Toward an Outcomes-Based Approach to RDM: Experiences from the CU Boulder Center for Research Data and Digital Scholarship

Andrew Johnson, Director, Research Data Management Initiative
Center for Research Data and Digital Scholarship
University of Colorado Boulder

Introduction

Many individuals working in research data management (RDM) support roles find themselves, at one time or another, in the position of presenting on the topic of "Introduction to RDM" or "Data Management 101". Indeed, a cursory web search for either of those titles returns a number of slide decks, many of which include introductory slides enumerating reasons why someone (typically a researcher) should care about RDM. These lists include everything from funder and journal publisher requirements to reproducibility and reuse, from the benefits of staying organized to the need for data security. What is often left unsaid is that these reasons, which can be thought of as the goals or outcomes of RDM processes, each require different approaches, skills, and infrastructure. In some cases, these outcomes can even be at odds with one another. When attempting to demonstrate the impact of RDM on science or comparing RDM practices across disciplines, it is important to understand that RDM processes will vary depending on the desired outcome (e.g., data access/sharing, secure data storage, long-term preservation of data, data reuse and reproducibility, etc.). Using the example of psychology/neuroscience researcher needs at our institution, we argue that an outcomes-based approach to RDM is often more productive and meaningful than treating RDM as a single process or an end in and of itself.

Security vs. Sharing

Among the many places a researcher at our institution might encounter the term "data management" during a single research project, two of the earliest in the research lifecycle are a data management plan requirement for a grant proposal and the data management section of a protocol for submission to the Institutional Review Board (IRB). Assuming there are cases where a researcher encounters and becomes familiar with the IRB process earlier than the data management requirements for a grant, a conflict in perception and understanding of the goals of RDM could conceivably arise. Throughout our IRB process, there are no mentions of data sharing or public access to data. In the data management section of the protocol form, researchers are asked only about data security (e.g., secure storage, limiting access, deidentification, unnecessary retention, etc.). Given the purpose of the IRB, this is perfectly reasonable; however, when a researcher who is already familiar with the IRB process then

encounters a data management plan requirement for a grant proposal, it is plausible that their instinct would be to craft a plan similar to their data management protocol for their IRB submission. We often see this conflict play out in practice when we receive requests for assistance with the data management section of IRB submissions (the purview of our IRB and Office of Information Security) to our RDM support service. We also see this from the other direction when reviewing data management plans for grant proposals from researchers who largely ignore the sections on access, preservation, and reuse in favor of an overly strong emphasis on secure storage. To the credit of many funding agencies, guidance for data management plans does tend to include references to honoring IRB protocols in discussions of data sharing, but perhaps it would be fruitful for IRBs to reference data sharing more explicitly as well. In contrast to the National Science Foundation and other agencies, the National Institutes of Health also take the approach of calling their requirement a "data sharing plan", which more clearly indicates the distinction between the purpose of this requirement and the purpose of something like the IRB's protocol section on data management. Without such distinctions, it is likely that misunderstandings about RDM outcomes will persist.

Aside from the potential for confusion based on the term "data management" being used to signify different processes throughout the research lifecycle, there is also a very real tension between security (including the protection of sensitive data) and sharing (including public access and potential reuse). Quite distinct infrastructure and expertise are necessary for these two functions, yet both certainly fall under the umbrella of RDM in the eyes of many funding agencies. It seems reasonable then for a researcher to expect that an RDM solution will include both of these functions, yet a review of various RDM tools would find this is often not the case (Dataverse Project, 2017). While working toward unified RDM infrastructure is certainly desirable, more precise language in defining and developing tools/services with regard to the specific outcomes they are intended to achieve (rather than always using the blanket RDM term) could lead to less confusion as well.

This tension between security and sharing is just one of the significant RDM challenges that cuts across disciplinary boundaries. All projects that collect sensitive data of any kind, yet are also required to make data publicly available, must deal with this tension. Thus, an outcomes-based approach to RDM would focus on infrastructure to meet this need regardless of discipline. For example, a project in psychology/neuroscience that collects human subject data could be more similar in terms of RDM needs to a cancer biology study than to another psychology project without human subject data. In other words, rather than suggesting that all studies should engage in "proper" RDM practices, we could specify that one desired RDM outcome for a particular study is to protect sensitive data while providing public access to as much data as possible. In such a case, it could be more effective to determine appropriate infrastructure and processes that could lead to that desired outcome rather than making a blanket recommendation that all data from psychology/neuroscience, for example, use one particular data repository or follow one set of guidelines.

Reuse and Reproducibility

Another area where it is potentially beneficial to take an outcomes-based approach to RDM is data reuse, which includes the specific case of reuse for the purposes of reproducing the results of a study. Unlike the tension between security and sharing, one of the major challenges with reuse in an RDM landscape where goals or outcomes are not explicitly made clear is that data can be managed perfectly with regard to funding agency or journal publisher requirements, but that same data can also be of no use to anyone other than the original researcher(s) if it is not documented with an eye toward reuse (Van Tuyl & Whitmire, 2016). Data can be stored securely, made available in a data repository (that also provides high-level metadata and a requisite Digital Object Identifier (DOI)), and migrated to a long-term preservation system, but as long as key information regarding the collection, processing, and/or analysis of data is missing from the documentation, it may not be sufficient to allow for reproducibility or other possible reuse. Thus, by some standards, this hypothetical data would be well-managed and even well-curated while still falling short of the intended outcome. This demonstrates the extent to which data reuse and reproducibility require specialized knowledge, skills, and infrastructure that are not required to meet other RDM outcomes (e.g., secure storage).

