Report from the

Research Data Workforce Summit

Sponsored by the Data Conservancy 

6 December 2010

Chicago, IL

Summit coordinated by:

Carole L. Palmer & Melissa Cragin

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

Bryan Heidorn

School of Information Resources and Library Science
University of Arizona

Report prepared by:

Tiffany Chao, Simone Sacchi, Virgil E. Varvel Jr., & Carole L. Palmer

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



# **Outline**

1. Overview ……………………………………………… 3
2. Summary ……………………………………………… 5
3. Cross-Cutting Themes ……………………………. 7
	1. Professionalization of data management ………………………. 7
	2. Communication and coordination across sectors…………….. 9
	3. Educational challenges ……………………………………………… 10
4. Future Directions …………………………………. 12
5. Presentation Briefs ………………………………. 14
6. Appendices …………………………………………. 19
	1. Meeting agenda ………………………………………………………. 19
	2. Participant information …………………………………………….. 21

# **Overview**

The 2010 Research Data Workforce Summit was held in Chicago on December 6th, 2010, in conjunction with the 6th International Digital Curation Conference (IDCC). It was sponsored by the Data Conservancy, one of the National Science Foundation’s Office of Cyberinfractructure, DataNet projects. The IDCC conference, co-hosted by the Graduate School of Library & Information Science at the University of Illinois and the Digital Curation Centre, provided the opportunity to bring together a group of data curation experts and educators for a one-day exchange on research data workforce development in the sciences. The 29 invited participants included representatives from government agencies and data centers, current NSF DataNet initiatives, 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 was of particular interest, in recognition of the greater data curation demands being put on government agencies and the lack of a well-trained professionals to meet the demand.[[1]](#footnote-1) 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.

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. Presenters were asked to address the following questions in the context of their work:

* + 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 discussions that emerged throughout the day were often applicable beyond government agencies, having broad relevance to the curation and use of data in research organizations more generally.

# **Summary**

Opening remarks from the summit organizers conveyed the importance of creating a shared platform to bring together insights and experiences on how data are managed, the roles of various stakeholder groups (i.e. information professionals, data scientists, scientists, students), and the skills required for each group. Government agency representatives from the Department of Energy and the Institute for Library and Museum Services emphasized the level of support being provided for the development of education and training opportunities to produce a workforce that is more conscious of data management responsibilities. They noted that for their agencies data management plans are becoming an increasingly vital component of proposals, similar to the trend at the National Science Foundation. A recent workshop organized by the Earth Science Information Partners Data Management 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 progress on articulation of long term archiving principles and the methodology of 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 initiatives, DataONE and the Data Conservancy, covered their project’s 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 provide training to prepare professionals that can create needed data resources and tools. Educators from iSchools (www.ischools.org) and other university based units provided overviews of their programs in data curation and data science, reflecting on the successes and barriers encountered to date. There was consensus that agreement in terminology around data management, and the various workforce roles such as curator and data scientist, would be an important step forward for these programs and contribute to more cohesive discourse within the community. (Note, since the various terms were used interchangeably throughout the summit, in this report when appropriate, we have used the single generic term, data professional, to capture the range of roles more generally.)

The summit moderator, Lucy Nowell, from the U. S. Department of Energy, provided integrative comments throughout the day and closed by drawing attention to the recent report by Shoshani, & Rotem (2010) [[2]](#footnote-2) stressing the need for curation services to scale-up to meet the demands of the current state of rapid data growth and the highly complex and variable organization and storage patterns of small science, which may well produce the majority of scientific data over all. She noted, in particular, the need in the current environment to accommodate parallel process computing with large sets of data.

## **Cross-Cutting Themes**

Over the course of the day, three themes were prominent across the presentations and discussion as a whole: professionalization of data management, communication and coordination across sectors, and educational challenges.

