ASIS&T Research Data Access and Preservation Summit: Conference Summary

by Reagan Moore and William L. Anderson

Reagan Moore is a professor in the School of Information and Library Science, University of North Carolina at Chapel Hill. He can be reached at rwmoore<at>renci.org

William L. Anderson is a co-founder of Praxis101, a firm based in Austin, Texas, that focuses on user-centered design, social networking and collaboration, software engineering practice and organizational learning. He can be reached at band<at>praxis101.com

The American Society for Information Science and Technology sponsored a Research Data Access and Preservation Summit in Phoenix, Arizona, on April 9-10, 2010. The Summit was chaired by Christine Borgman, professor in the School of Information and Library Science at the University of California, Los Angeles, and by Reagan Moore and Gary Marchionini, 2010 ASIS&T president. Moore and Marchionini are professors in the School of Information and Library Science at the University of North Carolina at Chapel Hill. The primary goal of the Summit was to build a broad perspective on the needs of scientific data management communities, the technology approaches going into production data management systems in support of research and the legal and social implications of sharing data. Objectives for the Summit included the characterization of large-scale data management needs, the identification of representative data management systems and the identification of interoperability practices.

An example of an emerging data management environment is the Australian effort to build a national data management infrastructure that supports all Australian researchers. In the United States, similar National Science Foundation funded efforts are exemplified by the DataNet projects. Infrastructure is being developed to support a wide range of science and engineering disciplines and all of the phases of the scientific data life cycle. By sharing data management ideas across research projects, hearing the actual mechanisms being used to implement institutional repositories and learning the social imperatives behind formation of shared collections, the Summit hoped to build a context within which a national data management infrastructure can be defined.

The Summit had 75 participants from academic libraries, federal repositories and private companies. Most of the participants were from the United States, but participants also came from Australia, Canada, the Netherlands and the United Kingdom. An advisory committee collaborated on the design of the Summit, proposing topics and perspectives for organizing the meeting. The advisory committee members are listed in the accompanying box.

Advisory Committee Members

William Anderson, Praxis101 and University of Texas at Austin
Christine Borgman, University of California, Los Angeles
Hsinchun Chen, University of Arizona
Sayeed Choudhury, Johns Hopkins University
Michael Lesk, Rutgers University
Gary Marchionini, University of North Carolina at Chapel Hill
William Michener, University of New Mexico
Reagan Moore, University of North Carolina at Chapel Hill
Art Pasquinelli, Oracle
Sudha Ram, University of Arizona
Stu Weibel, OCLC

The Summit was organized around a progressive presentation of the challenges facing groups that manage scientific data. To provide as much expertise as possible, six panels were created, each one focused on a specific topic. The expectation was that

NEXT PAGE >

each panel would define basic concepts, present the state of the art and identify dominant research questions. Each panel would then respond to questions from the Summit participants. A key motivation of this report is to track the issues that resonated with Summit attendees and identify new issues to indicate areas where a future conference might place emphasis.

Summit Program: Panel Topics and Discussion

The Summit panels were organized around six aspects of scientific data management. The program with links to the associated presentation files is found here: http://www.asis.org/Conferences/RDAP10/RDAP10Program.html

A summary of the topics and discussion follows.

- **Panel 1: Data Life Cycle Management**
  Mark McFarland, Texas Digital Library
  Erin O'Meara, Carolina Digital Repository
  Stacy Kowalczyk, Indiana University
  Cynthia Ghering, Michigan State University

  The panel experts were from universities implementing digital libraries or institutional repositories. Since most of the participants were from academic institutions, this representation provided a firm grounding in the digital data management approaches being pursued by university libraries. One long-term goal for this group is the integration of scientific data collections into reference collections housed within institutional repositories.

- **Panel 2: Promoting Re-use of Scientific Data Collections**
  Peter Wittenburg, Max Planck Society, Institute for Psycholinguistics
  Sudha Ram, iPlant Collaborative
  Roy Williams, International Virtual Observatory Alliance
  John Harrison, Sustaining Heritage Access through Multivalent Archiving
  Jonathan Crabtree, Odum Social Science Institute

  The panel experts were from national-scale research projects that face the challenges of organizing scientific data collections for use by multiple researchers. This application requires ingest of observational data, experimental data or simulation output and the organization of the data into sharable collections, as well as the publication of the data for use by the broader discipline and the preservation of the data as reference collection for use by future researchers. At each stage of the data life cycle, a broader community re-purposes the data for their specific uses. This perspective emphasizes social implications of use, technical challenges for managing massive amounts of data and expectations for how the data will be accessed and used.

- **Panel 3: Large-Scale Data Management Challenges**
  John Graybeal, Ocean Observatories Initiative
  Ken Galluppi, Renaissance Computing Institute
  Laurin Herr, CineGrid
  Philip Maechling, Southern California Earthquake Center

  The panel experts were from national-scale projects faced with massive data collections (petabytes to hundreds of petabytes of data), management of sensor data streams and management of highly distributed data. Since every data management environment needs to distribute data across multiple locations to minimize risk of data loss, the distributed data management approaches were expected to point to technologies that would be beneficial to everyone. The panelists represented projects that are assembling some of the largest academic data collections being created.

- **Panel 4: DataNet Federation**
  Sayeed Choudhury, Data Conservancy
  William Michener, DataOne
  Reagan Moore, Data Grids

  The panel experts included funded projects from the National Science Foundation (NSF) DataNet solicitation and current data grid software technology developers.

