

EDITOR'S SUMMARY

At the 2015 ASIS&T Research Data Access and Preservation (RDAP) Summit, panelists from Research Data @ Purdue University Libraries discussed the organizational structure intended to promote campus-wide collaboration among diverse data specialists. Amy Barton explained her role as metadata specialist for research data and the importance of being engaged with evolving standards and best practices. As digital user experience specialist, Tao Zhang evaluated the Data Curation Profiles Toolkit, including an interview protocol librarians use to explore researchers' data needs and a web application for researchers to create data management plans. Line Pouchard, an information specialist focusing on big data, explored the data curation lifecycle model in general and its unique aspects in the context of big data. Information specialist for molecular biosciences, Pete Pascuzzi addressed the importance of bioinformatics for data reuse and having librarians work in an embedded position with researchers.

KEYWORDS

academic libraries	collaboration
research libraries	big data
data curation	bioinformatics
data models	information reuse
research data sets	
metadata standards	

Research Data Integration in the Purdue Libraries

by Lisa D. Zilinski, Amy Barton, Tao Zhang, Line Pouchard and Pete Pascuzzi

In 2014, the Research Data group @ Purdue University Libraries developed and rolled out a new organizational structure, leading to increased collaborations across campus. This panel talk at the 2015 ASIS&T RDAP Summit gave an overview of the new Purdue model of data services and in-depth looks into how the specialists and liaisons have incorporated data services activities and education into their responsibilities and research. In this article, the panelists discuss their roles and how they integrate in the libraries' overall strategy.

Metadata in the Context of Research Data at Purdue University Libraries

Amy Barton, metadata specialist

The Role of a Metadata Specialist in Research Data @ Purdue Libraries [1]

While there are similarities between working with metadata in a traditional, technical services/cataloging environment – acknowledging that even this area of

expertise is rapidly changing – metadata in the context of research data has additional requirements for description, documentation, discoverability, access and reuse. A metadata specialist working in this area therefore takes on a different role.

Due to the relatively new prospect of incorporating research data services in the libraries, with metadata being a component of such services, understanding the current state of metadata for data and to developing standards and best practices in data documentation requires research. Research and services development in turn informs how to teach and train students, researchers and library colleagues effectively. An organic distribution of activities emerges that address metadata research, services and education needs within the libraries and the university as a whole.

Beyond addressing local expectations and needs, an additional role for a metadata specialist is to engage with national and international activities to promote and impact standards development and maintenance, best practice,

Lisa D. Zilinski is librarian, University Libraries Scholarly Publishing, Archives & Data Services, Carnegie Mellon University. She can be reached at ldz@andrew.cmu.edu.

Amy Barton is metadata specialist at Purdue University. She can be reached at hatfiea@purdue.edu.

Tao Zhang is digital user experience specialist at Purdue University. He can be reached at zhan1022@purdue.edu.

Line Pouchard is computational science and big data information scientist at Purdue University. She can be reached at pouchard@purdue.edu.

Pete E. Pascuzzi is molecular biosciences information specialist at Purdue University. He can be reached at ppascuzz@purdue.edu.

harmonization and interoperability. One effective strategy is to become involved in organizational metadata working groups. For instance, I am a member and have contributed to the deliverables of the DataCite Metadata Working Group and the MetaArchive Metadata Working Group. I have also recently become involved with the Research Data Alliance Metadata Directory Working Group. As an example of the activities of these groups, the ORCID-DataCite Metadata Harmonization Working Group has worked to align the schemas for future metadata exchange and also contributed to an ODIN (ORCHID and DataCite Interoperability Network) project report [2].

Metadata in the Data Repository

The Purdue University Research Repository (PURR) is a research collaboration and data management solution for Purdue researchers and their project partners. It is a dedicated data repository through which curated datasets are published. As metadata specialist, I developed a metadata scheme based on PURR requirements, which included descriptive, technical, structural, rights and long-term preservation metadata. To address these needs, I wove together several metadata standards. These include the following:

- Metadata Encoding and Transmission Standard (METS)
 - METS is the wrapper metadata scheme that encompasses all the other standards and provides file structure metadata.
- Dublin Core (DC)
 - DC is used to capture the submitter-contributed descriptive metadata entered during the dataset publication process.
- Metadata Object Description Schema (MODS)
 - MODS is used to document the primary person

responsible for the dataset, as well as the access condition (embargoed or publically available), which is important for disaster recovery, de-accession and, although unlikely, takedown notices.

■ Preservation Metadata: Implementation Strategies (PREMIS)

- PREMIS is used to capture technical, rights, agent and event metadata associated with each file within the dataset throughout the dataset's preservation.

User Experience Research and Design for Data Services

Tao Zhang, digital user experience specialist

Assessing the Data Curation Profiles Toolkit [3]

The Data Curation Profiles Toolkit (DCPT) was created at Purdue as an interview protocol for librarians to engage researchers in discussion about their data [4]. It has been widely applied in various contexts, but its usability as a tool had not been formally assessed. To address the need, my colleagues and I conducted a survey of DCPT users. The survey included quantitative measures of potential influencing factors for using the DCPT and its perceived usability, as well as open-ended questions.

Applying the Technology Acceptance Model (TAM), Zhang used factor analysis to identify underlying factors for users' perceived usability and intention to use the DCPT. Regression models of the factors and perceived usability showed that applicability, experience and share, as well as training and help are positive determinants of the DCPT's perceived usefulness. The DCPT's perceived ease of use is positively affected by its applicability to different contexts and negatively by its complexity and interviewee requirements.

