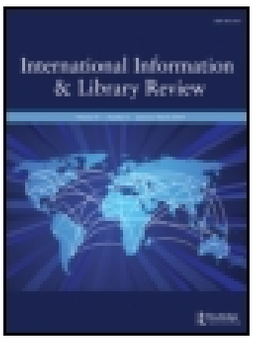
11-29-2017

Meeting Challenges in the Data World: RDAP 2017

Joshua Finnell

Brian J. Cain



International Information & Library Review

ISSN: 1057-2317 (Print) 1095-9297 (Online) Journal homepage: http://www.tandfonline.com/loi/ulbr20

Meeting Challenges in the Data World: RDAP 2017

Joshua Finnell & Brian Cain

To cite this article: Joshua Finnell & Brian Cain (2017): Meeting Challenges in the Data World: RDAP 2017, International Information & Library Review

To link to this article: https://doi.org/10.1080/10572317.2017.1383752

Published online: 29 Nov 2017.

INTERNATIONAL INFORMATION & LIBRARY REVIEW
2017, VOL. 49, NO. 4, 1–3
https://doi.org/10.1080/10572317.2017.1383752

Routledge
Taylor & Francis Group

GLOBAL POSTCARDS

Jacqueline Solis, Director of Research & Instructional Services, University of North Carolina, Chapel Hill; Robin L. Kear, Liaison Librarian, University of Pittsburgh

COLUMN EDITOR'S NOTES

The Global Postcards column is pleased to publish two contributions from Joshua Finnell and his colleagues. The first contribution with Brian Cane documents the themes and conversations of the Research Data Access and Preservation Summit (RDAP) in April 2017. The second contribution from Joshua with Stacy Konkiel documents the creation and sustainment of the Library Pipeline, a grassroots library organization. Finally, coeditor Robin Kear provides a personal synopsis of her attendance at the IFLA World Library & Information Congress (WLIC) in Wroclaw Poland in August 2017.

We always welcome contributions. If you would like to send a submission, please contact either of the column's coeditors: Jacqueline Solis, jsolis@email.unc.edu, and Robin Kear, rlk25@pitt.edu.

Meeting Challenges in the Data World: RDAP 2017

Joshua Finnell [a] and Brian Cain [b]

[a]University Libraries, Colgate University, Hamilton, NY; [b]Research Library, Los Alamos National Laboratory, Los Alamos, NM

On April 19, 2017, data librarians and researchers from academic, government, and private organizations descended upon the Renaissance Hotel in Seattle Washington to embark upon a sustained conversation about the continuing challenges in the world of data management. Now in its 7th year, the Research Data Access and Preservation Summit (RDAP) has become a crucial conference for data professionals grappling with issues concerning data reusability, interoperability, and metadata. With sponsors ranging from the International Association for Social Science Information Services and Technology (IASSIST) to DataCite, RDAP represents a broad cross-section of data community stakeholders and practitioners.

One of the Summit's sponsors, *Figshare*, recently conducted a survey of over 2,000 researchers about their issues and concerns related to open data practices and concerns. Though broad in context, the majority of concerns were classified into two major concerns: *structural* and *cultural* (State of Open Data, 3). Structural concerns arose around policy {"Do I have to make my data open?}, metadata, {"What does my data have to look like for me to share it?"}, to repositories {Where is the best place to deposit my data?"}. Cultural concerns stem from the competitive marketplace of ideas inherent to the scientific community, such as concerns around research being "scooped" if the data are publicly available for other scientists to build upon.

Structurally and culturally, the requirement of data management planning and deposit from major funders such as the National Institute of Health (NIH) and National Science Foundation (NSF) in the United States, as well as Europe's recent decision to make all publicly funded scientific research freely available by 2020, will certainly push researchers closer to the "how" of data sharing instead of the "why" in coming years. The programming at this year's RDAP Summit reflects this turning point in researcher's thinking. As a corollary, data professionals are beginning to build the infrastructure and schema necessary to curate and preserve the coming tidal wave of data.

