The Boilerplate Problem in Data Management Plans

Spencer D. C. Keralis, Ph.D.
Assistant Professor, Digital Humanities Librarian
University of Illinois at Urbana Champaign
spencerk@illinois.edu
@hauntologist

Elizabeth Grumbach
Assistant Director, Program Lead, Digital Humanities Initiative
Institute for Humanities Research
Arizona State University
egrumbac@asu.edu
@EMGrumbach

Sarah Potvin
Associate Professor, Digital Scholarship Librarian
Texas A&M University
spotvin@library.tamu.edu
@sp_meta

Published in *ResearchDataQ,* the publication of the Association of College and Research Libraries Digital Scholarship Section (ACRL DSS), October 21, 2019.

Citation

Keralis, Spencer D. C., Elizabeth Grumbach, and Sarah Potvin. “The Boilerplate Problem in Data Management Plans.” October 21, 2019. *ResearchDataQ.* https://researchdataq.org/editorials/the-boilerplate-problem-in-data-management-plans/

Abstract

While reviewing Data Management Plans from successful NEH-ODH, the Linked Open Futures project team (https://linkedopenfutures.net/) discovered significant problems in the way investigators were using the language of linked and open data interchangeably with open access, open source, and other terminology. These errors are exacerbated by boilerplate language describing digital libraries' data repositories, which render the data management plans illegible, incoherent, and virtually meaningless in terms of understanding what data will be preserved, and whether or not it will be publicly accessible. Boilerplate language often describes the digital library as a whole, rather than specifically addressing the preservation and access of research data; in some cases we’ve found the language appears to be simply copy/pasted from the About page of the digital library. Some of this may be related to the use of the DMPTool, which facilitates the use of boilerplate in lieu of compelling investigators to systematically

address questions of data preservation and access in their project proposals, which is arguably the point of data management plan mandates. While our research is focused on digital humanities projects, it is likely that the use of boilerplate in DMPs is widespread across agencies requiring this information as part of grant applications. In this editorial, we will offer recommendations for librarians supporting the development of data management plans to help develop standardized language that is specific to research data, and that does not obscure the investigators' actual plans to preserve and make open their research data.

In 2015, our research team began examining successful grant proposals from the National Endowment for the Humanities Office of Digital Humanities (NEH-ODH) to determine data preservation practices and the development and use of Linked Open Data (LOD) standards in humanities research. In 2019, through a Freedom of Information Act (FOIA) request, we obtained all of the Data Management Plans from funded projects through the 2018 grant cycle and began mining this text data for patterns of data preservation-related language. One surprising result of this analysis was the discovery of the prevalence of boilerplate language describing institutional repositories or digital libraries infrastructure and metadata schemas. Through both close and distant reading methods of qualitative analyses, we discovered some unintended consequences of the use of boilerplate language in Data Management Plans. In what follows we will describe these consequences as we see them, and offer a few recommendations for how librarians and others supporting proposal development by researchers can develop boilerplate that is meaningful.

Data Management Plans (DMPs) are short (usually two pages) documents required by federal funding agencies that describe researchers' plans to retain and share their research data. The NIH has had a data sharing mandate since 2003. The NSF and the

NEH-ODH followed suit in 2011. IMLS includes a questionnaire on the management of digital research products in its applications, but surprisingly does not have an explicit requirement for the retention and sharing of research data (Keralis, et al, 2013). In response to the DMP mandates, the California Digital Library, in collaboration with many other institutions developed the DMPTool to help researchers write DMPs (Meltzer, 2011). The DMPTool actively encourages the development and use of institutional boilerplate language to describe aspects of data management, in particular repository infrastructure.

This is great for grant applicants, and for the librarians and grants officers helping them develop plans. It streamlines the process of writing a bureaucratic document that is, more often than not, perceived as only tangentially related to the intellectual work of the research being proposed for funding. However, as Tomasz Miksa and his collaborators declare, “we need well-defined terms and precise identification of resources” to make DMPs legible to all stakeholders (Miksa et al, 2019). Unfortunately, boilerplate language as it is currently implemented more often than not serves to obscure the actual state of the data being preserved. For example, in our initial analysis of NEH-ODH DMPs, we were searching for evidence of the implementation and use of linked open data technology in digital humanities projects. In many instances, boilerplate language describing library infrastructure resulted in false positives for RDF, LOD, and other terminology. This is important because even if the repository itself implements an RDF schema for repository metadata, the actual data in the repository is not linked open data, and may not be discoverable, linkable, and actionable in the semantic web.

