SPECIAL SECTION

Research Data Access & Preservation
by guest editors Andrew Johnson, Jennifer Doty and Lizzie Rolando

12] RDAP14 Summit: Introduction
by Andrew Johnson

14] Research Data Management on a Shoestring Budget
by Margaret Henderson, Regina Raboin, Yasmeen Shorish and Steve Van Tuyl

18] Collaboration and Tension Between Institutions and Units Providing Data Management Support
by Stephanie Wright, Amanda Whitmire, Lisa Zilinski and David Minor

22] NSF DataNet Partners Update
by Dharma Akmon

26] Funding Agency Responses to Federal Requirements for Public Access to Research Results
by Wendy Kozlowski

31] Learning to Curate
by Jennifer Doty, Joel Herndon, Jared Lyle and Libbie Stephenson

35] Lightning Talk Summary
by Kristin Briney

38] Metadata Use in Research Data Management
by Christie Wiley
ASIS&T’s Research Data Access and Preservation (RDAP) Summits further the dialog among the diverse community of stakeholders in this highly interdisciplinary area. Practice and research in RDAP are becoming increasingly prominent and urgent as the United States government and others mandate the accessibility and reuse of research data developed with public funds. Many information professionals, including librarians and archivists, are trying to educate researchers about government-required data management plans and data archiving options as well as providing tools, training and repositories to facilitate data archiving and discovery.

The special section of this issue of the Bulletin contains overviews of the panel presentations and discussions, the lightning round and the poster session at RDAP14, held in San Diego at the end of March, with many thanks to guest editors Andrew Johnson, Jennifer Doty and Lizzie Rolando. In addition to the topics mentioned above, panels also address funding and research projects as well as the plans by some of the government agencies, such as the National Oceanic and Atmospheric Administration (NOAA) and the National Library of Medicine, who were tasked to publish new policies to improve data access shortly after the completion of the Summit. We are very grateful to all the editors, the many Summit participants and our authors, who do an outstanding job of briefing us on the latest developments.

If you are hungering for more excitement after the World Cup, consider entering the ASIS&T Annual Student Design Competition, which will be held during the Annual Meeting in Seattle. Look Inside ASIS&T for Candy Schwartz’s write-up of last year’s challenge, which was won by the home team in Montreal, and for information on how the contest will be run in Seattle.

Seattle is also the site for the ASIS&T Board of Directors’ Annual Retreat, which will be held just before we go to press. On his President’s Page, Harry Bruce previews that event, which allows the members to meet face-to-face to review reports from committees and assess projects and policies and, one suspects, have some well-deserved fun. Thanks to all our officers, Board and volunteers for the work they do that keeps us moving forward.
At the time of writing this column I am in the midst of preparing for the ASIS&T Board of Directors’ summer retreat. I had hoped that the timing of this column would allow me to provide a report on the outcomes of the retreat, but at this stage I can only let you know what your Board will be discussing. During the year, the Board meets several times via telephone conference. These meetings generally run for 2-3 hours. We also regularly engage with one another via the Board’s listserv with concerns brought forward by members of the Board or matters of business that need Board approval, presented by ASIS&T executive director Dick Hill. Our summer retreat is a wonderful opportunity for members of the Board to come together to discuss issues facing our association and to receive updates on a number of major projects that are underway. It is also an opportunity to discuss plans for the future of our Association. This year our Board retreat will be held in Seattle on Saturday and Sunday, July 26 and 27.

The agenda for the Board retreat includes reports from all hard-working ASIS&T committees and task forces. Members of the Board are liaisons to one or more ASIS&T committees; they will provide written or oral reports on the activities of the committees on which they serve. Each committee has been working to achieve goals set for this year. Board members will learn about the progress of each committee toward meeting their goals. Board liaisons will also communicate any issues or concerns arising or additional support or guidance required. This part of our agenda is intended as a celebration of the wonderful engagement and hard work of each committee. We also want to communicate our support and encouragement to committee members and chairs. The success of our association depends so much upon the input and commitment of members serving ASIS&T in this way.
This year the Board has had a particular focus upon membership, both in terms of increasing the number of ASIS&T members in all constituencies (academic, professional and student) but also in terms of the level of engagement by all members with the association, and the benefits that come with ASIS&T membership. At the Board retreat, we will therefore be discussing at length the insights gained from the membership committee’s survey of – and interviews with – ASIS&T members. The committee has been gathering data on the benefit members feel they receive (or don’t receive) from their affiliation with ASIS&T; why some members are not renewing their memberships; what the association needs to do to retain existing members and to attract new members from all parts of the world; and how we can encourage more professionals to join our association. Ultimately, the committee will be presenting the Board with a number of recommendations. We hope to put these recommendations into action this year and beyond.

In terms of member engagement, the ASIS&T Board has for some time pondered how we might enhance opportunities for all our members to interact across the year. We have a very successful webinar series, but believe that we have not fully exploited the potential of social media and other web-based communication technologies to help our members feel connected to their worldwide network of ASIS&T colleagues and collaborators. At the board retreat, we will take time to further elaborate and discuss this significant aspect of member satisfaction. Additionally, the Board will receive updates on the progress of work on two important initiatives that we have sponsored this year: the social media initiative and the web presence and website redesign initiative. We have commissioned a number of active users of social media to begin work on postings that will attract the interest of members. We have also hired a consultant to work on a redesign of the ASIS&T website. At the Board retreat, this consultant will be reporting on progress made on the website redesign and seeking feedback from the Board. I am sure you will agree that a complete overhaul of the ASIS&T website is due. We need to project a web presence that is leading edge. All members of the Board are excited about moving this initiative forward. We expect that we will be ready to make a public launch of the new website at the ASIS&T Annual Meeting in Seattle in November.

Finally, members of the Board will also use their face-to-face time discussing several issues related to the ASIS&T Annual Meeting. We are very fortunate as an association to have a highly regarded conference to which members choose to submit their best work and one where they come to network with colleagues and to meet new friends and share new ideas. That said, the Board is ever mindful of the need to pay close attention to the expectations of all members of ASIS&T when it comes to our Annual Meeting. There will always be ways that we can improve our conference and better match it with the needs of ASIS&T members. For this reason, the Board will focus some attention on issues such as the length and structure of the Annual Meeting, the expectations of SIGs and chapters, acceptance rates, the balance of papers and panels, and enhancing external and internal perceptions of the value and quality of the conference for professionals, students and researchers. This discussion will continue one begun at our retreat last year in Montreal. We aim to make progress on these issues in Seattle.

So the end of the month of July will see members of your Board coming together to work on these important issues facing your association. All your Board members are volunteering their time and talent to the role of stewarding your association. I am honored to have such a fine group of individuals to work with, and I look forward to welcoming them to the University of Washington iSchool later this month. I am also excited by the prospect of welcoming all ASIS&T members to our Annual Meeting in Seattle in November. I guarantee that you will all have a great time in my hometown. Please consider inviting a friend to join us for the Annual Meeting and encouraging everyone you know who works in our field to join ASIS&T.

Best wishes for the summer!
Summer Balloting Completed for Election of New Board Members

As this issue of the Bulletin of the Association for Information Science and Technology was being completed and prepared for uploading to its publishing platform, the polls were closing on the Summer 2014 election of officers and directors of the association.

While the outcome of the election will be noted on the ASIS&T website and will be reported in full in the next issue of the Bulletin, here’s a report on the candidates who agreed to serve if elected to maintain the international excellence and presence of your organization.

President-Elect

Two candidates running for president-elect in 2015 and succeeding to the presidency in 2016 are Nadia Caidi and Jens-Erik Mai.

Nadia Caidi is associate professor at the Faculty of Information, University of Toronto. Trained in linguistics and communication in France, she then obtained an MLIS and Ph.D. from UCLA. Nadia’s research interests focus on information policy and community informatics. Her studies of information post 9/11 and on diasporic communities have been funded by several grants from SSHRC (Canada). A longstanding and active member of ASIS&T, Nadia served as chair of SIG/III, co-chair of SIG/IFP and co-organizer (with Michel Menou) of the Global Information Village Plaza (2001-04). She received the James M. Cretsos Leadership Award in 2006. Nadia was the president of the Canadian Association for Information Science in 2010-11.

Jens-Erik Mai is professor at the University of Copenhagen, Royal School of Library and Information Science in Denmark. He was previously associate professor at the Faculty of Information, University of Toronto, where he also served as vice dean and acting dean. Prior to that he was assistant professor at the Information School of the University of Washington where he co-directed the Center for Human-Information Interaction. He earned his Ph.D. in library and information science from the University of Texas at Austin as a Fulbright Scholar and master and bachelor degrees from the Royal School of Library and Information Science, Denmark. Jens-Erik is currently an ASIS&T director-at-large; he has been chair of SIG/CR and a member of the SIG Cabinet Steering Committee.

Directors-at-Large

The top two vote-getters among the following four candidates will serve three-year terms as directors-at-large on the ASIS&T Board of Directors: Chirag Shah, Jamshid Beheshti, Fidelia Ibekwe-SanJuan and Boryung Ju.

Chirag Shah, assistant professor of information science and an affiliate member of computer science at Rutgers University, studies interactive information retrieval/seeking, especially in the context of online social networks and collaborations, contextual information mining and applications of social media services for exploring critical socio-political issues. He is also interested in theoretical and practical aspects of information as a dynamic construct and online information propagation. Shah holds degrees from University of North Carolina, Chapel Hill; IIT Madras, India; and University of Massachusetts, Amherst. An active ASIS&T member since his student years, he has held leadership roles in both student and geographic chapters; chaired Annual
Meeting tracks and sessions; served as reviewer for ASIS&T publications; and attended every Annual Meeting since he joined the organization. In recognition of his early active participation, he was awarded the 2013 ASIS&T James M. Cretsos Leadership Award.

**Jamshid Beheshti** holds degrees in mining technology, history and library and information science from various Canadian institutions. He has taught at the School of Information Studies at McGill University for more than 25 years, and is currently the Associate Dean in the Faculty of Education at McGill University. Jamshid has been the principal investigator and co-investigator on more than a dozen Social Sciences and Humanities Research Council of Canada grants, and in collaboration with colleagues has obtained several million dollars in research funds over the past decades. He is widely published in leading journals of information science and technology. His expertise in designing information systems for children and young adults has been the topic of his contributions in his recently co-edited books, *The Information Behavior of a New Generation: Children and Teens in the 21st Century* (with Andrew Large, 2013) and *New Directions in Children’s and Adolescents’ Information Behavior Research* (with Dania Bilal, in press).

