Report from the

Research Data Workforce Summit

Sponsored by the Data Conservancy

6 December 2010

Chicago, IL

Summit coordinated by:

Carole L. Palmer & Melissa H. Cragin

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

Bryan Heidorn

School of Information Resources and Library Science
University of Arizona

Report prepared by:

Virgil E. Varvel Jr., Carole L. Palmer, Tiffany Chao, & Simone Sacchi
Center for Informatics Research in Science & Scholarship

OUTLINE

OVERVIEW

The Research Data Workforce Summit (RDWS) was a one-day exchange on research data workforce development in the sciences, held in Chicago on December 6th, 2010, in conjunction with the 6th International Digital Curation Conference (IDCC). The RDWS was sponsored by the Data Conservancy, which is part of the DataNet initiative funded by the National Science Foundation, Office of Cyberinfractructure. The IDCC conference, which followed on December 6-8, was co-hosted by the Graduate School of Library & Information Science at the University of Illinois at Urbana-Champaign in partnership with the UK Digital Curation Centre and the Coalition for Networked Information (CNI). The theme of the conference was "Participation and Practice: Growing the Curation Community through the Data Decade,"[1] providing an excellent opportunity to bring together a group of data curation experts and educators around data workforce issues. The 29 invited participants included representatives from government agencies and data centers, the NSF DataNet projects, universities with active programs in data science and the curation of research data, and other schools that are actively training information professionals in digital curation, e-science, and related areas. See Appendix B for information on participants.

The summit provided a forum for sharing views on the research data workforce, with an emphasis on current practices and needs, projected changes in the future, and educational programs for advancing data expertise in the sciences. Challenges faced by governmental and affiliated organizations were of particular interest, in recognition of the increasing expectations for government agencies to curate and share data and the lack of well-trained professionals to meet the demand.[2] Governments have special concerns and needs for long-term information management for internal use, moreover many government agencies also have a mandate to create, gather, and disseminate data for the public.

---

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

[2] Interagency Working Group on Digital Data. (2009, January). *Harnessing the Power of Digital Data for Science and Society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council.* Washington, DC: Office of Science and Technology Policy. Retrieved from www.nitrd.gov/About/Harnessing_Power_ Web.pdf

The summit was organized in four sessions representing perspectives from the four interest groups: government agencies, national scientific data centers, DataNet partners, and university educators. Invited speakers provided short presentations, with the first two sessions covering workforce issues and the second two sessions covering current education efforts. Speakers were asked to address the following questions:

- What are the current research data activities in your organizations and how do they relate to the broader scientific domains?
- What are the immediate and near term workforce needs for research data in your organization and affiliated organizations and initiatives?
- What are the gaps in our current education programs? How do we stay ahead of the curve on state-of-the-art practices in our curriculum?

The speakers presented a range of perspectives from different organizational and educational contexts, and the discussion that emerged throughout the day reached beyond government agency contexts and was generally applicable to the curation and use of data in a broad range of research organizations.

SUMMARY

In their opening remarks, summit organizers conveyed the importance of creating a platform to share insights and experience on data workforce education and practical knowledge on how data are curated and managed in science. They noted the need to also assess current gaps in education and practice, as well as the expectations and needs of various stakeholder groups (i.e. information professionals, data scientists, scientists, students). Government agency representatives from the Department of Energy (DOE) and the Institute for Library and Museum Services (IMLS) emphasized the level of support being provided for the development of education and training opportunities to produce a workforce of data experts, and a new generation of scientists that are more conscious of data management responsibilities. They noted that data management plans are becoming an increasingly important component of funding proposals at their agencies, similar to the trend at the National Science Foundation (NSF). A recent workshop on data management organized by the Earth Science Information Partners (ESIP) Federation was identified as an example of a successful effort to address issues concerning government employees who work with data on a regular basis. That group has made solid progress on articulation of long-term archiving principles and methodologies for data management, sharing, access, and re-use across agencies (http://wiki.esipfed.org/index.php/_Data_Management_Workshop).

Data center representatives stressed their need to improve data services for scientists and other data users, as well as a desire for better relationships with educators in data curation and data management and other domain-related data experts. The DataNet projects, the Data Conservancy and DataONE, covered current activities in higher and continuing education, emphasizing the need to identify and implement metrics for assessing the needs of potential stakeholders and users, and to develop up-to-date training programs to prepare professionals to provide services, resources, and tools to meet those needs. Educators from iSchools (www.ischools.org) and other university based education programs provided overviews of their activities in data curation and data science, reflecting on the successes and barriers encountered to date.

The summit moderator, Lucy Nowell from DOE, provided initial framing on government contexts and shared integrative comments throughout the day to keep the group focused on current key

issues and problems. She closed the discussion by drawing attention to the recent report by Shoshani, & Rotem (2010)[3] and stressing the need for curation services to scale up to meet the current state of rapid growth in both data and computing, noting the particular need for education around parallel process computing with large sets of data. At the same time, she highlighted the significant size and importance of small science data, and the problems associated with its highly complex and variable organization and storage patterns. A post-summit update from Nowell drew attention to the America Competes Reauthorization Act of 2010, which requires science agencies to develop policies regarding data management and preservation by the end of 2011, suggesting an urgent need for communication forums on data management that promote exchange across communities invested in scientific data.

[3] Shoshani, A., & Rotem, D. (Eds.) (2010). *Scientific Data Management: Challenges, Technology, and Deployment*. Boca Raton, FL: Chapman & Hall / CRC Press.

CROSS-CUTTING THEMES

Over the course of the summit, three themes were prominent across the presentations and discussion: advancing professional education, coordination across disciplines and sectors, and key educational challenges.

