TECHNICAL SERVICES REPORT

This column will feature reports on what is going on in the rapidly changing, ever fascinating field of Technical Services. Each quarterly issue will consist of reports on new developments including integrated library systems, next generation catalogs, management of electronic resources; conference happenings and reports from professional meetings; what’s new in technical services publications; as well as reports from technical services professionals on their research and projects. Such reports, announcements, and brief articles for consideration for inclusion should be sent to: Barry B. Baker, Editor, “Technical Services Report,” Director of Libraries, University of Central Florida, P.O. Box 162666, Orlando, FL 32816-2666.

INVOLVING USERS IN PRESERVATION METADATA. A REPORT OF THE ALCTS PARS INTELLECTUAL ACCESS TO PRESERVATION METADATA INTEREST GROUP MEETING. AMERICAN LIBRARY ASSOCIATION MIDWINTER MEETING, PHILADELPHIA, JANUARY 2014

At the 2014 ALA Midwinter Meeting in Philadelphia, the Intellectual Access to Preservation Metadata Interest Group meeting centered on a case study of information professionals collaborating with content creators to generate metadata, in support of preservation. The meeting drew approximately 35 attendees.

After leading a brief business meeting, co-chairs Sarah Potvin and Chelcie Juliet Rowell introduced the speakers, Lorraine L. Richards and Adam Townes. Richards (Assistant Professor) and Townes (Doctoral Candidate) are part of a research group from the College of Computing and Informatics at Drexel University that also includes William C. Regli (Professor) and YuanYuan Feng (Doctoral Student). This group partnered with the Federal Aviation Administration (FAA) to develop a repository system aimed at data preservation and reuse, in service of the FAA’s mission “to provide the safest, most efficient aerospace system in the world.”

The presentation, “Decomposing Results without Burying the Body of Evidence: A Modus Operandi for Developing Metadata and Digital Preservation Requirements,” introduced work emerging from the Drexel–FAA col-
laboration. This research aims to develop a prototype data preservation environment that facilitates the reuse of datasets generated by labs performing simulations and experiments. Initial work has focused on three labs within the FAA’s William J. Hughes Technical Center, a scientific research facility charged with testing and developing aviation systems, procedures, materials, and equipment such as radar devices that detect aircraft. Still early in their partnership with the FAA, Richards and Townes focused on their information gathering and design processes rather than technical architecture during their presentation.

The FAA’s interest in a preservation environment was driven both by an internally recognized need and by broader trends. Scientific research has shifted toward discoveries reliant on collecting, curating, and analyzing vast quantities of data. Richards and Townes pointed to work by Jim Gray, who famously claimed that “The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.”

Additionally, agencies such as the FAA are subject to federal policies currently being formed around data sharing. In 2013, the White House Office of Science and Technology Policy issued a directive to federal agencies with over $100 million in scientific research expenditures, instructing them to develop an “approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research.”

Fulfilling federal mandates aimed at making data accessible, discoverable, and secure poses a challenge to federal agencies, and the FAA is no exception. Adding to these challenges, research has shown that for scientists to feel comfortable reusing data from other scientists they consider three main factors:

- Relevance—Do existing data map to the potential research needs? Do they use the same experimental parameters (in the case of the FAA, e.g., airspeed, type of airplane, nature of ground, and temperature)?
- Understandability—Is there enough documentation to ensure that scientists know the precise way the data are defined, created, and collected?
- Trustworthiness—Understanding how the data is produced increases trust, as does understanding how the previous scientists dealt with data-production problems.

The latter of these factors, understandability and trustworthiness, relate to a broader concept of reproducibility. By emphasizing the verifiability of results, the scientific method necessitates the reproducibility of data. As Richards and Townes explained, in addition to being able to locate data relevant to a
researcher’s information need, the FAA must also support understandability and trustworthiness, which underpin reproducibility. By extension, in order to demonstrate trustworthiness, provenance tracking is crucial.\textsuperscript{4}

With numerous research labs spanning a variety of science disciplines, the current data environment of the William J. Hughes Technical Center is complex and not yet “curation-centric.” Certainly the environment is a “big data” environment: one dataset for one experiment from one research lab may be more than 2.5 terabytes, and every research lab is continually producing data. However, very little sharing of data has been practiced in the past, and management of datasets produced as a result of research has often been siloed within individual research labs. Hand-to-hand data sharing and reuse requires a researcher first to become aware of unadvertised, existing data and second to personally contact original project investigators in order to gain access to and context for that data. Data sharing and reuse in this context relies on the deep institutional knowledge of scientists. Even at an organization with high retention, the turnover of principal investigators is inevitable, and de facto hand-to-hand policies for data sharing and reuse are no longer sustainable.