**Fidelia Ibekwe-San Juan** is professor of information and communication sciences at Jean Moulin University (JMU) in Lyon, France. Before joining JMU, she was associate professor at the University of Nancy, France. She earned degrees in French literature from University of Port-Harcourt, Nigeria, and Stendhal University, Grenoble, France, before embarking on a new career in information studies. She obtained a Ph.D. in information and communication studies from Stendhal University. Her research interests include using methods from computational linguistics, text analytics, information visualization and information retrieval to address information processing problems; the history and foundations of information and communication sciences; epistemology and the history of sciences. She has also been studying the impact of recent technological advances (open data, big data and web 2.0) on science and on the society. Since joining ASIS&T in 2009, Fidelia has served the association in many capacities, actively participating in SIGs and chapters, among other activities. She obtained the New Leader Award for 2012-2013.

**Boryung Ju** is an associate professor at the School of Library & Information Science, Louisiana State University. She received her Ph.D. in 2002 from Florida State University and a master’s degree from Indiana University. Her research areas include human-computer interaction and knowledge management. Some of her current research projects are on the impacts of collaboration technologies on knowledge sharing and interaction design for video digital retrieval systems. She co-chaired the 76th ASIS&T Annual Meeting and has actively participated in numerous scholarly and professional organizations. She currently is the managing editor of the *Journal of Information and Knowledge Management* and an editorial board member for the *International Journal of Information Science Theory & Practice* and the *Journal of the Korean Society for Library and Information Science*.

The International Calendar of Information Science Conferences (ICISC) is available through ASIS&T (Association for Information Science & Technology). Available in multiple languages. [https://www.asis.org/Conferences/calendar/](https://www.asis.org/Conferences/calendar/)
77TH ASIS&T ANNUAL MEETING

Two Keynote Speakers Help Shape Program for Upcoming Event

The 77th ASIS&T Annual Meeting, October 31-November 5, in Seattle, Washington, will be informed by the insightful presentations of two stellar speakers: Alessandro Acquisti, Carnegie Mellon University, and Kris Kutchera, Alaska Air Group. Operating under the general theme of Connecting Collections, Cultures and Communities, the conference committee, under the leadership of Jens-Erik Mai, selected the plenary speakers for their breadth of experiences in creating and managing connections between and among corporate units, academic specialties, customers and consumers.

Kris Kutchera, who will deliver the opening keynote address on Sunday, November 2, is vice president, information technology and strategy management for Alaska Air Group. She is responsible for managing and developing computer systems, information systems and voice and data communication for Alaska Airlines and Horizon Air. Alaska Airlines is the latest step in Kris’ career, following stints that began at Boeing Computer Services and Capgemini America and moved through leadership positions at Washington National Gas, Fluke Corporation and F5 Networks. Kris holds a bachelor of science degree from University of Wisconsin-Madison and an MBA from University of Washington, where she also serves on the Founding Board of the iSchool.

On Monday morning, November 3, Alessandro Acquisti, associate professor, information technology and public policy, Heinz College, Carnegie Mellon University, will address a plenary crowd. Alessandro is also co-director of the CMU Center for Behavioral Decision Research. His primary research interests are economics and behavioral economics of privacy and privacy in online social networks. His findings have been featured in numerous leading journals, edited books and conference proceedings, as well as in popular and business-related media outlets such as the Economist, NPR, CNN, New York Times and the Wall Street Journal. Alessandro holds a Ph.D. from University of California, Berkeley, and master’s degrees from Berkeley, London School of Economics and Trinity College Dublin. In a previous life, Alessandro was a classical music producer and label manager, freelance arranger, lyrics writer and soundtrack composer for theater, television and independent cinema productions.

Pre- and Post-Conference Workshops and Seminars

Information professionals once again have the opportunity to brush up on...
subjects that are changing at rapid paces or to learn something entirely new with an array of pre- and post-conference offerings at the Annual Meeting in Seattle. All workshops require separate registration and payment of fees; workshop attendees need not attend the Annual Meeting. See sidebar for the days, times and brief descriptions of courses that will be available this year. See the ASIS&T website for more details or look for the printed version of the preliminary program to arrive at your home or office in September.

**And for Students Only**

For the fifth year running, all students attending the ASIS&T Annual Meeting have an opportunity to participate in the student design competition in which groups of interested students are formed and given a design challenge to complete in less than 48 hours. Though the competition is intense and rapid, the reward is worthwhile: Each member of the winning design team receives free registration for the next ASIS&T Annual Meeting. For a report on last year’s exciting competition, see the article by Candy Schwartz and Rhiannon Gainor on page 10 of this issue of Inside ASIS&T.

**Conference Venue**

All formal conference sessions and events will be held at the Sheraton Seattle Hotel in Seattle, Washington. Special discount rates are available until October 9 or until the block is sold out, whichever comes first. Contact the hotel directly (206-621-9000) for more details or to book your room.

---

**FRIDAY, OCTOBER 31, CONTINUED**

9:00 a.m.-5:00 p.m.

**Digital Management** – Many library and information professionals are now being asked to plan, initiate and deliver digital projects. For those with limited experience in such projects, this session will provide a solid foundation upon which to build and help determine the right approach and methodology for a successful project.

9:00 a.m.-5:00 p.m.

**Trust & Information Policy in the Age of Data (Big or Small) (SIGs/III and IFP)** – Join researchers from an array of disciplines to address the moral, legal and policy questions raised by information creation, access and control in modern society.

9:00 a.m.-5:00 p.m.

**User Experience Research to Improve Access to Your Collections** – Librarians, archivists and other professionals with responsibility for providing digital access to collections will learn how to use methods and tools developed by the field of information architecture to improve access to those resources.

9:00 a.m.-5:00 p.m.

**Understanding Web Search Engine Users** – Researchers from academia and industry will discuss methods and results regarding research on web search engine use and user behavior, aiming for establishment of a research agenda for the future.

---

**SATURDAY, NOVEMBER 1**

8:30 a.m.-12:30 p.m.

**The Discipline of Organizing in iSchools: Collaborative and Digitally Enhanced Teaching of a Core Subject** – This half-day workshop focuses on the digital-enabled teaching of Information Organization and Retrieval, a core course in the information studies domain, using the book The Discipline of Organizing (TDO).

8:30 a.m.-12:30 p.m.

**10th Annual Social Informatics Research Symposium Connecting (Epistemic) Cultures and (Intellectual) Communities (SIG/SI)** – SIG/SI continues its research symposium to disseminate current research and research-in-progress that investigates the social aspects of information and communication technologies across all areas of ASIS&T.

8:30 a.m.-5:00 p.m.

**25th Annual SIG/CR Workshop: Universal Classification in the 21st Century (SIG/CR)** – Featuring research papers, position papers and lightning talks, this workshop will bring classification schemes, in all their history and complexity, into a fresh relationship with emerging information infrastructures and articulate new and exciting roles for these schemes to play in the years ahead.

Continued on following page
Pnina Fichman, previously director of the MLS program and co-director of the Rob Kling Center for Social Informatics at Indiana University (IU), is the new chair of the Department of Information and Library Science at IU. She takes over from Debra Shaw, who retired at the end of May.

Eileen Abels, dean of the Graduate School of Library and Information Science, Simmons College; Linda Smith, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign; and Lynne C. Howarth, Faculty of Information, University of Toronto, have received a grant from the Institute of Museum and Library Services for Educate to Innovate: Re-visionsing Library and Information Science Education. The project will focus on the importance of leadership and cutting-edge skills in LIS education.

Ali Shiri, formerly assistant professor, has been promoted to the rank of professor at the School of Library and Information Studies, University of Alberta. The promotion recognized his research, service and teaching contributions to the school, its students and the library and information field at-large.

The Simmons College Graduate School of Library and Information Science adds two new assistant professors to the faculty for this fall. Chaoqun Ni, whose research interests sit at the intersection of science policy, science and technology studies and scholarly communication in the context of data-intensive science, is completing her doctorate at Indiana University. Amber Stubbs, with a Ph.D. in computer science from Brandeis University, teaches courses in data structures and algorithms, programming and theory of programming languages and information retrieval.

Special Interest Group/Social Informatics (SIG/SI) seeks nominations for two best paper awards to be given this fall. The Best Student Paper Award honors the best paper written by a doctoral student on a topic informed by social informatics. Nominations may include a published paper in a peer-reviewed journal during 2013 or a conference paper presented in 2013. The Social Informatics Best Paper Award honors the best paper published in a peer-reviewed journal on a topic informed by social informatics during the 2013 calendar year. The winning authors will be invited to present their papers at the 10th Annual SIG/SI Symposium at this year’s Annual Meeting in Seattle. The deadline for nominations for both awards is August 15, 2014. For more information and full submission rules and guidelines, please visit the SIG/SI website at http://asistsigsi.wordpress.com
Soon it will be time for the 5th annual ASIS&T student design competition at the Annual Meeting in Seattle. I hope this recap of what happened in 2013 will encourage students to participate and others to come and cheer them on. The design competition is intense but fun. Each year teams are formed on the first day of the meeting (Sunday), and a design challenge is presented. Two days later, the teams come together to present their ideas to a distinguished panel of judges. Each member of the winning design team receives complimentary registration to the next ASIS&T Annual Meeting.

Last year, after several rounds of matching up (group speed-dating), four teams emerged:

- **The Fabs** (Rong Hu, Nathan Lowrance, Nada Naji, Esperanza Pacheco, Afroza Sultana)
- **Go Canada** (Rhiannon Gainor, Petronilla Muriithi, Zhen Yue, A Yusuf, Qing Zou)
- **Gryffindor** (Nicole Alemanne, Katie Chan, Guillermo Galdainez, Xiao Xie)
- **Flaming Squirrels!!!** (Kate Dillon, Wei Jeng, AJ Million, Philippe Mongeon, Fei Shu)

They were presented with the 2013 challenge, “All the Things”:

Today’s researcher has (synced) folders of full-text files; citation management systems packed with references (and more full-text); clipping tools containing yet more items to follow up on; and “read it later” and “favorite” collections stored by browsers or RSS readers. This personal collection and any public persona information (blogs, Twitter, etc.) are ignored during the process of discovery in the massive open or subscription-based scholarly databases.
we use to find new information. Coyle (2011) [1] points out that “[b]y having such large databases to search we are increasing our odds of finding everything in the world about our topic. Of course, we probably do not want everything in the world about our topic, we want the right books (articles, etc.) for us.”