ADVANCING PROFESSIONAL EDUCATION

One of the primary problems for advancing the field is the need to disambiguate and develop definitions for professional roles, such as "data manager," "data curator", and "data scientist," which tend to be used loosely in the science and information science communities. It will be important for education initiatives to work together to formulate agreement on terms and definitions and then to apply them consistently in programs across the country. A shared vocabulary is an essential step forward in professionalization and will promote more coherent and productive discourse in the community and in communicating with constituencies. Note, that since the terms related to professional roles were used interchangeably by participants throughout the summit, in this report we have opted to use the more general term "data professional" wherever appropriate to represent the range of roles in the data workforce

While there is much work to be done to fully articulate the responsibilities and skills associated with data professionals in the sciences, metadata standards were emphasized as an essential area of expertise. Appropriate and systematic application of metadata will provide the foundation for assuring the future accessibility and effective re-use of data, and therefore it is a core part of the curatorial work required for the deposit or ingest of data into repositories. However, since there is no formal organization for coordination and enforcement of standards, data professionals will need to develop and share their evolving metadata practices. Moreover, since data professionals need to be able to manage data across multiple disciplines, and recognize and enhance the value of data for reuse beyond its original intent, data professionals need to have knowledge of the range of data formats and the growing array of associated standards.

Four educators presented highlights of the programs at their universities: Rensselaer Polytechnic Institute (RPI) and George Mason University, and two iSchool programs at the University of

Arizona and the University of Michigan. The iSchool programs have been designed for training information professionals in data curation and data management, and RPI and George Mason are focused on training in data science, informatics, and data management for students primarily in the sciences. The Data Conservancy is actively extending existing masters level information science curriculum at Illinois and UCLA, while also providing continuing education at Illinois. DataOne education efforts are currently concentrated on community engagement rather than formal university based programs. Particular challenges identified by the group of educators included student recruitment, providing students with practical experience and mentorship, and learning objectives and curriculum development in this emerging and fast paced field.

As a new field of study, recruiting students is particularly difficult and requires new strategies for identifying potential students and building awareness of programs with advisors and other educators. Undergraduate computer science students were identified as a key target for recruiting, but they tend to be primarily interested in more traditional technology positions in industry and currently have little incentive to become involved in programs focused on data. Science laboratory courses at the undergraduate level were seen as an important channel for infusing sound data management practice into more general science curriculum and could also be used to attract students into the data profession.

Practical experience is considered essential for students to gain skills and knowledge, and programs are making progress developing internship and practicum opportunities. It was noted that there is consistent demand for interns from certain organizations that have a growing need for curation and management of digital data and little in-house expertise. While individual placements over the past few years have generally been successful, the long-term outcomes of these efforts have not yet been formally evaluated. There is a clear need for current practicing experts in the data community to get more involved in formal education, since they can be invaluable as mentors and instructors, bringing much needed front-lines experience to professional education.

In the area of curriculum development, participants emphasized computational science, statistics, and digital preservation as core areas that should be integral components of data programs. Theoretical concepts and interdisciplinary collaboration are important areas for graduate curriculum, where students are better prepared to engage with constructs at this level. Curriculum

development requires strategic structuring and sequencing of courses across undergraduate and graduate programs. And, while there has been considerable progress on curriculum for undergraduate and graduate students, there remains a clear and urgent need to adapt and deliver similar content for continuing professional education for the current workforce.

Three recommendations emerged for advancing existing professional education programs:

- Craft program and course descriptions to be more attractive to target pools of students.
- Reach out and recruit students from disciplines from departments across campus.
- Partner with data centers to facilitate internships and field experiences for students that include mentoring by data professionals.

Coordination across Disciplines and Sectors

The presentations by government and data center representatives generated serious discussion on the need for interdisciplinary and multi-disciplinary training approaches and the importance of cross-institutional solutions to data problems. To be effective, data professionals will need a high degree of cross-field awareness, but that will not be sufficient. There will need to be substantive communication and connections between science and data level operations in the organizations where data are generated and used, and in the repositories and centers that provide data services. Integration across knowledge bases is also fundamental to development of the data profession more generally. The blend of competencies, or the 'tridge'—the intersection of domain science, information science, and computer science—will be required to address the coming challenges involved in building functional, interoperable, cross-disciplinary networks of data resources. Knowledge and technologies will also need to be harnessed from public and private sectors. It was noted that educational programs aimed at scientists and especially those associated with data intensive science should also cover data sharing, access, and use across disciplines.

To be effective and innovative, data professionals will need the ability to learn and work across disciplines. Organizations that have successfully established practices for working across disciplinary boundaries and for engaging multiple fields in their data operations can provide highly instructive case studies and professionals with deep experience to share. At the same time, in some segments of the workforce there will need to be divisions of labor and specialization. For example,

data professionals in specialized research centers will require more extensive domain knowledge and specific computational expertise, while those in data repositories will require cross-domain understanding of science and curation, as well as cyberinfrastructure and interoperability expertise.

As cyberinfrastructure evolves, data workflows will increasingly cross the borders of a single institution. Strong communication and working relationships will need to be established between scientists who generate and use data and the data professionals who support their practices and contribute to the development of the global data network. As data professionals build a shared understanding of disciplinary data practices, it will be important to work with scientists to evaluate the impact of various curatorial and data management strategies on the conduct of science. Participants considered consortia an important approach to providing coordination and exchange in the educational arena, and for contributing to advancing the profession in areas of infrastructure, new services and data products, and for promoting the reuse of data. There was strong support for development of working groups that cross institutions and disciplines to make progress on such coordinated efforts. It was recognized, however, that this requires extensive community engagement to foster and consolidate stakeholder networks.

Two recommendations emerged for promoting coordination and integration across fields:

- Design training programs that are integrative and general yet allow for development of specialized expertise.
- Develop a consortium to support professional data education that crosses domain sciences, information science, and computer science, as well as institutional and international boundaries.

KEY EDUCATIONAL CHALLENGES

Much is changing in the current scientific environment with the rise in big data and computational approaches to analysis. Education has not kept up with this new paradigm, particularly the trend toward concurrent programming and parallel processing. For example, in some computer science faculty are resistant to the new methods needed for dealing with peta-scale or exa-scale data and

some departments are not providing training in true parallel processing. There is a sense that most faculty teach programming the way they learned it, within a single processor environment, and that they are prohibited by the time commitment required to shift to meet the demands of the new scales of practice. Government agencies, however, are providing funding for graduate fellowships and early-career programs for junior faculty to support research in parallel processing at scale and addressing the data challenges in this new paradigm. Making progress on these kinds of technical deficiencies in education is difficult, but as several participants stated, they are not as challenging as many social and cultural aspects of data production and use. Educational programs will need to train data professionals to function within the differing existing cultures but also to intervene and promote data awareness and best practices and ultimately change how data assets are valued and handled in the enterprise of science.

