Report on the Data Curation Research Summit

Sponsored by 

9 December 2010

Chicago, Illinois

Summit coordinated by:

Carole Palmer & Melissa Cragin

Center for Informatics Research in Science & Scholarship
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign

Scott Brandt

Distributed Data Curation Center
Purdue University Libraries

Report prepared by:

Nicholas Weber, Tiffany Chao, Carole L. Palmer, & Virgil E. Varvel Jr.

Center for Informatics Research in Science & Scholarship

TABLE OF CONTENTS

Table of contents 2

Summary and Context 3

Overview 5

Emergent Themes 7

- **Library and Archival Foundations** 7
- **Data Representation and Interoperability** 7
- **Scientific data practices** 9
- **Governance and Policy** 10
- **Trust and Publishing** 11
- **Closing Observations** 12

Directions for Future Research and Development 14

Presentation Briefs 16

Appendices 22

- **A. Meeting Agenda** 22
- **B. Participants** 24

SUMMARY AND CONTEXT

The Data Curation Research Summit was a one-day meeting, sponsored by the Institute of Museum and Library Services (IMLS). The objectives were to build awareness of current research projects and important research problems, foster stronger collaborations among researchers, and advance the Library and Information Science (LIS) research agenda in data curation. It was held in Chicago on December 9th, 2010, following the 6th International Digital Curation Conference (IDCC). The conference provided an excellent opportunity to bring together scholars and practitioners with a strong interest in advancing scholarship and practice in the curation of research data. The 35 invited participants, representing iSchools, research libraries, academic publishers, and funding agencies, are active in the growing research community and related areas of digital curation and archives. (See Appendix B for a complete list of participants.)

The Data Curation Research Summit (DCRS) was an opportunity to formally extend conversations that started at a workshop held after the Fourth Bloomsbury Conference on E-Publishing and E-Publications on June 24-25, 2010, in London. The theme of the Bloomsbury Conference was "Valued Resources: Roles and Responsibilities of Digital Curators and Publishers." The workshop that followed was organized and co-sponsored by the University College London (UCL), Department of Information Studies, and IMLS. It was attended by approximately 30 international leaders in digital curation and e-publishing and produced a report on next steps in research, education, and practice--*Advancing Research and Practice in Digital Curation and Publishing*.[1]

Convened six months after the Bloomsbury workshop, the DCRS assembled a mix of both returning and new participants. DCRS was more narrowly focused on the curation of research data and LIS contributions to the overall arena of data curation research. The preceding International Digital Curation Conference (IDCC) provided an informative backdrop for the summit and its participants. IDCC was co-hosted by the Graduate School of Library & Information Science (GSLIS) at the University of Illinois at Urbana-Champaign in partnership with the UK Digital Curation Centre (DCC) and the Coalition for Networked Information (CNI). The conference theme was "Participation and Practice: Growing the Curation Community through the Data Decade,"[2] and a

[1] http://ideaworkgroup.org/
[2] http://www.dcc.ac.uk/events/conferences/6th-international-digital-curation-conference

number of the topics and ideas posed by conference speakers resonated in the summit presentations and discussion, including: the role of community-based curation for growing data collections (Chris Lintott, Galaxy Zoo; Antony Williams, ChemSpider), rapidly changing models for sharing and publishing scientific results and claims (Barend Mons, University of Rotterdam and Leiden University Medical Center), and the emergent divisions of labor in the inter-institutional ecology of data curation (Kevin Ashley, director of the Digital Curation Center; MacKenzie Smith, Massachusetts Institute of Technology (MIT) Libraries).

The Data Curation Research Summit consisted of four sessions, which followed opening contextual remarks provided by the organizers of both the DCRS and the previous Bloomsbury workshop. The first session was focused on current directions in research from the perspective of LIS faculty. The second session was devoted to approaches and challenges studying scientific data practices and needs. The third session covered current directions in research from the perspective of research libraries, and the final session returned to the original themes of the Bloomsbury conference with perspectives from the publishing community. (See Appendix A for the summit agenda).

This report provides a synopsis of the presentations as well as the broader group discussion of the summit. More specifically, this report highlights key emergent themes and concludes with recommendations for strategic research directions for advancing the state of knowledge and practice in the curation of research data. Briefs of the individual presentations are provided at the end of the report.

