Report of the HEPAP Sub-Committee on the Dissemination of Research Results

June 3, 2011

Introduction

The importance of providing public access to the results of Federally funded research was highlighted in Sec. 103 of the America COMPETES Reauthorization Act of 2010. Inspired by this, Dr. Brinkman, Director of the DOE Office of Science, asked HEPAP to summarize the current practices of researchers funded by the Office of High Energy Physics (OHEP) for disseminating their results. The charge is included in Appendix A. In response the Chair of HEPAP, Professor Mel Shochet, formed a Sub-Committee to produce a report describing the criteria for and the methods of disseminating HEP results. The membership of the Sub-Committee is listed in Appendix B. For the purposes of this report, "dissemination" refers to the circulation of research results outside the originating person(s), institution, or collaboration while "research results" refers to both written findings – in the form of publications, presentations, or proceedings – and the digital data. Since there are significant differences between them, the standard dissemination practices of Experimental researchers are described separately from those of Theoretical researchers.

In describing the current practices the committee was asked to consider these aspects: the criteria for dissemination, the methods of providing access to the research results and whether or not that access is limited or provides additional functionality, who upholds the current policies regarding dissemination, whether peer review is a condition of dissemination, and whether long-term stewardship is accounted for in existing policy or practice. The committee was also invited to comment on which dissemination models, if any, "successfully maximize the potential benefit of research results in a way that is sustainable within the research community".

In the following the current practices for the Experimental research efforts are described first. It should be noted that all OHEP funded experimental research efforts involve multi-institutional collaborations of scientists, usually from all over the world. These large international collaborations have been a part of HEP research for many decades now and their current practices are by and large the same and are standards for the field. The typical practices for the Theoretical research efforts are described next. These efforts are usually fairly small in scale, typically involving maybe half a dozen scientists or fewer, but again most follow a similar set of HEP standards with regards to disseminating their results. As noted below, both the Experimental and Theoretical communities use the arXiv (http://arXiv.org), an automated electronic distribution system providing equal, open, and uniform global access to pre-publication material, to initially disseminate their results. The breadth of the arXiv is complemented by the functionalities and additional publication materials provided by the SPIRES/INSPIRE (http://inspirebeta.net) project. Together these two tools provide open access to the full breadth of HEP research articles and they play a key role in facilitating a global dialog among and between HEP theorists and experimentalists alike. The report concludes with a short summary.

Dissemination of Experimental Research Results

The committee contacted experiments across the spectrum of OHEP funded efforts. For the dissemination of research results in the form of papers and presentations, all of the experiments (ATLAS, BaBar, CDF, CDMS, CLEAN, CMS, D0, LHC-b) employ a very similar set of criteria and follow the same dissemination practices. The same is largely true regarding the dissemination of digital data. Since all OHEP funded experimental efforts are international in nature, their dissemination policies are affected by multiple funding agencies and contributing institutions. The policies described below are not formally upheld by any institution outside of the collaborations themselves and instead reflect long established practices within HEP.

Written Research Findings

Research results in the form of papers and conference presentations are not disseminated until they've undergone an internal (to the collaboration) peer review. These internal reviews typically consist of several steps, beginning with informal review by a sub-set of the collaboration sharing similar interests and expertise, followed by the appointment of a formal internal review committee that must vouch for the validity of the result as well as the quality of the manuscript, and ending with a solicitation for comments and criticisms from the collaboration as a whole. The timescales associated with each review period, the documentation required for each step, and the procedures for obtaining approval at each stage vary among the collaborations and are specified in their bylaws. For some of the smaller collaborations formally appointed internal review committees are not employed since a large fraction of the collaboration is engaged in performing the analysis and producing the manuscript anyway. Once the full set of internal review criteria are met, the result is disseminated in preliminary form either by posting a manuscript to the arXiv or by making a presentation at a conference. The manuscripts are also submitted to a journal for publication. It should be noted that the conference presentations are snap-shots of research results, all of which ultimately end-up in manuscripts posted to the arXiv and submitted for journal publication. The journals used are all peer reviewed, subscription based journals (e.g. Physical Review, Physics Letters, European Journal of Physics, Nuclear Physics, etc.). Formally the version of record is taken to be the published journal article stewarded by the publisher. In practice the collaborations also make available the published version of the manuscript from the arXiv and their own web pages – all of which are open access via the internet. It should be noted that many of the journals offer an "open access" option for an additional fee (per paper) ranging from about $1500 to $3000 depending on the journal. Until recently no collaboration has had a policy of exercising these open access options. However, all the LHC experiments have decided to publish results only in open access formats. Some additional functionality is provided beyond the manuscripts themselves. In particular the arXiv offers features that enable a quick literature search for all manuscripts and presentations posted to the arXiv. A more sophisticated functionality is offered by the SPIRES/INSPIRE search web pages, which include information for manuscripts published in all HEP related journals, links to the

