Fri 1/13/2012 1:26 AM
Personal comments on data preservation

I am Amber Boehnlein, the head of Scientific Computing at SLAC

SLAC is a DOE funded laboratory that hosts three user facilities (Linac Coherent Light Source (LCLS), Stanford Synchrotron Radiation Lightsource (SSRL) and Facility for Advanced Accelerator Experiment Tests (FACET). SLAC and Stanford staff participate as users of these facilities. Additionally SLAC has a particle physics and particle astrophysics program. The SLAC community of staff and users consists of representatives of many different scientific disciplines. This community participates in experimental research at other facilities as well. Our participation in data preservation and sharing include participation in the Study Group for Data Preservation and Long Term Analysis in High Energy Physics (http://www.dphep.org/) for BaBAR long term data analysis (http://today.slac.stanford.edu/feature/2010/babar-prototype-data-servers.asp) and the Joint Center for Structural Genomics "Structure Gallery" (http://www.jcsg.org/). We operate the Fermi-LAT production pipeline that produces data that is publically available. We also host the XLDB conference that brings together a broad community to discuss the issues of 'Big Data' (http://www.xldb.org/)

Our primary interest in data discoverability, preservation and access is to facilitate scientific advances and discovery within the domain communities affordably and with the appropriate level of controls to insure integrity of the results and appropriate attribution.

There are common issues for data accessibility across the domains represented by the SLAC scientific community. The common issues are typically social or technical—scientific challenges with data accessibility tend to be more unique.

Given the stated goal of data preservation accessibility to improve the productivity of the American scientific enterprise, one considers ways to improve the productivity of scientists. Scientists are motivated to transition to new practices when those practices increase their scientific productivity at a cost that they can afford. Stating expectations (rather than policies) for accessibility and preservation of publically funded scientific digital data coupled with Federal investment in reducing the technical challenges, burdens and costs to the research communities is likely to lead to faster change in the area of data preservation and accessibility than wide scale policy mandates will. This approach will be especially effective in domains where standardization of data formats and data is acknowledged to be useful, however, is a low priority for implementation for a variety of reasons. In order to facilitate scientific discovery, from the perspective of the scientists who produce the data and the scientists who would use data that they did not gather, mechanisms for data sharing, data preservation, performing operations on data and understanding the intellectual content of the data have to be intuitive and easy to use.

Many of the issues of digital data lifecycle and the associated technical challenges are legitimate topics of research. It is usually insufficient to preserve the data without also having mechanisms for the preservation and encapsulation of the knowledge and tools that produced the data and can operate on it-these are difficult problems without known technical solutions. Given the growth of 'big' data and associated analytics in private industry, there are untapped areas for public/private partnerships in this area.

As a template for potential Federal investments to reduce the technical burdens, achieving economies of scale, and the power of aggregating expertise and research to serve the needs of the diverse scientific communities, one could look to the success of advanced computation programs. Fifteen years ago, the state of unclassified advanced computation was similar to today's situation with scientific data management. Scientific programs that needed large scale computing built their own computers and developed their own algorithms for their own needs. This methodology was scientifically limiting. To address those limitations, programs to build capacity and capability machines as multiple science computational facilities and the associated research programs in applied math and computer science to enable the science communities to use those resources were developed. As a note, these facilities are now in use for some areas engineering and industrial design— infrastructure built for advanced scientific computation now is used in ways that directly grow the U.S. economy.

Aggregated mid-scale computing resources have been developed during this time within some disciplines. Programs such as Open Science Grid and Earth System Grid have been effective in building communities, providing user support, and aggregating the results of basic research into production solutions. Such programs lead to a "commoditization" that ultimately could lead to more science for fixed costs. Scientific databanks provide additional pilot programs that have been effective in promoting some degree of data preservation and data sharing.

Within the concept of building Scientific Data Management Facilities, existing user experimental facilities (where experimental scientific data is collected, cataloged and sometimes analyzed) and dedicated data centers that could serve aggregations of individual small labs would have a role in working with the scientific communities, provide expertise, developing technical solution, and carry out research projects in partnership with other academic institutions and private sector partners. This could determine informed practices and policies that met the scientific needs, understand the differences in data lifetime and advanced the scientific productivity within an affordable budget.