OVERVIEW

The DCRS began with a summary of the motivations for convening the summit, which emphasized the potential to forge new collaborations and develop a common research agenda, but also noted the role of research in advancing the information professions. In her welcoming remarks, Joyce Ray from IMLS discussed the agency's programs that have been supporting data curation education, training, and related research since as early as 2006. She acknowledged the leadership and inspiration provided by the Digital Curation Center in the UK and key meetings that have fostered the LIS curation community in addition to the 2010 Bloomsbury workshop. The series of events convened by the International Data Curation Education Action (IDEA) working group[3] are particularly noteworthy. However, with the exception of a 2007 post-IDCC Research Data Workshop sponsored by JISC and the Mellon Foundation, most previous meetings have been very broadly scoped, addressing "digital" curation generally, rather than the curation of research data specifically. Prior events also tend to focus on education and practitioner development rather than the research and researchers of fundamental issues in data curation.

Anthony Watkinson presented an overview of themes from the earlier Bloomsbury workshop, providing important continuity across the two meetings, especially for participants who had not attended the previous event. That workshop focused on value and trust in curation and publishing among the scholarly community, and Watkinson identified six "big words" to summarize the topics covered: value—how to represent the value of preserving and sharing data; impact—how to measure and account for data use beyond journal citations; cost—investment in managing and curating data throughout their lifecycle; connection—how do data and related scholarly research products relate to each other and how can these relationships be modeled; format—problems in propagating standards for archiving and preserving data; and organization—how those who curate data strategically align themselves in universities and repositories. Recommendations from the workshop included a call for broader collaboration among the various groups invested in scholarly communication and a better understanding of user needs and practices. In particular, Watkinson

[3] http://ideaworkgroup.org/

stressed the importance of broadening public understanding of the need to train the new generation of data curators and scientists.

In the final introductory segment, Carole Palmer, the lead DCRS organizer, talked about the high stakes for the LIS community in the current environment where scientific and scholarly information are deliberately being reshaped.[4] Palmer noted that this is a pivotal time for the field to assert its imprint on new "cyber" information infrastructures and services. She noted that LIS contributions are especially crucial in areas where the field has developed deep research based professional knowledge and principles, such as information organization, scholarly information use, and preservation. The summit was designed to clarify LIS contributions and to explore how to strategically move forward to have a significant, positive impact on the changing scholarly information landscape. Palmer noted that LIS will need to develop stronger partnerships with domain researchers, informaticists, and other stakeholders in the research enterprise, to succeed at making research data an integral and enduring part of the information assets retained for science and scholarship over the long term.

[4] Hine, C. (2005). Material culture and the shaping of e-science. First International Conference on E-Social Science. Manchester, UK. http://www.ncess.ac.uk/events/conference/2005/papers/papers/ncess2005_paper_Hine.pdf.

Emergent Themes

Over the course of the summit, five themes were prominent across the presentations and discussion: library and archival foundations, data representation and interoperability, scientific data practices, governance and policy, and trust and publishing.

Library and Archival Foundations

Principles, processes, and models in research librarianship offer a sound foundation for the development of curation processes and services for research data. For example, in research libraries the established role of the liaison librarian is proving to be an effective means of engaging with faculty about their data needs. Interestingly, the special collections departments of academic research library also offer a valuable model for the development of collections and services for research data as unique information objects. For example, data repositories will need to apply acquisitions criteria that provide for inclusion of "rare" or irreplaceable materials for future use rather than only prioritizing materials with potential for high use in the short term. Additionally, rare book cataloging and description techniques are highly applicable to the representation of complicated data sets.

While LIS principles can provide important guidance on the representation of research data, the limitations of library and document centered models need to be recognized and addressed. In this community, FRBR has received considerable attention as a viable conceptual model, especially the Group 1 entities in relation to levels of data transformation and the various derivative products. However, further investigation is required to determine where adaptation or alternative approaches are needed. In particular, semantic web and linked data approaches need much more attention and experimentation, with the understanding that ultimately hybrid or multiple models will need to be developed to accommodate the complex structures and features of scientific data.

Data Representation and Interoperability

Metadata and knowledge representation are established areas of expertise in LIS that is vital to data curation, and as such they should remain a focal point for future research investment. Current

efforts in metadata research are addressing the long-standing problem of how to generate optimal metadata, while also contributing to the further development of metadata theory. With the proliferation of metadata standards there is an urgent need for professional resources that review, compare, and evaluate standards, including their applications, roles, and adequacy for curation functions. One project of note, HIVE (Helping Interdisciplinary Vocabulary Engineering)[5] is making important progress on reconciling controlled vocabularies, developing techniques for dynamically integrating multiple discipline-specific vocabularies to support curators and catalogers in libraries, museums, and archives.

