Data Symposium 2012

SeWHiP (Southeast Wisconsin High Performance Cyberinfrastructure) and the Clinical and Translational Science Institute (CTSI) of Southeast Wisconsin present a one-day symposium to explore the challenges associated with storage, access, visualization, sharing, integration, and scaling of research data. Experts from U.S. agencies, national laboratories, and research institutes will present storage options for research communities, federated solutions and tools, and visualization highlights.

Data Symposium 2012 Program Committee:

Elizabeth Leake, Chair
Erik Bergman, BloodCenter of Wisconsin
Carlos De la Peña, Clinical and Translational Science Institute of Southeast Wisconsin
Jon Gladieux, Milwaukee Institute Inc.
Lars Olson, Marquette University
David Stack, University of Wisconsin Milwaukee

SeWHiP Founding Organizations:

BloodCenter of Wisconsin
Marquette University
Medical College of Wisconsin
Milwaukee Institute Inc.
Milwaukee School of Engineering
University of Wisconsin-Milwaukee

For more information about SeWHiP: www.sewhip.org.
For more information about CTSI: www.ctsi.mcw.edu.

March 1, 8:00 a.m. to 5:00 p.m.

Discover
Solutions
to the
Challenges
of Data
Intensive
Research

Blood Research Institute
8733 Watertown Plank Road
Milwaukee, Wisconsin

Sponsored by SeWHiP (Southeast Wisconsin High Performance Cyberinfrastructure) and CTSI (Clinical and Translational Science Institute) of Southeast Wisconsin

Agenda

8:00 Breakfast and registration.

8:30 Reza Shaker, Senior Associate Dean and Director, Clinical and Translational Science Institute, Medical College of Wisconsin, welcome and opening remarks.

8:45 Ian Foster, University of Chicago, Keynote.

What will global cyberinfrastructure look like in the future, and how will it enable highly-scalable and reliable data movement around the world?

Ian Foster is director of the Computation Institute (CI), a joint institute of the University of Chicago and Argonne National Laboratory. He is also an Argonne Senior Scientist and Distinguished Fellow, and the Arthur Holly Compton Distinguished Service Professor of Computer Science.

In 2001, Foster was a visionary behind the National Science Foundation's (NSF) TeraGrid project. He is now co-architect of the ExTreme Science and Engineering Discovery Environment (XSEDE), the NSF's five-year, $121-million project that replaces and expands on TeraGrid. Foster's team is developing XSEDE's User Access Services, which leverage cloud technologies to facilitate data movement between XSEDE's 16 supercomputers and high-end visualization and data analysis systems, as well as resources from non-XSEDE national and global sites. The Globus Toolkit grid software, whose development is led in part by the CI, was an integral component of TeraGrid, and continues to play a role in XSEDE infrastructure.

Foster's research deals with distributed, parallel, and data-intensive computing technologies, and innovative applications of those technologies to scientific problems in such domains as climate change and biomedicine. Methods and software developed under his leadership continue to underpin many large national and international cyberinfrastructures. In 2010, Foster's team identified the need for a powerful computational environment that would serve the growing resource requirements of the biomedical research community. *Beagle*, a 151 teraflops, Cray XE6 system, was made possible by a grant from the National Institutes of Health. Named after the ship that carried Charles Darwin on his famous voyage in 1831, *Beagle* is one of the fastest supercomputers in the world devoted to life sciences.

9:45 Brian Athey, University of Michigan—Ann Arbor, Keynote.

How could the emerging field of Translational Bioinformatics influence the way we store, retrieve, analyze, and visualize research data?

Brian Athey is a Professor of Biomedical Informatics and Chair of the Department of Computational Medicine and Bioinformatics at the University of Michigan (UM) Medical School. In addition to serving as overall Director of Academic Informatics and Information Technology of the School, he is the Principal Investigator (PI) of the National Institutes of Health (NIH) National Institute of General Medical Sciences Bioinformatics Training Grant made possible by a grant from the NIH National organization. Trained as a biophysicist, Athey is now recognized as one of the nation's experts in the new field of 'Translational Bioinformatics' and also research cyberinfrastructure.

In the Mid-1980s, Athey proposed the double helical crossed-linker model for the structure of chromatin, once quite controversial, it is now generally accepted. He established the first nationwide Internet2 Visible Human Project demonstration under contract with the National Library of Medicine (NLM), was PI of the DARPA Virtual Soldier Project, and he currently leads the NIH National Center for Integrative Biomedical Informatics (NCIBI.org), one of seven NIH Roadmap Centers for Biomedical Computing. Athey is also Associate Director of Informatics and IT in the Michigan Institute for Clinical and Health Research, home of UM's NIH Clinical and Translational Science Award, and has served as national co-chair of the Clinical and Translational Sciences Award, Informatics Key Function Committee. He also currently serves as a special advisor to the Chief Information Officer and Director of the Center for Information Technology of the NIH. In 2000, Athey was named a Peace Fellow of the Federation of American Scientists (FAS.org) for his work in the 1990s to combat biological warfare and terrorism. He is a founder and chairman of the board of the Washington, DC-based Scientists and Engineers for America (SEA.org).

10:45 Break.

11:00 Clifford Lynch, Coalition for Networked Information

Is there a standards-based approach to data-intensive scholarship and digital preservation that will help research libraries plan for the future?

Clifford Lynch has led the Coalition for Networked Information (CNI) since 1997. CNI, jointly sponsored by the Association of Research Libraries and EDUCAUSE, includes about 200 member organizations concerned with the intelligent uses of information technology and networked information to enhance scholarship and intellectual life. CNI's wide-ranging agenda includes work in digital preservation, data intensive scholarship, teaching, learning and technology, and infrastructure and standards development.

