Research Ideas and Outcomes 4: e30829
doi: 10.3897/rio.4.e30829

Project Report

Investigation of Non-Academic Data Management Practices to Inform Academic Research Data Management

Steven I Van Tuyl‡, Amanda Whitmire§

‡ Oregon State University, Corvallis, United States of America
§ Stanford University, Pacific Grove, United States of America

Corresponding author: Steven I Van Tuyl (steve.vantuyl@oregonstate.edu)

Reviewable v1

Received: 25 Oct 2018 | Published: 31 Oct 2018

Citation: Van Tuyl S, Whitmire A (2018) Investigation of Non-Academic Data Management Practices to Inform Academic Research Data Management. Research Ideas and Outcomes 4: e30829.
https://doi.org/10.3897/rio.4.e30829

Abstract

In recent years, the academic research data management (RDM) community has worked closely with funding agencies, university administrators, and researchers to develop best practices for RDM. The RDM community, however, has spent relatively little time exploring best practices used in non-academic environments (industry, government, etc.) for management, preservation, and sharing of data. In this poster, we present the results of a project wherein we approached a number of non-academic corporations and institutions to discuss how data is managed in those organizations and discern what the academic RDM community could learn from non-academic RDM practices. We conducted interviews with 10-20 companies including tech companies, government agencies, and consumer retail corporations. We present the results in the form of user stories, common themes from interviews, and summaries of areas where the RDM community might benefit from further understanding of non-academic data management practices.

Keywords

Research Data Management

Introduction

In recent years, the academic research data management (RDM) community has worked closely with funding agencies, university administrators, and researchers to develop best practices for RDM. The RDM community, however, has spent relatively little time exploring best practices used in non-academic environments (industry, government, etc.) for management, preservation, and sharing of data. Communication between academic to non-academic data management professionals is quite limited, generally, and we believe that the similarities in problem spaces and missions of these communities might offer opportunities to leverage knowledge across them. The primary goal of this project was to interview data managers at non-academic institutions about their data management practices and to compile this information in a way that might inform academic research data management practices. We present the results of a project wherein we approached a number of non-academic corporations and institutions to discuss how data is managed in those organizations and discern what the academic RDM community could learn from non-academic RDM practices. Specifically, we were interested in learning more about:

- What workflows are used for data acquisition, including assigning metadata or other documentation to the data during acquisition;
- How data are preserved for reuse at a future date, and how decisions are made to keep or otherwise make use of collected data;
- How data are shared inside of the organization and with potential users outside of the organization.

Project Description

We conducted interviews with individuals from six companies including individuals from three large technology companies, a metropolitan school district, an industrial research and development company, and a consumer retail corporation. Selection of interviewees was conducted haphazardly, with interviewees selected based on responses to "cold call" emails, contact with personal acquaintances, or with the help of professional connections to interviewees or their companies. Interviews were conducted in person using an interview script to guide conversation (Suppl. material 1) with both authors present and taking notes on the contents of the interview. After interviews were completed, the authors collaboratively created a summary of the interview based on notes and reflection on the discussion. All interviews were conducted in compliance with our Institutional Review Board (OSU IRB study number 6839). We present the results in the form of user stories, common themes from interviews, and summaries of areas where the RDM community might benefit from further understanding of non-academic data management practices.

Results and Discussion

We've distilled the problems presented by our interviewees and the solutions they propose for these problems into five broad categories: Workflows, Documentation, Priorities, Standards, and Service.

Documentation and Workflows - Be Intentional

"*We need to [be able to] recreate the decision-making process*"

Many of the problems raised by our interviewees relate to documentation and workflows for data and methods. Among the problems presented by interviewees were lack of clear ownership or responsibility for the data, *ad hoc* storage locations chosen by the user versus the project or data source, understanding the provenance of the data, reproducibility of methods and reasoning for data cleaning and management, and handling large quantities of data from a diversity of sources. Multiple interviewees also noted the single point of failure problem wherein one person holds critical information about a dataset or methodology and their departure from the group/company results in loss of said information.