NSF is pursuing development of infrastructure to support digital libraries, large scientific data collections, the full data life cycle and long-term sustainability. The expectation is that technology developed within the DataNet partners will be useful to institutional repositories and digital libraries, as well as large-scale data grids and preservation environments. The issue of interoperability is critical for linking data management solutions across all stages of the data life cycle. Data grids provide infrastructure for shared collections that span multiple storage environments.

- **Panel 5: Developing Assessment Criteria**

  Mark Conrad, National Archives and Records Administration
  Jane Greenberg, University of North Carolina at Chapel Hill and the Dryad Project
  John Graybeal, Ocean Observatories Initiative
  Steve Richard, Arizona Geological Society

  The panel experts represented national projects that are developing domain-specific descriptions for collection context, defining the properties that the collection should maintain and proposing criteria to validate the collection properties. Each community is tackling these challenges on its own – developing specific semantics, ontologies, policies and assessment criteria. A long-range goal is the identification of common properties that all collections should manage, common policies and shared semantics.

- **Panel 6: Legal and Social Implications of Shared Collections**

  Ann Zimmerman, University of Michigan
  Melissa Cragin, University of Illinois
  Noshir Contractor, Northwestern University

  The panel experts represented projects that are assessing motivations for sharing data, identifying the criteria under which collection sharing takes place and considering development of incentives for sharing data. This panel described the motivations for the construction of institutional repositories as well as those for disciplines to build shared collections. The social aspects of sharing strongly influence the success of attempts to re-purpose collections for new uses and develop reference collections for future researchers. A central question is the identification of incentives for researchers to share their data.

Two additional sessions were held to allow technology developers to present live demonstrations of their systems and give tutorials on use of the software. A poster and demonstration session was held on Friday evening, and a tutorial session was held on Saturday afternoon. The Friday evening session provided demonstrations of the LStore high-performance distributed data management system (Alan Tackett), the Sustainable Heritage through Multivalent Archiving (SHAMAN) preservation technology (John Harrison), the Fedora digital library middleware (Brad McLean) and the iRODS integrated Rule-Oriented Data System (Reagan Moore). The Saturday session included tutorials on Fedora (Brad McLean) and iRODS (Reagan Moore).

Summary Conclusions, Hypotheses and Questions

Conclusions

1. Institutional repositories are emerging within academia with the goal of building digital reference collections.
2. National-scale research projects are also assembling scientific data collections that represent the digital holdings for science and engineering disciplines.
3. The scale of digital collections is increasing beyond the capacity of institutional repositories, with major holdings housed in federal repositories.
4. The environments of institutional repositories, discipline-specific collections and federal repositories need to interoperate.

5. Initiatives are starting to drive formation of national-scale data infrastructure (NSF DataNet, NSF Teragrid, DOE Open Science Grid, DOE Earth Systems Grid, NOAA CLASS system, NASA DAACs). These systems need to interoperate.
6. Standards are being developed for digital data description (provenance, context), digital data representation (structure, processing mechanisms), digital repository trustworthiness (assessment criteria). These standards need to interoperate.
7. An understanding is being developed of the social motivations to encourage data sharing, formation of shared collections, collaborative research. The creation of a shared collection is a social process, requiring consensus on the properties that the digital objects within the shared collection will possess, the policies that will be used to manage the desired properties, the procedures that will be executed to enforce the properties and the assessment criteria that will be used to validate the original intent for forming the shared collection.

Hypotheses and other questions about the futures of scientific data collections

1. Will there be convergence among the multiple types of data management applications (data grid, digital library, persistent archive, data processing pipeline)? Can data management applications be characterized by the procedures and policies that control the collection?
2. Given that a social consensus drives the formation of a shared collection, what kind of mechanisms will emerge for communities to develop a consensus on collection sharing?
3. Given the massive size of emerging collections (tens to hundreds of petabytes), what sorts of computing and storage integration are essential? Will massive collections force the movement of the application to the data, instead of moving the data to the application?
4. Does the federation of collections drive long-term sustainability between institutions and the re-purposing of collections for use by new communities? What sorts of social agreements will emerge defining policies controlling re-use of collections? Can long-term sustainability be turned into a policy on collection re-use by requiring that the original use be sustained as an alternate set of procedures and policies within the new environment?
5. What forms of distributed data will become common across all data management applications to minimize risk of data loss, to ensure experts have a local copy and to facilitate data analyses?
6. Is a national data grid feasible? Could such a grid support research initiatives across federal agencies?
7. What kinds of federation of federal repositories are feasible? Can a unifying data management infrastructure support all federal activities?
8. How can social networks be formed to promote sharing and maintenance of data collections? What new mechanisms for social interaction are needed to promote development of consensus and agreements by communities?
9. Given that the multiple stages of the data life cycle correspond to re-purposing of a collection for use by broader communities, what are the social mechanisms for formation and maintenance of these broader communities?
10. Will collections become the standard for organizing data instead of file systems? The context provided by collections is essential for understanding how to discover, browse, access and manipulate data.

Participant Commentary: Summit Tweets

Summit participants used Twitter and the hashtag #rdap10 to share information about presentations and discussions. An online notebook of all the tweets so tagged is found here: www.twapperkeeper.com/hashtag/rdap10 ■