The Materials Data Facility: Data Services to Advance Materials Science Research

With increasingly strict data management requirements from funding agencies and institutions, expanding focus on the challenges of research replicability, increasingly complex challenges in multimodal analysis and data synthesis, and growing data sizes and heterogeneity, new data needs are emerging in the materials community. The Materials Data Facility (MDF) operates two cloud-hosted services, data publication and data discovery, to promote open data sharing, self-service data publication and curation, and encourage data reuse, layered with powerful data discovery tools. The data publication service simplifies the process of copying data to a secure storage location, assigning data a citable persistent identifier, and recording custom (e.g., material, technique, or instrument specific) and automatically-extracted metadata in a registry while the data discovery service will provide advanced search capabilities (e.g., faceting, free text range querying, and full text search) against the registered data and metadata. The MDF services empower individual researchers, research projects, and institutions to 1) publish research datasets, regardless of size, from local storage, institutional data stores, or cloud storage, without involvement of third-party publishers; 2) build, share, and enforce extensible domain-specific custom metadata schemas; 3) interact with published data and metadata via REST APIs to facilitate automation, analysis, and feedback; and 4) access a data discovery model that allows researchers to search, interrogate, and eventually build upon existing published data. In this talk, will describe MDF design, current status, future plans, and show a live demo of the MDF data publication service.

DR. BEN BLAISZIK is a Research Scientist in the Computation Institute at the University of Chicago, Ben obtained his Ph.D. in 2009 working with Nancy Sottos in the Department of Theoretical and Applied Mechanics (TAM) at the University of Illinois at Urbana-Champaign. He has over 12 years of experience in scientific research with focuses including: leveraging high-performance computing and big data techniques to meet the unique challenges faced by scientists and leading cross-disciplinary materials design and development efforts. His work has culminated in 5 issued patents (4 additional pending), and over 20 peer reviewed publications/book chapters and the associated research results have been featured by the BBC, Wall St. Journal, Popular Science, Business Week, and the Economist.

HUBzero as a Data and Simulation Infrastructure for Scientific Exploration

HUBzero is the underlying infrastructure behind at least 60 online scientific communities, receiving over 2 million visitors annually. nanoHUB.org is a major HUBzero hub has reached a point where it runs more than 800,000 simulations annually, and has a proven large impact in both research and education. The nanoHUB community has asked for nanoHUB to support not only simulation but also working with data resources. For several years we have embarked upon detailed usage analytics of activity on HUBzero sites with a variety of customized clustering tools. Some of these same activities are now being translated into a combined data exploration and simulation interface called IDENT. This talk will focus on issues faced with respect to handling data exploration on HUBzero sites, usage analytics performed on nanoHUB, the use of IDENT as an initial effort from nanoHUB to work with ill-structured and ill-behaved data, and IDENT as a driver for simulation through a data exploration graphical interface.

ZENTNER, MICHAEL As Director of the HUBzero Platform, Dr. Zentner is responsible for the daily operations and strategic vision for HUBzero, a software platform that is the infrastructure currently powering more than 60 online scientific communities with nearly 2 million visitors annually. Michael is also an Entrepreneur in Residence, where he helps Purdue faculty and students during the commercialization of their innovations. Michael's specific research focuses on studying data driven user behavior patterns on nanoHUB to determine the impact of nanoHUB on the international community in education and in advancing science, as well as on developing visualizations to illustrate this impact. Michael is also the CEO of SPEAK MODalities LLC, a Purdue startup with software that helps children with autism develop language skills. Prior to joining Purdue, Michael was founder/senior team member of several information technology startup companies, where he created innovative solutions for extracting patterns from data, collaboration, and constrained optimization. Michael has consulted with many Fortune 500 companies to apply these technologies for solving business problems including operations scheduling, strategic capital investment, process improvement, and new product innovation and creation. Michael holds a BS in Chemical Engineering from the University of Illinois, a MS and Ph.D. in Chemical Engineering from Purdue University, an MBA from Purdue's Krannert School of Management, and an MBA from the TIAS Business School of Tilburg University in The Netherlands.