Tracking and understanding a dataset through its lifecycle and documenting roles and responsibilities, methods used, and reasoning behind decision making are critical for understanding where a dataset has been or what should happen when capturing new data. Interviewees discussed the need to create a culture of documentation, clear workflows, and retention policies for datasets. We also heard from interviewees that knowing who made changes to datasets or documentation can be critical to helping fill out the picture of why changes were made to a dataset or methodology. Last, some interviewees identified centralization of datasets and documentation for their groups as key to bringing transparency to the use of common datasets. Centralized documentation and transparency across the group also helps resolve or at least alleviate the single point of failure problem.

RDM practitioners already spend a fair amount of time working with researchers in these areas, though we wonder, as have others, about how effective our modes of communicating workflow and documentation needs are. Two points in this area stand out from our conversations with interviewees. First, building a culture of documentation has been critical for many of these companies to improve data management and workflows. Even in companies where one might expect documentation to be of ultra high priority, entire areas of work have gone undocumented or poorly documented leading to a host of problems, but centralization, transparency, and consistency of documentation and procedure almost always help. Our impression that the challenges of implementing effective RDM can be solved by technology is belied by the fact that even documentation focused technology companies cannot solve the problem without cultural change. Second, in most cases, the workflow and documentation practices discussed with interviewees were focused at the work/research group level, rather than at the project or individual dataset level. This higher level view of how to define documentation and workflow standards may be helpful in the context of academic RDM practices where many projects are parts of a

whole research program that is much larger and longer lived than the individual project or dataset. Cultural change within research groups bridges the gap between individual habits and group-level behavior. Vertically integrated culture of documentation is key. We mostly train students (who have time and are more adaptable), but the culture must be supported by, ideally driven by, the PI to enable consistency and long-term integrity.

Priorities - Solve Real Problems

A second major category of problems identified in our interviews was the need to identify what problems one is trying to solve in order to facilitate prioritization. It was not uncommon for us to hear that, even in well resourced corporations, teams were overloaded with work, inundated with large amounts of data from multiple sources, and could easily lose focus on what problems they were trying to solve. Building in a process for identifying what projects the team is working on, and knowing when to say "no" to a project, was an important part of setting priorities for more than one of the interviewees. In addition, multiple interviewees either explicitly or implicitly indicated that it was important in their work to solve "real problems," which connects us to the next major area of importance - service.

Academic RDM service providers are often beset with a variety of problems to solve and it is impossible for most providers to solve them all, given limited resourcing. This in mind, it is important for service providers to identify, for any given service or initiative, what value the service provides to the overall RDM program, and how applicable any lessons learned from the activity might be to other end users (either the researchers you support or your colleagues). Saying “no” to special projects can be challenging, but with a process in place to help define what projects are of highest value for the program, the projects one says “yes” to can have large impact. RDM service providers should take advantage of the experience of colleagues to help evaluate potential services and prioritize high-impact work.

Service - Connect With Users

Almost all of the companies we spoke with have a service oriented approach to their work and most identified communicating with end-users as a key to their work. This communication comes in three main forms: user experience, user expectations, user enlistment. First, user experience is a way of communicating with users that many of us are familiar with, and is crucial to providing service - we engage with users to find out what they need and how or if the service is meeting those needs. All of the organizations we spoke with engage in some form of user experience work, though some more explicitly than others. Next, user expectations is the process of setting expectations with the user for what the service will provide. This can come in many forms, but often it is as a formal or informal project plan or memorandum of understanding signed with the end user to make it clear such elements of a project as roles and responsibilities, end products, and other requirements. Last, user enlistment, is an area of service that we had not heard about or considered prior to discussing it with a few interviewees. Essentially, user enlistment is the process of identifying not what makes a service or experience easy, but identifying what actions a user takes that draw them further into the service. An example of this is that one

of our interviewees discussed how the core technology provided by his company “kind of sucks” and has a high bar to entry. This company, though, has identified that after a certain amount and type of use, users of the service suddenly find the service much easier to use - essentially, they’ve hit the point in the learning curve where it is no longer challenging to use the service. The goal of user enlistment, then, is to make it easy for users to get to that point in the learning curve with as much help as possible.