The first panel of the summit focused on data reusability and underscored the challenges to both the "how" and the "why" of research data. Amy Pienta from the ICPSR social sciences data repository relayed the disconcerting fact that less than 15% of NSF and NIH funded data have been archived for long-term use. Yet, for Pienta, dataset availability is only half the story, curation services are an essential ingredient for data reuse. ICPSR provides high curation investment with the goal of making data highly usable in the long tail. An emphasis is also placed on quality metadata to ensure use, credit, and impact of deposited datasets. Pienta also pointed to analysis that indicated publically available studies in ICPSR on average show peak usage in the second year after release while still keeping high

CONTACT Joshua Finnell jfinnell@colgate.edu Colgate University, Hamilton, NY 13346, USA.

usage even seven years after release. She attributed this prolonged usage to the enhanced curation of ICPSR datasets, with the clear point that datasets with more curation see more use/citation.

Lisa Federer from the NIH provided an overview and update of NIH's data sharing policy/guidance and their efforts to support data reuse. For discoverability, the new website DataMed (currently in beta) serves as the PubMed for data. In terms of usability, the NIH is encouraging the use of Common Data Elements to improve the quality and harmonization of clinical data across studies and for electronic health records.

Presenter Ixchel Faniel of OCLC discussed the DIPIR project, a joint IMLS funded project with University of Michigan, which is investigating how important contextual information about research data supports reuse and how it can be created and preserved. Looking at the disciplines of quantitative social science, archaeology, and zoology, Faniel discussed the confluence and disparities of researchers' data reuse needs. She outlined how researchers' in these disciplines vary in the importance they place on types of contextual information. For instance, trust factors for data in the archaeologists' view hinged most importantly on data transparency whereas social scientists viewed the reputation of the colleague/institution as the greatest concern. Data reusers' satisfaction was also discussed, with data accessibility, completeness, and documentation quality being the most important attributes.

Thomas Padilla from UC Santa Barbara rounded out the panel with his presentation on data reuse in the humanities. Padilla made the point that much of the conversation and infrastructure for data reuse is in an embryonic stage. He offered the provocation that the data reuse value proposition could be considered "underdeveloped, possibly misaligned." Scholarly works rarely offer their underlying data, even fewer in an actual data repository, while our systems (whether they be storage, retrieval, or communication) are not designed to foster data reuse. At the same time, Padilla makes the point that data may have a particular rather than generalizable potential overall. Somewhat converse to the prevailing zeitgeist, the goal of data reuse may be more appropriately aimed toward repeatability than reproducibility.

The panel session on managing and preserving complex data reflected how quickly the heterogeneity of data are outpacing both best practices and existing models. Amanda Whitmire, head librarian at the Harold A. Miller Library at Stanford, discussed the challenges in both the reproducibility and preservation of historical, observational data. Between 1951 and 1974, The Hopkins Marine Station in collaboration with the California Cooperative Oceanic Fisheries program collected oceanographic data from the Monterey Bay. These raw data were collected and stored in analog form—mostly in the form of handwritten logs. Whitmire underscored the challenge in formulating a curation strategy, ranging from issues of metadata granularity to formatting. While scanning reams of paper into PDF format would help preserve the data it would, ultimately, not facilitate data reuse in a statistical software program without an additional conversion to a tabular format. Moreover, in thinking through the question of data reuse, questions arose about the most appropriate metadata schema, given the extensive set of variables captured in the dataset. Through the lens of a historical collection, Whitmire demonstrated the dizzying complexity involved in making 43 binders of data both accessible and reusable to the scientific community.

Whereas Whitmire discussed the importance of data conversion to facilitate reusability with R or Python, Fernando Rios, a CLIR postdoctoral fellow at Johns Hopkins University, discussed the challenges of preserving software for a future generation of researchers. Whereas data are facts or observations, utilized as evidence in formulating an argument, software is a creative work that manifests in a tool to analyze data. As a corollary, software is executable and data are not. Yet the dependencies between software and data are inherently evident, as data are often formatted for use in specific software packages. As a developing field, software archiving is just beginning to establish best practices, in terms of licensing, version control, and citation. Rios pointed to the Software Preservation Network and the Journal of Open Source Software as key stakeholders in broadening the discussion and establishing best practices for researchers.

