Meeting Data Needs on Today's Campus: The Digital Commons Data Pioneers Report

Promita Chatterji, bepress
Meeting Data Needs on Today’s Campus

The Digital Commons Data Pioneers Report

Promita Chatterji
Introduction

During 2014-2015 bepress partnered with sixteen Digital Commons subscriber schools to launch the bepress Data Pioneers program. The program convened representatives of each of these schools to share their experiences in launching data support services at their schools throughout the year. Each of the schools was at different stages of building a data program — some had already been advising their faculty on data management plans and hosting data in their repositories, others participated in the program alongside exploratory taskforces on their own campuses. The program offered schools a chance to discuss some of the specific questions around data management and to connect with others using Digital Commons to support data programs. For bepress, these use cases helped to inform system-wide decisions about storage infrastructure and future directions in development.

What did we learn? After a year of regular meetings and shared research and experiences, there were a lot of stories and bits of information to digest related to failures, successes, and laying groundwork. Overall, the Pioneers found that participating in a community of users left them with a feeling of confidence that they had the tools to support their faculty’s data needs, whether through the institutional repository or other channels. This report looks at some of most compelling topics that emerged during our meetings and shares stories of individual projects on which the Data Pioneers schools spent their time.

Staffing

Many Pioneers were concerned about finding adequate staffing to support the needs of a data program. Some worried that once they opened the floodgates, they would be confronted by a deluge of data needs that the IR could not handle. Others were concerned about the technical challenges that come with data curation and wondered if they needed to hire a staff member with a background in data in order to provide sufficient data support. A few schools dropped out of the program due to losing their science or data librarian, or not having a staff member in place.

“I feel that we are now in a much better place to support this work (even with reduced staff) than we had when we started.”

- Marilyn Billings, UMass Amherst

However, others found that not only was the “data deluge” not a problem, but that when a few unexpected data sets did come in, they were able to provide support more quickly than other solutions available to the researcher. One example came from University of Massachusetts, Amherst. Prior to the Data Pioneers program, they had done quite a lot of work on understanding campus data needs through an exploratory committee. However, they lost a staff member and were in the position of having to re-invest
in EZID for DOI services. Marilyn Billings, head of the Office of Scholarly Communications at University of Massachusetts, Amherst said, “This made me realize that we had to step back and examine our research support services writ large, to make sure we had the components in place to meet our best practices goals in the long term. Hence even though we didn’t participate in creating a lot of examples for the Data Pioneers group, I feel that we are now in a much better place to support this work (even with reduced staff) than we had been when we started. It was also excellent to have colleagues with whom to brainstorm as we went through this iterative process. We have now hired a science librarian who has data experience.”

Finding Data on Campus

For most Pioneers, it was not hard to identify who was working with data on campus: there was data all over the place. What was challenging was finding researchers who a) had data that was in shareable condition, b) understood the benefits and requirements behind sharing, and c) felt that the library was the right partner to work with in managing their data.

Two strategies seemed to produce the best results. The first strategy was working in-depth with one or two campus champions. As Lisa Palmer from the University of Massachusetts Medical School put it, “all you need is one (researcher)” to get a model for a collection up and running. She worked with one of her researchers who needed to share his data as a prerequisite of publishing with PLOS. The workflow and publication structure that she created with this researcher, then replicable for future scholars, was also a model she could apply to datasets that were already in the repository.

The second strategy involved partnering with the Office of Research, or Faculty Senate. At Chapman University, some of the earliest collections came out of a presentation that Kristin Laughtin-Dunker gave to the faculty senate. A faculty member had recently published an Art History monograph and was in need of a platform to house and preserve his data for the long-term (including spreadsheets and GIS data). Becky Thoms at Utah State University has an ongoing partnership with the Office of Research where she is a featured speaker at one of their series of workshops for faculty.