Since professional responsibilities will vary and require navigating and supporting multiple disciplines, it is difficult to determine in advance the level of subject expertise required for an entry-level data professional. Masters level education or comparable experience in science was seen by some to be essential for data professionals to manage the issues related to domain-specific practices and needs, and the related terminology. However, this is a long-standing issue in the information professions, where training in a single discipline does not provide adequate breadth or the technical and theoretical foundation needed to provide information services to diverse user communities.

A number of problems were raised around providing students with practical field experience. Within data centers there are no reward structures in place to encourage involvement in the education and training of students. At present, data programs often provide practical experience through internships and fellowships within scientific research operations, however some students could benefit more from field experiences with data practitioners at data centers and repositories. These organizations need to be more proactive in the education arena by better publicizing their activities and opportunities for students and providing incentives for data professionals to serve as mentors. Apprenticeship approaches were praised as an effective approach for transferring current skills and competencies. At the same time, it was recognized that actual practice in the workplace is not always best practice or may not represent the state-of-the-art.

It is expected that the professionalization of responsibilities and skills for research data will be uneven, taking hold in some disciplines but not in others. In a number of fields, scientists are assuming data management roles—developing competencies as needed, perhaps with no expectation of allocating of these duties to trained or experienced data professionals. While these scientists can benefit from initiatives and resources in the developing profession, they also should be recognized as a functioning part of the community and a source of knowledge, and need to be part of coordinated education initiatives and development of best practices.

Recommendations for addressing key education challenges:

- Begin substantive curriculum revision to address the current gap in concurrent programming and parallel processing.
- Promote recognition of data workforce education and research within reward structures in universities, research organizations, and data centers.
- Support documentation of emerging best practices, as well as areas where new or improved practices need to be developed, for dissemination in the education community.

FUTURE DIRECTIONS

In the wrap-up session, participants identified three priorities for continued discussion and collaboration among the summit group:

1) **Differentiate and establish definitions for professional data roles.**

Across the schools, programs in data curation, data management, and data science address similar topic areas and problems, but they also have important variations in emphasis. Clarification and branding is critical for strengthening the identity of academic programs, but also for development of job titles and position descriptions within scientific organizations. Terminology for professional roles, perhaps building on those defined in the 2009 Interagency Working Group report,[4] could provide a unified base of understanding for scoping education programs. It would also benefit employers recruiting research data professionals and help them provide career paths in their organizations. It was suggested that a common certification might be developed among the iSchools or some other organized group, but any such effort would need to accommodate the need for specializations within the emerging profession and the distinct contributions cultivated by individual programs.

2) **Continue to build the data education community and promote awareness of the different activities at iSchools and other departments and institutions.**

Several activities were identified as potential first steps in building the education community. First, there is interest in developing a web presence that serves as a knowledge base for current initiatives and makes potential opportunities for teaching and learning visible to faculty and students. The "education hub" currently under development by the Data Conservancy may provide an initial platform, but there will need to be a mechanism that allows for coordination and growth in response to the community. The initial release will include this report and a database of courses and programs in data curation and closely related library and information science curriculum in the

[4] See, in particular, Appendix C, Exhibit C-4 in Interagency Working Group on Digital Data. (2009, January). *Harnessing the Power of Digital Data for Science and Society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council.* Washington, DC: Office of Science and Technology Policy. Retrieved from www.nitrd.gov/About/Harnessing_Power_ Web.pdf

U.S. In addition, it was suggested that the summit group establish ties with international organizations, such as the Digital Curation Centre, and coordinate with university libraries to allow better interaction among science data efforts and more traditional library science.

3) **Determine workforce needs across different environments to inform development of existing and new programs.**

Workforce roles and needs should be assessed in a number of ways, such as surveys and interviews, analysis of job descriptions, and additional summit events focused on gathering this kind of information. It is important to note that new studies should build upon prior and ongoing work in this area, but more research will be vital for advancing programs in higher and continuing education. This is a new and dynamic field, and keeping curriculum current will be an ongoing challenge for educators who must train professionals to support the all the sciences producing data of value, ranging from individual small science projects to very large-scale, highly computational data operations.

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.

Government Agencies

Lucy Nowell, U. S. Department of Energy

- Complex data environment – increase in scale, diversity, and complex uses of data
- Need for stronger, focused education programs
  - Intellectual paradigms
  - Programming at concurrency of larger scale
  - Knowledge representation across disciplines to support data integration
- Need to facilitate data manager career path
  - Coherent data manager definition
  - Analysis of data is critical to science as is visualization of data, but these should not necessarily be separate jobs.

Joyce Ray, Institute of Museum and Library Services

- Progress with programs targeting training the next generation of librarians and data curation specialists
- Important recent step with IMLS now requiring a data management plan with research grant applications
- Researcher role as domain specialist is distinct from role of data manager
- Data managers need to accommodate scientists while supporting use across disciplines
- Data management requires collaborative approach involving data discovery, data standards, and cross-disciplinary communication.