relevant arXiv posting (if applicable), links to the references, and a linked index of all citations. The journals also provide cross-publisher searches of the citations used in a manuscript. Additionally collaborations occasionally make available more detailed information about the result obtained or, if the event sample is quite small, about the events used in the data analysis itself. This additional information can take the form of a table listing in more detail the properties of each event used in the data analysis, or providing functional forms that describe the shape of the data and the associated uncertainties, or likelihood or chi-squared distributions that characterize consistency of the data with a given set of hypotheses. This additional information is disseminated via the collaboration's web page and/or via the HepData Project (http://durpdg.dur.ac.uk) or the Particle Data Group (http://pdg.lbl.gov), all of which are open access via the internet. By default the journal publishers provide long-term stewardship of the published results although the arXiv and the collaboration web sites also provide long-term access.

Digital Data

Historically, the digital data collected by HEP experiments are not disseminated for analysis by people outside the originating collaboration. There are several reasons for this. The raw data set acquired for most of the collaborations is large, measuring from hundreds of tera-bytes to tens of peta-bytes in size. In order for the data to be useful a good deal of high level processing must be performed and an understanding of the details of that processing is important in extracting physics results. In addition, corrections for resolution and acceptance effects as well as background contributions must be accounted for using dedicated Monte Carlo simulation samples. Thus, there are a number of technical challenges that need to be overcome in order to make dissemination of the digital data a useful thing to do. Besides addressing the data-handling and database challenges associated with these large data sets, a great deal of additional functionality would also need to be made available. Implicit in all of this is the availability of clear and thorough documentation. Moreover, any potential user must also have access to a large scale computing facility in order to generate and process the necessary Monte Carlo samples and in order to perform the analysis itself. While there is a general consensus that these issues probably all could be addressed, significant additional personnel and capital resources would be required to do so.

In many cases a small sub-set of the data is made publicly available for outreach and education purposes. In those cases some limited additional functionality is also provided allowing users to easily visualize and/or manipulate the data. The data set that's released is most always composed of a well understood sub-set of the data and the questions posed have well known answers.

In recent years there has been a growing recognition in HEP of the importance of long-term stewardship (>5 years past the end of data collection) of the data and the associated software and documentation as well as the associated challenges. In 2009 the International Committee for Future Accelerators (ICFA, http://www.fnal.gov/directorate/icfa), a working group of the International Union of Pure and Applied Physics (IUPAP, http://www.iupap.org), endorsed a study group for Data

Preservation and Long Term Analysis in High Energy Physics (DPHEP). This group is charged with detailing the technical and governance challenges, surveying current practices, and studying proposed solutions associated with long-term stewardship of the data and the relevant analysis functionality. The ultimate goal is to produce a common set of specifications that would form the basis of the data preservation and long-term analysis policy of future collaborations. There are regular DPHEP workshops and the work is well documented (http://www.dphep.org). A sub-group of DPHEP studies the challenges associated with providing open access to the peta-bytes of digital data generated by modern HEP experiments as well as supplying the necessary software functionality and associated documentation to make use of the data. The preservation of HEP data and its dissemination requires organized action from a variety of stakeholders including the experimental collaborations, the laboratories that host the data storage and associated computing services, and the funding agencies.

It is worth noting that since the middle 1990's all large-scale HEP experiments use ROOT, a publicly available software package (http://root.cern.ch), to format and analyze their data. Within the HEP community, this standardization is purposeful since it facilitates and simplifies collaboration among the contributing laboratories and institutions from across the world. In addition it is recognized that this standardization facilitates the work of DPHEP and offers the possibility of providing broader access to the digital data.

