The Relevance of Research Data Sharing and Reuse Studies

by Nicholas M. Weber

The hyperbole of “big data” [1] and the surprising backlash for “small data” advocacy [2] have been well documented in popular journalism, but research data are definitely experiencing a cultural moment. Much of this excitement turns on the potential that increased accessibility, interoperability and computing power can offer to the exploration of loosely related datasets (for example, tweets and the fluctuations of a stock index like NASDAQ) [3].

Research data’s cultural moment should also be one for the field of information science. Our field has traditionally studied some of the most difficult problems in the use of large-scale information resources, including the meaningful organization, access, management and storage of scholarly products in all of their formats and encodings. But, thus far, our field has struggled to make its expertise in this area well understood, and more importantly, we’ve been slow to demonstrate the relevancy of our work to the vital issues that we face as an intellectual community and, more importantly, as a society [4].

Our struggles stem in part from the fact that this space is already crowded with sociologists, economists, computer scientists and statisticians, to name a few of the disciplines involved. These disciplines all play an increasingly important and insightful role in building information systems, developing standards and creating services to support the meaningful use and preservation of research data. While each of these fields faces similar dilemmas with respect to meaningful engagement with research data, information science should be well equipped to handle systems-based problems. In short, we must better apply what we’ve traditionally known about citation behavior, document retrieval and information seeking to a data-intensive paradigm, while simultaneously avoiding generic simplifications such as “publications are just like datasets.”

Another part of our struggle is that we have a poor conception of our
problems space – as a colleague more elegantly put it, complications in data sharing and reuse are often due to the fact that we study “poorly bounded” digital objects [5]. At the most basic level, we have no idea how much data actually exist within an institution, department or even research group, let alone the entire enterprise of science [6]. This point is an especially important one for our field to address earnestly; it is tempting to indignantly accept headlines like “75% of research data is never made openly available” [7], but how much or how little of research data are “made available” is a fundamentally unknowable number. Irresponsible statements such as this headline hurt our cause much more than they help. We have no idea how much research data are produced or stored, let alone how much of it is shared. The goal of information science shouldn’t even include the search for such numbers. Instead, we should seek ways to meaningfully define our research subject so that we can make reliable statements about what is knowable – a perfect example is a recent study indicating the citation advantage for publications with archived, openly accessible data [8].

**RDAP’s Promising Results**

Over its short period of existence, the Research Data Access & Preservation Summit has become a crucible for data sharing and reuse studies. It provides a forum for those working in research institutions to share early results and take stock of or note gaps in our current understanding of these two related issues. Two particularly promising areas of research in sharing and data reuse were presented this year at RDAP – I highlight them both because their preliminary results are exciting and to combat any negativity that might be inferred from my earlier comments.

**Data Management Planning.** The first area is the rapid and unbridled success of research data management planning. Most formalized data management planning services and tools have emerged only as a result of requirements from funding agencies for grant applicants to explicitly document how research data will be stored and made available for future use. As these mandates were handed down, many institutions were quick to adopt templates and develop tools to help grant applicants submit competitive data management plans. Emerging from the data management planning work is a thread of evaluation research that looks at how these policies have shaped grant applicants’ behaviors and points out the sometimes subtle gap between the expectations of funding agencies and potential best practices for institutions supporting basic research activities. The RDAP13 lightning talks presented by Katherine G. Akers and Jennifer Doty, Heather Coates and Martin Donnelly all address this research question.

The RDAP community has been especially active in this sphere, and this year a quarter of all accepted Summit submissions had data management in the title or subject. Some of these presentations included general overviews, while others offered a more detailed look at data management for a more narrowly defined field, such as Konkkel’s lightning round talk, “Bootstrapping Library Data Management Services for Epidemiology.”

As part of the panel Data Use and Reuse – Sharing Open Data Success Stories, Renata Curty shared preliminary results from a survey study of successful NSF grant applicants. The approach taken in this pilot study was to sample awardees that recently created data management plans for NSF – asking about attitudes and opinions on both the process of creating and the prospect of sharing the data produced in the course of their funded research projects. Though this survey was largely a proof of concept study, its results can serve a dual purpose. On the one hand it helps in the gathering of valuable demographic data so that the RDAP community can better understand questions regarding who shares or reuses which data; on the other, it provides valuable insight as to what is difficult about this process and what can be improved in the near future.

This study is a promising first attempt to look at how sharing and reuse are affected by data management planning policies from the perspective of data producers. We should also think about operationalizing this type of study across different funding agencies and awardee types – as there are many public policy implications to this type of work. To put this work in some perspective, the U.S. government contributes about 59% ($32.6 billion) of the $54.9 billion in annual spending on research and development in higher education. Of that $32.6 billion, six agencies (Departments of Agriculture, Defense and Energy, NASA, NIH, NSF) provide 97% of all research and development funding [9], and NSF has by far the smallest...
Understanding differences among awardees of NIH, NSF or DoD is an important next step to understanding sharing and reuse in a broader research data context.

**Data Sharers, Data Reusers.** A second exciting thread of research focuses more closely on the behavior of researchers in sharing and reusing data, often from a practice-based perspective. This vein of research is similar to traditional use and user studies in information science, but also includes issues of appraisal, valuation and quality that cut across archival science and information systems literature more broadly.

Ixchel Faniel presented novel findings from the DIPIR (Dissemination Information Packages for Information Reuse) project. In this iteration, she largely focused on data quality indicators for reuse satisfaction among quantitative social scientists. Early analysis of an Inter-university Consortium for Political and Social Science Research (ICPSR) survey indicated that, not surprisingly, data documentation, as well as completeness and accessibility, were important in determining data quality; however, survey results also indicated that producer reputation was not a significant factor in reuse satisfaction. Of particular value in this analysis is Faniel et al.’s framework of data quality indicators. This framework drew upon diverse information systems literature, and it will be interesting to see how it evolves as it is reused and tested by the data curation community in future applications.

Similarly, Dharma Akmon offered early analysis from a dissertation that tackles large questions about how time-scales and accessibility affect perceptions of value in research data. The implications of this work may affect not only systems development, but also our understanding of data practices more generally. Akmon’s field work and observations of stream ecologists emphasized an evaluation of data in an instrumental capacity, which resonates with Bernd Frohmann’s plea for information science to more seriously consider material practices in user studies [10]. This type of ethnographic work plays a mapping role similar to the surveys discussed by Curty and Faniel, but we get a narrower, yet richer account of how data use and reuse might be better supported in field campaigns. This perspective is especially valuable as data curators begin to embed themselves in research centers and individual small science laboratories.

Slides from the Summit for many of these talks are available on slideshare at [www.slideshare.net/asist_org/tag/rdap13](http://www.slideshare.net/asist_org/tag/rdap13).

### Future Directions for RDAP Research

At a time of sequestered budgets and shrinking economies, our field should be capable of providing policymakers and legislators with reliable information about the value of research data from one domain to the next. As the RDAP community begins to coalesce and clearly articulate its research agenda, it is similarly worth considering whom the audience of this research might include. As studies of data sharing and reuse expand in scope and sophistication, so, too, will their impact. While ASIS&T may not have the capacity to lobby Capital Hill on our behalf, the research of this community increasingly plays an evaluative role for agency mandates and funding at a federal level, so it seems natural that we should point our research results, however preliminary, in that direction.
Resources Mentioned in the Article