Data Centers

Bruce Wilson, Oak Ridge National Laboratory, DataOne – Finding and Making Bridge Builders for Research Informatics

- Data management requires interaction among domain science, information science, and computer science, with data managers working as 'tridge' builders and walkers among all three.
- Critical for combined practical experience to come together across communities
- Educational gaps and challenges require:
  - Multidisciplinary team approach
  - Experienced instructors
  - Support of research and sociological aspects
  - Better defined interdisciplinary units on par with departmental silos
  - Strategies for recruiting students and the faculty

Don Collins, National Oceanographic and Atmospheric Administration (NOAA)

- Already massive number of data collectors and amounts of data, and facing tremendous increases production
- NOAA embraces OAIS reference model with emphasis on submission agreements and automated systems working with standards.
- Immediate and near-term workforce needs:
  - Subject matter expertise
  - Ability to handle data in large quantities in many formats
  - Metadata skills
- Current gaps in educational programs:
  - Communication
  - Connections with the community
  - Professionalization of data management role

Bob Downs, Center for International Earth Science Information Network (CIESIN) – Developing the Data Center Workforce for Long-Term Management of Scientific Data

- Scientific work is increasingly in a multidisciplinary environment, with need to enable the use of research data by diverse scientific disciplines.
- Data center knowledge and skills:
  - Appraise, prepare, and describe data while promoting and enabling discovery
  - Identify systems, tools, and applications that reduce costs and improve quality
  - Develop new products and services based on existing and new data that are useful on the global scale for the long-term
- Data providers knowledge and skills:
  - Prepare, document, and identify data
  - Recognize authors and contributors
  - Identify rights holders and restrictions
- Data users knowledge and skills:
  - Locate, identify, and access data
  - Acquire appropriate rights to that data
  - Provide attribution

DataNet Education Initiatives

Data Conservancy
Sayeed Choudhury – Data Conservancy: A Blueprint for Research Libraries
Carole Palmer – Leveraging Data Conservancy R & D to Advance Data Curation Education

- Three pillars of libraries—collections, services, and infrastructure—all relevant for data
- Data potentially new kind of special collections, however curation needs to start further up-stream rather than at end of data life cycle
- Data curation as a means and not an end.
- Data scientists are human interface between domain scientists and data management professionals—interoperability more difficult among humans than machines

- Data professionals need to provide access to a broad landscape of information across scales, disciplines, institutions, and generations.
- Illinois building on established data curation masters specialization, summer institute for series for professional development
- New program integrating field experience into masters and doctoral programs
- Need to build awareness and share resources among existing programs
- Greatest challenges:
  - Documenting best practices
  - Informing curriculum with research-based knowledge
  - Recruiting students with background in science
  - Managing internships
  - Monitoring the employment market

Data ONE
Bill Michener – Changing Community Practice and Transforming the Environmental Sciences
Amber Budden – Advancing DataONE Outreach and Education Initiatives

- New cyberinfrastructure needs to build on existing structures and support communities of practice.
- DataOne components include member nodes at diverse institutions and coordinating nodes, and an investigator toolkit with commonly used tools, to be extended worldwide.
- Purpose is to support data-intensive science through seamless access, use, reuse, and trust of data.
- Categories of challenges: metadata, interoperability, data integration, representation, reasoning, trust and quality assessment, reward system through citation and use statistics, and education and training.
- Baseline community assessment shows that many scientists interested in sharing data, but often with conditions. Barriers can be reduced through education, training, and the proper tools.

University Educators

Kirk Borne, George Mason, Data Science Program – Informatics in Education and an Education in Informatics

- Everyone will encounter data—all professions, in citizen science, and everyday life.
- All need skill sets and to understand ethics of data usage
- Problems with terminology used in hiring data managers
- Need for common language around data management
- To develop further as a profession, university administration needs to recognize the emerging profession and faculty contributions.

Peter Fox, Tetherless World Constellation, Rensselaer Polytechnic Institute, Earth and Environmental Sciences – Curriculum Development at the Tetherless World Constellation

- Multi-disciplinary science program at RPI: web science, data sciences, x-informatics, semantic e-science
- Data science requires recognition of how science interacts with information in terms of the scientific, policy, computation, and social aspects.
- Focus on clear teaching and learning objectives in instruction
- Learning objectives for future scientists and future technologists
- Successful programs require buy-in from employers and industry.

Bryan Heidorn, University of Arizona, School of Information Resources and Library Science, Digital Information Management

- Important to leverage previous work experience of incoming students
- Current education program is focused on practice and knowledge acquisition.
- Need to articulate positive ways that data management skills and competencies can improve job operations and influence organizational change
- Can develop new programs without additional funding by exploiting synergies across existing programs

Margaret Hedstrom, University of Michigan, School of Information, Integrative Graduate Education and Research Traineeship (IGERT)

- Integration of data curation and data science for educating scientists to manage data and data managers, and educating computer and information scientists to understand domain practices and needs
- Need to identify generic and specific principles, methods, and tools across disciplines
- Selection of data to be kept essential for investing in valuable data rather than worthless, high-volume data
- As with other forms of communication, need for data peer review tied to the perceived value of data in the designated community

APPENDIX A

MEETING AGENDA

9:00 – 9:20 – Welcome

- **Background & Objectives**
  Carole Palmer, Director, Center for Informatics Research in Science & Scholarship, University of Illinois

- **Summit Overview**
  Bryan Heidorn, Director, School of Information Resources and Library Science, University of Arizona

9:20 - 10:15 - Government Perspectives

- Lucy Nowell – Department of Energy
- Joyce Ray – Institute of Museum and Library Services (IMLS)

10:15 – 10:30 – Break

10:30 – 11:30 - Perspectives from Data Centers and Initiatives

- Bruce Wilson – Oak Ridge National Laboratory
  Finding and Making Bridge Builders for Research Informatics
- Donald Collins – National Oceanic and Atmospheric Administration
- Bob Downs - Center for International Earth Science Information Network (CIESIN)
  Developing the Data Center Workforce for Long-Term Management of Scientific Data

11:30 – 12:00 - Panel discussion on gaps in workforce

12:00-1:00 – Lunch

1:00 – 2:00 – Current DataNet Education Initiatives

- **Data Conservancy**
  - Sayeed Choudhury – Johns Hopkins University
    Data Conservancy: A Blueprint for Research Libraries
  - Carole Palmer – University of Illinois at Urbana-Champaign
    Leveraging Data Conservancy R & D to Advance Data Curation Education
- **DataOne**
  - Bill Michener – University of New Mexico
    DataONE: Changing Community Practice and Transforming the Environmental Sciences
  - Amber Budden – Director for Community Engagement and Outreach
    Advancing DataONE Outreach and Education Initiatives