Develop a preliminary design for a system that makes use of a researcher’s personal collection and online persona to enhance discovery in large databases.

Teams were sent off to work on their designs, with a reminder to also attend sessions and enjoy the conference. Two days later we gathered again. Each team had seven minutes to present its design and five minutes to field questions from the distinguished panel of judges: Jamshid Beheshti, McGill University; Nick Belkin, Rutgers University; Sandra Erdelez, University of Missouri; Jim Jansen, Penn State University; Liz Liddy, Syracuse University; and Howard Rosenbaum, Indiana University. Each judge scored the presentations for creativity, impact on solving the problem, feasibility of the solution, contribution to humanity and quality of presentation, including the question and answer session.

Each of the teams contributed interesting ideas to the mix, including apps entitled Definitely Nut Mendeley (Flaming Squirrels!!!, of course), Harmonious (Team Canada) and EurekaPortal (Gryffindor). The judges submitted their score sheets, and after a few minutes of furious addition by the design committee, Team Canada emerged from a close contest as the winner (appropriately for a conference held in Montreal). Member Rhiannon Gainor was kind enough to describe her team’s design for this report:

[They] interpreted “enhanced discovery” as having two aspects: personal information management and information retrieval. Their goals were to propose a way to improve (simplify, make more relevant) large database searches, improve user time management (automate searching, make retrieval and discovery quicker) and decrease information fragmentation of search results. Several user features were proposed for the interface, including click-and-drag tagging and graphical clustering of files by assigned tags and scraped subject terms assigned by the scholarly databases, to reduce re-finding. A user-controlled search display provided side-by-side comparison of the enhanced searches with standard searches of scholarly databases. However, the most innovative aspect of the system design was the attempt to combine a recommendation algorithm for learning the researcher’s online persona and personal collection and a relevance ranking algorithm for bringing disparate information elements together in a way that could be controlled by the end user for building and tailoring not only external database searches, but also internal searching of the user’s own computer.

First, to capture the researcher’s online persona, Yahoo! Pipes was recommended since it provides a graphical user interface for building data mashups of Twitter feeds, blogs, CiteULike, etc. The personal collection of the researcher would be captured with elasticsearch, an open source search and analytics engine that permits the indexing of text items such as Portable Document Format (PDF) and HyperText Markup Language (HTML) documents. Lastly, the system would allow users to enter personally developed user factors, such as weighting certain subjects or terms more heavily than others, or filtering out results that already exist in the personal collections. These three elements, the online persona, the user-built personal collection and user factors, would inform the recommendations and relevance rankings of the proposed system. The team members were thrilled by their win and would like to express their appreciation to the organizers and the judges and to ASIS&T for sponsoring their awards.

For my part, I would like to give my thanks to all the students, the judges and especially my fellow committee members Michelle Kazmer, Florida State University, and Rong Tang, Simmons College, with assistance on the opening day from judge-to-be Sandra Erdelez, University of Missouri. Drs. Kazmer and Tang will carry the competition on in 2014.

Resource Mentioned

The 2014 ASIS&T Research Data Access & Preservation Summit (RDAP14) was held March 26-28, in San Diego, California. In this special section of the Bulletin, for which Jennifer Doty, Lizzie Rolando and I served as guest editors, RDAP14 session leaders and presenters reflect and expand on their experiences at the Summit and the research data issues that were explored in their respective sessions. Special thanks to all of the presenters, session leaders and planning committee members as well as to our sponsors:

- ASIS&T SIG/DL (Special Interest Group/Digital Libraries) for sponsoring and providing refreshments for the poster session.
- ASIS&T SIG/MET (Special Interest Group/Metrics) for sponsoring a keynote presentation and a coffee break for attendees.
- Cray, Inc. for sponsoring a coffee break for attendees.

This year’s Summit focused primarily on practitioners working with research data (and often the researchers who produce that data) at various types of institutions, including university libraries, data centers and repositories, and research funding agencies. We heard from individuals operating in a number of roles ranging from teaching courses on data management to assisting with the creation of complex data management plans to developing policies for research data. Many of the discussions and themes that emerged from the Summit revolved around defining these roles (and who should play them) along with the need for collaboration between and within institutions.

In her opening keynote, MacKenzie Smith provided a perfect example of the complex collaboration needed to solve major research data challenges through her discussion of the SHARE project (www.arl.org/focus-areas/...
Similarly, Maryann Martone kicked off the second day of the Summit with a keynote that covered a disciplinary collaborative effort aimed at enabling discovery of and access to research data: the Neuroscience Information Framework (http://www.neuinfo.org/).

In addition to the two keynote presentations, RDAP14 featured a number of panels on topics of great interest to those of us working with research data, including how to begin to build services with few resources, how to develop policies for research data and how to collaborate both with other institutions and with other units within our own institutions. We also heard updates on important national efforts like the NSF DataNet program and the White House Office of Science and Technology Policy’s requirements for public access to federally funded research data.

The lightning talk and poster sessions provided additional opportunities to hear about the full range of work underway at universities, data centers, disciplinary repositories and other institutions. For the first time ever at RDAP, two post-conference workshops gave attendees the chance to develop skills that they could bring back to their institutions. The discussions at the Summit were lively, and feedback from attendees was overwhelmingly positive. We hope to continue these conversations next year at RDAP15!

For additional information on the RDAP14 Summit and for plans for the 2015 Summit, please refer to the following resources:

- View slides from presentations and a selection of posters from the Summit are available on slideshare: www.slideshare.net/assist_org/tag/rdap14.
- View the full RDAP14 program, with additional information on Summit themes and presentations: www.asis.org/rdap/program/.

Join the RDAP community’s on-going conversation and stay updated on RDAP15 event information by signing up for our email listserv at http://mail.asis.org/mailman/listinfo/rdap.
Research Data Management on a Shoestring Budget
by Margaret Henderson, Regina Raboin, Yasmeen Shorish and Steve Van Tuyl

EDITORS SUMMARY
In response to mandates from the federal government, many academic librarians face a new challenge to become service providers for research data management. A panel at RDAP14 convened representatives from Virginia Commonwealth University, James Madison University and a collaboration of seven New England libraries to discuss their strategies in response to the regulations when faced with limited resources. A common theme was to take advantage of work already done by others, including faculty surveys and existing data management resources such as DMPTool and Data Curation and Profiles Toolkit. Spreading the word on data issues to spark collaboration and gaining the support of library administration are key. Positive response from stakeholders can provide momentum to request additional resources and provide instruction on data management planning.

KEYWORDS
academic libraries funding
data curation collaboration
research data sets strategic planning

Changes to the regulatory environment around research data management over the past few years, specifically the mandates to manage research data from the National Institutes of Health (NIH) and the National Science Foundation (NSF), have resulted in a great deal of research and discussion about roles, responsibilities and infrastructures for research data curation. Results include a large body of tools such as DMPTool, Data Curation and Profiles Toolkit, training programs (at various MLIS granting institutions) and reports. One major result, however, is that academic libraries, research administration offices and IT units at colleges and universities in the United States are stepping into the role of service providers for research data management.

In February 2013, moreover, the White House Office of Science and Technology Policy (OSTP) changed the landscape of the discussion around research data management by hinting strongly at more widespread requirements [1]. The consequence, in the communities of academic librarians and affiliates who are trying to assess how to approach this problem, is a broad uptick in activity around research data management at institutions that were previously not engaged on the topic.

We convened this panel on shoestring budgets to discuss the experiences of a variety of institutions standing up research data management services on a relatively short timeline and with limited resources. Panelists represented data services providers at an assortment of institutions (a research university, a master’s comprehensive university and a multi-university collaborative) and discussed a range of issues related to creating functional research data management programs with few resources.
The Research University

Virginia Commonwealth University (VCU) is a large, urban research university with over 30,000 students on two campuses, with a library on each campus. Less than a year ago, the libraries hired a director of research data management (DRDM) – a new position to both the libraries and the university.

Given the size of the research data management problem, the DRDM and an associate university librarian created a plan that relied on the second mover advantage – the idea that VCU could learn from the successes and mistakes of other institutions in order to build its program. They decided to forego one element of data collection that is very common at many institutions initiating research data management programs – the faculty survey. There were already too many surveys running at the university, including an Educause survey by the IT group, and the Ithaka S+R Faculty Survey, which was going to be sent out by VCU libraries in early 2014. By reading articles and other resources, such as the Purdue Data Curation Profiles [2], they were able to build a picture of what was going on in our institution.

VCU’s final plan consisted of these elements:

1. Create a web presence/communication plan. Set up research guides with lots of links and tools and work with the library public relations (PR) department to create handouts. The PR department also promoted the creation of the DRDM position online and in some newsletters.

2. Conduct an environmental scan for data and data management resources. Com through websites, call people to see what they can help with and compile a resource list for research data management. Through this process, they found that some individuals felt territorial, and it was necessary for the library to ensure them that the program was there to fill in the service gaps, not take over.

3. Pick the low hanging fruit. Many data services have started by helping with NSF data management plans (DMP), just like many relationships have started with NIH Public Access Policy discussions, and more will start with the OSTP memo.

4. Talk to researchers, students and others involved in data. Identify faculty, students and administrators who are interested in data issues and people who are looking for help. Through these interactions the library has formed collaborative relationships where possible. One key set of relationships is with technology services and a couple of divisions of the University’s Office of Research.

5. Educate everyone you can. Seek to engage colleagues, administration, faculty, staff and students on the topic of research data management. The staff spoke to and taught classes for as many groups as possible.

Lastly, the backing of library administration was very important. Starting any sort of new service is hard, and knowing the administration is supporting these efforts makes things much easier.

