REPORT ON CURRENT POLICIES AND PRACTICES FOR DISSEMINATING RESEARCH RESULTS IN FIELDS RELEVANT TO THE ADVANCED SCIENTIFIC COMPUTING RESEARCH PROGRAM

Reported on behalf of the Advanced Scientific Computing Advisory Committee

30 June 2011

ASCAC Report Panel Members

Dr. James J. Hack (Report Chair), Oak Ridge National Laboratory
Dr. Marsha Berger, Courant Institute of Mathematical Sciences
Dr. Jackie Chen, Sandia National Laboratories
Dr. Jack J. Dongarra, University of Tennessee
Dr. Roscoe C. Giles (ASCAC Chair), Boston University
Dr. Susan L. Graham, University of California Berkeley
Dr. Anthony Hey, Microsoft Research
Dr. Thomas A. Manteuffel, University of Colorado at Boulder
Dr. John Negele, Massachusetts Institute of Technology
Dr. Linda R. Petzold, University of California, Santa Barbara
Dr. Vivek Sarkar, Rice University
Dr. Larry Smarr, University of California, San Diego
Dr. William M. Tang, Princeton University
Ms. Victoria White, Fermi National Laboratory

1. Introduction and Approach

The recently passed America COMPETES Reauthorization Act of 2010 highlights the importance of public access to research results, particularly in the forms of scholarly publications and digital data. The DOE Office of Science has charged its ASCAC advisory committee with describing current policies and practices for disseminating research results in fields that are relevant to the Advanced Scientific Computing Research program.

There is a growing recognition of the central role that data play in science and society. Modern computer and communication technologies enable a level of dissemination of static and dynamic research results that is unprecedented in the history of science. There are multiple audiences for dissemination – other researchers and professionals, as well as the general public. Access to publications and raw data alone may not be enough for all users. The possibilities of how to determine the kind of access that is needed and to provide it effectively are currently being studied by the whole scientific community with new practices being pioneered within disciplinary communities.

The time scale allowed for the preparation of this report precludes an examination of all the relevant issues with the depth and scope that is deserved for a topic of such growing importance. As one example, the National Science Foundation Office of Cyberinfrastructure recently convened a task force (NSF-OCI Task Force on Data and Visualization: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf) which engaged dozens of researchers in the scientific community over an extended period to examine the issues of data and visualization. The NSF exercise, like similar investigations that have been undertaken by others (e.g., NRC Committee on Building Cyberinfrastructure for Combustion Research: http://www.nap.edu/openbook.php?record_id=13049&page=R1; Interagency Working Group on Digital Data report *Harnessing the Power of Digital Data for Science and Society*: http://www.nitrd.gov/About/Harnessing_Power_Web.pdf) underscore the importance and complexity of the data challenge. Therefore, we regard this response to be a limited examination of current policies for disseminating data, that will hopefully lead to more comprehensive explorations and recommendations for how to evolve sustainable mechanisms to deal with the challenges of data curation and distribution.

The Advanced Scientific Computing Research (ASCR) program supports the DOE Office of Science mission by delivering forefront computational and networking capabilities to scientists nationwide. This support enables the research community to extend the frontiers of science, answering critical questions that range from the function of living cells to the power of fusion energy. ASCR supports its mission through the publication of peer reviewed scientific results in mathematics, high performance computing, and advanced networking, and through the application of computers capable of quadrillions of operations per second to the modeling and simulation of phenomena, achieving breakthroughs that cannot otherwise be achieved via traditional laboratory experiments, observations, or theoretical investigations. Effective scientific utilization of this high-end capability computing requires dynamic partnerships among application scientists, applied mathematicians, computer scientists, and facility support staff. This multi-disciplinary type of work is most frequently referred to as computational science.