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 both a past president and recipient of the Award of Merit of the American Society for Information, and a fellow of the American Association for the Advancement of Science and the National Information Standards Organization.

Lynch currently serves on numerous advisory boards and visiting committees, including the National Digital Preservation Strategy Advisory Board of the Library of Congress and Microsoft Corporation's Technical Computing Science Advisory Board. His work has been recognized by the American Library Association's Lippincott Award, the EDUCAUSE Leadership Award in Public Policy and Practice, and the American Society for Engineering Education's Homer Bernhardt Award.

11:45 John Cobb, Oak Ridge National Laboratory

Is progress being made toward a federated data management solution for multidisciplinary research?

John W. Cobb's research focus has been evangelism for high performance computing (HPC), data science, and cyberinfrastructure (CI) to leverage scientific research. His area of professional preparation was in theoretical physics from the University of Texas at Austin specializing in large-scale

computational plasma physics modeling and simulation applied to alternative fusion energy concepts. He has worked in several areas, all with the unifying theme of applying advanced computation to solve research problems. His experience includes: developing simulation algorithms for semiconductor device equipment; lab-wide strategic planning for scientific computing at the Oak Ridge National Laboratory (ORNL); CIO, cyber security, and other IT roles for the Spallation Neutron Source construction project; ORNL TeraGrid resource provider principal investigator; and NSF's DataONE DataNet project co-investigator.

Cobb will provide examples of the use of advanced CI in the form of HPC and data science to augment scientific exploration to enable scientific discovery, which would be impossible without CI. Cobb will also explain the difference between CI and IT and why that distinction is important. Finally, Cobb will describe efforts such as the National Science Foundation's DataONE collaborative project where the main thrust is to develop effective means for interoperability between different data archives and data formats and to use that interoperability for new scientific discovery.

12:30 Lunchtime Viz-it!

Network with the experts, see examples of scientific visualizations, and learn about the resources and methodologies used to produce them. Catering by Panera Bread, Inc.

1:30 Serge Goldstein, Princeton (remote presentation)

Is there a way to fund long-term storage of data generated by a two or four-year grant funded projects?

Serge Goldstein will present Princeton's DataSpace funding and operational model. The "Pay Once, Store Forever (POSF)" concept implies that it is possible to fund long-term data storage with one-time payments that cover the current costs of storage, and leave enough excess to cover on-going replacement and management of that storage. POSF banks on the steady decline of the cost of physical data storage over time, as well as the steady increase in the amount of storage that can be managed by a given number of personnel.

As Director of the Academic Services Department in Princeton University's Office of Information Technology, Goldstein is responsible for coordinating the University's instructional technology support, including classroom media support, web-based instructional support, and research computing support. Goldstein co-authored Princeton's first home-grown course management system, and led the effort to replace that system with a commercial product (Blackboard). He serves on the OKI advisory board, and has been an active participant in the evolution of Learning Management Systems. Goldstein earned a B.A., M.Phil. and PhD in Anthropology from Yale University, and has held a number of positions in Information Technology, including VM and Unix System's Programmer and Web Services Manager.

2:15 Rachana Ananthakrishnan, University of Chicago

Are there collaborative research tools that simplify the sharing of distributed resources and data?

Rachana Ananthakrishnan is a principal software development specialist at Argonne National Laboratory and a staff appointee at the University of Chicago's Computation Institute (CI). She will provide an overview of the Globus Online Platform, and demonstrate how it is used to move data across multiple and disparate security domains.

Ananthakrishnan's presentation will reflect her involvement as a security specialist in several national and multinational research collaborations. She is co-chair of the Biomedical Informatics Research Network (BIRN) Security Working Group and Argonne's project coordinator for the U.S. Department of Energy-sponsored Earth System Grid. She is also the CI project lead on an effort to build secure solutions for the sharing of traumatic brain injury research data. As a member of the Globus Online design team that is developing "software as a service" and data transfer solutions as well as a research collaboration platform, her focus is the secure management of distributed resources across multiple and disparate security domains.

3:00 Break.

3:15 Michael Huerta, National Institutes of Health

What is being done to help manage the nation's biomedical data for the future?

Michael Huerta is Associate Director of the National Library of Medicine (NLM) and Director of the Library's Office of Health Information Programs Development. He works on numerous trans-NIH and trans-government groups and initiatives developing standards, technologies, practices, and policies to more widely, efficiently, and meaningfully share data generated in biomedical research. In addition to supporting vibrant research programs in informatics and biomedical computation, NLM is the largest health science library in the world and provides trusted, authoritative online information on a large number and wide range of topics, including: consumer health, disaster preparedness and response, toxicology, and clinical effectiveness research. The Library also develops and maintains a large number of bibliographic and informatics standards and tools, and hosts many extensive interconnected data resources filled with data from genetics, genomics, molecular biology, biochemistry, clinical trials, and the health science literature. Michael coordinates efforts to make these considerable resources known to librarians, researchers, healthcare providers, and the general public. He also oversees the Library's international efforts, as well as NLM's evaluation and strategic planning activities.

4:00 Sylvia Spengler, National Science Foundation (remote presentation)

How is the NSF preparing for the data deluge of the future? What should I consider when developing a data management plan for my NSF grant proposal?

Sylvia Spengler is a program director in the NSF Directorate for Computer & Information Science & Engineering, Division of Information and Intelligent Systems. Spengler has been at NSF since 2000 and has also served in the Biology directorate. Her interest in data policies extends back to the formation of the Human Genome Program at the U.S. Department of Energy. She has recently served on the Interagency Working Group on Digital Data and is the co-chair of the NSF Data Working Group.

4:45 panel discussion

5:00 Lars Olson, Associate Professor of Biomedical Engineering, Marquette University. Closing remarks.

5:15 Adjourn