The Master’s Comprehensive University

James Madison University (JMU) is a predominantly undergraduate institution of nearly 20,000 students in the Shenandoah Valley in Virginia. This master’s comprehensive university used existing resources to create data management services, built on that success to advocate for additional resources and laid the groundwork for future data curation services. Prior to the fall of 2011, JMU lacked any formal infrastructure to assist researchers in meeting the National Science Foundation’s (NSF) data management plan (DMP) requirement. With the hiring of a new science librarian, the library began to offer support to researchers via DMP consulting, albeit in an ad hoc manner given the lack of additional resources to offer a more comprehensive, systematic approach. Despite the lack of additional funding or release time, the library began investigating cross-campus collaborations and outside resources to create a more robust support system. The library initiated coordination with a variety of campus partners to determine researcher needs on campus and the limits of the support available to respond to those needs.

While the NSF requirement was the impetus for starting these conversations, it became clear that there were other areas where offering data management support would be appreciated. While all parties on campus recognized that there was a need that required a systematic response, there were no additional funds or personnel available to create a product to meet that need. Utilizing existing tools, such as the DMPTool,
allowed JMU to quickly address researcher concerns, while investigating what measures could be taken to build additional support.

The library also used information learned from hosting the ACRL Scholarly Communication Roadshow in 2012 to perform an environmental scan of the campus. A taskforce of library faculty met with every college and leadership organization on campus to discuss scholarly communication needs, including data management and curation. Undeniably, constituents wanted the library to take a leadership role in providing support for created digital works, including research data. The positive response by faculty to these actions positioned the library to ask for additional resources to expand services. JMU is currently rolling out an institutional repository and has recently hired a metadata librarian and digital collections librarian to help support any future data curation efforts.

These positions meet current needs and have the potential to be involved in building infrastructure for additional, more curatorial, efforts like RDF and linked data. While it is possible that the library will grow data services in the future, there is still a need to improve the existing infrastructure and to continue to process digital project backlogs. Could the increase in service stimulate enough growth to warrant additional positions, such as a scholarly communications librarian? Or additional staff to aid in digital collections? These considerations are currently unknown, although it may be advantageous to have a single, discipline-agnostic champion in the library who can speak to nontraditional scholarly products such as data.

The Multi-Institution Collaborative

A new service and role for librarians is teaching research data management (RDM) best practices to undergraduate and graduate students and researchers. Funded by the National Library of Medicine (NLM) with the principal investigator and project coordinators from the University of Massachusetts Medical School Lamar Souther Library, an open, case-based, modular course for librarians to use to teach RDM was developed by seven libraries in the New England region. The New England Collaborative Data Management Curriculum (NECDMC) [3] is an open curriculum, with a CC-BY license. NECDMC is an instructional tool for teaching data management best practices to undergraduates, graduate students and researchers in the health sciences, sciences and engineering disciplines. Each of the curriculum’s seven online instructional modules aligns with the National Science Foundation’s data management plan recommendations and addresses universal data management challenges. Each module provides a lesson plan, PowerPoint activities and resources that can be used out-of-the-box or customized to fit an institution’s or audience’s needs. In addition to the workshop materials, the NECDMC site provides additional information and training content [4].

After launching the online course, the researchers conducted a webinar to introduce librarians to RDM and writing data management plans (DMPs). Following the webinar, the researchers held an in-person train-the-trainer workshop [5] for librarians who had attended the webinar from the United States and Canada. The workshop introduced participants to the course materials and provided instruction on teaching the curriculum. Post-workshop, the NECDMC training team identified a sample of librarians interested in piloting the course in their settings. After completing the pilot course at their libraries, the participants conducted a post-course self-evaluation, describing their participants, setting and methods used. They also conducted post-course evaluations of the students’ perceptions of the course’s content and of the methods and conduct of the course. The researchers analyzed these results and then conducted follow-up qualitative interviews with the pilot instructors. The website also has Community Bulletin Board and Join the Collaboration sections that provide information on how participating libraries are using NECDMC and links to resources used and developed by the partners and pilot institutions.

Eighteen institutions have enlisted to use NECDMC, and of these, nine are in the process of piloting the curriculum. Examples of these pilots are a semester-long, for-credit course; librarian professional development; workshops for graduate and undergraduate students, professional organizations and conferences; library-school for-credit course; and one module per week for researchers. Recently, Module 1 was translated into French and presented at the University of Montreal [6].
Conclusions

The institutions represented on this panel are just a few among many that have initiated new services around research data management with little financial or personnel support. This lack of support is not insurmountable, and the members of this panel have outlined a number of service points that can be provided under these conditions. Broadly categorized they include the following:

1. **Understand the research data landscape.** The VCU and JMU examples, while different in execution, highlight the importance of surveying the landscape of research data support needs when planning research data services. This environmental scan can be internal or external, based on a formal survey or on meetings with stakeholders, can include formal or informal reporting. Regardless of how the scan is conducted, this step is crucial to targeting services to the right audience and at the right level.

2. **Take advantage of existing resources and tools.** Each of the panelists in this session identified existing resources and tools provided by external entities (e.g., DMPTool, the Data Curation Profiles Toolkit, etc.) or units within their organizations (e.g., public relations offices, local administration). While lack of resourcing may be a long-term issue, there are many tools at one’s disposal to springboard research data services at one’s institution.

3. **Collaborate to create new tools.** The NECDMC case is a great example of how a group of collaborating institutions has identified an internal need (the need for a data management curriculum) and has, in turn, provided tools to help others with that need. The success of NECDMC is a testament to the value of collaborating to create tools and services. Some of the aforementioned resources and tools, such as DMPTool, were created in this same spirit of collaboration.

While the discussion on this panel was helpful for framing the challenge of providing research data services in a resource-scarce environment, it was clear from the questions during the session that these problems were common among panel attendees, regardless of institution size and type. It is crucial that communities in academia tasked with providing research data management services continue to share tools, ideas and inspiration to the benefit of all.

---

**Resources Mentioned in the Article**


[4] NECDMC Slides from Workshops: http://library.umassmed.edu/necdmc/slides_from_workshops


Collaboration and Tension Between Institutions and Units Providing Data Management Support
by Stephanie Wright, Amanda Whitmire, Lisa Zilinski and David Minor

EDITOR’S SUMMARY
A panel of speakers from three universities explored their challenges and progress in building programs to support research data management, whether within the library system or with research offices or computing groups. Since 2012 Oregon State University has partnered with its research office and graduate school, helping students prepare data for preservation and sharing and developing a graduate course for credit in research data management. Based on needs identified through an environmental scan, the University of Washington hired a data services coordinator to promote the services provided and to increase collaborations, visibility and support. Purdue University pairs data services specialists with subject liaison librarians to reach disciplinary faculty and researchers. The connections identify champions, lead to successful collaborations and, most importantly, provide the opportunity to show data services specialists as peers and collaborators. With basic services established, each institution looks forward to strengthening relationships and expanding services, skills and staffing.

KEYWORDS
data set management collaboration
research data sets academic libraries
strategic planning

Many academic libraries across the country are building programs that address data management issues on their campuses. These programs vary in size and complexity, from single-person operations to entire departments. The specific services offered also vary, often involving consultation services, training, software support and help with writing grants. Regardless of the specific characteristics of these efforts, they all reside within larger spheres. Whether this sphere is the library organization (with multiple departments or programs interacting with research data) or the larger campus (with organizations such as research offices and computing groups), there are impacts on priorities, implementations and ownership.

This panel looked at how three different universities are approaching these issues. Each presenter addressed her own environment, laying out the challenges and successes she has had. Here the presenters continue the discussion, focusing specifically on two emergent themes from the panel: the importance of campus champions and thinking long-term.

Identifying Campus Champions
OREGON STATE UNIVERSITY (OSU). OSU began developing research data...
services relatively recently – in the fall of 2012. As a new campus service, we are still working to raise awareness among faculty and administrators regarding the spectrum of assistance that we offer. We reached out to the research office (RO) and the graduate school (GS) to initiate conversations about faculty and student needs on campus and to look for areas of collaboration. The RO is very supportive of our efforts to develop data services and acknowledge that we have not only the relevant expertise but also the strong relationships with faculty across campus that enable us to provide a centralized, well-utilized service. The RO is also delighted to have someone to refer faculty to when they have questions about data-related funder requirements. The RO has made it clear that they expect the library to take the lead on the continued development and expansion of research data services, and they see us as a partner in supporting the OSU research enterprise.

The graduate school has become an unexpected source of support and partnership in the development of data services for graduate students. We already work closely with the GS to facilitate the mandated self-deposit of electronic theses and dissertations (ETD) to our institutional repository, a requirement for graduation. During a conversation with the GS about requiring concurrent deposit of the data that underlies an ETD, we had to acknowledge that most students likely do not have the skills necessary to adequately prepare and document their datasets for long-term preservation and sharing. This realization prompted OSU’s data management specialist to develop a two-credit, graduate-level course [1] in research data management and offer it under a GS course designator (at the time of course development, the library lacked a designator and thus the ability to offer a credit-bearing course).

UNIVERSITY OF WASHINGTON. The University of Washington Libraries began its foray into forming a data services program by charging a planning committee to perform an environmental scan of current research data management (RDM) support services available on campus. This scan made it clear that, while several groups on campus were offering data management support, these services were a) scattered across the university campuses and departments, b) narrow in scope of population served and c) frequently not well-known. The report submitted by the planning committee in 2010 recommended (among other suggestions) hiring a data services coordinator to act as a central resource for communication about and referral to data services available for university researchers. In short, this person was to be a data concierge. The position would also bear responsibility for identifying where gaps in RDM support existed and for developing services to address those needs.

The obvious challenges in meeting those goals were related to serving a large, decentralized institution and the fact that most researchers do not automatically think of the library when they need help with data management. To address these problems, the data services coordinator spent a lot of time reaching out to the groups identified in the environmental scan as providing RDM support services. She used invitations for coffee instead of requests for meetings to insert herself into busy schedules and have one-on-one conversations with key personnel in groups around campus regarding their issues and challenges and to get more information on services they provide. These folks became the initial names on our list of champions. In fall of 2012 the data services team distributed a survey to PIs and co-PIs on research projects across all three campuses. Over 300 researchers responded and, of those, over 100 agreed to participate in hour-long follow-up interviews. Those who were interviewed were added to our growing list of champions.