2:00-3:00 – Educator Perspectives

- Kirk Borne – George Mason Data Science Program
  Informatics in Education and an Education in Informatics
- Peter Fox – Rensselaer Polytechnic Institute
  Curriculum Development at the Tetherless World Constellation
- Bryan Heidorn – University of Arizona
- Margaret Hedstrom – University of Michigan

3:00-3:15 – Break

3:15 – 4:00- Discussion on areas for development and collaboration

APPENDIX B

PARTICIPANT INFORMATION

SUZIE ALLARD

Associate Professor and Assistant Director, **School of Information Sciences, University of Tennessee at Knoxville**

Background Statement: Suzie Allard is an Associate Professor and Assistant Director of the School of Information Sciences at the University of Tennessee. Her research focuses on science information and communication, particularly the full life cycle of earth environmental information, and how scientists use and communicate information to improve science data practices. Dr. Allard's work includes studies conducted at labs across the U.S. and in India.

CHRISTINE L. BORGMAN

Professor & Presidential Chair in Information Studies, **Graduate School of Education and Information Studies, University of California at Los Angeles**

Background Statement: What's missing from the data curation curriculum? The Data Conservancy (http://www.dataconservancy.org) embraces a shared vision: scientific data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society.

Viewing curation as a ***means*** rather than as ***ends*** requires a research data workforce with deep knowledge of the scientific process. In developing a two-course sequence entitled "Data, Data Practices, and Data Curation" for the graduate curriculum at UCLA, it became apparent that the LIS approach tends to begin not at the beginning of the data life cycle but near the end, once data have been transferred to librarians and archivists for safekeeping. Our research on data practices in multiple scientific domains reveals that much essential knowledge about the data may have been lost by this stage. We have devoted the foundational UCLA course to addressing the ways in which curation can serve as a means to facilitate knowledge discovery. With this background, the second course is devoted to handling data on behalf of scientific communities and collaborations.

KIRK BORNE

Associate Professor of Astrophysics and Computational Science, **Computational and Data Sciences, George Mason University**

Speaker Bio: Kirk Borne is Associate Professor of Astrophysics and Computational Science in the Department of Computational and Data Sciences at George Mason University. He has 30 years of research experience in astrophysics, but his research took a turn into Data Sciences about 10 years ago. Since then, he has contributed to several large data projects, including NASA's Astronomical Data Center, the National Space Science Data Center, the National Virtual Observatory, the Zooniverse Citizen Science project, and the future Large Synoptic Survey Telescope (LSST), which will produce one of the world's largest scientific data collections. He is chairman of the LSST Informatics and Statistics research collaboration team, a member of the LSST

education and public outreach (EPO) Advisory Board, and a major contributing scientist to the LSST EPO program. In these roles, he advances the science of Discovery Informatics (which focuses on achieving big science discoveries from big data), and he promotes the use of informatics research experiences with big data in the STEM education pipeline at all levels.

Educational Interests and Activities: About 9 years ago I discovered the incredible science of data mining, which is the application of mathematical algorithms to the problem of discovering hidden (sometimes surprising) knowledge within large databases. I soon realized that skills in Data Science are absolutely critical for every future scientist and (in fact) for every future citizen. This is because science, government, and industry are all generating massive (and exponentially) growing quantities of data. Without training in the skills of Data Science, science disciplines and societal organizations will never reap the full benefits (scientific or otherwise) from their enormous data collections.

PETER BOTTICELLI

Assistant Professor of Practice, **School of Information Resources and Library Science, University of Arizona**

Background Statement: Dr. Peter Botticelli, Assistant Professor of Practice, directs the Digital Information Management (DigIn) program at the University of Arizona School of Information Resources and Library Science (SIRLS). Dr. Botticelli teaches courses in the certificate program as well as in SIRLS' master's program, focusing on data curation, digital librarianship, scholarly communication, and digital preservation. He is currently a PI for two IMLS-funded research projects, one of which is focused on the development of virtual labs and authentic technology learning methods for online courses on digital curation. A second grant is investigating best practices for presenting culturally sensitive data on the Web.

GEOFFREY BROWN

Professor, **Indiana University**

Background Statement: My primary research focus as well as current graduate teaching is on technologies to support long-term access to born-digital materials. A key component of this work, and the dissertation topic for recent PhD Kam Woods, involves the nearly 5000 CD-ROMs published by the United States Government Printing Office. Our work includes capturing bit-faithful disk images, enabling web-based browsing of the image contents, on-the-fly migration of obsolete data formats, and emulation to support data-sets requiring obsolete software. Recently we have begun the study of risk-assessment for migration of scientific data -- specifically tools to reduce the cost of quality assurance.

AMBER BUDDEN

Director for Community Engagement and Outreach, **DataONE**

Speaker Bio: Dr Amber E Budden is Director for Community Engagement and Outreach at DataONE. In this role she engages the community as Vice-Chair of the DataONE Users Group, through participation in Education and Outreach Working Groups and directly via the organization and co-instruction of data management and best practices training sessions. She has a joint BSc in Psychology and Zoology from the University of Bristol and a PhD in Behavioral Ecology form the University of Wales,

Bangor. Prior to joining DataONE, Dr Budden engaged in ecological and sociological research as a postdoctoral fellow at the University of California Berkeley and at the National Center for Ecological Analysis and Synthesis. Her ecological research focussed on avian parental care and parent-offspring conflict and her other research explored the use of bibliometrics in research evaluation, bias in publishing, and scientific workforce composition. Dr Budden has been involved in postdoctoral representation and was president of the Berkeley Postdoctoral Association and member of the UC Council of Postdoctoral Scholars from 2002 to 2003, chaired the National Postdoctoral Association Publications committee from 2003 to 2007 and served on the Board of Directors of the National Postdoctoral Association during 2005 and 2006.

