/data-availability-statements

Many Springer Nature journals including the Nature and BioMed Central journals have policies on the sharing and reporting of software/code (e.g. http://www.nature.com/news/code-share-1.16232).

Funding and research organisations and publishers should collaborate in the development of standards and policies for data/software citation, via specific fora within the Research Data Alliance (https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation) and Force 11 (https://www.force11.org/group/dcip/eg3publisherearlyadopters) communities, for example.

Publishing technology can also support bidirectional linking between publications and repositories, such as embedding data viewers in publications, and initiatives such as Scholix (http://www.scholix.org/) to make data-article link exchange more widespread and interoperable.

Identification of the authors of data/software

The policy of Springer Nature’s BioMed Central journals and of the journal *Scientific Data* do not just encourage (or, at some of these journals, require) the sharing of software but furthermore encourage the deposition of software in repositories that can assign persistent identifiers (DOIs) to software to enable citation and provenance tracking e.g. *Scientific Data* policy (http://www.nature.com/sdata/policies/editorial-and-publishing-policies#code-avail): “authors are encouraged to archive their code in a public repository that can assign it a DOI, such as figshare...we recommend using an open control version system (CVS), such as GitHub, in combination with a DOI providing repository...Code with an assigned DOI may be formally cited and listed in the References section of the manuscript.” BioMed Central policy (http://www.biomedcentral.com/getpublished/editorial-policies#availability+of+data+and+materials): “include a link to the most recent version of your software or code (e.g. GitHub or Sourceforge) as well as a link to the archived version referenced in the manuscript. The software or code should be archived in an appropriate

repository with a DOI or other unique identifier. For software in GitHub, we recommend using Zenodo."

By encouraging sharing and citation of software via repositories that provide DOIs, as a consequence authors must be identified as part of deposition.

Granularity of data citations

In our experience as an early adopter of data citation principles and practices, simplicity is very important in communicating data citation expectations. The journal *Scientific Data* recently shared guidance on data citation. A basic principle of data citation is to "Cite what you used" (http://blogs.nature.com/scientificdata/2016/07/14/data-citations-at-scientific-data/). If a researcher/author used associated datasets, especially data archived outside of a journal article and its supplementary material, then they should cite the data. Often it will be appropriate to cite both: the paper and any datasets used.

The focus of Springer Nature's research data policies is on citing datasets that are assigned DOIs in reference lists (e.g. Nature's policy: www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf), but other types of dataset identifier can also be cited - either as in-text references or links or as formal citations. Very data focused journals, such as *Scientific Data*, that support the most stringent of Springer Nature's data policies, require any stably archived datasets mentioned in a publication to be formally cited with its persistent identifier, regardless of identifier type.

In terms of what data should be reported and cited, this will vary by research community. The research data policies of Springer Nature in general concern the minimal dataset that supports the central findings of a published study (http://www.springernature.com/gp/group/data-policy/faq).

Unambiguously identifying and citing digital repositories for data/software

As described previously, following the standards of DataCite and including authors, title, publisher (repository name), identifier in citations of/references to data and software is important. It is also important that those citing data and software resources do not "invent" metadata. If for example "authors" or "title" for data or software are not clear from the information available from the source/repository, they should not be included. In our experience titles of data and software can be more difficult to report uniformly (http://blogs.nature.com/scientificdata/2016/07/14/data-citations-at-scientific-data/).

Additional routes by which NIH might strengthen and incentivize data and software sharing

Possibilities include:

- Encouraging or requiring the reporting and citation of data and software in scholarly publications (journal articles, monographs, preprints etc.) as well as RPPRs
- Require provision of consistent, standardised data and software availability/accessibility statements in RPPRs and publications

- Monitor and encourage inclusion of information on data and software reuse in RPPRs:
    - This could take the form of metrics such as “pull requests” for software in Github and citations and downloads of datasets, where this information is available from repositories.
    - This could also be achieved anecdotally, by researchers providing case studies and examples of data reuse, such as the number of requests to share data they received.
- Commit to working with publishers and other stakeholders to share information on data-article links and discuss policy standardisation for example via the Scholix framework (https://blogs.openaire.eu/?p=1589) and Research Data Alliance (https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation), respectively
- Encouraging researchers to, where appropriate, publish data papers and software papers in journals to promote reuse of data and software and submit their data and software for consideration by peer reviewers of traditional papers
- Encourage NIH funded repositories that provide accession IDs, to standardize data citation guidance for researchers in collaboration with publishers. http://identifiers.org/ may be a useful resource for this.

Acknowledgements

For comments on the first draft of this response thanks to: Grace Baynes, Marketing and Development Director, Open Research Group, Springer Nature; Varsha Khodiyar, Data Curation Editor, *Scientific Data*; Sowmya Swaminathan, Head of Editorial Policy, Nature Research Group, Springer Nature. We are happy to be contacted for more information on this submitted response at researchdata@springernature.com or iain.hrynaszkiewicz@nature.com