Thanks to these activities, the services offered by our data services program became more widely known much more quickly than if we had stuck to our traditional methods of marketing. With our increased visibility in the data management arena, the libraries have collaborated on several related projects and events with various groups around campus. Among these are co-hosting events with the iSchool and the College of the Environment, drafting a data repository proposal with a champion from UW-IT and being invited to be on a working group for a multi-year, multimillion dollar grant related to data science with the eScience Institute. There are plans in the works for more collaborations: with the Office of Research on a project with ORCID IDs; with the Office of Sponsored Research on a project to customize the new version of CDL’s DMPTool and integrate data management curriculum into new researcher orientations; and with the Institute of Health Metrics & Evaluation to provide access to global health datasets.
PURDUE. It was important for Purdue’s data services specialists to partner internally with the subject liaison librarians when reaching out to disciplinary faculty. Accordingly, training and support for the liaisons have been important aspects of the data services specialist role. The data group has acted as the bridge between liaison librarians and various campus stakeholders, such as disciplinary faculty, information technology, the Office of the Vice President of Research (OVPR) and research centers. In doing so, the data services specialists were able to identify campus champions who further the data services initiative on campus by talking about the services with their colleagues, offering to pilot new potential services and collaborating as partners in research. Examples of successful collaborations with stakeholders include librarians partnering with faculty as principal investigators on grants; the collaborative and collective development of an institutional research data repository; and co-teaching with faculty in areas of data management, data information literacy and data curation. Since there is no clear line between where the liaison librarian ends and data services begins, data services specialists and repository specialists often complement the liaison librarians and are available to attend the consultations with disciplinary faculty.

One issue that we occasionally run into is that some disciplinary faculty may not see us as peers or collaborators, but rather as just support services. However, the more testimonials that faculty can give, the more we are asked to speak, teach, consult and collaborate. Just in the first quarter of 2014, the data group at the Purdue libraries has been able to deepen our relationships with the OVPR by participating in a Rolodex of Campus Research Services and the research expo and with the College of Agriculture by presenting a poster and lightning round talk at the Big Data in Agriculture Symposium. Participating in such events across campus creates opportunities for the data group and the libraries to continue to develop existing relationships, as well as foster new ones that will lead to creating new champions on campus.

Thinking Long-Term
OREGON STATE UNIVERSITY. As a newer service on campus, those of us in research data services have spent most of our time thinking about short-term goals and approaches to meeting them. We started small, with only one FTE dedicated to providing data services support. It was important for us to recognize that we could not be all things to all people. We had to decide what was within the scope of our abilities and our resources and focus on doing those things really well. These things now include the following:

- providing information on data management best practices (via a website, http://bit.ly/OSUData);
- offering in-person consultation support for activities like writing data management plans, planning data management strategies for projects already underway, finding appropriate methods to share data and so on;
- providing education and training via workshops and our credit-bearing course; and
- expanding the scope of our institutional repository to include the preservation of datasets.

Now that we have basic services established, we are looking ahead to areas of expansion. We are continuing to explore the feasibility of supporting graduate students in preserving and sharing the data that underlies their theses or dissertations via deposition of the data to our institutional repository. Preparing graduate students to deposit their data would not only provide them with lifelong practical skills in proper research data stewardship, but would also result in the preservation of critical data at a time when it is vulnerable to being lost. We will soon begin launching an online survey of OSU graduate students to better understand the nature and extent of the data that they are generating and what their current level of data management skills is.

We are also investigating how to implement targeted cultivation of data skills for liaison librarians who show interest or inclination and who serve in units where it especially makes sense to do so. In order for our services to continue to grow, we will need a capable front line to head off basic questions about common data management concerns or topics in the liaison area. For example, it makes sense for liaison librarians to have a basic understanding of the data types common in their areas of disciplinary purview or to know the data-sharing policies of the most popular journals for OSU authors in their discipline(s). A reasonable scope of what data-
related knowledge and understanding should be expected from liaison librarians will be fleshed out through conversations directly with them.

**University of Washington.** The data services program has gone from one FTE dedicated librarian to 1.5 FTE librarians (with the addition of our data services communication and curriculum librarian), along with recent approval for two graduate students at 0.3 FTE each. We have also requested approval for a geospatial and numeric data services librarian in the coming year. Our proposal for an institutional data repository in February included budgeting for a repository librarian, as well. This request brings the potential of dedicated data services support up to 4+ FTE in the next biennium.

If the data repository is approved, it will be a large undertaking that will take much of our focus in terms of service expansion, training and outreach for the next year. We continue to slowly introduce new data management skills to our subject liaisons, including presenting a data management workshop at the upcoming Western Science Boot Camp for Librarians and plans for a Software Carpentry Boot Camp in late 2014. With the appointment of our dean of libraries to the position of vice provost for digital initiatives last year, we hope to use this opportunity to be able to engage our champions in development of some institutional policies related to research data management.

**Purdue.** Working with graduate and undergraduate programs has also led to some interesting partnerships and has uncovered unexpected champions. For example, we collaborated with the summer undergraduate research fellow (SURF) program at Purdue to find ways to introduce undergraduates to the concepts of data management. We have created a research data management LibGuide for undergraduates ([http://guides.lib.purdue.edu/undergraddata](http://guides.lib.purdue.edu/undergraddata)) and have uncovered a new champion on campus. This is an area of future growth for Purdue.

It has been important for us to become the resource rather than just point researchers to resources. By continuing to collaborate and participate with disciplinary faculty and PIs on grants, we strengthen our relationships and grow our voice on campus. This expansion cannot be done in a vacuum, and the support and collaboration between the data services liaisons and disciplinary liaisons is critical for creating and maintaining these relationships. A liaison may face challenges while trying to balance subject-specialist responsibilities with data management opportunities, but with additional professional development opportunities, support from library and university administration and continued testimonials from disciplinary faculty champions, this team-based model should be successful. Taking the team perspective will allow us to continually overcome challenges and provide a higher level of sustainable service.

---

**Resource Mentioned in the Article**

In 2007, responding to a scientific research context that was increasingly data-intensive, the National Science Foundation (NSF) Division of Advanced Cyberinfrastructure created the DataNet funding program. The purpose of this program was to create a set of “exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning” [1]. The original plan called for a $100 million initiative. The NSF would make five awards of $20 million each, to be distributed over five years, with the possibility of continued funding after the initial five years. In the first review cycle (proposals due in March 2008), the NSF awarded two DataNets: DataONE and the Data Conservancy. The second round of DataNet funding was delayed as a result of the 2008 global financial crisis. Ultimately, awards in the second cycle were reduced from $20 million down to $8 million per project, and three additional projects were awarded DataNet funding in 2011: SEAD, DataNet Federation Consortium and TerraPopulus.

The DataNet Partners share the goal of developing cyberinfrastructure that advances science; however, they differ in the specific communities and needs they aim to serve as well as in the particular problems they seek to address. Representatives from each of four DataNet Partners came together on an RDAP14 panel to present an overview of their projects and to update RDAP participants on important developments and plans going forward.

DataONE

As a product of the first round of DataNet funding, DataONE – a collaboration of multiple institutions based at the University of New Mexico – is the most established of the projects represented on the panel. In her talk, Amber Budden (DataONE’s director for community engagement and outreach) highlighted DataONE’s mission to enable universal access to
environmental and earth science data. DataONE (www.dataone.org) carries out this mission through three main sets of activities: community building, data discovery and interoperability solutions, and development of scientific tools and services.

Budden highlighted a number of DataONE’s community-building efforts, in particular the establishment of several working groups and the development of a rich set of educational outreach activities and resources. DataONE’s working groups serve as the foundation for the project’s research and education activities. These groups include data preservation and metadata, distributed storage, and usability and assessment working groups (among several others). Educational outreach programs are another DataONE venue for building community around data curation issues. Specific education resources and programs include a best-practices database that provides users with recommendations for working with data at each stage of the data lifecycle; a set of education modules on data management that users can download and incorporate in their own lessons; and a librarian outreach kit aimed at alerting librarians to DataONE’s most relevant products. DataONE actively enlists the participation of multiple stakeholder groups – through the DataONE Users Group and other outreach mechanisms – in disseminating and using these resources.

DataONE has also worked to enable data discovery and interoperability, developing a network of what it terms member nodes and coordinating nodes. Member nodes – of which SEAD, Dryad and USGS (U.S. Geological Survey) are three examples – are preservation-oriented repositories that have agreed to expose their data through the DataONE API, allowing the data to reach a larger audience. The three coordinating nodes at the University of New Mexico, the University of Tennessee Oak Ridge Campus and the University of California, Santa Barbara retain a complete metadata catalogue for the data, index those metadata for search and ensure content availability.

At the time of the presentation, DataONE had over 20,000 users, 34 production and in-development member nodes, 462,000 data objects, 13 tools and an active and diverse community of collaborators and partnering projects. Looking forward, Budden reported that the DataONE team is particularly focused on organizational sustainability planning to ensure that the project continues beyond the current NSF award.

**SEAD**

SEAD (Sustainable Environment Actionable Data) is a collaborative project between the University of Michigan, Indiana University and the University of Illinois to build cyberinfrastructure for sustainability science – a multidisciplinary area of research that addresses human impacts on the environment.

SEAD manager of outreach and education, Dharma Akmon, emphasized that SEAD supports active and social curation of data by embedding data curation within tools that support data creators’ early work with data. SEAD (sead-data.net) allows scientists to more easily manage data as they work and then leverages scientists’ early data curation work to facilitate the data’s long-term access.

SEAD is made up of three sets of connected tools and services: the Active Content Repository (ACR), a secure area where project teams can manage their data; the Research Network and the Virtual Archive (VA), which is the data preservation layer for SEAD.

The ACR is intended to support data creators’ work with data. Using the ACR, scientists can preview data in a web browser; annotate and describe datasets; automatically extract metadata; and create collections. Additionally, researchers are able to add geographic metadata to their datasets and view map overlays of geospatial data – features that are particularly welcome in sustainability science.