All three of these elements of service can be employed by RDM practitioners in academia. Setting and communicating user expectations may be somewhat less common in academic RDM contexts, though may be a helpful way to approach interactions with potential end-users and in outreach materials for RDM programs. Anecdotally speaking, we have interacted with researchers who arrived at a meeting expecting a “magic bullet” to address their RDM frustrations. Finding effective methods to adjust their expectations on the fly was critical in establishing credibility for ourselves as RDM professionals and for our services. Actions to help set reasonable user expectations range from explicitly communicating learning outcomes during formal training sessions to emphasizing the iterative and progressive nature of improving individual RDM habits during one-on-one consultations. More challenging might be the implementation of user enlistment for RDM, and we encourage the RDM community to continue to engage with user enlistment by working to systematically understand what is difficult about data management for researchers and consider focusing resources in those areas. Active listening is an easy and obvious place to start (Rinehart 2015), but deeper, more meaningful engagement with researchers provides data services providers with actionable understanding. The ability to move beyond offering general data management best practices advice to providing targeted recommendations or hands-on assistance are much more likely to draw people back to your services for more help. Where immersion in research groups or even departments is not feasible, any experience that augments an RDM service provider’s understanding of the researcher perspective and the process of their work is valuable.

Standards - Use Them or Make Them

Some of the companies we interviewed have data management practices that are strictly governed by federal, state, and/or local regulations and standards. In these cases, it was interesting to note that these organizations generally saw these regulations as helpful for defining the types of work they could do with their data, defining workflows and processes for data management, and defining documentation needs for record keeping and reporting. In a few cases, these companies also identified gaps in standards and regulations and created their own regulations to help provide consistency of data formats, storage, and preservation.

Defining standards for researchers to use has been somewhat challenging in the work of RDM service providers - primarily for those working in academic libraries that serve researchers in a breadth of research domains. That said, the RDM field has generally engaged with the idea of setting standards and with some success, including some broad-level agreement on outreach and education topics for researchers (cf Carlson and

Johnston 2015, Kafel et al. 2014), as well as some domain-specific documentation standards (cf Wilkinson et al. 2016).

Conclusions

Results of the interviews provided new and expected insights into data management practices in non-academic environments. To an extent, the data management practices discussed by our interviewees were reminiscent of the types of practices and challenges experienced by academic researchers and those providing support to the research community. We found that data management problems arising in industry were almost exclusively regulated, if at all, by government requirement (in the case of the industrial R&D company) or by requirements of a multinational corporate entity (in the case of the multinational retail company). These examples were the exception, and even within these organizations, parts of data workflows and management regimes were highly disorganized. Other interviews highlighted new approaches that research groups might take to work through specific types of data management issues. In the end, it is clear that there are many parallels between academic research data management and data management in non-academic settings. While many, if not all of the companies we interviewed with were well resourced, the prevalence of common data management problems suggests to us that resourcing can only go so far to solve RDM problems. Defining process, workflows, and documentation is critical, as is engaging with end-users in a meaningful way.

Author contributions

Van Tuyl and Whitmire contributed equally to conception, implementation, analysis, and write up of this project.

Conflicts of interest

No conflict of interest reported.

References

- Carlson J, Johnston L (2015) Data information literacy: Librarians, data, and the education of a new generation of researchers. 2. Purdue University Press
- Kafel D, Creamer AT, Martin ER (2014) Building the New England Collaborative Data Management Curriculum. Journal of eScience Librarianship 3 (1): 60-66. https://doi.org/10.7191/jeslib.2014.1066
- Rinehart AK (2015) Getting emotional about data: The soft side of data management services. College & Research Libraries News 76 (8): 437-440. https://doi.org/10.5860/crln.76.8.9364

- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Bouwman J (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

Suppl. material 1: Data Interview Protocol doi

Authors: Van Tuyl, Steve, Whitmire, Amanda
Data type: document
Brief description: Interview protocol used for the project, "Investigation of Non-Academic Data Management Practices to Inform Academic Research Data Management"
Filename: VTWDataInterviewProtocol.pdf - Download file (117.14 kb)