In addition to these core strategies, the Pioneers group brainstormed a number of strategies that seemed to yield positive results:

- Following up with researchers who worked with the library for help with Data Management Plans.
- Sending targeted letters to authors who publish in PLOS, BioMed Central, Nature, or other journals that require shared data as a prerequisite of publication.
- Working with a taskforce on digital humanities.
- Working with researchers in Environmental Sciences (this field has data from hybrid sources, which benefit from the flexible publication structure Digital Commons can provide).
Educating Faculty about Data

At the Data Pioneers kick-off meeting, this topic was a top concern for many participants. The release of the 2013 OSTP memorandum, and subsequent growth of funder mandates for sharing data related to federally funded research, planted seeds for a major changes in how researchers need to manage the data related to their research. In many cases, libraries found themselves in a position where they were more aware of what changes researchers would be facing than those in the fields themselves. Education involved a number of activities, including talking to faculty about the funder mandates driving the changing landscape around data sharing, working with faculty to make sure their data were in a shareable state, and letting them know what kinds of repositories were available for use. The Data Pioneers developed a number of materials to support faculty education, including LibGuides, presentations, guides to funder mandates, and numerous data surveys and interviews. Two schools also developed plans to set-up more elaborate data-curation profile-type interviews.

Defining the Scope of the Data Program

Many institutions are in the middle of conducting investigations into the evolving data landscape. As such, libraries are often asked to be part of a data taskforce or other exploratory committee charged with understanding what roles various campus units could play in supporting data needs. Members of the group found that participating in the Data Pioneers program helped to clarify where the library fit in with the overall scope of data needs on campus.

What about the IT department?

When many researchers think of getting campus support for their data needs, they turn first to the IT department. One of the big topics of discussion early on in the Data Pioneers’ meetings was how to distinguish between what the IT department can provide and the library’s role. Numerous conversations between library, IT, and individual researchers clarified these distinctions for the group quite a bit during the program. For the most part, while IT services can work with scientists and researchers to help them use, process or store large amounts of raw data, there is still a real unmet need for finished data that needs to be published with individual records, citations, and metadata.

Moreover, IT support tends to focus on hard sciences, and with the growing trend of data-driven research in a large variety of fields, there are scholars with data needs in social sciences and humanities who have emerging needs for storage, sharing, and publication. Overall, data needs on campus are growing and there is a real need for campus-wide partnerships. While IT is crucial for handling needs of “big data” processing or dark archiving raw datasets, researchers can turn to the library for support with published data, and for assistance with data in disciplines not traditionally supported by IT.

“Researchers can turn to the library for support with published data, and for assistance with data in disciplines not traditionally supported by IT.”

Compliance with funder policies

Digital Commons meets storage, sharing, and preservation requirements for nearly all funder policies. Major funding agencies, including NIH, NSF, NEH, and the DOE, all require that applicants submit a data management plan that describes the data that will be collected, as well as a plan for dissemination, with the understanding that the varied nature of research projects will call for different kinds of support. A few of the Pioneer schools developed boilerplate language that could be used for Data Management Plans and other institutional publications. Examples of these policies and LibGuides, as well as sample interviews, can be found in the appendices or the Digital Commons Data Resources page.
Disciplinary repositories or DIY solutions?

Many researchers work in fields with prominent disciplinary repository solutions for sharing and preserving their data: GenBank, ICPSR, and Protein Domain to name a few. Disciplinary repositories offer searchability, as well as tailored and detailed metadata. One of the most interesting findings of the Data Pioneers program was understanding scenarios where these disciplinary repositories were not fully able to meet researcher needs. As mentioned above, it became clear that data-driven research is becoming more common and that there are many disciplines without their own data repositories. But even ones that do have great solutions often cannot capture the related content that might be associated with a dataset and cannot offer flexibility for researchers who would like to try experimenting with the ways they are sharing their data. For example, Utah State University developed a solution for a researcher working on genome data that needed to be shared in a different format than is allowed by GenBank.

Some schools were also in the process of exploring what it would take to build their own data repository. Reasons for this solution included standardizing metadata and creating security for publications that were not being shared. Two of the pioneer schools came into the program with plans in place for their own data repository, but found that getting all the stakeholders together was complicated and that it took a long time for these solutions to move out of the planning phases.