The SEAD Research Network contains over 1,800 profiles for researchers working in sustainability science. This network allows researchers to link their data and publications to a public profile that provides a record of their output. Users can also view visualizations of a researcher’s co-author network and disciplinary affiliations.

Lastly, the VA leverages existing institutional repositories to ensure long-term access to data. Researchers can deposit data by marking them “ready for publication” from the ACR, or they can bypass the ACR and submit a collection directly to the VA. The VA matches the collection to an
appropriate repository and hands it off for curator review. Published data are assigned a DOI, and the metadata are indexed to facilitate discovery.

SEAD is expanding outreach efforts through workshops that introduce scientists to SEAD’s tools and collect feedback for future developments. SEAD is also currently making its user interfaces more intuitive and refining a research-object lifecycle model that will aid in future iterations of the VA and ACR.

**Terra Populus**

Terra Populus (known as “TerraPop”) is based at the Minnesota Population Center and focuses on enabling the integration of data on population and the environment. David Van Riper, director of spatial data at TerraPop, highlighted the project’s overarching goal to lower barriers that make it challenging to conduct interdisciplinary human-environment interactions research. To that end, the TerraPop team is building tools that allow users to combine seemingly disparate data from multiple sources into new, customized data extracts.

Scientists’ understanding of population-environment dynamics could be significantly enhanced by the combination of different types of data. For example, a researcher might want to combine census microdata that describe socioeconomic/demographic information about individuals, raster data that characterize rainfall and vector data that depict land use. Normally a time-consuming process to carry out, with the tools TerraPop is developing, researchers will be able to select the particular variables from the different datasets of interest, choose the format they would like their data to be in, and create a new data extract.

TerraPop is at a fairly early stage of development. Going forward, the team will expand TerraPop collections to include more census, survey and global environmental data as well as GIS mapping files. This expansion will be accomplished, in part, by developing data curation protocols and mechanisms that will make it easier for scholars to deposit their data and metadata with TerraPop. Shortly after our panel, TerraPop (www.terrapop.org) released a beta version of its customized data extraction and creation service. In addition to collecting feedback from users about the beta version, TerraPop is working on becoming a DataONE member node, ramping up outreach efforts, and forming collaborations with the other DataNet Partners.

**DataNet Federation Consortium**

The DataNet Federation Consortium (DFC) – based at the University of North Carolina at Chapel Hill – aims to assemble a national data infrastructure in support of collaborative scientific research. Mary Whitton, DFC project manager, underlined three main areas of DFC activity: federating iRODS (Integrated Rule-Oriented Data System open-source data management software)-based grids and making those grids interoperable with other systems; enabling reproducible science through iRODS workflow data objects and extending the iRODS policy-based data system to better support creation, use and management of federated collections. In her presentation, Whitten focused primarily on describing the DFC’s work to facilitate interoperability and long-term data management.

The DFC (datafed.org) serves data producers, data users, curators/archivists and data center managers from any discipline. Targeting data producers and users, the DFC is developing client-side tools for creating metadata templates and bulk-uploading data to iRODS collections. The project team is also implementing iRODS policy-based access controls to enforce access rules – such as IRB (Institutional Review Board) restrictions – and allow a researcher to enable data access for a defined group of users in one step. Plans include providing integrated access to data analysis tools (for example, R and MATLAB) from iRODS workflow objects. With data curators and archivists in mind, the DFC is writing iRODS rules to implement archiving best practices such as ISO 16363 (repository trustworthiness) and ISO 14721 (open archival information system). The DFC’s development of data-center collection management tools promises to streamline functions such as integrity checking and replication. Future iRODS rules and policies for machine-verifiable best practices and standards will help establish iRODS grids as trustworthy repositories. Additionally, iRODS can now automate execution of important repository tasks such as copy, backup and checksum.
Whitton also highlighted iRODS version 4, which was released shortly after RDAP14. This version embodies current software engineering practices, easier installation and a plug-in model for future extensions. Significant interoperability efforts are ongoing between iRODS grids and DataONE, and between iRODS grids and the Dataverse Network. As with the other DataNet Partners, outreach will be an important component of the DFC’s work going forward.

**Conclusion**

The DataNet Partners have developed and are continuing to develop infrastructure to support 21st century science. Each of the four panelists emphasized the vital role that outreach will play as they cultivate communities, solicit feedback, promote their offerings and work together. The RDAP14 panel was a timely opportunity to update data librarians, curators and archivists on DataNet Partner activities and to invite feedback on the ways in which the partners can support these constituencies’ work with scientists and their data.

**Acknowledgements**

I would like to thank Amber Budden, David Van Riper and Mary Whitton for their participation on the panel and for their review of relevant sections of this article. The projects discussed are funded by the NSF under cooperative agreements #OCI-0830944 (DataONE), #OCI-0940824 (SEAD), OCI-0940818 (Terra Populus) and OCI-0940841 (DFC).

---

**Resource Mentioned in the Article**

Funding Agency Responses to Federal Requirements for Public Access to Research Results

by Wendy Kozlowski

On February 22, 2013, John P. Holdren, director of the federal Office of Science and Technology Policy (OSTP), released a memorandum for the heads of executive departments and agencies about increasing access to the results of federally funded scientific research. In the memo, the administration reiterated its commitment to ensuring, to the greatest extent possible, availability of federally funded scientific research results to the public, industry and the scientific community. The OSTP memo directs “each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government” [1]. It specifies a preference for agencies to work together to develop these plans and goes on to lay out objectives for access to both scientific publications and scientific data. Key drivers for this push toward increased access include maximization of “the impact and accountability of the Federal research investment” and enhancement of “innovation and competitiveness by maximizing the potential to create new business opportunities…” [1].

Affected agencies had until August 22, 2013 (six months) to submit a proposal. The OSTP, together with the Office of Management and Budget (OMB), reviewed and provided a first round of feedback on the drafts in February of 2014. Agencies had 90 days from that point to revise and resubmit final plans for review; it was in that window of time that this panel was held at the 2014 RDAP Summit.

The speakers on this panel shared updates on their agencies’ efforts toward addressing the Public Access to Research Results (PARR) memo. The speakers were Jeff de La Beaujardière, data management architect at the National Oceanic and Atmospheric Administration (NOAA); Amy Friedlander,
staff associate in the office of the assistant director, National Science Foundation (NSF), directorate for social, behavioral and economic sciences; and Neil Thakur, special assistant to the National Institutes of Health (NIH) deputy director for extramural research. Neal Kaske, chief, central library and information services division of the NOAA central and regional libraries, joined the group for the question and answer portion of the session.

**NOAA Plans for Improving Public Access to Science Research**

De La Beaujardière began by describing the NOAA research domains as diverse and complex, with a wide variety of data types and formats; methods of data production, gathering and dissemination; and data collection purposes. Adding to the complexity, NOAA comprises five major units, each with established but sometimes incompatible and entrenched methods for data management (for example, the tornado warning system). Despite these challenges, interconnections and interdependencies among them exist, allowing NOAA to fulfill mission goals. Long before the OSTP memo was released, NOAA’s vision was that all their environmental data should be discoverable, accessible and usable and preserved for all types of users across a variety of applications. The new federal mandate applies to NOAA-funded data and has provided guidance toward meeting these goals.

De La Beaujardière noted that there are two recent federal initiatives involving open data: the 02-22-2013 OSTP PARR memo [1] and the 05-09-2013 OMB Open Data Policy memo [2]. Both apply to NOAA, and, while there is some overlap, the scope is different between the two, with the former affecting the grantees and the latter the NOAA employees and projects. Existing guidance for NOAA grantees requires data sharing in a timely fashion (no more than two years). In addition, NOAA projects must already plan how they will document, preserve and distribute their data, and NOAA requires ISO19115 metadata for discovery, use and understanding. New to NOAA will be sharing publications; they have a library, but NOAA repository efforts have focused on data archiving. NOAA’s exact response to the PARR memo was still under review and could not be shared at the time of the panel, but de La Beaujardière did mention some potential impacts on their grantees, the largest of which will be in the area of data sharing.

Some of these possible outcomes include requiring data management plans (DMPs) from grant-issuing NOAA programs; requiring or encouraging use of the NOAA National Data Center for long-term preservation; more specific templates for grantees DMPs; inclusion of data sharing costs in proposal budgets (at the discretion of the grant-issuing agency); clarification or standardization of how funding sources are indicated in papers (for example, FundRef [3]); and adoption or establishment of methodology to ensure publications are broadly available after a suitable embargo period. Regardless of specific details, it was stressed that there would be an attempt to minimize additional effort on the part of the grantees, and that in general, feedback from OSTP/OMB has been positive for NOAA’s proposed plan.

Finally, the NOAA presentation detailed recent data-related activities. De La Beaujardière talked first about the Dataset Identifier Project, which uses DataCite DOIs to link National Data Center data and metadata to publications (21 DOIs had been issued at the time of the RDAP conference, 66 at the time of preparation of this publication). Second, we learned about the NOAA Data Catalog beta-version website (http://data.noaa.gov), which is a CKAN-based catalog of datasets harvested from NOAA data centers (~48,000 datasets at the time of the RDAP conference, ~54,000 at the time of preparation of this publication). Finally, de La Beaujardière described NOAA’s recent Big Data Partnership Request for Information, with the goal of increasing return on investment by using commercial enterprise to make NOAA’s large body of datasets quickly available, at a large scale, to public users (see also [4]).

**The NSF Public Access Initiative – Where Are We?**

Amy Friedlander began by talking about not only the role of communication in maintaining transparency to taxpayers, but also its importance for progress in science and technology. Scholarly communication allows for vetting, validating and reproducing scientific findings, allows others to build on previous results and encourages innovation and the transition to products and services, a key goal of these recent data access initiatives. Friedlander went on to review the NSF’s long-standing policy on data sharing, which includes the expectation that investigators will “share
with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants” [5, 6]. Building on this base, NSF began requiring data management plans (DMPs) as supplemental material for all proposals in 2011 [7], and those DMPs are evaluated as part of the merit review process. In 2013, the NSF began allowing datasets to be cited as relevant work products in biographical sketches [8] independent of the publication(s) based on them. In addition, the NSF allows request of funds to support publication and preparation of data for deposit as part of the budget proposal, and the http://nsf.gov and http://research.gov sites allow the public to search awards, research descriptions and publications (but not data).