## **Professionalization of Data Management**

As noted above, there is a need to disambiguate and develop definitions for ‘data manager’, ‘data curator’, ‘data scientist’, to help establish and sustain the range of professional roles within the future workforce. The agreement on terminology and definitions needs to start with education initiatives and be applied consistently in various programs across the country. In general, data professionals need to be able to manage data across disciplines and recognize the value of data for reuse within and outside the original domain. Metadata standards were emphasized as an essential area of expertise, since the appropriate and systematic application of metadata will provide the foundation for assuring future accessible of data. Since there is no formal organization for coordination and enforcement of standards, data professionals will need to develop and share their evolving metadata practices, which will necessarily cover application of various standards to data in a broad range of formats and associated curatorial processes required prior to repository deposit or ingest.

Educators from four schools presented highlights of their programs at Rensselaer Polytechnic Institute (RPI), George Mason University, and two iSchools—University of Arizona and University of Michigan. The iSchools’ programs are 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 in the sciences and beyond. The Data Conservancy is actively extending masters level curriculum at Illinois and UCLA, as well continuing education at Illinois, and DataOne’s education efforts are currently concentrated on community engagement rather than formal university based education. Student recruitment, practical engagement and mentorship opportunities for students, and curriculum development were themes stressed by the group of educators.

As a new field of study, recruiting students is particularly challenging and requires strategies for exposing programs to groups of potential students. Computer science students were identified as a key undergraduate pool, but currently there is little incentive for them to become involved in data programs, since they are still primarily drawn to more traditional technology positions in industry. Science laboratory courses at the undergraduate level were seen as an important channel for infusing sound data management practice into science curriculum.

Practical experience is considered essential for students to gain skills and knowledge, and several participants had made progress developing internship and practicum opportunities. It was noted that there is high demand for interns from a number of organizations that recognize their growing need for data 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. A critical question posed was to what extent the current practicing experts in the data community are getting involved in formal education, since they can be valuable as mentors and provide an essential, practical component of professional education.

In the area of curriculum development, participants emphasized the need for computational science, statistics, and digital preservation to be integrated into current programs. Theoretical concepts and interdisciplinary collaboration are important areas addressed in courses for graduate students, who are better prepared to engage with constructs at this level. Discussions addressed appropriate structuring of courses across undergraduate and graduate programs, but there is also a clear and urgent need for provision of continuing professional education opportunities for the current workforce.

The following three recommendations emerged for improving professional education:

* Craft program and course descriptions to be more attractive to students.
* Reach out and recruit students from disciplines in other departments.
* Partner with data centers to facilitate internships and field experiences for students that include mentoring relationships with data professionals .

## **Communication and Coordination across Sectors**

The presentations by government and data center representatives generated discussion on the need for interdisciplinary and multi-disciplinary training approaches and cross-institutional solutions to data problems. Data professionals will need much more than cross-field awareness. Substantive communication and connections will need to be established at both the scientific level and the data management level. The blend of competencies, or the ‘tridge’, at the intersection of the domain sciences, information science, and computer science is required to address the coming challenges in data management, and the management of science more generally. It was noted that educational programs that focus on scientific discovery should also encompass development of policies for data sharing, access, and use across disciplines, and that knowledge needs to be harnessed from both the public and private sectors.

Integration of skills from diverse disciplinary backgrounds is an important step towards the creation of effective, broadly trained data professionals. There is much to learn from professionals that have successfully established practices for working across disciplinary borders and for engaging multiple fields in their data and research operations. At the same time, in some organizations there will need to be divisions of labor and specialized roles. For example, data professionals in specialized research centers will require a higher degree of domain expertise, while those in data repositories will require a higher level of cross-domain understanding and general curation, infrastructure, interoperability expertise.

Since data workflows generally cross the boundaries of a single institution, strong communication and working relationships need to be established between data providers and data professionals, supported by a shared understanding of the disciplinary data practices and the implications of various curatorial and management strategies on the conduct of science. Consortia were seen as an important way of providing coordination and for building on existing infrastructure to provide new services and data products and promote reuse of data. Participants expressed strong support for working groups that cross institutions and disciplines for development of consortia and coordinated initiatives, but recognized that this will require a high level of community engagement to foster and consolidate stakeholder networks.