Data Publishing: Technical Questions

A particularly successful part of working together as a group was to identifying and finding solutions for some very specific questions having to do with data publishing. Here are some of the main concerns and what we learned.

- **Storage Capacity** Many Pioneer schools came into the program with this concern at the top of their minds; they wanted to be able to advise researchers based on the size and complexity of their files. The message that there were no fees and no overall limits was a powerful tool for the library to use as they began marketing their data services to their campuses. Anecdotally, the vast majority of data files were not any larger than files the libraries were already storing in their repository.

- **Citation and Discoverability** Datasets on Digital Commons receive an individual record, persistent URL, and suggested citation that can be used to ensure preservation and discoverability. Like other content on Digital Commons, data ranks high on search engines including Google and Google Scholar. For research that is not ready to be shared, there are a variety of options for levels of access control. While there is a growing movement to assign DOIs to datasets, most Pioneers found that this is a far from universal practice, largely dependent on the nature of the faculty’s research and field of study. Some schools were able to sign up for DOI services, either through their own CrossRef accounts or through CDL’s EZID service. Other schools found that assigning DOIs was of secondary importance to creating individual records for datasets and offering
the ability to publish data in a way where it could be contextualized with related content.

- Dataset with DOI from University of Massachusetts Medical School
- Dataset without DOI from Yale University

**Metadata** Depending on field and discipline, datasets may have individualized metadata requirements. In most cases, metadata was less problematic than anticipated. Librarians found that a) researchers knew what fields they wanted included for metadata and they could use the custom metadata fields on Digital Commons, b) they were able to use the Digital Commons default metadata schema for datasets, or c) they could conduct a short interview to capture the necessary information for a thorough set of metadata.

- Simple metadata schema from Southern Illinois University
- Extensive metadata schema from University of Denver, Center for Orthopaedic Biomechanics

### Data Publishing: Enhanced Presentation

What does a data publication look like? Pioneers found that the library could offer their researchers a valuable service by providing a variety of showcasing options that enhance both the content and context of the data. Researchers can maximize the discoverability of the data by linking it to related content, capturing metadata, and supporting streaming media – needs that might not be met otherwise be met by IT or other campus units. At bepress we’ve found that three common data publishing needs prevail for researchers. The following examples show how these needs are being met; and we’ve included a 4th example to illustrate how IR’s are collecting data together to demonstrate institutional impact.

#### Example 1: Supplementary data related to an article or monograph

This dataset published at Chapman links directly to a related article. The record features a button directly underneath the download button that connects to the related article.

Paleontologists at UPenn use these 3D scans of Dinosaur skulls to conduct morphometric analyses. Users can download the scans and/or explore them with an interactive interface on the record page. The data is linked to a related article published in PLoS ONE.

#### Example 2: Supporting Grant Requirements

From the Data Management Plan, to publishing datasets, to connecting to related content and supporting materials, support for grant-funded research requires a detailed workflow and a variety of publishing needs.
Marquette University’s Dr. Doolittle Project is a great example of a collection that makes it easy to navigate between articles, datasets, conference proceedings, and other materials that were generated as a result of a grant from the NSF.

At the University of Denver, a NIH grant funded research for a collection of experimental, image, and computational modeling data from the testing of natural knee biomechanics. Because it is an unusually comprehensive collection, the researchers wanted a repository solution that would satisfy grant requirement and also be easily accessible by the research community.

Example 3: Project-Specific Data

Data does not exist in isolation; prod a little bit and you are bound to uncover an exciting initiative or long-standing research question. Yet, too often, data collections are gathered together in lists or formats that leave these contexts unexplained or unappealing to explore. During the course of the Pioneers program, libraries experimented with telling the story of data collections in ways that complemented the related research question, which resulted in a beautiful final publication. Many of these projects include “long-tail” data not affiliated with a grant or publication, archival data, or smaller sets that weren’t supported by IT or their own disciplinary repositories. However, the growth of digital humanities and data-driven research in social science disciplines is another factor impacting this fast growing category.