Friedlander went on to describe the NSF’s approach to public access as one that must consider the needs of researchers and investigators that study a broad diversity of science, come from a range of institutions, work with various publishers and often have the support of multiple funding streams. This diversity inherently impacts researcher behavior, and the NSF will deploy its plan in phases, allowing ongoing innovation as implementation occurs. As with NOAA, plan details could not yet be discussed, but the initial focus will be on sharing publications. NSF currently has no internal repository for content, nor have they shared plans to build one, so expectations are that content will go in discipline repositories. This practice will require interaction with other federal agencies and the research communities, as well as integration of internal systems within the enterprise architecture.

Looking forward, NSF expects to make no large changes to their existing requirements and practices because of the PARR memo. Once approved, the NSF will post its plan to the http://nsf.gov website, and any changes to current procedures will be announced with nine-month windows for notice and comment. More specific guidance may also be released at the program, division or directorate levels.

Public Access: NIH’s Update

Keeping with the pattern of including a description of the agency’s current open access efforts, Neil Thakur began his talk with a brief review of the NIH data sharing policy that has been in effect since 2003 [9]. These are among the key points: 1) data sharing plans are required for funding applications seeking $500,000 or more in direct costs in any year, or researchers must state why data sharing is not possible, 2) reviewers do not factor the proposed data-sharing plan into determination of scientific merit or priority scores and 3) specific program announcements may request data-sharing plans for proposals with less than $500,000 in direct costs.

NIH does not expect great changes to publications policy because of the PARR memo. Current requirements mandate that institutions and investigators are responsible for assuring copyright agreements are consistent with submission to PubMed Central (PMC). Upon acceptance for publication, authors are responsible for depositing the paper to PMC and then properly citing those articles with the PMC identification number in applications, proposals and reports as evidence of compliance. Thakur did touch on ways that institutions can help to ensure compliance with NIH policies, including training and author support for policy awareness, submission of manuscripts to PMC and proper preparation of citations. Also mentioned was offering support on understanding policies surrounding publishing agreements, communicating with publishers on behalf of researchers and the possibility that institutions might take a role in ensuring (or at least monitoring) institutional compliance.

Finally, Thakur stressed that preparation is key to avoiding funding delays. Authors need “plans that can withstand” forgetfulness and miscommunication among authors and between authors and publishers [10]. NIH encourages researchers to use My NCBI’s My Bibliography service [11] to track their own compliance, associate papers with awards as soon as possible (don’t wait until it’s time to write a final report or apply for a renewal or new grant) and think about compliance plans as they write their papers, not at the last minute.

Commonalities, Differences, Impacts and Responses

Data management planning guidance is currently vague. Both de La Beaujardière and Friedlander acknowledged the lack of concrete guidance for writing and evaluating DMPs. NOAA discussed the possibility of
offering additional guidance as they move forward, but the NSF believes it is important for disciplines themselves to define what is important. Perhaps some combination of these two will shake out eventually, with communities helping to concretize best practices for planning. At least the initial stages of this process already seem to be happening within certain disciplines, as seen in documentation prepared for the marine science community [12], the social science community [13], the ecology community [14] and others. Twitter discussion during the panel included whether or not this lack of specific guidance might actually lead to more anxiety on the part of grantees, as well as the reality that researchers do not necessarily know where to start designing DMPs [15].

There is a common desire to reduce impacts on researcher workload. Acknowledging the inevitability of some level of increased burden, all three agencies talked about the importance of making access to research results as least onerous as possible. Currently, NIH policy puts responsibility for publication upload to PMC in the hands of authors, but does offer several discipline-specific repositories for data. NSF discussed the possibility of using institutional and/or discipline repositories for research outputs, which, depending on implementation, may be a complicated and expensive alternative to (in the eyes of the researchers and institutions) agency-hosted repositories. NOAA has automated the move of project data from its repositories to its data catalog for increased discovery, but has not committed to that for grantees data, nor have they detailed how they will address sharing of publications. Given the top-down directive that plans be implemented using resources from within existing agency budgets [1], it will be interesting to see how well agencies are able to minimize administrative burden on researchers. Audience skepticism, based on ensuing twitter comments, seemed high, with specific concerns about library involvement in compliance monitoring and questioning where burden will actually fall (“if not the researchers, then who?”) [15].

Libraries have a potential role in supporting researchers. Friedlander and Thakur mentioned several places that libraries might play a useful role in this migration to openly shared research results. Proper citation of publications, increased citation of datasets and linking of research outputs to grants were suggested as important places to focus. Everyone needs to consider who will actually do data management. One option is to increase embedded librarian efforts with researchers to provide or facilitate these activities. Other places that libraries are a likely fit are in training and outreach roles with the goal of increasing awareness and, when appropriate, researcher skills.

Despite the fact that panelists were not able to divulge details of their PARR memo plans, it was encouraging to hear that after over a year of waiting, there is indeed movement in this arena. Given the range of current positions and priorities of impacted agencies, it was also clear that there is no panacea for sharing research results. Nevertheless, it would appear that we can expect to see collaborative, iterative solutions that will evolve over time. As research and library professionals, one of our tasks will be to watch this space and communicate emerging requirements to the constituents we serve.

Resources Mentioned in the Article


Resources Mentioned in the Article, cont.


Learning to Curate
by Jennifer Doty, Joel Herndon, Jared Lyle and Libbie Stephenson

EDITOR’S SUMMARY
Three data specialists reviewed their experiences learning about and applying the Inter-university Consortium for Political and Social Research (ICPSR) processes and tools for curating research data. A small virtual community discussed curation theories on data acquisitions, review, processing, metadata and dissemination and shared progress implementing the ICPSR workflows and tools. Curators at Duke University, dealing with data on political donors, found processing obstacles from incomplete and mismatched data and faced confidentiality questions. At Emory University, gaps in coded data on home schooling practices revealed problems arising from lack of preplanning for long-term archiving and research and the need to clarify data needs for later re-use. By applying ICPSR processes, curators at UCLA’s Social Sciences Data Archive were able to improve their workflow and understand the work necessary for open archiving. All participants gained from the opportunity to practice ICPSR curation practices, realized the resource demands and saw the value librarians can provide by consulting with faculty on data management and preservation.

KEYWORDS
data curation
research data sets
archival science
barriers
training

Jennifer Doty is research data librarian at Emory University. She can be reached at jennifer.doty<at>emory.edu.
Joel Herndon is head of data and GIS services at Duke University Libraries. He can be reached at joel.herndon<at>duke.edu.
Jared Lyle is director of curation services at ICPSR. He can be reached at lyle<at>umich.edu.
Libbie Stephenson is director of the UCLA Social Science Data Archive. She can be reached at libbie<at>ucla.edu.

Research Data Access & Preservation

While many venues exist to meet and discuss data curation topics – from listservs to conferences – few opportunities arise for data curators to engage in personalized, but collaborative, hands-on work using the tools of an established domain repository. During the latter half of 2012, a working group of data specialists sought to address this need by meeting regularly (via online meeting) to discuss data curation topics and apply them to actual processing of data using the internal workflows and tools of the Inter-university Consortium for Political and Social Research (ICPSR). Working group members included Ron Nakao, Stanford University; Jared Lyle, ICPSR; Rob O’Reilly and Jennifer Doty, Emory University; Joel Herndon, Duke University; Libbie Stephenson, UCLA; and Jon Stiles, UC Berkeley.

Virtual sessions combined hands-on work and discussions of the theories behind the practice. Participants curated their own data and shared their own novel methods for improving the data curation experience. The following topics were included:

Acquisitions
■ Gathering and collecting information from the data producer (how much should/will they contribute?)
■ Legal agreements
■ Appraisal
■ What to keep, and for how long?

Review
■ Quality review - are the data complete, accurate, and well documented?
■ Disclosure review - is there sensitive or private information?
■ Creation of a plan of attack
Processing

- Data cleaning
- Insuring data integrity
- Quality review - is the final package self-contained?

Metadata

- Standards overview
- Variable level metadata
- Study level metadata

Dissemination

- Final packaging and review
- Workflows
- Preservation policies
- Web delivery

Following are the perspectives from three of the working group participants.

**Learning to Curate @ Duke - Presidential Donor Survey 2000-2004**

At Duke, the ICPSR Learning to Curate project provided an opportunity to learn best practices in social science data curation from ICPSR while evaluating the feasibility of providing data curation as a service in Duke libraries. By participating in this project, we hoped to get a better sense of the amount of effort required to identify, process and publish data collections created by Duke researchers using ICPSR’s curation standards. As a result of the curation project, we have created a section of our repository explicitly for datasets and now have a much better sense of the opportunities and challenges inherent in incorporating faculty created data collections into the library’s institutional repository.

Data for this project came from a team of political scientists wishing to share their survey data on the characteristics of presidential donors during the 2000 and 2004 presidential elections (http://hdl.handle.net/10161/7882). The research team that conducted the study had always intended to share the data and had already documented the content of the associated data files. Additionally, we had direct access to one of the principal investigators, Alexandra Cooper, who frequently provided invaluable context for processing the study.

Despite the initial work by the research team to document their work, the curation team encountered a range of challenges in processing the files. First, the documentation occasionally omitted information necessary for secondary use and occasionally did not match the values that we found in the dataset. We expect that these alignment errors are probably present in most data projects produced by large research teams since it is difficult to document for secondary usage when your goal is to produce the primary research. Another challenge we encountered on the project was dealing with issues of confidentiality. The IRB requirements prohibited directly revealing the respondents of the survey, yet the dataset initially contained a large amount of demographic and geographic information on respondents. After some negotiation, the principal investigators agreed to provide some of the demographic information as long as it could be aggregated at a level that would make disclosure much less of a concern.

Overall, the Learning to Curate project benefited our data curation workflow in three ways. First, the experience allowed us to see the curation process at ICPSR firsthand, providing a much higher appreciation of data curation at ICPSR and the efforts of their staff to produce quality data sets. Second, it provided us with a much better sense of the resource implications of providing quality data curation services to researchers on campus. At the end of the project, we realized that providing human mediated data curation proved extremely resource intensive. However, the project also raised many questions about the quality of data collections that do not receive a high level of processing/screening as they are archived. Finally, the project reaffirmed that the libraries could play a valuable role consulting with faculty on the best ways to manage and preserve data as a scholarly object.

