UNIVERSITY OF CALIFORNIA, IRVINE

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO   SANTA BARBARA • SANTA CRUZ

JOHN C. HEMMINGER
VICE CHANCELLOR FOR RESEARCH
OFFICE OF RESEARCH
160 ALDRICH HALL

IRVINE, CALIFORNIA 92697
Phone 949-824-5796 Fax 949-824-2095
JCHEMMIN@UCI.EDU
HTTP://SURFSCI.PS.UCI.EDU

August 23, 2011

Dr. William F. Brinkman
Director
Office of Science
U.S. Department of Energy
1000 Independence Avenue
Washington, D.C. 20585

Dear Dr. Brinkman:

On behalf of the Basic Energy Sciences Advisory Committee (BESAC), I am forwarding to you the response to the Research Data Public Access Charge that you presented to the committee at our March, 2011 meeting. The sub-committee met on this issue and their report was presented and accepted at the August, 2011 BESAC meeting. Dr. John Tranquada of the Brookhaven National Laboratory chaired this sub-committee.

The issue of public access to research data is of great interest to the membership of BESAC and Basic Energy Sciences community in general. I hope we will have continued opportunities to provide input as the interagency discussion moves forward.

I would like to thank you for the opportunity to involve BESAC in this very important process.

Sincerely,

John C. Hemminger

Digitally signed by John Hemminger
DN: cn=John Hemminger, o=University of California, Irvine, ou=Office of Research, email=jchemmin@uci.edu, c=US
Date: 2011.08.23 15:10:18 -07'00'

John C. Hemminger
Chair, Basic Energy Sciences Advisory Committee
Vice Chancellor for Research, UC Irvine

cc: John Tranquada, BNL
Patricia Dehmer, SC-2
Laura Biven, SC-2
Harriet Kung, SC-22
Katie Perine, SC-22

BESAC Report on Public Access to Research Results

The America COMPETES Reauthorization Act of 2010 highlights the importance of public access to federally-funded research results, particularly in the forms of scholarly publications and digital data. This report describes current policies and practices for disseminating research results in the fields relevant to the Basic Energy Sciences (BES) program.

To begin with, we note that the Basic Energy Sciences program covers a diverse range of fields, with researchers spanning the areas of physics, chemistry, and materials science. To quote the BES web page, BES "supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security." Much of this research is performed by small groups, typically involving fewer than ten people. BES also provides sophisticated tools at major user facilities, such as x-ray and neutron sources, electron microscopy centers, and nanoscience centers. While these are large scientific facilities, most of the users of these facilities are small groups.

Central to the basic research enterprise is the dissemination of research results. The primary mechanism for recording and communicating results is the publication of research reports in scholarly journals. Other forms of dissemination include reviews, books, theses, conference proceedings, conference presentations, patents, and, increasingly, preprint servers and university library servers on the Internet. The criteria for what gets published, where and when, are effectively controlled by individual researchers, their collaborators, and peers. The quantity and quality of peer-reviewed, published research articles are principal, though not exclusive, criteria for hiring and promotion of research scientists at universities and national laboratories. Publications are also an essential criterion used by funding agencies in the peer-review process of selecting grant proposals for funding. Thus, it is in the self-interest of researchers to publish their results in a fashion that makes them widely accessible to all interested parties. For research funded by BES, this process satisfies the policy requirement that project results be promptly disseminated, as expressed in 10 CFR 605.20. It also satisfies the policies of BES user facilities, which obligate users to publish their non-proprietary results in the open literature. (For proprietary work, researchers must reimburse the facility for associated operating costs.)

The criteria for what can and should be disseminated are controlled by the research community through the peer review process. At scholarly journals, the review process is managed by editors, who have typically been selected from the research community. An editor will send a submitted manuscript to one or more research peers for anonymous review. These individual reviewers use their expertise to judge whether a submitted manuscript contains new and substantiated research results, and whether it presents them in an appropriate context of the associated research field. Publishers and editors generally establish the standards for acceptance, as well as the scope of the specific journal, but they rely on the reviewers to help in enforcing them. Hence, the community of active researchers effectively sets the standards of what can and should be published.

Researchers are also judged on the number of times their published articles are cited by others. This provides a strong incentive for making results public in a timely fashion, so that priority for new discoveries can be established. Perhaps the only exception is when the research results lead to a patentable idea. In that case, submission of results for publication will occur after a patent application has been submitted.

In present-day experimental research, raw data are inevitably collected in digital form. The amount of data can vary considerably depending on the experiment and/or experimental technique. In any case, it is important to acknowledge that extensive raw data by themselves will generally neither be submitted nor accepted for publication. Part of the research art is to extract useful information from the raw data, and to present it in a reduced form that provides new insight and benefit to other researchers. The results are generally presented in the form of graphs and tables. Peer reviewers provide their judgment on what amount of data presentation is adequate to substantiate research conclusions. They also judge what amount of data may be useful to others for future research. Effective communication benefits from economical use of space, and many journals have length guidelines, if not page limits. As it may be useful (or required) to make extra results available, most research journals now provide a way to make supplementary information available to readers. In addition, the broad availability of opportunities to post items on the Internet makes it possible for researchers to make sets of subsidiary data available to the scientific community and to the public at large, even when these data and analyses are not of sufficient novelty to satisfy the peer-review process. For example, it is possible to make use of manuscript servers such as the arXiv (to name but one) for rapid but informal dissemination, or to submit more formal reports or data sets to the Office of Science and Technical Information (http://www.osti.gov/).

