Lightning Talk Summary

by Kristin Briney

Research Data Access & Preservation

EDITOR'S SUMMARY

Eight brief talks at RDAP 2014 gave an intriguing introduction to a variety of projects in data curation. EarthCube, an effort to build a cyberinfrastructure to manage and share data in the geosciences, and GlueX, a massive scale data management project, both face the challenge of building strong collaborations of diverse stakeholders to direct project development. Training and implementation are the focus for an intense team effort by Stanford University librarians to handle a repository backlog as well as for the University of Cincinnati's training for trainers to help researchers adopt good data management practices. Establishing a deposit workflow and creating and assigning consistent metadata are key to linking reports for Purdue University's Joint Transportation Research Program and to coordinating the National Snow and Ice Data Center's multiple databases. Data management and preservation extends to data related to the process as well as the end product in game design at the University of California, Santa Cruz. Speakers from the University of Illinois at Urbana-Champaign discussed the disheartening finding that the National Science Foundation's funding for research proposals did not correlate with their having data management plans.

KEYWORDS

data curation
information infrastructure
research data sets
metadata
access to resources
digital object preservation
training
grants

Kristin Briney is data services librarian at the University of Wisconsin-Milwaukee. She blogs about best practices for managing research data at dataabinitio.com. Kristin can be reached at briney<at>uwm.edu.

RDAP 2014 showcased a lot of great work in the field of data curation, and nowhere was this more evident than during the conference's lightning talk session. This session squeezed eight talks into 75 minutes with room to spare for questions. Here are highlights of the 2014 RDAP lightning talks. Slides from the 2014 RDAP lightning talks are available on Slideshare at www.slideshare.net/asist_org/tag/rdap14. Tweets from the session have been Storified and are available at https://storify.com/KristinBriney/rdap-2014-lightning-talks.

Rachael Black of the Arizona Geological Survey gave the first lightning talk, discussing the governance of the data management and sharing network EarthCube. EarthCube is a project sponsored by the NSF to build cyberinfrastructure in the geosciences. Part of Black's role in the project is to help develop a governance system for the project, a particularly challenging task due to the project's diverse stakeholders and EarthCube's mission to be community-led. Black described EarthCube's current efforts to assemble an advisory council, which will help guide the project and serve as a platform for community concerns. This project is an ambitious endeavor and it will be interesting to watch how it tries to grow from the bottom up.

Amy Hodge of Stanford University came next with an enjoyable talk on training her peers to deposit data into the institutional repository. Her solution? A deposit-a-thon. This two-hour event involved 15 librarians and 162 technical reports needing deposit. With a little help from a subject expert (and a few snacks), the librarians were able to deposit 137 items into the repository, all while growing more comfortable with the depositing process. Hodge cited the ability to have everyone in the same room to ask and answer questions as being key to the success of the event. Obviously, combining hands-on training with a real task to accomplish is a useful strategy to get our peers comfortable with data-related services.

Continuing on the theme of preserving reports in a university repository, Lisa Zilinski of Purdue University described a different approach: a direct collaboration between the library and members of the university's Joint Transportation Research Program. The key challenge in this project was linking reports with the, often multiple, datasets that support them. To address this challenge, Zilinski and her partners created a detailed deposit workflow that included data management planning, the creation of proper metadata and assigning DOIs to each dataset. Linking reports with the corresponding data through permanent identifiers brought value back to the researchers by enriching the original reports and improving information dissemination. Those interested in disseminating both reports and their corresponding datasets would do well to study Purdue's model for the process.

The University of Cincinnati library recently implemented the New England Collaborative Data Management Curriculum (NECDMC) (http://library.umassmed.edu/necdmc/index), and librarian Kristen Burgess described efforts to evaluate the program in her lightning talk. Burgess showed that researchers responded well to the concepts in the curriculum, though there is room for improvement. In particular, trainers need to provide solid case studies to work through and ample time to ingrain researchers with the concepts. Burgess also recommends that, while we may focus on teaching data management planning to our researchers, librarians and other data services members should operationalize good data management practices in our day-to-day work. By "practicing what we preach," we become better supporters for data management overall.

You can't have a conference on research data access and preservation without a talk stressing good metadata practices, and Lynn Yarmey was only too happy to fill this required role. Yarmey, who works at the National Snow and Ice Data Center, discussed the importance of data sharing and discovery, especially in light of the fact that her agency trades in data that monitors the effects of global climate change. Much of Yarmey's work focuses on the Arctic Data Explorer (http://nsidc.org/acadis/search/), which provides a federated search for arctic datasets by pulling from several related databases. Translating metadata among the different systems is only one part of the challenge of running this system, with funding issues, natural language handling and inconsistent data policies among the groups all being hurdles for the project team to overcome. Yarmey cited continual communication, staying focused and dedicating adequate resources to data sharing as key ways to keep this system running. And, of course, creating good metadata.

For those interested in taking on large data management challenges, the talk by Steve Van Tuyl of Carnegie Mellon University was of particular interest. Van Tuyl discussed the challenge of creating a data management plan for the GlueX experiment, which spans over 30 institutions and one national lab. The particular challenges of this task were handling the project's 15 petabytes of data per year, the fact that there is no one leader of the project (rather, there is a spokesperson who cannot make ultimate decisions) and dealing with university lawyers leery of sharing important data outside of the institution. Given this situation, Van Tuyl and collaborators' efforts have focused on developing a data management plan for GlueX members to adopt. I hope we continue to hear more about this project in the future, particularly if the adoption of a central data management plan is successful.

The talk that generated the most buzz came from William Mischo and Mary Schlembach of the University of Illinois at Urbana-Champaign (UIUC) and Megan O'Donnell from Iowa State University. This talk outlined a forthcoming paper that examines the differences between data management plans for NSF grants that were funded and those that were not funded. The researchers categorized over 1,200 plans from UIUC by storage and sharing practices and the university service being used. Ultimately, Mischo and collaborators found no significant difference between plans in grants that received funding and those that did not. This finding is unfortunate for those of us who would like to see better management of data and suggests that we can do more to help both researchers and grant reviewers understand best practices in data management. We are all looking forward to the publication of this paper later this year.

The session wrapped up with an interesting talk from Christine Caldwell of the University of California, Santa Cruz. Caldwell described an interesting problem in preserving games, which are increasingly being used in the digital humanities. Taking the game Prom Week (http://promweek.soe.ucsc.edu) as their case study, Caldwell and her collaborators recognized that

researchers value the process of game design in addition to the content of the game itself, which has important considerations to data management and preservation. The team therefore came up with a draft data management plan that addressed ways to better capture the creation process through code comments, consistent file names, sufficient documentation and good storage and organization practices. Data management in the field of gaming is still a fairly new prospect and Caldwell hopes that the community will soon develop some unified practices.

All together, the 2014 RDAP lightning talks presented a snapshot of interesting projects in the field of data curation. While talks discussed both finished and ongoing projects, it was valuable to see how real progress is being made on a wide range of data issues. ■