One of the central aims of curation is to maintain and preserve the valuable connections between scientific data and related information, especially the research results reported in publications. The Dryad project[6] is demonstrating the value of strong collaborations with professional societies and publishers in building systems that support linking data with publications. Dryad is partnering with journal publishers and scientific societies to provide access to the data underlying journal articles, while also advancing workflows and production efficiencies for extracting and generating metadata for these data sets. At Johns Hopkins University, a workflow for publishing data, based on the OAI-ORE model, is being tested for arXiv.org and AAAS publications. Techniques for automatic generation of metadata continue to be a high priority for research, but work to date shows that it is imperative to capture as much metadata as possible from authors when they submit data to a repository, or even earlier in the scientific research process.

Many of the challenges of coordinating data services between repositories and publishers are escalated in large, cross-institutional storage and access systems. As the networks of repositories scale, effective curation will be key to functionality and interoperation. The LukII project[7] is aiming to connect numerous repository nodes in Germany, addressing problems related to multiple formats and duplication of content across the network devoted to long-term preservation of digital information. The complexities of national interoperability involve technical, social, business, political, and legal challenges involved in achieving data exchange across organizations, domains, and borders.

[5] http://ils.unc.edu/mrc/hive/
[6] http://datadryad.org/
[7] http://www.lukii.hu-berlin.de/

At a fundamental level, the ability of data models and infrastructures to interoperate and scale beyond local instances will require the systems development community to have a consistent conceptual understanding of data as information objects. Representation of meaningful units of data, and the ability to identify duplicate or functionally equivalent data are hindered by the variation in terminology and definitions of data used among domain and development groups. The Data Concepts team at the University of Illinois GSLIS is developing a formal framework for data concepts, making progress toward a more precise and shared understanding of terms such as "data", "data set," and related concepts including format, encoding, file, and derivation. The framework will guide consistent identification of data objects and their parts, transformations, grouping, and relations.

Scientific data practices

How researchers generate, manage, and share data has been a recent focus of research in LIS, and will continue to be important as academic libraries become more involved in data services for local and distributed research communities. As with other types of information work, data practices are influenced by disciplinary norms and cultures, as well as other organizational and collaborative arrangements. The differences across disciplines have important implications for professional data curation processes and services. To add value to the research process, curators need a sophisticated understanding of how researchers currently work with their data, as well as the potential for innovative re-use of data within and across disciplines. Both DataNet projects, DataOne[8] and the Data Conservancy[9], are informing development of their respective data initiatives with studies of scientific data practices and needs, with particular attention to the socio-cultural dimensions of data sharing.

DataOne is examining all stakeholders in their distributed virtual organization in the ecological sciences. Their baseline assessment of data practices and sharing will allow them to track changes in practices over time. Initial results show wide variation in data practices and

[8] https://www.dataone.org/.
[9] http://dataconservancy.org/.

very limited metadata production by scientists. At the same time, there is strong interest in sharing data, with advances dependent on systems for attribution, specific conditions for re-use, and data management and metadata production support for scientists.

Data Conservancy's research on scientific data practices is being conducted at the University of Illinois and UCLA. The Data Practices team at Illinois is examining variation in data practices and curation needs across the disciplines served by the Data Conservancy, with a particular focus on the small sciences. The UCLA team is conducting an ethnographic case study of projects and researchers in astronomy to investigate how data management and sharing has evolved in this exemplary data community. Trust in data has been a key to establishing the existing culture of data sharing in astronomy, and has required vetting of data and services by both institutions and individual researchers. Reconciling data from different instruments has been a major challenge. While astronomy is one of the most mature data communities, there is still need for improvement in management and curation, especially in certain sub-disciplines, and development of tools to support new and innovative uses of astronomy data to advance the science.