Dissemination of Theoretical Research Results

The dissemination practices of the theoretical community are significantly different than those of the experimental community and reflect long established practices in HEP.

Written Research Findings

Typically theoretical research results are produced by a small number of people, less than half a dozen and often times just two or three individuals. By-and-large the dissemination of these results in the form of written manuscripts is not governed by a formal set of criteria. Instead, the decision to disseminate a result is made by consensus among the collaborating individuals, although the ultimate responsibility for determining whether or not a particular piece of research is ready for dissemination formally lies with the group leader, either the university principal investigator or the senior laboratory staff scientist involved. Completed manuscripts are disseminated posting them on the arXiv, which is open access via the internet. Usually, but not always, the manuscript will also be submitted for publication in a peer-reviewed journal. Prominent examples include Phyical Review D and Physical Review Letters (APS), Journal of High Energy Physics (run by SISSA and published by Springer-Verlag), Nuclear Physics B, Physics Letters B and Computer Physics Communications (Elsevier), and European Physics Journal C (with a scientific advisory committee representing European physical societies, and published by Springer-Verlag). These are all subscription-based journals, although, as mentioned above, many of them offer

an open access option for an additional fee. The use of the open access arXiv, along with the associated SPIRES/INSPIRE tools, is so pervasive in the theory community that the journal-provided open access options are rarely used. In the past the traditional Version of Record has been the journal-published articles. This has changed in recent years, and for theoretical papers written since the late 1990s the arXiv version is generally considered the Version of Record, and the arXiv provides long-term stewardship of the result. Theorists and experimentalists turn first to the arXiv and SPIRES/INSPIRE when consulting the literature or when looking for the latest iteration of a particular piece of research. Since the tradition in the theoretical community is to disseminate research first by posting to the arXiv, formal peer review is not technically a condition for dissemination. However, peer review through journal submission remains an important step in validating and improving research results, and as a matter of practice experimental papers tend to refer to theoretical results appearing in peer-reviewed journals.

Digital Data

Although not technically "digital data", it's important to note that some theoretical research produces results besides the published articles. Examples include simulation programs (e.g. lattice gauge theory simulations like USQCD or MILC and Monte Carlo simulation programs like PYTHIA, HERWIG, SHERPA, or ALPGEN), computation programs (e.g. MCFM or MadGraph), and global fits to a large corpus of data (e.g. CTEQ, ZFITTER, or CKMFITTER). Typically the computer code itself is disseminated in an open access manner via the internet. The release of the computer code is usually accompanied by a publication in a peer reviewed journal describing the functionality of the code and, if relevant, specific results obtained using the code. Since these endeavors usually involve a larger collaboration of theorists (and sometimes experimentalists too), the criteria for dissemination typically include some set of cross-checks that verify the validity of the computer code and/or the results being released. The Version of Record is taken to be the latest version available from the relevant URL, which also provides additional functionality by providing versioning, documenting the relevant differences among versions, producing a User's Manual, and referencing the related articles in peer reviewed journals and/or posted on the arXiv. The long-term stewardship of these results is provided by the collaborations themselves via their web pages. It's worth noting that the HepForge (http://www.hepforge.org) project offers a common repository for many of these computer codes.

Summary

The dissemination practices in HEP reflect long established practices in the field. For experiments, internal review is a condition of dissemination and manuscripts describing research results are always published in peer-reviewed journals. For theorists, manuscripts describing research results are disseminated after consensus is reached among the contributing individuals, usually numbering less than half a dozen. Nowadays, both experimental and theoretical manuscripts are also routinely posted on