Open responses revealed several themes of users' concerns: time requirement, structure and format, and alignment with particular contexts. I correlated these themes with the quantitative results to suggest improvements, including developing a lighter and more adjustable version, improving its comprehensiveness and coverage of typical data management scenarios in research, and providing additional support in customization.

User Interface Design for Data Management Planning Tool

The DMPTool is a web application that guides researchers through the process of creating data management plans for funding agencies [5]. Its latest version has a number of new features, including plan co-ownership; customizable plan templates, guidance, and resources; individual and institutional profiles; and different user roles and privileges. These features created significant challenges for user interface and workflow design. I designed the interface prototype (wireframes), which involved multiple iterations of a three-step process: 1) convert the use cases and functional requirements into interface designs, factoring in user experience standards and best practices; 2) share the wireframes with stakeholders and evaluate whether the design meets their needs and expectations; and 3) based on feedback, identify elements of design that need to be refined or further defined. The wireframes were used as blueprints during the later implementation process.

Data Curation for Big Data

Line Pouchard, computational science and big data information scientist

Preserving big data for the long term is about preserving many series of processes that are interconnected and may be repeated several times during the research lifecycle. We examine the following activities below: plan,

prepare, analyze, preserve, describe and assure. While these activities are similar to those of other data curation lifecycle models, the tasks involved and the questions raised at each phase are specific to big data due to its characteristics.

1. At the *planning* stage, because of potential data volume and growth, the selection of data for preservation must be discussed, such as the decisions to keep raw data (some experiments are too costly to reproduce) or not (cheaper and easier to run a simulation or a sequencer again to obtain the raw data than preserve it).
2. *Preparing* datasets and staging them for analysis is time-consuming and its complexity is often overlooked. Data wrangling involves reformatting, cleaning and integrating datasets so that they can be analyzed and visualized.
3. The *analysis* activity is the domain of the scientists performing research. Recording and preserving the parameters of experiments is needed for the reproducibility of results. Objects that have not traditionally been part of data curation such as software and source code may be considered for preservation.
4. The *preservation* activity should include the creation of workflows that track dependencies between data and code. Preservation activities should aim to capture data transformations in order to address these challenges.

The describe and assure activities are encountered at every step of the life cycle.

1. *Describe*: Describing the data and processes used in the analysis at every step is crucial for big data.

Workflow tools exist, but they do not capture all the necessary metadata and documentation. Semantic tools that derive metadata from annotations and ontologies can help.

2. *Assure*: Acquiring data from multiple sources presents quality-related challenges. Sources may have different levels of quality resulting in a combined dataset with the lowest common denominator for quality.

The description of these activities and the questions they raise provide librarians with introductory material for the data interview and help structure the tasks required for big data curation.

Facilitating Research Data Management and Reuse as an Embedded Librarian

Pete E. Pascuzzi, molecular biosciences information specialist

I joined Purdue as part of a cluster hire in systems biology. My Ph.D. is in biochemistry and with postdoctoral research experience in bioinformatics. For this panel, I gave an overview of my research-data-related activities that emphasized my work in the areas of education and research. One aspect that unifies my work is the idea of data reuse.

I developed a graduate-level introductory course in bioinformatics. Discussions with faculty and graduate students in the biochemistry department identified the need for this course. It uses the statistical programming language R as well as Bioconductor, an associated bioinformatics project. The course has no prerequisites so it is often the first computer programming class for students, but, more importantly, it is often the first course in which students wrestle with complex datasets.

From the student's perspective, they are acquiring skills that enable them to answer complex questions on gene expression, DNA sequence composition and genome organization. However, the course has a hidden agenda on research data management, including use and organization of a data project on PURR (Purdue University Research Repository), acquisition of published datasets from public repositories, directory structure and file organization, use of metadata and standard file formats and interconversion.

I also collaborate directly with faculty, postdocs and graduate students on research data needs. Generally, this work encompasses data reuse projects. In two cases, these collaborations have resulted in co-authorship on peer-reviewed research publications. These collaborations lead to a greater understanding of researcher needs and have led to the development of workshops on web resources for cancer research and the development of a web application to assist researchers with data acquisition and visualization.

In collaboration scenarios, the question sometime arises: How embedded is too embedded? The consensus seemed to be the more embedded the better. Embedding leads to greater opportunities for the librarian and can take the guesswork out of liaison duties such as collection development. From the perspective of research data management, embedding is essential to fully appreciate the problems faced by researchers.

Conclusion

It takes a village is the theme for research data management at the Purdue Libraries. In addition to the panelists, the Purdue Libraries has many other data participants, including liaison librarians, the geographic information systems (GIS) specialist, the information

literacy specialist, the scholarly communication specialist, the digital data repository specialist and the data specialist. It is important that the libraries also work with other units

on campus, such as the Office of Research, Sponsored Programs and Computing Services, when integrating research data services. ■

Resources Mentioned in the Article

- [1] Purdue University Libraries. Research Data Services: www.lib.purdue.edu/researchdata
- [2] ODIN Consortium. (April 10, 2015). D4.2: Workflow for interoperability. Retrieved from <http://dx.doi.org/10.6084/m9.figshare.1373669>.
- [3] Zhang, T., Zilinski, L., Brandt, D. S., & Carlson, J. (2015). Assessing perceived usability of the Data Curation Profile Toolkit using the Technology Acceptance Model. *International Journal of Digital Curation*, 10(1), 48-67.
- [4] Data Curation Profiles: <http://datacurationprofiles.org/>
- [5] Data Management Planning Tool: <https://dmptool.org/>