Governance and Policy

Articulating the necessary conditions for open data licensing among interoperable repositories will require a well-developed and clearly defined data governance. Guidelines for data sharing need to be made clear and consistent among the myriad stakeholder groups in research data production. While it is anticipated that standards for data citation will enable large scale sharing, many problems remain unresolved around promoting attribution practices and recognition for data sharing in academic reward systems. Moreover, licenses and ownership rights for data generated by federally funded research are not yet well understood and may impose serious barriers to sharing and re-use. Other obstacles include application of identifiers that are persistent over time and capture and representation of accurate provenance information. Sustainable identifier models are beginning to emerge, as in Dryad's use of DataCite to issue DOI's for the datasets attached to journal articles. But, Dryad's success attracting depositors has been

partly due to policies adopted by a number of evolutionary biology and biodiversity journals to make the deposit of data associated with an article mandatory for publication.[10]

To facilitate open and legal sharing, trusted repositories will be needed for data and metadata, but trusted registries will also be needed for content standards and ontologies to support aggregation, and discovery. Long-term preservation is foundational to building trusted repositories and can be greatly informed by a community of research libraries with a strong base of experience in preservation policy development. As a research library based initiative, for example, the Data Conservancy specifies preservation objectives and options in its collection and service policies.

Trust and Publishing

In keeping with the earlier Bloomsbury workshop, the role of publishers as trusted partners in scholarly communication was a key message from presenters representing the publishing community. Much of the value publishers contribute to scholarly information lies in the quality assurance they provide for the products they disseminate, most notably through the coordination of peer review for potentially publishable articles and manuscripts. However, publishers have yet to expand this quality assurance role to data, or supplemental materials accompanying a publication. This is due in part to the fact that they currently do not have the resources or access to the expertise required for systematic, expert review of data. Publishers are gradually beginning to take on some of the responsibilities of curating data, but have not yet come to terms with the impact on their workflows and the costs that these activities will require them to absorb. Complications are arising, for example, with journals that accept supplementary data but lack policies appropriate for multimedia content or clear commitments to the long-term preservation and access of this material. Data challenges will be most difficult for smaller publishers that are not yet engaged with archiving and preservation.

Publishers that currently provide access to large aggregated archives of publications may move toward delivery of all kinds of materials, including primary data and metadata. However, they recognize the advantages of metadata being packaged further upstream and the importance of

[10] http://datadryad.org/jdap

partnering with academia and research libraries, which are more familiar with the difficulties of metadata generation. The major contribution of large publishers and aggregators to curation research and development may be in advancing general and efficient curation processes, beginning with converting their vast stores of text into more useable data sources.

The current discourse on data curation and sharing often implies what publishers refer to as a "publisher-free zone," suggesting a more limited role for publishers in the scholarly communication environment than they actually anticipate. Part of this dynamic is related to the market shift with data; historically libraries have been the primary market for traditional scholarly and scientific publications, but with data the demand is linked more directly to the academic research communities than a formal institution. New products, such as data journals, may be viable in this new environment, but such products will likely emerge to meet disciplinary practices of a specific discipline rather than being uniform models from multiple publishers, like the traditional article based scholarly journal. In the short term, however, publishers are likely to host primary data in very special cases where there is a clearly defined user market that publishers can sustainably support. Moreover, there is strong interest in working with the data curation community around common interests in metadata and identifier problems associated with linking data and publications, regardless of their storage location.

Closing Observations

Cliff Lynch provided a closing commentary, first remarking on the surprising amount of discussion that had strayed from the summit's focus on research to address practical issues of repository systems and service provision. He stressed the need for leaders in the LIS field to work on "evolving" the relationship and integration of publishing into the curation lifecycle. In particular, Lynch identified the need to "air and resolve" publisher roles, especially in regard to the peer review of research data, suggesting that some data will likely be reviewed using on-going editorial approaches. To investigate the effectiveness of publisher peer review, he called for a commissioned study on what peer-reviewed data currently exist, how and why the peer review was done, and the usefulness of these data and the peer review process. He also noted his surprise that

software issues had not come up in the summit discussion, since data-metadata-software form a "tripartite" for curation.

Lynch asserted that current "on-the-ground social research is fundamental" but "does not go far enough," recommending the model of clinical trial case studies over the past 100 years, which have produced rigorous, systematic, and high-impact results of great value to medical practitioners. He also reminded participants that curation is largely about facilitating research processes in the short term and persistent access to data in the long-term. This implies that more user studies will be needed now to inform data aggregation efforts by repositories and services that facilitate future re-use by researchers. Data resources, for example, should support more complete meta-analysis. Long-term demands and implications of wide-scale data curation and sharing are, of course, still being played out and will become more apparent as the curation community matures.

