Data&Society RESEARCH INSTITUTE

Data Management Plan: A Background Report

DRAFT VERSION

by Jacob Metcalf / February 24, 2015

Produced for Council for Big Data, Ethics, and Society[1]

Introduction

At the November 2014 meeting of the Big Data Council, a number of possible trajectories for incorporating a more robust engagement with data ethics into big data basic research agendas and policies were discussed. The group's consensus was that the best leverage point for incorporating data ethics into basic research projects was to be found at the NSF, the likeliest target is a revision of data management plans (DMP's). Follow-up conversations with staff at the NSF indicated that preliminary discussions about reforming DMP's were already under-way, indicating that the Council indeed had an opportunity to effect a meaningful policy change.

This report addresses some background information about DMP's that may be useful as the Council proceeds and looks for more specific opportunities to reform DMP's.

History & Context

DMP's were first required for all NSF grant applications in January 2011. A DMP is a "supplementary document [that] should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results." It is supplementary in the sense that it is not a component of the NSF's primary review criteria, Intellectual Merit and Broader Impacts. The justification for DMP's derives from the long-standing policy in the Award & Administration Guide (Chapter VI.D.4) that researchers should make specific plans to disseminate and share research data within the body of their grant application. This requirement was often ignored by researchers, peer-reviewers and program managers alike prior to the implementation of DMP's as a required supplement for all grants.

The NSF offers very general guidelines for DMP's, and each directorate and/or program offers more specific guidelines. A universal guideline was deemed inappropriate because each discipline, and even sub-discipline, may have very different criteria for what counts as research data and different expectations about how to store and share data. The NSF therefore empowers peer-reviewers and program managers to set the standards for data management within the

[1] Funding for the Council was provided by the National Science Foundation (#IIS-1413864).

appropriate “community of interest.” Insofar as it is impossible to provide universal criteria for what counts as data, it is impossible to develop universal standards for managing that data. Likewise, for our purposes, it would be inappropriate to try to propagate a universalist model of data ethics. As discussed more below, there are options for pursuing a more flexible, distributed model of data ethics that differs from prior approaches.

The history of DMP's in their current incarnation is multiply determined. In an informal interview, Christine Borgman (UCLA) stated that “the history of DMP's has not been written—it should be—but it is unlikely,” because there isn’t a single, coherent force or reason behind their creation. Borgman claims that DMP's are significantly a response to external pressure to provide more avenues for accountability to the public. The National Science Board (which jointly governs the NSF) received pressure from Congress, and GOP members in particular, to demonstrate more value to taxpayers from basic research projects. The core sentiment of DMP's—that the results of publicly funded basic science should be shared widely—was already present in the requirements for every grant, so it was not a significant institutional feat to make that requirement more explicit.

Additionally, the 2007 AMERICA COMPETES ACT contains provisions related to data sharing. The act required federal agencies to create plans for public data- and results-sharing for any results produced by their employees, with a particular focus on spurring innovation through efficient use and re-use of data. NSF also held a series of workshops and regular internal meetings focused on squeezing more value out of data that had already been generated. Borgman also suggested that academic publishers have lobbied for such policies to foster opportunities for marketing data repository services alongside academic journals.

Other agencies, especially the NIH, began requiring specific plans for data storage and sharing prior to the NSF, and helped set the stage for the NSF’s efforts. The NIH’s “Data Sharing” policy applies to all projects with a budget over $500M. As we might expect with medical data, certain ethical requirements are more explicit and robust than the NSF’s. The work funded by the NIH is also much less diverse than that funded by the NSF and therefore needs to accommodate less ambiguity. The Department of Energy’s DMP policy is very similar to the NSF’s—it is designed to accommodate a broad range of data types and disciplinary preferences for data storage. Agencies, sub-agencies and directorates that deal with a narrower set of disciplines (e.g., genomics) have the benefit of directing grantees to submit data to a specific set of databases that are standard within the discipline (e.g., GenBank). That might help explain why genomics data management is relatively well funded—it has a robust infrastructure that everyone is expected to make use of.

The initial requirement for DMP's at NSF can thus be situated as an early waypoint in ongoing conversations about the values attached to data—innovation, profit, accountability, community, and efficiency.

DMP's immediately ran into some challenges related to their somewhat haphazard implementation. Paul Uhlir, formerly the head of the National Academy's Board on Research Data and Information (BRDI), noted in an informal interview that the NSF was cautious about offering specific guidelines for data retention and sharing because its leadership felt that they did not know enough about how different sciences would respond to common requirements. This left PI's and university research offices with the responsibility to divine what the NSF expected from them, instilling the sense that the requirement was both burdensome and weakly enforced. BRDI proposed to systematically evaluate how well DMP's were achieving their goals in order to develop best practices for NSF directorates. The NSF rejected this proposal and it appears that there is as of yet no rigorous study of how well DMP's have lived up to their goals. Uhlir argues that even though there is good reason for avoiding a universalist approach to data management, the lack of a common structure across disciplines, or even a simple series of checkboxes, means comparing and contrasting approaches is impossible. Because of this unstructured approach to implementation and assessment there is a dearth of systematic knowledge about DMP's.

In early 2013, the White House Office of Science and Technology Policy issued a memo outlining data-sharing policy priorities in the coming years. The memo instructed all large federal funding agencies to develop plans for data-sharing that meets certain benchmarks. Included are priorities to "Ensure appropriate evaluation of the merits of submitted data management plans," and "Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies." These goals are consonant with a common complaint against DMP's in their current formulation—they often are treated as boilerplate and there is little accountability for following through on the proposed plans.

Presently, informal conversations are circulating at the NSF about reforming DMP's. Notably, several AAAS science policy fellows in residence at NSF have recently begun a more rigorous study of what PI's have proposed to do in their DMP's. Conversations with these fellows did not indicate a formal plan for an overhaul is yet on the table, although program managers may have a better sense of that. The timing indicates the plausibility of including the Council's recommendations in an overhaul.

Recommendations for changing DMP's

Several themes were apparent in discussions about DMP's with people involved in their formation and people currently at the NSF.

Researchers need guidance, not prescriptions and new obligations. DMP's are odd in the sense that critics describe them as simultaneously unstructured and prone to being treated as boilerplate. Usually one would assume having to work out the details on one's own would mitigate against the boilerplate treatment. This points to an important dynamic: we need to find a sweet spot between 'check-box' compliance and unstructured open-endedness. Paul Uhlir suggests that PI's should encounter some hypothetical examples appropriately tailored to their disciplines, rather than a set of requirements. Showing PI's a set of common technical and ethical problems, along with plausible solutions, would be more engaging than an open-ended requirement alone.

The diversity of scientific data matters, but the lack of common standards hampers effectiveness. More guidance about what questions and criteria are held in common across disciplines would lead people to better understand what is expected of them. This could be coordinated with developing databases to track and learn from responses.

Recommendations need built-in mechanisms for assessment and iteration. The NSF has found itself in a low-information trap when trying to assess and change DMP's. There is currently little follow-up and enforcement of DMP plans.

Data ethics education is often ill-timed and scattered. Renata Afi Rawlings-Goss, one of the AAAS science policy fellows working on DMP's at the NSF, noted that her only education in data management and ethics came in the first month of grad school, well before she had any data to manage or ethical questions to raise. The Council should include recommendations about education best practices with any proposal for overhauling DMP's.

Data management efforts are scattered within universities (much like other research ethics), but research offices and libraries are beginning to get serious about offering a set of tools and services to facilitate grant writing and meet mandates. Many R1 universities are now offering researchers the opportunity to outsource some aspects of their DMP's for the NSF (and other agencies) to staff and infrastructure affiliated with their university libraries. The most prominent example is the University of California's DMPTool, hosted by the California Digital Library. In most universities, however, there is insufficient connection between now-familiar data *literacy* education often hosted by libraries and data *management* education that could address data ethics in the context of proper scientific practice. Of the little scholarly literature addressing data management plans directly, much of it is located in the library sciences. This suggests that librarians may be a critical component of propagating out new standards for data ethics.

Prospects for Change & Some Skepticism

From a pragmatic perspective, DMP's are an ideal target precisely because they already exist. Reforming them does not require new mandates, and parallel efforts of reform are already underway at the NSF.

However, as discussed above, DMP's have a multiply-determined history with a variety of implementation problems that may make integrating data ethics challenging. The primary purpose of DMP's is to facilitate the sharing and re-use of data. However, depending on contextual details, in many cases data ethics could proscribe sharing and re-use. The Council would need to carefully recommend how PI's should address conflicting values. As we should expect, the NIH has explicit guidance on when to restrict data sharing, and so there are models for navigating the balance between sharing and restricting that could be adapted for our purposes. But the types of data encountered in NIH datasets is much more homogenous and medical research infrastructures are experienced at managing data restrictions.

The many variations in how different disciplines treat research data could multiply the ways in which we need to describe research data ethics—it will be undesirable and likely impossible to develop a single framework of data ethics. Also, because DMP's are presently treated as boilerplate by many PI's, **we risk turning nascent data ethics efforts into something that looks like compliance.** Of the individual directorates' DMP instructions, only the Biology directorate mentions ethics, so this recommendation to include ethics in DMP's will represent a cultural shift.

We might look to professional research organizations for ideas about how to avoid this. For example, the Association of Internet Researchers (AoIR) has some thoughtful materials about professional ethics, including a substantial list of ethics questions that should be asked at the outset of a project. These questions could be adapted for a DMP overhaul, much along the lines that Paul Uhlir recommended above—trying to find a sweet spot that accommodates diversity—and AOiR also hosts an ethics wiki, which might be a useful model to maintain engagement with PI's.

It should be noted that recent policy changes elsewhere impact the decision to target DMP's as the likeliest target for integrating ethics into big data research. Among the other targets discussed at the prior meeting was finding opportunities to have more big data research fall under human subjects protections, especially IRBs. While this was mostly dismissed at the meeting, it is still worth noting that the National Research Council recently proposed changes to the Common Rule governing human subjects protections. If adopted, these revisions would seem to entirely exempt most research methods used in big data from IRB oversight, reducing the niches that could plausibly host data ethics.