Professor Justin St. P. Walsh is an art historian whose work looks at the circulation of vases to examine questions about consumerism and identity in the ancient world. The research data for his monograph consists of ArcGIS files, layered PDF maps, and a number of databases, all of which are stored together.
The data from University of Montana’s Flathead Lake Seismic Survey was languishing in a filing cabinet for nearly 40 years. Recovered by Professor Robert Lankston, the collection of bathymetry recordings, has provided a resource for new scholarship as well as a teaching tool for students learning about research practices.

**Example 4: Institution-wide Discoverability and Organization**

In addition to meeting the needs of individual researchers or research teams, many libraries showcased the breadth and depth of research data across campus by collecting datasets into a top-level collection. Combined with the dashboard, which shows download counts and readership distribution, these top-level collections are an extremely useful tool for the Office of Research, Sponsored Programs, or other offices looking for tools to measure research impact.

**Top-Level Data Collections**

This top-level data collection at University of Massachusetts, Medical School pulls together data from a variety of departments. Datasets are displayed in multiple relevant locations using the collection tool.

Departmental Data Collections

At Pepperdine, datasets for an entire department are uploaded to a series as individual objects. Metadata can be customized for individual disciplines.

**Conclusion**

The experiences, examples, successes, and failures of the Data Pioneers provided a wealth of information about what it takes to get a data program up and running, and what it looks like to work with datasets on campus. Most Pioneers had little to no background working with data, and one of the most valuable things the group learned as a whole was that technical training with data is not a necessary prerequisite to launching a successful program. We were able to learn about issues in the data landscape, including wording for DMPs, practices around citation and metadata, and storage needs. Most importantly, the Pioneers program clarified that the library is distinct from IT or other campus groups in supporting research data management needs for published or finished data and for smaller projects that fall outside the purview of these groups.

Since the launch of the Data Pioneers, data content in Digital Content repositories continues to grow. We’ve seen that small steps like putting together a LibGuide, or providing a proof-of-concept collection by working with one researcher can reap valuable results. It is clear that data needs are only going to grow on campuses nationwide. The Data Pioneers program illuminates how effectively and easily the library can meet the wide variety of needs related to campus data support.
Appendix One: Data Resources

Digital Commons Resources
- DC Community Library Data Page
- Digital Commons for Data Handout
- Data Toolkit
- Data Interviews
  - Bepress Data Interview
  - Conducting a Data Interview by Michael Witt and Jake Carlson

LibGuides - LibGuides on Data Management
- University of Kentucky
- UMass Amherst
- UMass Medical
- James Madison University
- Missouri University of Science and Technology

Appendix Two: Metadata

Digital Commons provides a recommended metadata schema which is based on, in part, Data Management and Publishing recommendations from MIT, the DataCite schema, which is a recommendation of an international group of data services and libraries, Dublin Core, and existing data structures that Digital Commons’ users have created.

We review and update the schema every year based on changes in the field. Bepress Consultants are available to work with clients to identify the crucial fields necessary for each collection.

Here are a few examples of the metadata schema in action:
- University of Massachusetts Medical School
- Yale University: Tree Density Data
- University of Denver, Center for Orthopaedic Biomechanics: Natural Knee Data

Appendix Three

Boilerplate language for DMP and Digital Commons:

Many schools advise researchers on writing data management plans, including language about the role the repository can play in helping fulfill the open access obligations related to funding. Sample language on data storage and preservation is included here:

The IR is hosted on the Digital Commons platform, a high-availability hosted service that utilizes a distributed file system to provide fast and reliable access for content upload and delivery worldwide. The service offers unlimited storage with offsite backups of files on Amazon Glacier as well as quarterly full-site archives. Long-term accessibility to resulting, peer-reviewed content will be facilitated by bepress’ commitment to the provision of long-term stable URLs and the preservation of all objects uploaded to the repository in the original format. Beyond that, the service is committed to making PDF documents web-accessible on a permanent basis.

We’d love to hear more about data needs at your campus. If you have questions, stories to share, or other feedback, please let us know at outreach@bepress.com.