Accountability to funding agencies will be a major force in how data are treated by scientists and scholars in the immediate future. While research practices generally evolve based on the needs and demands of a discipline, the directives from funders on data management and sharing are meant to change current unacceptable conduct. This is an opportune time for the LIS research community to study the effects of the new requirements on the research process, while also devising methods for measuring the outcomes of data management policy initiatives. The "reproducibility problem" in computational science, as articulated by Victoria Stodden,[11] will be a driver for curation efforts around both data and the associated code. In the social and medical sciences the conflict between data sharing and human subjects requirements enforced by institutional review boards will need to be resolved, and our community should be working to advance policies for sharing data on human subjects. Following on his comments on human subjects data, Lynch acknowledged that some problems are so hard to solve that the community will need to move ahead without solutions, identifying standards as a primary example.

Additionally, we would add to Cliff's summarization that perspectives on standards were quite divergent among summit participants. Some voiced concerns about the proliferation of standards as a serious problem facing smaller institutions and projects, and others were confident that new standards

[11] See for example: "Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science," with Yale Roundtable Participants, Computing in Science and Engineering, vol. 12, no. 5, p. 8-13, Sep./Oct. 2010, doi:10.1109/MCSE.2010.113.

are still needed for work to be automated and consistently carried out in the curation community, including but not limited to issues of infrastructure, metadata, and data formats. It was also suggested that there might be too many standards because the right ones have not yet been developed. Clearly this is an area where there is need for more general analysis and an articulation of a sound and actionable research agenda. Curation processes and repository development needs to proceed in a way that allows systems and services to respond to future advances in research.

Directions for Future Research and Development

In addition to the research directions suggested by his closing observations on peer review, case studies, policies, and standards, Cliff Lynch offered two promising areas of "pure research" vital for the field to succeed in preserving data as a substantial part of the scholarly record:

- Formal representations of the intellectual content of research data.
- Provenance and chain of custody for complex digital objects.

Primary areas under investigation in current projects and topics needing further research were suggested in the above discussion of summit themes, including problems around the capture and generation of metadata and the applicability of document-centric data models to scientific data. Additional general research questions raised or suggested through in the presentations or dialogue included:

- Where and how do the curation contributions of university libraries, data services, and publishers intersect?
- How can publishers fit data curation into their existing workflow and assist scholars in managing their data?
- What organizational and business models best apply to data curation operations?
- How much does each stage of the data lifecycle cost and which stages can be subsidized by various revenue channels?

- How can curation, as a profession, attend to the many dimensions of trust necessary for successful curation systems and processes?

Specific recommendations for future research included:

- Identify and evaluate multiple strategies for curating, identifying, and linking data to publications.
- Investigate the range of issues involved in curating supplemental data (especially multi-part and multi-media materials), including identifiers, integrity, format, metadata, and access.
- Assess the range of intellectual property and policy issues surrounding the sharing and reuse of data.

A number of developmental priorities were identified, of which several were aimed at advances that would expose content to larger audiences and broaden data accessibility. All of these efforts would need to be underpinned by research, and some could benefit from substantive cooperation between publishers and libraries.

- Cross-linking supplemental data using standard formats, identifiers and protocols, and supporting metadata.
- Providing richer and more granular linking.
- Supplying multiple formats of data to enable meaningful reuse.
- Build on successes of the linked data community.

Segments of the exchange moved into related topics in the field, including research library capacity to meet data curation demands and LIS education for data curation, which were also highly suggestive of productive research directions. Workforce development issues arose in relation to infrastructure development and curation services. The European Commission's "Riding the Wave"[12] report was suggested as important background reading on the topic. Specific areas identified for future work included:

- Empirical research to solidify a knowledge base for future practitioners.

[12] cordis.europa.**eu**/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

- Development of metrics of success for education and training (What constitutes a successful program?)

There was a general consensus that there needs to be more investment in retooling and professional development to move digital libraries and librarians into skilled data curators competent in providing service to data creators. Moreover, the need for professionals to have both domain knowledge and curation skills was raised, as was the need for computer science expertise. Thus, a question that has long been debated within the information professions, especially in regard to research librarianship, still stands for the curation of research data: What is the optimal balance of domain knowledge, information science, and computer science needed for professional work in the field?

PRESENTATION BRIEFS

Key points from each presentation are outlined below, following the sequence of speakers on the program (provided in Appendix A). Most slide-sets are also available online at the Data Conservancy, Research Data Workforce Summit, website, at http://cirss.lis.illinois.edu/SciCom/DC/index.html.

The first panel included LIS faculty members discussing current research projects and future directions for curation research.

Jane Greenberg, University of North Carolina- Chapel Hill.