The Mathematical, Computational, and Computer Sciences Research subprogram develops mathematical descriptions, models, methods, and algorithms to describe and understand complex systems, often involving processes that span a wide range of time and/or length scales. The subprogram also develops the software to make effective use of advanced networks and computers, many of which contain many thousands of multi-core processors with complex interconnections, and to transform enormous data sets from experiments and simulations into scientific insight. The High Performance Computing and Network Facilities subprogram delivers forefront computational and networking capabilities and contributes to the development of next-generation capabilities through support of prototypes and test beds.

2. Findings

FINDING 1:

One of the primary mechanisms through which Computational Science investigators make research results available to the public is publication in peer reviewed scholarly journals and conference proceedings. Access to these published materials, the peer reviewed Versions of Record (VOR), is determined by the policies of the publishers in whose journals the articles appear (for example, as described in a separate letter to Dr. Brinkman's office dated 20 May 2011, prepared by Dr. H. Frederick Dylla, Executive Director & CEO, American Institute of Physics). Some publishers are adopting dissemination mechanisms that make articles immediately available upon publication, although in many cases publishers disseminate published materials through subscriptions or through single article purchase mechanisms. In practice the vast majority of researchers have access to most of the online scholarly literature via institutional subscriptions to journals and/or personal memberships in scholarly societies publishing scientific and technical journals. Rapidly evolving technology and dissemination mechanisms are leading to explorations of broader public access channels that must also preserve the sustainability of the publishing enterprise.

FINDING 2:

The application of high-performance scientific computing facilities to the modeling and simulation of physical and biological systems leads to the generation of large datasets that are used to reveal new scientific understanding. These datasets exist in a variety of forms ranging from pre-publication datasets arising from an investigator's explorations of experimental configurations to final reference datasets that are used in the preparation of scholarly publications. As with other experimentally-derived data, the expertise required to utilize these data is generally beyond the ability of individuals unfamiliar with the modeling frameworks, simulation configurations, and issues like the particular format of the data. Distribution of these raw datasets is generally governed by the principal investigator and initially modulated by data storage and access policies at the high-performance computing centers. Since all ASCR computing facilities have policies that limit archival to a finite period of time following the completion of computational projects (e.g., see https://wiki.alcf.anl.gov/index.php/Data_Policy, http://www.olcf.ornl.gov/kb_articles/storage-policy), long-term stewardship of these data frequently become the responsibility of the investigator. Exceptions to these data practices can be found in communities who have developed organized experimental programs with established policies for the synthesis, organization, and long term curation of reference simulation datasets (see Finding 4 for examples). In general, although these practices work reasonably well, they are inadequate for ensuring the long-term curation and distribution of computer-generated data, particularly for extremely large and complex data sets.

FINDING 3:

The application of high-performance scientific computing facilities to the modeling and simulation of multi-scale phenomena using mathematical models leads to the need for new and novel computational methods, related algorithms, and the associated computer software. As discussed in **Finding 1**, disseminating developments of new mathematical and computational methods is done through publication in peer reviewed scholarly journals and conference proceedings. These algorithmic techniques, however, can have widespread utility in the computational science community, and frequently are embodied in the form of portable computer codes, often referred to as computational libraries. ASCR investigators are engaged in the creation of novel software libraries that allow the sharing and exchanging of computer software and data in a modular fashion. These software efforts can have disciplinary-specific targets (e.g., science application-specific software), or have wider-spread utility such as in the form of commonly used numerical solvers. All software developed under ASCR support is subject to a longstanding policy that the software is to be designated and distributed to the public as Open Source Software or designated as unrestricted releasable software to the public by delivering the software to DOE's Energy Science and Technology Software Center (see http://science.energy.gov/~/media/ascr/pdf/research/docs/Doe_lab_developed_software_policy.pdf) . This policy provides for more restrictive licensing of such software, with ASCR approval, when it is demonstrated that extraordinary circumstances exist such that commercialization of software through restrictive licensing is necessary, or the software is subject to export control, classification or contractual requirements. In practice, most of this software is made publically available through a variety of Open Source Licenses. These Open Source Licenses are copyright licenses that make the source code available, allowing end users to modify and redistribute the source code for their own needs, where some may permit only non-commercial redistribution. Examples of such Open Source Licenses include the Berkeley Software Distribution (BSD), and the GNU Lesser General Public License. Although the ASCR distribution and archival policy is well intended, in the longer term this is an inadequate practice since software artifacts must be maintained and supported if they are to be generally useful. This is currently done within the labs for software adopted as part of library suites, but in general is otherwise left to the investigators, placing the software investment in jeopardy when an investigator moves on or retires.

