Addressing the Gaps: Recommendations for Supporting the Long Tail of Research Data

August 2017

WRITTEN BY Wolfram Horstmann, Amy Nurnberger, Kathleen Shearer, Malcolm Wolski

REVIEWED BY Kirsten Elger, Francoise Genova, Varsha Khodiyar, Tim Smith

Other Acknowledgements: Thanks to the various contributions of members *The Long Tail of Research Data Interest Group* of the Research Data Alliance

Major societal challenges such as health, climate change, energy, food availability, migration and peace depend on the contributions of a distributed and diverse international network of researchers and subject experts. The aim of open science is to improve the accessibility of research outputs, including articles, data and other research objects, so that researchers, industry and the public can make use of, build on, and ensure the validity of these research outputs..

Among research outputs, research data are often the most diverse - as diverse as the international network of experts that perform research. Datasets may be small or large, simple or complex, structured or unstructured. Data may stem from hundreds of different subjects, may be produced by numerous methodologies, and exist in a plethora of different formats. The diversity of data is also characterized by a variety of data management practices, of varying quality and comprehensiveness. Historically, large structured datasets in well-established disciplines are more likely to adopt unified and standardized formats that are disciplinarily defined and accepted. Similarly well established disciplines tend to have common and understood workflows, where as in the long tail of research it is not unusual for researchers to use a variety of tools and to develop ad-hoc data workflows. Long tail datasets, on the other hand, which vary radically in source, discipline, size, subject, provenance, funding, format, longevity, location and complexity, are less likely to adhere to common standards. The wide distribution and diversity of long-tail data means that ensuring such data is discoverable and stored in appropriate formats with relevant curation and metadata to facilitate reuse is challenging, and that these data have received less attention historically. Furthermore, the terms used to refer to long tail data, e.g. ‘small data’, ‘legacy data’ or ‘orphan data’ have contributed to diminishing the perceived importance of such data.

Considering that a large portion of research datasets (and associated research funding) are found in the long tail, it is paramount that we address the specific and unique data management challenges for this data. The risks of neglecting long-tail data are real and significant. These include both limiting the reproducibility, transparency, and verifiability of research results, and

unnecessary costs associated with the duplication of research data. Moreover, the potential benefits for reuse are significantly reduced.

The Research Data Alliance (RDA) "Long Tail of Research Data Interest Group" has been assessing the situation of long tail data over the last three years, and urges the broader community to consider the risks and opportunities related to long-tail data. This document provides seven recommendations for a variety of stakeholders, including governments, funders, research institutions and researchers to help improve the current approach to managing long tail data. We call on the community to work together to create necessary and sufficient conditions to ensure we are able to properly steward these valuable research outputs for future generations of researchers.

Seven Recommendations for Supporting the Long Tail of Research Data

1. Recognize and understand the diversity of data created at your organization, or through your funding support, and develop appropriate frameworks for managing those data.

Given the varying dimensions of data sets (e.g. by size, subject, provenance, funding, format, longevity, location or complexity of research data), dealing with them is highly context-sensitive. When drafting policies, designing funding programmes, producing data or building technical infrastructure it is paramount to understand the nature of data being produced, along with the inherent opportunities and limitations of the data being generated. The use of data management plans, along with local institutional support for data management will contribute to ensuring that long tail data are managed and shared appropriately.

2. Scale existing funding mechanisms to support research data management for small research projects

Funding for data management is often available for large research activities, but much less so for the data produced through smaller scale research projects. Additionally, some disciplines have subject-specific data-services, but these are not available to less well-established fields. There is a need to allocate funding for data management across all fields and project scales in order to support the management of long tail data.

3. Expand and strengthen the institutional role in managing research data.

Many long tail datasets are at risk of being lost because they are not managed appropriately.[1] Local support for researchers generating data will increase the adoption of standards and best practices earlier on in the research process improving the likelihood that data are preserved, understood, and reused by others. We encourage universities and institutions to offer support

[1] *Dealing with Data: Challenges and Opportunities*. Science 11 Feb 2011: Vol. 331, Issue 6018, pp. 692-693 DOI: 10.1126/science.331.6018.692

services for research data management (RDM). In particular, RDM services should become part of the standard service provision of research libraries, where libraries supply expertise in issues of information management from the initial stages of data management planning, through active data management challenges, to careful consideration of the requirements for longer term data management, such as repositories.

4. Develop and apply common standards across institutions and domains to ensure greater interoperability across datasets.

The integration of disparate datasets offers tremendous potential for new discoveries. A distributed network of research data management services has many advantages including greater support for local needs and requirements, more comprehensive coverage and increased resilience against loss. These advantages, however, come with corresponding challenges around the coherence and integration of research data, one of the major objectives of open science. Many of the current standards for research data are discipline specific, and therefore are not immediately applicable for interoperability and/or integration for the diversity of long tail data. We recommend the development of common, high level metadata elements that will support data integration across diverse types of research data and disciplines.

5. Support reproducibility and transparency of research by linking data, software, and literature.

One of the great opportunities in the digital environment is the improved capacity to use research data and methods to reproduce research findings. Reliably linking the literature to the underlying data and tools, such as software and code (as well as the physical samples that are the sources of data) supporting research conclusions, will make it easier for others to verify claims, whilst also facilitating greater reproducibility of research. We encourage the community to work together to identify best practices for linking research data with related literature and associated tools.

6. Establish governance structures that reflect the diverse dimensions of research data.

In order to ensure the appropriate mechanisms are in place to support long tail data, RDM governance should reflect the diversity of data. We need to ensure that the diversity of long-tail data, both in terms of scope and discipline, are well represented in the evolving RDM governance structures. This can be accomplished by ensuring greater involvement by subject specialists from both novel and well-established disciplines, technology experts, and research data managers from diverse institutions.

7. Develop coherent principles and policies for the collection and preservation of long tail data.

In the context of the long tail, not all data may have value for future use or there may be budget restrictions around collecting and preserving all data. Institutions and funders need guidance to determine good practices for assessing the potential value of research data, and data repositories need to develop policies for the selection, collection, curation, and stewardship of

data and for evaluating which data have long term value. Related to this, there are also need to be better established tools for calculating costs of long-term data stewardship and curation.

In addition to the other stakeholder communities, there are also a number of existing RDA groups that could be mechanisms for moving recommendations forward:

- *Data Citation Working Group*
- *Libraries for Research Data Interest Group*
- *Long Tail of Research Data Interest Group*
- *Metadata Interest Group*
- *RDA/WDS Publishing Data Interest Group*
- *Research Data Collections Working Group*
- *Research Data Repository Interoperability Working Group*
- *Archives and Records Professionals for Research Data IG*