- What are the optimal metadata generation equations?
  - How do we determine if we've implemented an optimal metadata generation workflow? How can we confirm an optimal ROI for metadata generation? Target for future work should be optimal metadata generation.
  - We need the capability of automating routine activities. This is possible only by leveraging resource creator knowledge base. Save the effort and labor of a curator for metadata requiring human subjective qualities.
- Too many standards are as much a technical hurdle as they are a social problem.

- Theory is integral to our field and crucial for advancing knowledge in any discipline- LIS should continue to articulate a philosophical and theoretical understanding of metadata and its place within the context of digital curation.
- Currently there are very few, if any, data repositories, data sharing mechanisms and platforms supporting data curation researchers.

Michael Seadle, Humboldt University. "LuKII Projeckt- LOCKSS und KOPAL: Infrastructure and Interoperability."

- Too often discussions concerning digital archiving rely on marketing claims and insufficient data. No system has all of the answers, and curation is currently being undermined by industries that claim preserving content forever is feasible, and inexpensive. Interoperability is a much stronger end product, and should be the immediate intention of curation community.
- In LIS we need to focus on preserving specific types of data and what we can demonstrate with these efforts, both empirically for own research and in collaboration with scholars re-using our most successful preservation efforts.
- If bit rot is a serious problem, how do we test and address the problem?

Allen H. Renear, University of Illinois, Urbana-Champaign. "Data Concepts."

- What does it mean to 'use the same data'? How can we reliably tell what 'the same' means with respect to both the data and its use?
- The definition of dataset is defined differently across the various scientific disciplines. In the scientific and technical literature there are generally four classifications made: content, grouping, purpose, and/or relatedness (particularly with respect to levels of abstraction and ontological status)
- Data curation requires precisely defined and shared concepts for key notions such as data and dataset.
- Data Conservancy seeks to develop a formal framework of data concepts with both conceptual and operational identity conditions for data and dataset.

The second panel moved on to discuss the approaches and challenges of studying data needs and uses.

Suzie Allard, University of Tennessee, Knoxville. "Research in the Socio-Cultural Aspects of Data Curation, Measuring and Improving Data Practices."

- DataOne is focused on the user side of the user-technology continuum, specifically the socio-cultural aspects surrounding data management. Curation research needs to look at issues of motivation, practices, and usability of both immediate data user groups and potential groups (including those responsible for curation and access as well as those focused on data creation and re-use). There is also a need for a baseline assessment.
- Need to take a holistic view of the data lifecycle as it fits into the research lifecycle.
- According to their survey of scientists
  - Data practices vary but most data are not well described.
  - Many scientists are interested in sharing data but few participating.

Christine Borgman, University of California, Los Angeles. "Curators to the Stars."

- Astronomy data is unique in that it is highly attractive to the public, the research community is small, the data has little commercial value, data is collected in real time and well documented, instruments for collection are diverse and distributed, and the volume of data is very large.
- Studies of data practices need to expand beyond a single project in order to more generally capture sub-disciplines and discipline level practices.
- Research questions for work with Data Conservancy:
  - What are the data management, curation, and sharing practices?
  - Who shares what data when, with whom, and why?
  - What data are most important to curate, how, and for whom?
- Data management can vary widely and should act as a caution for infrastructure development, but successful data management has yet to be clearly articulated or defined by this community.

The third panel discussed current directions in research from the perspective of research library projects.

Michael Furlough, Penn State University. "An Administrator's Perspective on Publishing and Data Curation."

Four main questions need consideration for the future of curation research:

- How are partnerships forged between publishers and various actors they serve, particularly libraries and librarians?
- What cost / sharing models for sustainability of data services are there? In particular there is a need to understand how these types of issues are restrained by the current economic climate of most higher education institutions. We should look at shared infrastructure for common services.
- How can curators demonstrate that upstream involvement for librarians is beneficial not just to individual researchers but to the institution as a whole?
- Which kinds of services fit into which phases of the research lifecycle?

MacKenzie Smith, MIT. "The Role of Policy in Data Curation: Aspects for Future Research."

- The importance or need for automated policy enforcement and "architecting" policies into digital archive (e.g. policy stores, policy capture, policy injection).
- Future research should include tools for archivists to specify policy and network protocols for policy sharing (Peer-to-Peer).
- Data interoperability has to be international, interdisciplinary, large scale, single access, and sustainable; this will need to address technical, social, business, political, and legal dimensions.
- There is a need for data governance research, including how to integrate data with differing licenses- When and how to apply deposition/creation/derivative work licenses.
- Best practices for providing credit for sharing scholarly data continue to go missing from scholarly communications. This includes both an identifier system and method for citation and attribution, including how to best guarantee the persistence of identifiers.