A Local Example: Psychology/Neuroscience

At our institution, we have attempted to delineate our RDM services by the outcomes each are intended to support. For example, we offer a large-scale storage service (PetaLibrary) in addition to our institutional repository (CU Scholar), which is intended for data publication and some level of preservation. Similarly, we have found a significant and growing demand from researchers in psychology/neuroscience for infrastructure to support reproducibility of findings, which is a separate RDM outcome from either managed storage or data publication/preservation. To help meet this need, we have entered into an institutional partnership with the Center for Open Science's Open Science Framework (OSF). OSF provides features that allow for data collection and analysis to be more transparent and well-documented, including the ability to freeze and timestamp a project at various stages of the research lifecycle (called "registration" in OSF terminology). These are features that most repositories and data storage systems do not offer, so we provide this additional infrastructure in order to support reproducibility as the intended RDM outcome for many of our psychology/neuroscience researchers. While we are working toward integrating all of our RDM infrastructure in order to create a more seamless user experience, we recognize that an overemphasis on, say, storage alone can lead to RDM needs like reuse and reproducibility going unmet. Likewise, infrastructure without an accompanying set of education/training/consulting services can also jeopardize the potential for data reuse and reproducibility because researchers themselves play such a large role in providing the

documentation that enables those outcomes (Becker, 2017). For example, if a researcher has not been made aware of a tool like OSF and how to utilize it for the purposes of reproducibility, or does not even appreciate that reproducibility is a desirable goal in the first place, then it is highly unlikely that this RDM outcome will be met. Indeed, at our institution we have hosted outside speakers on reproducible research, held workshops on OSF for the purposes of reproducibility, and offered consulting services on these topics as complementary efforts to our infrastructure offerings. At present, psychology/neuroscience researchers have been the primary users of our tools and services that are intended to support reproducibility, but other disciplines grappling with similar issues might find these resources valuable as well.

Conclusion

There are numerous sources of guidance on RDM that vary significantly and sometimes conflict with one another. While IRBs rightly focus on data security, some journal publisher policies focus solely on data access, typically via a list of recommended repositories. Funding agency requirements tend to cover everything under the RDM umbrella, but usually only with regard to a data management plan submitted during the grant application process. All of these approaches can be problematic if the desired RDM outcomes are not explicitly defined and the recommended RDM processes and infrastructure do not align with those outcomes. Failing to do either or both adds to confusion among researchers and makes it difficult to measure the impact of RDM. For example, if data access is the only desired outcome, then requiring a statement regarding the location of the data is probably sufficient to ensure that outcome has been met; however, proof of data access should not be used as a stand-in for data quality or an indicator of the trustworthiness of the repository that houses the data. In practice, we see the latter scenario occur with regard to journal policies that require an identifier (e.g., DOI) as evidence of data availability in one of a list of recommended repositories. Since compliance with journal policies is quite a strong motivator for researchers, those of us working on the front lines of RDM services often observe these data availability policies being equated with the entirety of RDM in the minds of researchers. Thus, acquiring a DOI for a data set becomes the sole concern of the researcher, and options that provide high quality RDM services that are not included on recommended repository lists are ignored (DataCure, 2015). Given the complex and sometimes conflicting landscape of journal policies, funder requirements, and guidance from many organizations and institutions with regard to RDM practices, well-defined outcomes (e.g., reproducibility in psychology/neuroscience) can lead to more effective RDM support and infrastructure. As we discovered on a small scale at our institution with regard to data management plan support, when outcomes, guidance, and support services align, it is possible to not only assess RDM services effectively, but also to demonstrate positive impact on actual RDM practices (Johnson & Knuth, 2016). RDM is only a means, albeit an important and multifaceted one, so we must be clear about our desired ends.

References

Becker, E. (2017). How better training can help fix the research reproducibility crisis. Retrieved July 27, 2017, from https://www.insidehighered.com/blogs/rethinking-research/how-better-training-can-help-fix-research-reproducibility-crisis

DataCure. (2015). Open letter to PLoS: Libraries role in data curation. Retrieved July 27, 2017, from https://datacurepublic.wordpress.com/open-letter-to-plos-libraries-role-in-data-curation/

Dataverse Project. (2017). A comparative review of various data repositories. Retrieved July 27, 2017, from https://dataverse.org/blog/comparative-review-various-data-repositories

Johnson, A. & Knuth, S. (2016). Data management plan requirements for campus grant competitions: Opportunities for research data services assessment and outreach. *Journal of eScience Librarianship*, *5*(1). https://doi.org/10.7191/jeslib.2016.1089

Van Tuyl, S. & Whitmire, A. L. (2016). Water, water, everywhere: Defining and assessing data sharing in academia. *PLOS ONE*, *11*(2), e0147942. https://doi.org/10.1371/journal.pone.0147942