If we define research results as those that the research community deems worthy of dissemination, then, based on current publishing practices, it is reasonable to conclude that raw digital data do not qualify as research results. The common practice is for researchers to maintain their raw data for a period following publication of associated research results. The length of time varies from field to field, may be set by the policy of the research institution, and may also depend on the ease or difficulty of acquiring new sets of comparable or relevant data. Policies that require data retention for more than one year, but less than ten, are common. Data retention allows any timely questions that would require access to the raw data to be answered. Requests for data from other researchers are frequently handled in an informal fashion. It should be noted that data have a finite lifetime for usefulness. As techniques and materials improve, new measurements supersede the old. Any important new results that are reported will be tested for reproducibility by other groups.

There can be situations in which parameters obtained from the analysis of raw data are worth saving in the form of digital data. A significant example occurs in the field of protein crystallography. In an x-ray or neutron diffraction study of a protein crystal, one seeks to determine the relative positions and elemental identities of all of the atoms within a protein molecule, which may contain on the order of 10^3 to 10^4 atoms. The detailed structural information is the key outcome of such a study, and is of value to others in the community. As a result, a large group of researchers has developed (and obtained funding for) the Protein Data Bank (http://www.pdb.org/pdb/home/home.do). When a paper describing a new protein structure is submitted to a journal for publication, the authors are generally required to upload the digital data containing the structural parameters to the PDB. The PDB provides software tools that can be used to image and analyze the structural information. The PDB is a bottom-up solution that was developed in response to a recognized need of the research community.

A related example is in the area of computational chemistry, where it is common practice to provide digital data to other researchers. One method is to include the data in the Supporting Information of published papers. Such data include the input and output files from calculations, the coordinates of calculated structures, and the parameters used in calculations. In other cases, the sharing of digital data is more informal. For example, a scientist will send a colleague an e-mail request for digital data, such as an input or output file, a structure of a molecule or protein, or the data from molecular dynamics trajectories. Typically, computational chemists are happy to share their data with colleagues in the community. Thus, the computational chemistry community has developed a culture of sharing digital data.

Access to research reports is excellent from virtually anywhere, as all relevant research journals are now available through the Internet. Powerful free search engines make finding reports and informal presentations ever easier. It is generally free to search a journal's content and to obtain titles, author lists, and abstracts of papers. To read the text of a paper may require payment of a fee. Researchers at universities, national laboratories, and commercial companies typically have full access to papers through institutional journal subscriptions (although financial stress can lead institutions to reconsider the number of journal subscriptions that they can maintain). In the past, when journals were only available in paper form, access to a subscription was the only way to monitor the full content, but with publishing on the Web the situation is vastly improved. Not only can one electronically identify and locate papers of interest---many journals allow one to purchase access to the text of individual articles. In addition, many publishers are offering the option for authors to pay an article-processing charge, in exchange for which an accepted manuscript is made freely available to all (i.e., open access).

Research journals perform a valuable function by having each submitted paper reviewed by one or more anonymous referees, selected from the pool of active researchers. The journal must retain editors who are aware of the associated research field and the active researchers. The review process not only screens out papers of questionable validity or insufficient novelty of method, result, or insight, but can also induce authors to improve their presentation, so that the nature and significance of the reported results are more clearly communicated.

Electronic access to journal articles typically includes additional functionality. For example, each research paper must have at the end a list of cited references. Many journals now provide hyperlinks for these references (even for papers in competing journals), so that finding a selected reference can be as simple as the double-click of a computer's mouse. Another example involves listing references that cite a given paper. Obviously, this information does not exist when a paper is first published; it grows with time. Such lists may be provided through the journal's web site or by scientific search engines, through a link that appears when one displays the title and abstract of a particular paper. Such lists can be quite useful in tracking the progress on a particular topic, and in locating the latest research connected to an older report.

When a research report is published, the Version of Record (VOR) is the edited and typeset manuscript that appears in the journal. (Increasingly, the version that appears on the Web takes precedence over the printed version, as post-publication corrections, when necessary, can be made directly to the VOR.) In addition to the VOR, it is common practice in some fields to post manuscripts on freely accessible websites. For example, in condensed matter physics and materials science, researchers commonly post their manuscripts on http://arxiv.org/, often at the same time that they are submitted to journals. In the latter case, posting would precede peer review; at the same time, it makes it possible for any interested reader to send comments to the authors, which might lead to improvements in the manuscript prior to publication. While many journals openly accept this practice, some have restrictions on independent posting until some period (typically six months) after the paper appears online in the journal.

For written findings, peer review is the standard for formal dissemination. As discussed above, manuscripts may be made available online prior to peer review, but this is generally with the anticipation of peer review. Public dissemination of digital data generally occurs only in association with a research report, and hence is covered by the peer review process of journals.

Archival stewardship of research reports is largely accounted for by existing practice. It is in the economic interests of publishers to ensure that their holdings remain accessible and in readable formats as software evolves. It is telling that many journal publishers have chosen to make back issues (previously available only in printed form) available on the Web. This has greatly increased accessibility to the research literature. This literature should remain available into the future, as long as the economic model for scholarly publication remains viable. The current system has strong competition among publishers both from scientific research societies and for-profit firms. Competition has driven innovation, especially in terms of improved accessibility to papers, but also in terms of simplified presentations of highlights that are intended for a broad audience. Public archives, such as arXiv, rely on voluntary support from major institutions to underwrite their operating costs (http://aps.arxiv.org/new/#jan2010).

The current system of disseminating research results works reasonably well for the research community. Of course, there are possibilities to enhance public access.

Importantly, any proposed changes to the existing system should be evaluated for the possibility of unintended consequences.

Finally, we note that a related set of responses, available at http://www.aip.org/aip/DOEadvisory.html, has been prepared by four representatives of research-society publications and endorsed by at least 22 academic publishers. That document contains, in particular, valuable information on the variety of access models that different publishers are testing and further examples of the added features that publishers provide when readers access articles.