Q&A Panel: Future Directions

DR. ZACHARY TRAUTT will moderate this panel. Dr. Trautt holds the title of Material Research Engineer in the Materials Measurement Science Division at the National Institute of Standards and technology (NIST). Dr. Trautt plays a leadership role in the NIST Material Genome Initiative and has particular interest in the development of modular data models in materials science. Prior to taking on this position, Dr. Trautt was a Research Assistant Professor at the School of Physics, Astronomy, and Computational Sciences at George Mason University. He completed his Ph.D. in Engineering Systems with a minor in Material Sciences at the Colorado School of Mines in 2009. He also completed a B.S. in Engineering Physics at the Colorado School of Mines in 2004.

CyberFab

Coordinated Science Laboratory, Room 301

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MAY 24, 2016

MNTL micro+nanotechnology lab

MRL FREDERICK SEITZ MATERIALS RESEARCH LABORATORY

CSL: COORDINATED SCIENCE LAB

CyberFab

CyberFab Agenda

7:30 AM	**Registration**
8:00 AM	**Welcome/ CyberFab Workshop Objectives**
8:15 AM	**4CeeD - Trustworthy and Timely Data Capture and Curation** *Klara Nahrstedt, Steve Konstanty, Timothy Spila et al., University of Illinois*
9:20 AM	**National Data Service (NDS)** *John Towns, National Center for Supercomputing Applications, University of Illinois*
10:25 AM	**Break**
10:40 AM	**The Materials Data Facility: Data Services to Advance Materials Science Research** *Ben Blaiszik, University of Chicago*
11:45 AM	**Lunch**
12:45 PM	**HUBzero as a Data and Simulation Infrastructure for Scientific Exploration** *Michael Zentner, Purdue University*
1:50 PM	**Break**
2:05 PM	**Q & A Panel: Future Directions** *Moderated by Zachary T. Trautt, National Institute of Standards and Technology*
3:20 PM	**Closing Remarks**
3:30 PM	**END**

CyberFab Speakers and Abstracts

Welcome Remarks

CAMPBELL, ROY – Associate Dean for Information Technology, College of Engineering

CUNNINGHAM, BRIAN – Director, Micro and Nanotechnology Laboratory

SARDELA, MAURO - Director of the Central Research Facilities, Material Research Laboratory

NAHRSTEDT, KLARA – Director, Coordinated Science Laboratory

4CeeD – A Timely and Trusted Curator and Coordinator of Scientific Data

Materials and Semiconductor Fabrication research heavily depends on experimental laboratories equipped with scientific instruments such as Scanning Electron Microscope (SEM), Transmission Electron Microscope (TEM) and others. In many laboratories these microscopes are stand-alone devices, producing extensive digital data and metadata and yielding high volume, high variety and high velocity data. For many of these instruments, data is collected in a largely manual fashion. In this presentation, 4CeeD, a networked microscopes cyber-physical infrastructure, running a distributed service for real-time acquisition, streaming, and cloud-based real-time analysis of materials and semiconductor fabrication data will be presented. 4CeeD is a distributed infrastructure-as-a-service (IaaS) with an optimized real-time acquisition, transmission and analysis service architecture to break the ineffective process of scientists at the start of their scientific discoveries. 4CeeD is being implemented to enable scientists from Material Research Laboratory (MRL) and Micro and Nanotechnology Laboratory (MNTL) at UIUC to speed up in multiple ways scientific discovery. We will discuss 1) the unique cyber-physical requirements of networked microscopes, 2) digital data and metadata capture and representations, and 3) present an analysis and architectural design of cloud storage to enable timely processing, communication and delivery of data to scientists. This will be accompanied with live demonstrations of the 4CeeD prototype.