- Development of mechanisms for documenting data provenance on the Web
- Need for "open and legal sharing of relevant metadata to enable discovery, re-use, aggregation, and long-term preservation of referenced data."
- Additional needs for standards and ontologies that will help solve higher-level interoperability issues.
- Role of "trusted" registries needs assessment and further discussion within curating community, e.g. Unified Digital Format Registry (UDFR).

Sayeed Choudhury, Johns Hopkins University. "A Curation System for Linked Publications and Data."

- Illustrated an approach for data set capture during the publishing workflow, so that relationships among data and text can be captured, maintained, and communicated as they are submitted into a publication system pathway.
- Ultimately in this process data needs to be transmitted to an external archive for long-term preservation and citation for search and discovery.
- Professional societies should take on responsibility of advocating for more responsible data practices, and should facilitate sharing, discovery and preservation through their relationships with repositories.
- Many of the problems encountered early on with the Data Conservancy were a matter of the differing terminology used among the various domains they were interacting with, "Vocabularies are a problem. Scholars are the means of mediation."
- Similarly, data authors are often know the most appropriate metadata for discovery and re-use of their data within their own domain; curation researchers need to understand how to better extract this knowledge from depositors, as well as determine how to represent the same data to other communities of interest.

The final panel of the summit leveraged the expertise of publishers who have worked across the various stakeholder groups represented in first three panels of the summit.

Anthony Watkinson, University College London. "Publishers, the Article, and Data – Relationships and actions"

- Publishers and publishing are rarely represented in an ecosystem of scholarship surrounding data intensive disciplines. But data publishers are important and growing more important in terms of infrastructure and capacity to manage interoperable access platforms.
- Central to the process that inhibits data publication is peer review. There are not good models for peer-reviewing data, and publishers currently do not know how to organize these activities.
- Publishers currently lack incentive to facilitate re-use of data; they aren't convinced it will foster new publications, and have little guarantee that the massive investment can be sufficiently monetized.

John Burns, JSTOR. "A View from the Archives: User Needs Driven Curation"

- Publishers are good at dealing with large quantities, but not in knowing what to do with that content: "Generality trumps specialization in curation."
- Archives like JSTOR will start with text, and transform content to be more "data-ish." Archives will, in the future, have stronger role in the smaller sciences.
- Data support and curation will evolve as a result of user demands; right now archives are not experiencing those demands. There is a publisher need to collaborate with libraries and curators to better articulate user needs with respect to accessing data, especially legacy data.

APPENDICES

A. MEETING AGENDA

Meeting Agenda	
9:00 - 9:10am	Welcome
	Joyce Ray – *IMLS*
9:10 - 9:30	Key themes from Bloomsbury & workshop
	Anthony Watkinson - *University College London*
9:30 - 9:45	Objectives for the morning
	Carole Palmer (moderator) - *University of Illinois at Urbana-Champaign*
9:45 - 10:30	Current directions in research – LIS faculty projects and perspectives
	9:45 - 10:00 - Jane Greenberg - *University of North Carolina-Chapel Hill* 10:00 - 10:15 - Michael Seadle - *Humbolt University* "LuKII Projeckt- LOCKSS und KOPAL: Infrastruktur und Interoperabilitat" 10:15 - 10:30 - Allen Renear - *University of Illinois at Urbana-Champaign* "Data Concepts"
10:30 - 10:45	BREAK
10:45 – 11:15	Studying data needs & uses: Approaches and challenges
	10:45 - 11:00 – Suzie Allard - *University of Tennessee* "Research in the Socio-Cultural Aspects of Data Curation: Measuring and Improving Data Practices." 11:00 - 11:15 - Christine Borgman - *University of California at Los Angeles* "Curators to the Stars"
11:15 – 11:45	Audience/Panel discussion
11:45 – 1pm	LUNCH
1:00 - 1:15	Objectives for the afternoon
	Scott Brandt (moderator) - *Purdue University*
1:15 – 2:00	Current directions in research- Research librarian projects and perspectives
	1:15 - 1:30 - Mike Furlough - *Pennsylvania State University* "An Administrator's Perspective on Publishing and Data Curation"