Finding 4

There are a growing number of discipline-specific communities of researchers adopting more comprehensive strategies with regard to the management of software, data, and publications. One example is the U.S. lattice quantum chromodynamics collaboration (USQCD) which has substantial Office of Science support through ASCR, nuclear physics, and high-energy physics. USQCD software is described and made freely available to the international community via a public web site. The site provides a beginning graduate student who has knowledge of quantum field theory but no prior experience in the field of QCD with all the software tools required to perform state-of-the-art calculations on any of the high-performance computing platforms supported by the DOE. The USQCD community's data is made available as part of the International Lattice Data Grid (ILDG), an international organization that provides standards, tools, and lattice data to lattice theorists around the world by uniting regional data grids. Finally, this community has adopted the practice of posting preprints of major publications on the e-print arXiv, described at the URL: http://arxiv.org/ . The arXiv updates the status of each preprint, so one can eventually see the date and journal in which it has been published. Another community that continues to evolve mechanisms for the distribution of software, data, and publications is the Community Earth System Modeling effort, a jointly supported activity by the Office of Science and the National Science Foundation (http://www.cesm.ucar.edu/index.html). This effort is part of the broader international climate simulation enterprise which has moved to adoption of common metadata standards and the creation of new distribution infrastructures (e.g., the Office of Science supported Earth System Grid, http://arxiv.org/ftp/arxiv/papers/0712/0712.2262.pdf) to make large datasets more readily available and useable by non-experts. These activities appear to be examples of best practices in the dissemination of a wide variety of "data" within disciplinary fields, as well as to the broader external community

3. Summary

There are a wide range of mechanisms and policies in place for the dissemination of research results for ASCR-related programs. At the moment there does not appear to be a set of uniform policies and procedures governing the distribution of the wide range of research related data, but instead a collection of standards and practices that have evolved along with the technologies available for making the products of research investments widely available. The committee believes that this issue transcends the ASCR Program, and transcends DOE at the agency level. Many independent groups are actively engaged in investigating this issue which leads this committee to conclude that the topic would benefit from a much broader, coordinated and comprehensive study.

It appears that one factor playing an important role in the adoption of more aggressive data dissemination practices is the resourcing for the required infrastructure. One example is the long-term maintenance and dissemination of software. Currently, there is neither the obligation nor the resources to induce ASCR investigators to invest the effort to make software developed through DOE research funding more general-purpose and production-ready. Nor are there mechanisms available to fund investigators to continue to evolve and maintain this software in a way that keeps pace with evolving computer architectures. Similarly, the costs of archiving and providing mechanisms for the distribution of computer-generated data, which is growing exponentially, are outside the scope of existing programs, often leaving these types of activities to the good will of investigators. In practice this approach still works reasonably well, but not by design. This points to the need to define a sustainable data infrastructure that balances the costs of sharing research results, particularly in the form of large datasets, with the science enterprise that produces these data and the accompanying discoveries. An exploration of the elements of such a “knowledge distribution infrastructure” would greatly benefit from the experiences of some of the discipline specific community efforts who are in the process of defining what might be called best practices for communicating within their respective disciplines as well as with the general public. An important element of this exploration will be to explicitly tackle the challenge of resourcing such an infrastructure to ensure sustainable long-term stewardship, curation, and distribution of scientific data products.