General directions for communication and coordination of professional data management:

* Develop coordination structures that cross domain sciences, information sciences, and computer sciences, as well as institutional and international boundaries.
* Design training programs that are integrative and general yet allow for development of specialized roles and expertise.

**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 aspects of this new paradigm, particularly the trend toward concurrent programming and parallel processing. It was suggested that in some computer science departments faculty are resistant to the new methods needed for dealing with peta-scale or exa-scale data and are not providing training in true parallel processing. It was acknowledged that most faculty teach programming the way they learned it, within a single processor environment, and may consider the time commitment prohibitive for shifting to meet the demands of the new scales of practice. Government agencies, however, are providing funding to support research activities in parallel processing at scale, with graduate fellowships and early-career programs for junior faculty available for those involved in addressing the challenges of data processing in this new paradigm.

Problems were also raised with regard to providing incentives for data professionals to serve as mentors. There are no reward structures in place to encourage involvement in the education and training of new students or in-service professionals. At present, efforts tend to target provision of practical experience through internships and fellowships within research operations. The field could benefit from more support directly targeting field experiences with data practitioners, but data centers could also better publicize their activities and be more pro-active in the education sphere. Apprenticeship approaches were considered effective for transferring existing skills and competencies. At the same time, it was recognized that the apprenticeship model is complicated by the fact that current practice may not be optimal or state-of-the-art.

As noted by several participants, the social and cultural challenges associated with data production and use are more difficult than the technical demands. Since the work requires navigating 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 barriers related to domain practices and processes, and the related terminology. However, this is a long-standing issue in the information professions, where training in a single discipline may not provide the breadth needed to serve diverse user communities.

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 efforts and resources around the developing profession, they are also a functioning part of the community and a source of knowledge, in areas such as data discovery and access issues, and need to be part of coordinated engagement on best practices.

Recommendations for addressing these challenges:

* Begin substantive curriculum revision to address current gap in concurrent programming and parallel processing.
* Promote the value of data workforce teaching and research within university and research organization reward structures.
* Support documentation and dissemination of emerging best practices and identification of areas where new better practices need to be developed.

# **Future Directions**

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

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

Across the schools, programs in data management, data curation, 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. A standard terminology, perhaps building on the roles defined in the 2009 Interagency Working Group report[[3]](#footnote-3), could provide a unified base of understanding for scoping education programs while benefiting employers in that wish craft positions to attract research data professionals. 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 distinct contributions by individual educational programs .

1. **Continue to build the data curation community and promote awareness of the different activities at iSchools and other departments and institutions.**

Several activities were identified as first steps in building the education community. First, there is interest in developing a web presence that serves as a knowledge base on current efforts and makes potential opportunities for teaching and learning visible to faculty and students. The “education hub” currently under development by the Data Conservancy can 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 for data professionals in the 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.

1. **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 there any new studies should build on prior and ongoing work in this area, but clearly new studies will be vital for advancing programs in higher and continuing education. This is a new and dynamic field, and keeping curriculum current and reflecting the growing base of knowledge and best practices will be a tremendous challenge for educators, especially in providing training for the high level of curatorial support that will be required for highly variable small science data and organization and preservation for very large-scale, interdisciplinary 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)

* 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 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.
* Building on established data curation masters specialization and summer institute for professional development at Illinois, and new program integrating of 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

* Data science requires recognition of how science interacts with information in terms of the scientific, policy, computation, and social aspects.
* Focus on clear objectives in instruction
* 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 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 ofScientific 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
* 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**

**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**

**Principle Investigator**

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

**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**

**Center for Informatics Research in Science and Scholarship
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.

**Educational Interests and Activities:** My research interests are in scientific informatics, particularly enabling scientific collaboration and data-intensive science through applying information technology and advancing data storage, curation, distribution, analysis, and visualization technologies and practices.

1. 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 [↑](#footnote-ref-1)
2. Shoshani, A., & Rotem, D. (Eds.) (2010). *Scientific data management: Challenges, technology, and deployment*. Boca Raton, FL: Chapman & Hall / CRC Press. [↑](#footnote-ref-2)
3. 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 [↑](#footnote-ref-3)