Including boilerplate with this language creates confusion - perhaps even on the part of researchers - about the actual disposition of the data stored in the repository.

Further, boilerplate language exacerbates a profound misrecognition of the labor, infrastructure, and maintenance required to preserve and share research data; the social work of data management, maintenance, and preservation is erased in DMPs that focus exclusively on the technical systems that enable long-term access and persistence. As the volume of research data increases, it becomes increasingly important that the growth in corresponding costs is clear to researchers and administrators. As the term “library” is often used to hide the human labor of *librarians*, repository has come to mask the labor of many, and to render those individuals invisible to stakeholders and decision makers.

While our research has focused on the humanities, these problems are not limited to DMPs from NEH proposals. In a 2015 analysis of DMPs from NSF proposals conducted at the University of Minnesota - Twin Cities Library, researchers found broad inconsistencies in what constituted data sharing, and concluded that scholars need a better understanding of how to make data open (Bischoff et al, 2015). There is now movement at the NSF towards Machine-Readable DMPs that incorporate compliance checks, require updates, and specify roles and actions (NSF 2019), but those standards remain aspirational.

So what can librarians, and others supporting the development of funding applications, do to make DMPs more meaningful? We have a few recommendations.

1. **Name names.** Boilerplate language provided to researchers should clearly identify who in the library provides the labor to support data preservation. If this role is subject to turnover (which may be the case as repository librarians are often early career professionals), use the title. Do whatever is necessary to put a human face on the function of data preservation.
2. **Describe the data.** Ensure that the DMP accurately describes the condition of the data in the repository, not repository infrastructure. If all data in the repository is rendered as .CSV files, this should be clear. Using terminology in boilerplate that describes the repository metadata schema, but does not reflect the actual state of data stored in the repository is not helpful. Unless researchers' data is explicitly formatted as Linked Open Data, terms like LOD and RDF should not appear in the DMP.
3. **Count the cost.** Since data preservation remains an unfunded mandate from federal agencies, researchers are highly unlikely to dedicate grant funds to long term preservation and access, assuming, when they think of this at all, that those expenses will be absorbed in library budgets. To ensure that the actual costs of data preservation are not invisible to administrators or funders, librarians and repository managers should calculate and communicate clearly how much it costs in terms of time, labor, and infrastructure to preserve research data. This could be based on individual file deposits, or on a per-terabyte basis. Include this information in institutional boilerplate. Ask PIs and grants officers to include this number as an in-kind contribution from the library on grant budgets; or ask this

cost to be included in grants. Factor these costs in calculations of institutional matching funds for awards. This is particularly important for institutions whose libraries are funded by student fees - effectively supporting research with student debt. Anything that renders these costs visible to researchers and administrators is important.

We recognize that boilerplate language in DMPs can be very helpful to researchers and the librarians and grant officers working to facilitate research proposals. Likewise the DMPTool has been valuable in simplifying the development of DMPs for thousands of grant applicants. But DMPs should accurately reflect the state of data being preserved, and account for the social work of data management. What we are asking here is that, when we use boilerplate, we should make it mean something. We must be intentional about ensuring that librarians are given credit for their labor, and that we accurately communicate the costs of the labor and infrastructure necessary for sustainable data preservation.

References

Bishoff, Carolyn, and Lisa Johnston. “Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University.” *Journal of Librarianship and Scholarly Communication*, vol. 3, no. 2, 2015. *Crossref*, doi:10.7710/2162-3309.1231.

Keralis, Spencer D. C., Shannon Stark, Martin Halbert, and William E. Moen. “Research Data Management in Policy and Practice: The DataRes Project.” Spencer D. C.

Keralis, editor. R*esearch Data Management: Principles, Practices, Prospects*. Report #160. Council on Library and Information Resources, 2013.

Meltzer, Ellen. “Data Management Tool Launched by University of California Libraries and Partner Institutions.” California Digital Library, November 1, 2011. https://www.cdlib.org/cdlinfo/2011/11/01/data-management-tool-launched-by-university-of-california-libraries-and-partner-institutions/

Miksa, Tomasz, Stephanie Simms, Daniel Mietchen, and Sarah Jones. “Ten principles for machine-actionable data management plans.” PLOS Computational Biology, March 28, 2019.

National Science Foundation. “Dear Colleague Letter: Effective Practices for Data.” NSF 19-069. May 20, 2019. https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp