Supporting Long-Tail-Data in Research

WRITTEN BY

Wolfram Horstmann, Amy Nurnberger, Kathleen Shearer, Malcolm Wolski

Major societal challenges such as health, climate change, energy, food, migration and peace depend on a distributed and diverse international network of researchers and specialized experts. Reproducibility and transparency in science is a fundamental underlying principle, and this means that research data must be accessible, managed and preserved over the long term. Research data are extremely diverse - as diverse as the international network of experts that perform research. Datasets can be small or large, simple or complex; they can stem from 100s of different subjects, and can be produced from numerous methods and exist in a plethora of different formats.

For data, as for many other things, there is a ‘head’ and a ‘long tail’. That is, in terms of size of the dataset there are some data resources that contain very high volumes of data, while a significant portion of research data sets are medium to small in size and fall into the so-called long tail‘’[1]. Most of the research being undertaken in the world falls into the long tail category and projects that have large amounts of funding and produce large volumes of data are rather the exception. Furthermore, big data resources usually apply unified and standardized formats, but most of the data created through research varies in size, subject, provenance, funding, format, longevity, location or complexity. Therefore, the ‘Long-Tail-Data’ phenomenon is not reducible to a single X/Y coordinate system but requires careful consideration across multiple dimensions. It is noteworthy that this notion can seem counterintuitive. The long tail is often assumed to be the small minority of data, and therefore has been somewhat neglected in value. The terms used to refer to the long tail data such as ‘small data’ or ‘legacy data’ or ‘orphan data’ help to diminish the importance of this data.

We maintain, however, that the opposite it true and argue that a disproportionate amount of resources are devoted to a rather small portion of research data in the ‘head’; the research data that arises from big projects, with well-funded infrastructure. Considering that a large percentage of research data can be found in the long tail, addressing the challenges resulting from this data is paramount. The wide distribution and diversity make it more challenging for ‘Long-Tail-Data’ to be discoverable and stored at in an appropriate environment where it is well curated and documented with metadata (as opposed to saved on a USB stick or researchers hard drive).

[1] *Science* 11 February 2011: Vol. 331 no. 6018 pp. 692-693 *DOI:* 10.1126/science.331.6018.692

This is not an insignificant matter. The risks of neglecting ‘Long-Tail-Data’ are tangible and significant. Besides limiting the reproducibility and verifiability of research, additional costs can be engendered when research is duplicated. Moreover, the potential benefits and impact of research is greatly limited when it is not available for re-use and integrated with other related research data supporting new discoveries.

Our ability as a community to support the data in the long tail will depend greatly on the types of policies, infrastructure and practices we adopt in the broader scholarly community. The Research Data Alliance Interest Group “The Long Tail of Research Data” has been analyzing the situation of the long tail over the last three years. This document urges governments, research institutions and others in the research community to consider the risks and opportunities related to Long-Tail-Data, and calls on the community to ensure that we create the necessary and sufficient conditions to ensure we are able to steward these valuable research products.

Recommendations for supporting Long-Tail-Data

1. Recognize and understand the diversity of data created in your organization and develop appropriate methods 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 writing policies, designing funding programmes, curating data or building technical infrastructure it is therefore paramount to understand what you are dealing with.

2. Scale existing funding mechanisms to support research data management in small research projects (as well as large ones).

Including data management funding in large research projects can be easier than in small research projects. In addition, large disciplines often have subject-specific data-services while innovative interdisciplinary research in less well-established fields may not have dedicated infrastructure in place. Therefore, in order to curate and preserve data in these research projects, funds are required.

3. Provide support for the Long-Tail-Data at the location of their origin.

Even though data can be managed anywhere in the world, researchers typically have an institutional home and would benefit greatly from local support. Larger research communities often have their own external data services, or data scientists working in their research team, however, this is often not the case for long tail data. Universities and institutions should increase

their capacity for supporting RDM for researchers that have no external support. Libraries, IT services and research offices at the institution are asked to collaborate regionally, nationally and globally in order to build these networks that support diverse data in the long-tail.

4. Expand and strengthen the existing network of data libraries

Libraries have been the guardians of research publications and other research outputs for centuries. In order to support open science and we need a comprehensive network of data libraries that collect and provide access to data. This network of data libraries is especially important in the context of Long-Tail-Data which is often not served by the large data centres. Managing research data should become part of the standard service provision for research libraries around the world.

5. Develop and apply common global standards in order to maximize value of data of distributed datasets.

A distributed network of research data management has many advantages including support for local needs and requirements, greater global capacity for RDM and strengthening resilience against loss. However, it comes with extra challenges around the coherence and integration of research data, a major objective of open science. Standards, such as persistent identifiers, common formats, and metadata with provenance information that allow traceability and citation of research data are needed. The global research community, should adopt more rigid standards for interoperability of research data so they can be harvested, exchanged and interlinked that apply common standards, while also still recognizing the varying needs of different disciplines and data types.

6. Support reproducibility and transparency of research by linking data to literature.

One of the great opportunities in the digital environment is the improved capacity to use research data and methods to reproduce research findings. In the analogue world with printed literature, written descriptions of how to reproduce results were provided. But the tools to actually reproduce results were complicated to put in practice or even missing. Linking the literature to the underlying data that supports conclusions will make it easier for others to verify claims and also facilitate greater reproducibility of research. Data practitioners, system designers and publishers are asked to agree on best practices for linking data and literature.

7. Implement control of technical infrastructure at the lowest level possible.

The virtualization of research infrastructure allows for the support of a distributed and localized approach that is required to enable data diversity and nurturing long tail activities. One of the risks of trying to centralize infrastructure and expertise is that it restricts the ability of researchers to create novel systems and develop innovative and experimental practices. Funders and system designers as well as operators are requested to consider this and develop systems that balance centralized versus distributed approaches so that the benefits of both can be achieved.

8. Establish governance structures reflecting the diverse dimensions of research data.

Based on the global principle of academic freedom, research is managed and developed by means of autonomous and self-organized principles. Research infrastructures should reflect these principles in governance structures such as advisory boards, review panels or consultancy for funding programmes. We need to ensure that data diversity and Long-Tail-Data are well represented in the evolving governance structures for RDM, by ensuring greater involvement by subject specialists from novel disciplines, technologists and emerging fields or managers from small institutions.

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