SAYEED CHOUDHURY

Associate Dean for Library Digital Programs, Hodson Director of the Digital Research and Curation Center, Data Conservancy Principle Investigator Sheridan Libraries, Johns Hopkins University

Speaker Bio: G. Sayeed Choudhury is the Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University. He is also the Director of Operations for the Institute of Data Intensive Engineering and Science (IDIES) based at Johns Hopkins. He is a Senior Presidential Fellow with the Council on Library and Information Resources, member of the ICPSR Council, DuraSpace Board and the Digital Library Federation advisory committee. He has been a Lecturer in the Department of Computer Science at Johns Hopkins and a Research Fellow at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign.

Choudhury serves as principal investigator for projects funded through the National Science Foundation, Institute of Museum and Library Services, and the Andrew W. Mellon Foundation. He is the Principal Investigator for the Data Conservancy, one of the awards through NSF's DataNet program. He has oversight for the digital library activities and services provided by the Sheridan Libraries at Johns Hopkins University. Choudhury has published articles in journals such as the International Journal of Digital Curation, D-Lib, the Journal of Digital Information, First Monday, and Library Trends. He has served on committees for the Digital Curation Conference, Open Repositories, Joint Conference on Digital Libraries, and Web-Wise. He has presented at various conferences including Educause, CNI, DLF, ALA, ACRL, and international venues including IFLA, the Kanazawa Information Technology Roundtable and eResearch Australasia.

Educational Interests and Activities: I am most interested in capacity building for the human side of infrastructure particularly in the form of data scientists. These individuals would act as the human interface between domain scientists and data managers. Currently, it seems that most scientific projects choose one of their own researchers for this role but it would be important to consider the potential roles for library and information science professionals.

DONALD COLLINS

Oceanographer, **National Oceanographic Data Center National Environmental Satellite, National Oceanic and Atmospheric Administration**

Speaker Bio: Don Collins has been an oceanographer at the NOAA National Oceanographic Data Center (NODC) for more than 20 years. His early work activities included several years of providing reference services to a diverse constituency of research scientists, business people, engineers, teachers and the public. Current activities are working on improving and documenting NODC archival practices policies, in addition to improving controlled vocabularies and data cataloging practices for the data collections maintained at NODC.

Don is a member of the Joint Committee on Oceanography and Marine Meteorology Expert Team on Data Management Practices, which is co-sponsored by the World Meteorological Organization and the Intergovernmental Oceanographic Commission. He is the project manager for the NOAA Coral Reef Information System (CoRIS) and the NOAA alternate representative to the Inter-agency Working Group for Scientific Digital Data (IWGDD).

MELISSA CRAGIN

Research Assistant Professor, **Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign**

Speaker Bio: I am a Research Assistant Professor on the faculty of the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign (Illinois), and affiliated with the Center for Informatics Research in Science and Scholarship (CIRSS), where I lead the Data Practices team for the Data Conservancy. I am co-Investigator on an IMLS-funded National Leadership grant investigating data sharing and curation requirements for institutional repositories, and PI for the Data Curation Education Program (DCEP) grant, funded by IMLS.

Educational Interests and Activities: As part of the DCEP, we work to maintain awareness of scientists' needs for data management support, and curation assistance and services. To stay at the forefront of this emerging field, we have to make regular assessments of our curriculum. Today's meeting is a perfect occasion for me to consider both of these objectives. One aspect of our DC program at GSLIS is developing sustainable internship sites for our students, and I am interested in learning about possible opportunities today.

BOB DOWNS

Senior Digital Archivist, **Center for International Earth Science Information Network, Columbia University**

Speaker Bio: Dr. Robert Downs is the Senior Digital Archivist and acting head of cyberinfrastructure and informatics research and development at CIESIN, the Center for International Earth Science Information Network at Columbia University, where he has been employed for over ten years. He has over twenty-five years of experience in information systems management, holds the Ph.D. in Information Management from the Stevens Institute of Technology, has taught courses in management and computer science, and conducts research on the development and management of information systems to support research and scholarship.

WENDY DUFF

Director of the Digital Curation Institute, Associate Professor, Faculty of Information, University of Toronto

Educational Interests and Activities: My main personal research interests focus on the use of cultural heritage material, predominantly archival records. I am also interested in, and currently studying, the convergence of libraries, museums in the digital environment. Finally I am investigating aspects of educating the education of museum studies and information studies students. The DCI's research interests, however, are much broader and include the preservation of databases, blogs and other types of data.

PETER FOX

Professor and Chair of Tetherless World Research Constellation, Rensselar Polytechnic Institute

Educational Interests and Activities: Peter Fox is Tetherless World Constellation Chair and Professor of Earth and Environmental Science and Computer Science at Rensselaer Polytechnic Institute. Research interests: Sun-Earth system science, computational/computer science and distributed semantic data frameworks addressing the full life-cycle of data and information within and among science and engineering disciplines. Fox chairs the International Union of Geodesy and Geophysics Union Commission on Data and Information, is associate editor for the Earth Science Informatics journal, and editorial board member for Computers in Geosciences. Fox serves on the International Council for Science's Strategic Coordinating Committee for Information and Data. Education interest: Data science, Xinformatics, Semantic eScience. http://tw.rpi.edu/wiki/Peter_Fox

JOSH GREENBERG

Program Director, Alfred P. Sloan Foundation

Educational Interests and Activities: I'm building the Alfred P. Sloan Foundation's program in Digital Information Technology and Dissemination of Knowledge, which is turning toward a more explicit focus on data-driven, computationally-intensive research across the disciplinary spectrum. I'm particularly interested in issues of data openness, data interoperability, and the division of labor among researchers, traditional institutional structures like libraries and computing centers (as well as new configurations of skills and professional identity).