Operating within a fourth paradigm of science and meeting federal mandates for data sharing are not achievable as a one-time effort; they require ongoing and iterative work. Human factors pose a significant barrier to robust data curation. Richards and Townes described how, in their work with the FAA, they set out to understand the work of scientists in the labs—through close observation, semi-structured interviews, and gathering feedback on process diagrams—so that they might intervene earlier in the data lifecycle, shaping processes, and encouraging a culture of data curation. In service of this goal, they continually emphasize the value proposition of preservation and metadata. However, the challenge of encouraging a culture of data curation is heightened by the fact that, while the FAA administration forged the partnership with Drexel, the buy-in of scientists at the practitioner level of the agency must be earned. This latter group must be convinced that incorporating new tasks into their processes will prove beneficial to scientific discovery and reduce the need to generate new data from scratch by making more viable the options to reuse an existing dataset or to derive a dataset from an existing one.

The Drexel researchers envision a “curation-centric” data environment at the William J. Hughes Technical Center. Crucially, this environment would enable scenario creation with reusable data, rather than requiring scenario creators to generate new data for each scenario. For example, a scenario might consider a pilot who hits turbulence in Kansas City and how the alteration of a flight instrument in the cockpit might impact his reaction to environmental conditions. In order for scenario creators to be able to reuse existing data, it must be relevant, understandable, and trustworthy. An existing dataset intended for use in a scenario might be appropriate
except for one differing parameter: conditions were sunny whereas a dataset with cloudy conditions is needed. In this case, a new dataset that differs in terms of this particular variable could be derived from the existing dataset. Another cultural change that the Drexel researchers hope to affect relates to the recognition of many communities of potential reusers. While FAA personnel anticipate that reusers will be fellow FAA scientists, in time the Drexel researchers suggest that FAA personnel will also see other federal agencies, academia, private industry, and the public as potential reusers of FAA research data.

This “curation-centric” data environment would facilitate automated search and discovery; recognize which data are most valuable for reuse by a diverse set of scientists; incentivize sharing and reuse; ensure trust in the data; and meet preservation requirements. Therefore, a repository system implemented by the FAA must be able to verify that datasets have gone through reliable processes that ensure that any actions taken upon them—be they to create, interpret, transform, transfer, or in any way modify—are well-documented and maintained alongside the datasets themselves. The repository system envisioned by the Drexel researchers and the FAA would describe not only the data but also the wider context in which it was created, including:

- Results of analysis,
- Goals of the experiment or simulation,
- How data was created and modified throughout the scientific process,
- Experimental parameters,
- Intermediate results,
- Logs,
- Final results, and
- Problems encountered when creating the data for use.

The partnership between Drexel researchers and the FAA offers many takeaways for libraries who are wrestling with the challenges of preservation metadata related to facilitating data sharing and reuse. In fact, this partnership with the FAA is informing the design of Drexel’s institutional repository for digital research data that is being planned jointly by Drexel University Libraries and Drexel’s Applied Informatics Group. Richards and Townes underscored the fact that early intervention in the research lifecycle has implications for preservation metadata. Understanding the work processes of researchers can lead to better support for versioning and provenance tracking—both of which underpin trustworthiness and, by extension, the reproducibility of data. The information gathering methods of the Drexel researchers provide a model for how other libraries can seek to understand the work processes of their researcher communities. Finally, as the partnership between Drexel and the FAA moves forward from information gathering
to implementing a technical architecture for the research data repository, they may make design decisions that are well worth considering by other institutions.

Sarah Potvin  
Texas A&M University  
College Station, TX

Chelcie Juliet Rowell  
Wake Forest University  
Winston-Salem, NC

NOTES

1. Jim Gray, “Jim Gray on eScience: A Transformed Scientific Method,” edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Research, 2007. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf
2. John P. Holdren, “Increasing Access to the Results of Federally Funded Scientific Research.” White House Office of Science and Technology Policy, 2013. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
3. Faniel, Ixchel M. and Trond E. Jacobsen. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues’ Data. Computer Supported Cooperative Work (3-3/4): 355–375. http://link.springer.com/article/10.1007%2Fs10606-010-9117-8
4. Muniswamy-Reddy et al. (2009). Making a Cloud Provenance-Aware. Proceedings of TAPP’09, the First Workshop on Theory and Practice of Provenance. Berkeley, CA: USENIX Association. http://dl.acm.org/citation.cfm?id=1525944

CATALOGING IN RDA AND LINKED DATA ENVIRONMENTS. A REPORT OF THE ALCTS CATALOGING & METADATA MANAGEMENT SECTION, CATALOGING AND CLASSIFICATION RESEARCH INTEREST GROUP MEETING. AMERICAN LIBRARY ASSOCIATION MIDWINTER MEETING, PHILADELPHIA, JANUARY 2014

The ALCTS Cataloging and Classification Research Interest Group meeting at ALA’s 2014 Midwinter Meeting was well attended with 75 people in the audience for three presentations on the topic “Cataloging in RDA and Linked Data environments.”

Peter Spyers-Duran, University of Central Florida Libraries, presented “RDA Transcribing TOC.” In his presentation Spyers-Duran discussed a cataloging workflow that he developed to enrich records in OCLC Connexion by copying and pasting tables of contents (TOC) directly from publishers’ websites into bibliographic records using RDA rule 1.7.1. Under this rule catalogers are allowed to accept metadata without modification when: (1)