**Learning to Curate @ Emory – Home Schooling Policies in the US 1972-2007**

Emory’s participation in the ICPSR Data Curation Working Group coincided with the library having completed the ARL/DLF E-Science Institute in 2012 and subsequently having hired a new specialist for research
data management. We anticipated that we would learn quite a bit from the experience of working with ICPSR's processing pipeline and tools for data archiving. Long recognized for being the gold standard among social science data archives, ICPSR has set an example of best practices in data curation across many disciplines, with references to their data management guidelines appearing in multiple federal agency data sharing plans. As we were considering the development of appropriate services at Emory to assist researchers in preparing and depositing data for long-term access and preservation, we also felt it could prove useful to determine the implications of providing a premium level of service for both staffing and resource allocation.

The dataset we used for the ICPSR project consists of coded data on home-schooling policies in the United States from 1972-2007. Data was taken from a variety of publicly accessible and/or freely available sources (National Conference of State Legislatures, Census Bureau, National Center for Education Statistics), so we had no issues with proprietary data or sensitive, human-subject research to consider while preparing the data for archiving and sharing. However, the data had been assembled for a particular project and was not documented with long-term archiving and research reuse in mind.

Concurrent with the virtual group meetings, we worked through the dataset and identified issues requiring further clarification and adequate documentation for potential reuse. There was further back-and-forth with both the principal investigator (PI) and the graduate student researcher who had done much of the actual documentation and assembling of the data. It was to our advantage that we already had a long-standing relationship with the PI and that one of us had actually been involved in helping her locate relevant data sources for the project. Even with that prior knowledge, there was still a lot of clarification involved to make the data fit for long-term archiving.

This clarification did create a challenge in terms of using the ICPSR's Secure Data Enclave (SDE). Much of our processing work took place outside the SDE, since we were getting follow-up datasets and further documentation from the grad student who was our primary contact on the project (the PI had left Emory for another school). The amount of work involved just for this dataset raised valid questions of how to handle potential issues with datasets from other investigators (for example, sensitive information or lack of prior knowledge). What is required to provide such levels of service to researchers, and what are the resulting implications in terms of personnel, budgets and similar factors?

Our participation in the project was worth the time and effort involved. It was both impressive and informative to peer behind the curtain of the ICPSR data archiving process and see how the sausage is made, so to speak. And it aided us in answering some of those questions about how to provide assistance to researchers in preparing data for deposit. With existing staffing levels and resources, we could not realistically allot this degree of our time to curating individual datasets. But with the knowledge we have acquired about data archiving best practices, we feel comfortable providing consultations and guidance to PIs, especially if they have resources to bring to the table, such as graduate research assistants and funding support. And last but not least, we have the satisfaction of assisting this particular dataset in finding a good home for the benefit of both the original investigators and future researchers.

**Learning to Curate @ UCLA – The Los Angeles County Social Survey**

At the UCLA Social Sciences Data Archive (SSDA), we identified three goals we hoped to achieve through our participation in the project and use of the ICPSR curation tools:

1) develop a better understanding of the use of tools in data curation processes;

2) compare our local workflow with the ICPSR pipeline process; and

3) provide enhanced processing for legacy files.

Established in the mid-1960s, the SSDA facility at UCLA is a small domain-specific archive of data used in quantitative research, and the collection consists primarily of surveys, enumerations, public opinion polls and administrative records, many of which have been deposited by UCLA researchers.
SSDA carries out a data quality review of all files, creates detailed DDI-compliant metadata using tools from Colectica, carries out format migration as needed and processes data for online analysis with Survey Documentation and Analysis (SDA). SSDA is a member of the Data Preservation Alliance for Social Sciences (DataPASS) and the SSDA holdings are shared with DataPASS partners. For the ICPSR project, the Los Angeles County Social Survey (LACSS) was chosen because it is a uniquely held study, is one of the SSDA legacy files and would benefit from enhanced processing.

The learning curve created some initial barriers to effectively using the ICPSR curation tools. However it was immediately obvious that the ICPSR tools helped to improve the overall quality of the study for long-term preservation. By studying the processing pipeline at ICPSR, we were able to refine and streamline local workflows at the SSDA. The data quality review aspects showed how essential these steps are if the goal is to carry out the OAIS (Open Archival Information System) defined process for providing that data will be independently understandable for informed reuse. In contrast to the hands-on aspects of gold standard curation, we still have some qualms about the use of self-deposit tools where there is no data quality review, because it has already been shown that relying on researchers to do this deposit will result in significant data loss.

Some final thoughts in conclusion: there are new needs for institutions to work collaboratively with units external to libraries and archives, develop institution-wide policies for what to preserve, determine how to sustainably fund this work and develop and hire a trained workforce. We reiterate the need to recognize the kind and scope of commitment the preservation of usable research data will require. Institutions will need to evolve in terms of having a workforce with new responsibilities, in building and using data management tools beyond the institutional repository and in reallocating funding to support these endeavors.

Acknowledgements

We would like to thank our fellow participants in the ICPSR data curation working group: Ron Nakao, Rob O’Reilly and Jon Stiles.
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 access to resources
information infrastructure digital object preservation
research data sets training
metadata 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.

Research Data Access & Preservation

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.
Metadata Use in Research Data Management
by Christie Wiley

EDITOR’S SUMMARY
Prompted by federal requirements, academic institutions and librarians have become more involved in data management planning. An e-research implementation group was formed at the University of Illinois to develop and promote data initiatives in a variety of ways, including updating a data management website to provide education and awareness to students and faculty. Part of the site focuses on documentation and metadata, describing traditional and nontraditional data types and sources, including lab notebooks, data files and database contents, workflow and operating procedures. The site details the importance of data documentation for the immediate research project and for the institution, explains metadata and provides links to metadata standards and resources.

KEYWORDS
metadata
data set management
research data sets
strategic planning
academic libraries

The National Science Foundation and other agencies require researchers to develop data management plans as part of their grant applications. These data management plans can identify types of data being collected, use of metadata and data gathering procedures, as well as policies and mechanisms for sharing data. However, researchers have generally not concentrated on the organization, access, reuse and preservation of data in their day-to-day research. Given these gaps, libraries and universities have been actively discussing the role and participation of librarians in research data management.

Many institutions have formed initiatives, committees and groups to provide support for a wide variety of research data activities: depositing data in institutional and external repositories or data archives; finding relevant external datasets; developing data management plans; creating tools to assist data management; bringing together available technology, infrastructure and tools; and data literacy education/training. To support researchers, some universities – such as Cornell, Johns Hopkins, MIT and Purdue – have established formal scholarly data services, and many other institutions are in the process of developing similar programs [1].

The University of Illinois created an e-research implementation group to bring together subject specialists, research data librarians and functional specialists to advance the library’s data initiatives. These are among the initiatives created since the group’s formation:

- hosting webinars on the data management plan tool;
- a pilot using EZID to assign persistent identifiers to datasets;
- creating a research data management interest group to provide awareness of data management to the campus and
- starting a research data management blog to provide information on data and data management topics.
In order to support the broader goals of 1) educating local researchers about the research data services that are available to them, 2) discussing ways that people think about data and 3) offering tools to meet their data needs, a smaller group of librarians within the e-research implementation group began a project to update the data management website. The updated website www.library.illinois.edu/sc/services/data_management/index.html now includes information and education regarding the definition of data, intellectual property, data sharing, funder requirements, file formats, privacy considerations, data documentation and metadata, and preservation and storage information. Users can access template guides and website resources needed for the data management plan tool. The goal of this website is to provide education and awareness regarding data management to students and faculty within and outside of the University of Illinois community. Since its launch in March 2014, the site has been accessed at least 129 times.

The website section featured at the 2014 Research Data Access and Preservation Summit focused on documentation and metadata. As libraries become more responsible for more data, they are being called on to support data preservation, discovery and analysis. Data types and sources include, but are not limited to, the following:

- bibliographic records,
- digital library collection metadata,
- website resources and
- research data.

The range and variety of researcher needs further complicate the process of creating and publishing data. Librarians often work with researchers to help them make their data more accessible, just as they have done for more traditional bibliographic materials.

Library bibliographic data includes information about printed and manuscript textual materials, computer files, maps, music, serials and continuing resources, visual materials and mixed materials. Bibliographic data commonly includes titles, names, subjects, notes, publication data and information about the physical description of an item [2].

Research data is collected, observed or created for purposes of analysis to produce original research results. Research data categories can include observational, experimental, simulation, derived, compiled and reference. The types of research data may include text, word documents, laboratory notebooks, questionnaires, audiotapes, photographs, slides, data files, database contents, collections of digital objects, models, methodologies and workflows, and standard operating procedures and protocols.

The data documentation and metadata section of the website provides users with information about why data documentation is important, as well as the types of data level documentation that is needed for research data. The metadata section provides a definition of metadata and lists website links for general/bibliographic, sciences, social science and humanities metadata standards.

Multiple metadata standards exist within the various subject disciplines, although many standards collect similar information. Therefore, it is important that researchers carefully consider which metadata standard will best suit their data and their research needs. Factors to consider when choosing a metadata...
Metadata helps researchers avoid duplicating data, better share information and promote their work in various fields of study. It provides users with the ability to search, retrieve and evaluate datasets. Metadata allows users to find data and decide if data meets a particular information need. It also allows users to discover, process and use a dataset. Metadata provides value to an organization and institution because it helps protect the organization’s investment in the data. It creates an institutional memory and advertises an institution’s research efforts, thus creating partnerships and collaborations through data sharing [3].

The data management website is useful because it provides education to students, faculty and the community within and outside of Illinois about data management and importance of metadata and documentation to proper data management and data sharing. Reactions to this website have all been positive. Librarians, subject specialists and individuals from various organizations and institutions have stated that they would review the website or recommend it as a resource for others to use. Future plans are to continue assessing the use of the data management website and other campus research data management resources and to survey faculty to determine how they use the website in order to determine and possibly implement additional improvements.

---

### Resources Mentioned in the Article