MARGARET HEDSTROM

Associate Dean for Academic Programs and Professor, School of Information University of Michigan

Speaker Bio: Margaret Hedstrom is Associate Dean for Academic Programs and Professor at the School of Information, University of Michigan where she teaches in the areas of archives, electronic records management, and digital preservation. She is PI for a NSF-sponsored traineeship (IGERT) called "Open Data" that is investigating tools and policies for data sharing and data management in partnership with faculty and doctoral students in bioinformatics, computer science, information science, and materials research. She was project director for the CAMiLEON Project, an international research project that investigated the feasibility of emulation as a digital preservation

strategy. Her current research interests include digital preservation strategies, sharing and reuse of scientific data, and the role of archives in shaping collective memory. She is a member of the Board for Research Data and Information, National Research Council, National Academy of Sciences. She has served on the National Digital Strategy Advisory Board to the Library of Congress, and the Advisory Committee on Historical Diplomatic Documentation, U.S. Department of State, and on the ACLS Commission on Cyber-Infrastructure for the Humanities and Social Sciences. Hedstrom is a fellow of the Society of American Archivists and recipient of a Distinguished Scholarly Achievement Award from the University of Michigan for her work with archives and cultural heritage preservation in South Africa.

BRYAN HEIDORN

Director, **School of Information Resources and Library Science, University of Arizona**

Speaker Bio: P. Bryan Heidorn holds a degree in biology, and a PhD in Information Science. We was an owner of a software company specializing in chemical tracking and environmental monitoring. He was an associate professor for 12 years at the University of Illinois Graduate School of Library and Information Science and served two years as a program manager in the National Science Foundation Division of Biological Infrastructure where he served in several programs including Advances in Biological Informatics, Cyber-enabled Discovery and Innovation and the Data Working Group. He is now Director of the School of Information Resources and Library Science at the University of Arizona and a Director of the JRS Biodiversity Foundation.

Educational Interests and Activities: I am interested in the management of information particularly biodiversity data. We currently have an IMLS funded program called DigIn that offers a certificate in digital records management. That program as well as our masters program requires constant revision to support data management issues. In addition I currently chair a campus committee on research data management that tackles some of the issues addressed in this workshop. In my role as a member of the board of directors of the JRS Biodiversity Foundation we are attempting to support methods for long-term data preservation and access for biodiversity projects across Africa.

MIKE LESK

Chair, **Department of Library and Information Science, Rutgers University**

Speaker Bio: Michael Lesk is a professor of Library and Information Science at Rutgers University, after previous work at Bell Labs, Bellcore, the National Science Foundation, and part time at Google. He is best known for work in digital libraries, and his book "Understanding Digital Libraries" was published in 2004 by Morgan Kaufmann (second edition of a 1997 book). His research has included the CORE project for chemical information, and he wrote some Unix system utilities including those for table printing (tbl), lexical analyzers (lex), and inter-system mail (uucp).

Educational Interests and Activities: I'm doing two new courses, data stewardship and data preservation. Our real question is what should we be teaching students to prepare them for data curation; how much subject matter, for example? Are we training them to work in a library IT department or to work with such a department? (Research Interest) Open access to scientific data. The internet breaks business models, and it's also broken the model for academic research. Everyone exploiting

large data files to do research is enthusiastic; but the barriers to expanding them throughout science are technical, economic, legal, and most seriously, cultural. We need to focus on reducing curation costs and providing career rewards for data sharing. And we need to remember "anything worth doing is worth doing badly" (G. K. Chesterton).

CLIFFORD LYNCH

Director, **Coalition for Networked Information**

Educational Interests and Activities: Clifford Lynch has been the Director of the Coalition for Networked Information (CNI) since July 1997. CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and intellectual productivity. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation. Lynch, who holds a Ph.D. in Computer Science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science and the National Information Standards Organization. Lynch currently serves on the National Digital Preservation Strategy Advisory Board of the Library of Congress and Microsoft's Technical Computing Science Advisory Board.

MARY MARLINO

Director of e-Science and the NCAR Library, **National Center for Atmospheric Research**

BILL MICHENER

Professor and Director of e-Science Initiatives for University Libraries, DataONE Principle Investigator, **University of New Mexico**

Speaker Bio: William Michener is a Professor and Director of e-Science Initiatives for University Libraries at the University of New Mexico. In this role, he serves as Principal Investigator for DataONE—a large program focused on supporting discovery, analysis and visualization, and preservation of biological, ecological, and environmental data. He also directs the New Mexico EPSCoR Program—a statewide program designed to enhance competitive research through strategic investments in research infrastructure, cyberinfrastructure, and education and outreach. During the past decade he has directed several large interdisciplinary research and cyberinfrastructure projects including the Development Program for the U.S. Long-Term Ecological Research Network, the Science Environment for Ecological Knowledge, and various NSF- and USGS-funded cyberinfrastructure programs that focus on developing information technologies for the ecological and environmental sciences. Prior to joining the University of New Mexico, Michener managed the Biocomplexity and Ecology Programs at the National Science Foundation. He has published extensively in the ecological sciences and information sciences and presently serves as Data Archives Editor for the Ecological Society of America and as Associate Editor of the Journal of Ecological Informatics.

LUCY NOWELL

Program Manager, The Office of Advanced Scientific Computing Research, Department of Energy

CAROLE PALMER

Professor and Director, Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Speaker Bio: Carole L. Palmer is Director of the Center for Informatics Research in Science and Scholarship (CIRSS) and Professor in the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign. Her research investigates problems in scientific and scholarly information work, with a particular focus on barriers to interdisciplinary inquiry and the changing nature of research collections in the digital information environment. She is a co-PI on the Data Conservancy, an NSF DataNet award, and PI on the IMLS Digital Collections and Content project. Her other recent funded projects include investigations of data curation needs across sciences, high-impact information in brain research, scholarly annotation, and institutional repository development, as well as IMLS and NSF funded projects to develop educational programs in data curation and biological informatics. She has helped lead the school's development of data curation education programs: a specialization in data curation within the Master of Science in Library and Information Science (MSLIS) since 2007; a biological information specialist program offered as a concentration in the campus-wide MS in Bioinformatics since 2005; a summer institute for professional development since 2008; and a new doctoral initiative that includes on-site training at NCAR. As part of the Data Conservancy, she is enhancing these programs and developing mechanisms for coordinating and sharing educational approaches, methods, and materials among DataNet partners and other educators active in training for curation of research data.