KLARA NAHRSTEDT is the Ralph and Catherine Fisher Professor in the Computer Science Department, and Director of the Coordinated Science Laboratory in the College of Engineering at the University of Illinois at Urbana-Champaign. Her research interests are directed toward trustworthy power grid, 3D teleimmersive systems, mobile systems, Quality of Service (QoS) and resource management, Quality of Experience in multimedia systems, and real-time security in mission-critical systems. She is the co-author of widely used multimedia books `Multimedia: Computing, Communications and Applications' published by Prentice Hall, and 'Multimedia Systems' published by Springer Verlag. She is the recipient of the IEEE Communication Society Leonard Abraham Award for Research Achievements, University Scholar, Humboldt Award, IEEE Computer Society Technical Achievement Award, and the former chair of the ACM Special Interest Group in Multimedia. She was the general chair of ACM Multimedia 2006, general chair of ACM NOSSDAV 2007 and the general chair of IEEE Percom 2009. Klara Nahrstedt received her Diploma in Mathematics from Humboldt University, Berlin, Germany in numerical analysis in 1985. In 1995 she received her PhD from the University of Pennsylvania in the Department of Computer and Information Science. She is ACM Fellow, IEEE Fellow, and Member of the Leopoldina German National Academy of Sciences.

DR. TIMOTHY SPILA currently works as a Senior Research Scientist in Materials Characterization in the Materials Research Laboratory at the University of Illinois, where he is the local expert in Secondary Ion Mass Spectrometry. This includes responsibilities for sample analysis, maintenance of the instruments, and training of researchers in the technique. He is also responsible for a Thermogravimetric Analysis tool and the collection of facility usage data for billing and usage statistics. He has also taken on the role of programmer to enhance and maintain the proposal, registration, reservation, and billing system for the MRL Research Facilities. Dr. Spila has twice served as Interim Director of the MRL Research Facilities and represents the MRL on the NSF Sponsored T2C2: Timely and Trusted Curation and Coordination project. Dr. Spila received his Ph.D. in Materials Science and Engineering in 2001 from the University of Illinois for his work examining the reaction pathways for self-organized surface roughing of Si1-xGex alloys during hydride gas-source molecular beam epitaxy.

STEVE KONSTANTY currently works as a Senior Research Programmer in the Coordinated Science Lab, where he is leading the team that is developing the 4CeeD Curator. In his 15 years on campus he has developed multiple full stack web solutions for NCSA, the Institute for Genomics Biology, and Student Affairs. Most recently he worked on the XSEDE resource allocation system (XRAS). His interests lie in project management, user experience, and web front end design and development. Steve graduated from the University of Illinois with a Bachelor of Philosophy in 1997.

National Data Service

The National Data Service is an emerging vision of how scientists and researchers across all disciplines can find, reuse, and publish data. It is an international federation of data providers, data aggregators, community-specific federations, publishers, and cyberinfrastructure providers. It builds on the data archiving and sharing efforts under way within specific communities and links them together with a common set of tools. Simply put, the National Data Service intends to make it easy to find, use, and publish data. Plans call for providing a common set of services that can work across communities and disciplines, but that also build on top of existing infrastructure already put into place by those communities. NDS services are expected to have a strong connection to publications and publishing process to ensure that robust links between published literature and the data they discuss. Through these links, it should be possible to make data as citable as literature. This presentation will provide an overview of the NDS and demonstrate some of its key features.

JOHN TOWNS is Deputy CIO for Research IT at the University of Illinois at Urbana-Champaign and the Executive Director for Science & Technology at NCSA. He is also PI and Project Director for the Extreme Science and Engineering Discovery Environment (XSEDE) project, Director of the National Data Service (NDS), and Director of the Illinois Campus Cluster Program. He provides leadership and direction for the enhancement of technology resources and services in research, particularly in the areas of high-performance computing, high-speed networking, big data, visualization, and other cutting-edge research technologies. His background is in computational astrophysics utilizing a variety of computational architectures with a focus on application performance analysis. He earned M.S. degrees in Physics and Astronomy from the University of Illinois and a B.S. in Physics from the University of Missouri-Rolla.

CyberFab
MAY 24, 2016