While Whitmire and Rios discussed the challenges researchers and librarians face in both preserving and using research data and software, Timothy Norris and Genevieve Podleski discussed the curricular challenges related to data information literacy. A CLIR postdoctoral fellow at the University of Miami, Norris discussed the hurdles he faced in developing a data curation course, from inception to assessment. Though his course garnered student interest, finding a home department to adopt the course was a challenge.

Moreover, Norris surfaced a discussion on where a data carpentry course would be most impactful, at the undergraduate or graduate level. Though many disciplines, from the social sciences to the humanities, are becoming more data-intensive, no formal course in data curation is currently required at many institutions of higher learning. In line with many of the other presenters, Norris pointed to best practices being developed at Purdue, the DataCure Data Instruction Materials Repository, and the University of Massachusetts Medical School, the New England Collaborative Data Management Curriculum.

Podleski, senior digital projects librarian at the Federal Reserve of St. Louis, discussed the diverse user group of their data, from research economists to elementary and middle-school teachers. Because educators primarily access the Federal Reserve Economic Data

(FRED) database, Podleski has begun creating data literacy lessons specifically for elementary school teachers. Exposing students at a young age to how data are both collected and shaped to make arguments creates a foundation upon which students can build their knowledge and expertise as they matriculate through the education system. Moreover, the development of the FRED app allows students to access economic data on their preferred mobile device. Collectively, both Norris and Podleski surfaced a conversation about the role of data literacy in educating a new generation of researchers and citizens.

The last session of the conference, Data Publishers: A Variety of Perspectives Panel, was a discussion between the audience and Meghan Byrne, Senior Editor at PLOS, Anita de Waard, Vice President of Research Data Collaborations at Elsevier, and Mark Hahnel, the found of Figshare. There was no shortage of questions/concerns directed toward the publishers/providers, with the librarians amiably holding the panelists' feet to the fire. A theme of natural conflict between data openness within structures that are not incentivized to be open emerged throughout the course of the discussion. Challenges to open data and open science are no doubt present in the current environment, the tenure system and profit-motive two notable examples called out during the session. There was no shortage of opinions for reconciling researcher, institutional, disciplinary, and publisher needs. Key points centered on partnerships between publishers, libraries, and the communities of practice and what these might look like moving forward. General agreement that enhanced communication and cooperation between all the stakeholders is needed, leveraging organic feedback, building self-correcting systems, and enabling community-driven change to produce tools and services that foster a productive and sustainable open data ecosystem.

Conclusion

As mentioned at the outset, the first RDAP summit convened in 2010. The topics of discussion in 2010 are strikingly similar to 2017. For example, at the first RDAP conference, Peter Wittenburg from the Max Planck Society and Roy Williams from the International Virtual Observatory Alliance were part of a panel that discussed the need to promote the re-use of scientific data collections. Seven years later, Ixchel Faniel from OCLC and Lisa Federer from NIH formed part of a panel underscoring the need for data reusability in research communities. Though the topics remain consistent, what has changed is the scope and breadth of voices involved in the conversation. This year's summit was the most attended in RDAP's history, reflecting both the growing role of data management and curation and the dedication of the program chairs, Brianna Marshall and Yasmeen Shoorish.

Underscored in RDAP Summit's mission statement is its emphasis on community. "*RDAP supports an engaged community of information professionals committed to creating, maintaining, advancing, and teaching best practices for research data, access, and preservation.*" This feeling is also reflected in its members who identified *people* as the "meaning" of RDAP when polled by the RDAP Future Vision Task Force. Naturally, the growth of the Summit the last few years leads to questions of both sustainability and scalability. Will RDAP continue to incubate under the Association for Information Science & Technology (ASIS&T) or become its own organization? What will RDAP's relationship be with parallel organizations with overlapping interests such as Research Data Alliance or Force11 or the Research Data Management Forum? This year's Summit may very well be a pivotal moment in the history of the RDAP, as its success and growth over the last few years suggest an expansion in both scope and structural governance.

ORCID

Joshua Finnell http://orcid.org/0000-0002-0816-1381