Educational Interests and Activities: I conduct research on fundamental problems in the use of scientific and scholarly information and teach courses on information behavior, scientific information practices and problems, and user study design. My program of research is about mobilizing information for researchers, and it focuses on two interrelated areas: information work in the research process and context-rich digital research collections.

JIAN QIN

Associate Professor, School of Information Studies, Syracuse University

Educational Interests and Activities: Jian Qin developed and taught the Scientific Data Management course as part of the scientific data literacy project funded by NSF. She currently leads an eScience Librarianship (eSLib) curriculum development project, which includes three core courses – scientific data management, cyberinfrastructure and scientific collaboration, and data services – and a series of activities (such as the monthly eScience Lab facilitating learning and research sharing among students and researchers, collaborating with data librarians in a mentorship program) and short courses specializing in eScience workflows and data publishing. A rubric is being developed to perform outcome-based assessment for student learning achievements and effectiveness.

JOYCE RAY

Deputy Director for Museums and Director for Strategic Partnerships, **Institute of Museum and Library Services**

Speaker Bio: Since 2003, IMLS has provided support for master's and doctoral students in graduate schools of library and information science through its 21st Century Librarian program. The program has supported more than 3,000 master's students and approximately 200 doctoral students, as well as continuing education programs for more than 30,000 current library staff. In 2006, IMLS began inviting proposals to develop programs and courses of study in digital curation and digital archiving and has since funded a number of successful projects, many of which are represented in this summit. In 2011, they are evaluating the first five years of the program and are developing plans for the next five years. Particularly interested in the role of LIS education and research, and the development of library services, to support the management, preservation, presentation and reuse of digital data, Joyce Ray directs competitive grant programs that award approximately $40 million annually through programs including National Leadership Grants for Libraries and Museums; Sparks! Ignition Grants for Libraries and Museums; and the 21st Century Librarian Program, which funds education, professional development, workforce research, and Early Career Development grants in library and information science.

ALLEN RENEAR

Associate Professor and Associate Dean for Research, **CIRSS, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign**

Speaker Bio: At GSLIS Renear teaches courses in information modeling and digital publishing and leads research on the application of logic-based formal ontologies to problems in data curation and the foundation of information systems. He is the author or co-author of over 50 academic publications, including articles in *Communications of the ACM*, and *Science*. As Associate Dean for Research he is responsible for strategic planning for GSLIS research activities and oversees the School's $16M research portfolio. Renear has been President of the Association for Computers and the Humanities, a Distinguished Visiting Fellow at the Oxford University Computing Unit, and participated in a number of standards development efforts, including serving on the Advisory Board of the Text Encoding Initiative, and as first chair of the Open eBook Publication Structure Working Group (now IDPF/ePUB). Prior to joining GSLIS he was, from 1992 to 2000, Director of the Brown University Scholarly Technology Group, an applied R&D group focusing on digital publishing and research computing, primarily in the humanities. Renear received an AB from Bowdoin College and a PhD (1988) from Brown University. His research is focused on developing a logic-based formal ontology for the fundamental concepts that are important in information systems and data curation, such as data, dataset, file, preservation, derivation, encoding, and so on, with applications to both scientific and cultural information. The context of much of his current work is the NSF-funded Data Conservancy, where he co-leads the Data Concepts group. As Principal Investigator on an IMLS-funded project to extend the Illinois data curation specialization to the humanities he is using the research findings to shape the content of the GSLIS data curation curriculum, for both scientific and humanities data.

HELEN TIBBO

Alumni Distinguished Professor, **School of Information and Library Science, University of North Carolina at Chapel Hill**

VIRGIL VARVEL

Research Analyst, Data Conservancy Education Coordinator, **CIRSS, Graduate School of Library and Information Sciences, University of Illinois at Urbana-Champaign**

Educational Interests and Activities: With the Center for Informatics Research in Science and Scholarship, Virgil serves as both a research analyst and education coordinator. He heads projects researching library use data as well as needs analysis research in the data curation education program. Before joining CIRSS in September 2007, Virgil worked for eight years with University Outreach and Public Service at the University of Illinois. There he performed Web design, database programming, instructional design, online course teaching, online research, program evaluation, educational consulting, and other tasks. He is still administering a longitudinal survey of online learners. Among the honors he earned during this time are a WebCT Exemplary Online Course award and a Center for Transforming Student Services Best Practice award. He has numerous publications on a wide-range of policy issues and educational research. His research has also included various aspects of distance education including the pedagogical assumptions of socially organized versus independent study instructional design in distance education.

BRUCE WILSON

Group Leader, Environmental Data Science & Systems, **Environmental Sciences Division, Oak Ridge National Laboratory**

Speaker Bio: Bruce Wilson is the Group Leader for the Environmental Data Science and Systems Group, the Manager for the ORNL Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC), and an Adjunct Professor of Information Sciences at the University of Tennessee.

After receiving his Ph.D. in Analytical Chemistry from the University of Washington under the direction of Bruce Kowalski, Wilson joined Eastman Chemical Company, where he worked in a variety of roles over 11 years. His work at Eastman included studies of cellulose acetate production, polyester production, thermotropic liquid crystalline polymers as rheology modifiers, chemical information management, and computational chemistry applied to partial oxidation chemistry. Wilson moved to Dow Corning for a year working on improved understanding of silicone sealant production. He then worked for Dow Chemical for five years, eventually becoming a Technical Leader, responsible for informatics support to high throughput research in catalysis, materials and formulations. He joined ORNL in June 2006 as the Systems Engineer for the ORNL DAAC, and he was promoted to the Group Leader position in late 2007. Wilson is a co-inventor on 4 US patents, an author or co-author on over 20 peer-reviewed publications, and an author or co-author of over 120 corporate technical reports. Wilson serves on the Core Cyberinfrastructure Team for the DataONE (Observation Network for Earth) project, the Board of Directors for the USA National Phenology Network (USA-NPN), and the Finance Committee for the Federation of Earth Science Information Partners. He also serves as a peer-reviewer for several journals and for grant programs at NSF, NASA, and DOE.