	1:30- 1:45 - MacKenzie Smith - *Massachusetts Institute of Technology* "The Role of Policy in Data Curation: Aspects for Future Research" 1:45 - 2:00 - Sayeed Choudhury - *Johns Hopkins University* "A Curation System for Linked Publications and Data"
2:00 – 2:30	Perspectives from the publishing community
	2:00 - 2:15- Anthony Watkinson - *University College London* "Publishers, the Article, and Data - relationships and actions" 2:15 - 2:30 - John Burns - *JSTOR* "A View from the Archives: User Needs Driven Curation"
2:30 – 2:45	BREAK
2:45 – 3:15	Audience/Panel discussion
3:15 – 3:30	Priorities for future research
	Clifford Lynch - *Coalition for Networked Information*
3:30 -3:45	Wrap-up & next steps

B. PARTICIPANTS

Name	Position	Institution / Agency
Alexander, Patrick	Director	University Press, Pennsylvania State University
Borgman, Christine	Presidential Chair & Professor of Information Studies	Graduate School of Education and Information Studies, University of California at Los Angeles
Brandt, Scott	Associate Dean for Research and Professor	Purdue University Libraries
Burns, John	Director of Platform Research	ProQuest
Carlson, Jacob	Data Research Scientist	Purdue University Libraries
Choudhury, Sayeed	Hodson Director of the Digital Research and Curation Center	Sheridan Libraries, Johns Hopkins University
Cragin, Melissa	Research Assistant Professor	CIRSS, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
Duff, Wendy	Director of the Digital Curation Institute and Associate Professor	Faculty of Information, University of Toronto
Faniel, Ixchel	Post-Doctoral Researcher	OCLC
Frick, Rachel	Director	Digital Library Federation
Furlough, Mike	Assistant Dean for Scholarly Communications and Co-Director of the Office of Digital Scholarly Publishing	University Library, Pennsylvania State University
Giannini, Tula	Dean and Professor	School of Information and Library Science, Pratt Institute
Green, Ann	Digital Information Strategic Analyst	Office of Digital Assets and Infrastructure, Yale University
Greenberg, Jane	Professor and Director	SILS Metadata Research Center, School of Library and Information Science, University of North Carolina at Chapel Hill
Hodson, Simon	Programme Manager, Managing Research Data	JISC
Johnston, Lisa	Physics, Astronomy and Geology Librarian	Science/Engineering Library, University of Minnesota
Kozbial, Ardys	Technology Outreach Librarian	University of California, San Diego
Lagoze, Carl	Associate Professor	Information Science, Cornell University
Lesk, Michael	Chair	Department of Library and Information Science, Rutgers University
Lynch, Clifford	Director	Coalition for Networked Information
McDonald, Robert	Associate Dean for Library Technologies and Digital Libraries, Associate Director Data to Insight Center, and Executive Director Kuali OLE	University Libraries, Pervasive Technology Institute, Kuali Foundation, and Indiana University

Palmer, Carole	Professor and Director	CIRSS, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
Qin, Jian	Associate Professor	School of Information Studies, Syracuse University
Ray, Joyce	Associate Deputy Director for Library Services	Institute of Museum and Library Services
Renear, Allen	Associate Professor and Associate Dean for Research	CIRSS, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
Ross, Seamus	Dean	Faculty of Information, University of Toronto
Seadle, Michael	Dean, Professor & Director	Faculty of Arts, Berlin School of Library and Information Science, Humboldt University
Smith, MacKenzie	Associate Director for Technology	Massachusetts Institute of Technology Libraries
Steinhart, Gail	Research Data & Environmental Sciences Librarian	Cornell University
Tenopir, Carol	Professor, Director of Research, Director of the Center for Information and Communication Studies	School of Information Science, College of Communication and Information, University of Tennessee
Thomas, Chuck	Senior Program Officer	Institute of Museum and Library Services
Tibbo, Helen	Alumni Distinguished Professor	School of Information and Library Science, University of North Carolina at Chapel Hill
Walters, Tyler	Associate Dean	Library and Information Center, Georgia Institute of Technology
Watkinson, Anthony	Senior Lecturer	Department of Information Studies, University College London
Watkinson, Charles	Director	Purdue University Press, Purdue University
Winget, Megan	Assistant Professor	School of Information, University of Texas at Austin
Witt, Michael	Interdisciplinary Research Librarian & Assistant Professor	Distributed Data Curation Center, Purdue University Libraries
Zimmerman, Ann	Research Assistant Professor	School of Information, University of Michigan