the arXiv. Researchers regularly employ the search functionality provided by arXiv and SPIRES/INSPIRE to perform literature searches and to obtain the latest results. Together these tools provide equal, open, and uniform global access to the full breadth of HEP research articles and play a key role in facilitating (and expediting) a dialog among the international HEP community for experimentalists and theorists alike. This model has served the HEP community well over the last two decades and owes its success, in part, to the strong collaboration between the community driven efforts (ie. arXiv and SPIRES/INSPIRE) and the publication journals. There is a consensus in the community that any open access policy should build upon this model. Long-term stewardship of the written research results is provided both by the journals and the arXiv. In many instances theoretical research produces more than manuscripts – usually in the form of computer code that provides simulation, computation, or compilation functionality. This code is usually disseminated after the collaboration has verified its validity and robustness and documented the functionality and user interface. The code is disseminated in an open access manner via the internet on dedicated web sites stewarded by the collaborations themselves. Much of the available code is also available via the HepForge repository. The dissemination of the digital data collected by HEP experiments has so far been restricted to limited releases for outreach and education purposes. These releases are typically open access via the internet and stewarded by the collaborations and their host laboratories. They come with some limited additional functionality in the form of software that enables some simple visualization or manipulation of the data. To date no HEP experiment has provided large-scale open access to its raw form digital data, although limited access to processed data has sometimes been granted upon request. The size and complexity of these datasets present significant technological, governance, and support challenges. The IUPAP and ICFA sanctioned DPHEP Study Group is an international effort working to develop solutions to these challenges and to provide common guidelines for use by future collaborations. The preservation of HEP data and its dissemination requires organized action from the experimental collaborations, the participating laboratories, and the funding agencies.

Appendix A: Charge

Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

February 25, 2011

Professor Melvyn J. Shochet
Chair of HEPAP
Enrico Fermi Institute
Mail Station: 318 (CDF BUILDING 327)
University of Chicago
Chicago, IL 60637

Dear Professor Shochet:

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. A copy of the relevant section, Sec. 103, of the COMPETES Act is appended to this charge letter for your information.

As a first step in assessing the policies for researchers funded by the Office of Science, I am requesting your assistance. Please submit to me, no later than July 1, 2011, a report describing current policies and practices for disseminating research results in the fields relevant to the High Energy Physics program. For the purposes of this report, "dissemination" refers to the circulation of research results outside of the originating institutions or scientific collaborations; "research results" refers to both written research findings (scholarly papers, presentations, reports, etc.) and digital data; and "practices" refers to accepted practices within a scientific discipline. Policies from DOE and other federal and non-federal agencies, including foreign institutions and international scientific collaborations, should be considered within the scope of this report provided that these policies have notable impact on the dissemination of research results in your fields. Examples of relevant government policies include provisions in grants and contracts as well as overarching guidance as set forth in federal regulations and DOE orders[1].

Although your report should be sensitive to the differences between written findings and digital data (and, indeed, differences among each of these), you may find many of the

[1] See, for example, 10 CFR 605.20 (http://law.justia.com/us/cfr/title10/10-4.0.1.3.13.html#10:4.0.1.3.13.0.59.20) and DOE O 241.1B (https://www.directives.doe.gov/directives/current-directives/241.1-BOrder-b/view).

same considerations useful in describing the existing policies, practices, and procedures:

- ***The criteria for dissemination and who makes this determination.***
- ***How access is provided and controlled.***
  Access could be provided through commercial or not-for-profit publishers or databases including archives, websites, and agency repositories.
- ***Whether access is limited in any way.***
  For both written findings and digital data, the distribution could be limited by, for example, subscription fees, technological barriers, by request only, or limited to the members of a particular research group. Furthermore, access may be exclusive for a limited period of time.
- ***Whether the access comes with any additional functionality.***
  For written material, this could be interoperable, cross-publisher searches or federated search and discovery tools; links to data or other supplementary material used in the research (particularly if this ensures reproducibility of the research result); or multimedia; etc.
  For digital data, this could be the ability to reference the data as entered (or as part of a larger dataset), additional metadata or software interfaces for meaningful data mining by people outside the field, or interoperability with other data sets.
- ***The version of the written material or data provided.***
  For example, for written findings, the Version of Record is usually considered to be the manuscript published and stewarded by the publisher; however, internal university or laboratory drafts may also be disseminated.
  For digital research data, a distinction may be drawn between data sets that are statically preserved and those that are continually updated; whether the data are considered "raw" or "analyzed"; and whether the data that support a particular finding can be referenced, for example, by a persistent identifier.
- ***Whether peer review is a condition of dissemination.***
  For written findings, a distinction could be drawn between external peer review, as usually happens with published articles, and an internal peer review as might happen within a Laboratory, university, or scientific collaboration for draft articles to be submitted for publication or conference proceedings.
  Any comparable review process for digital data should be described in the report.
- ***The institution, DOE user facility, or other body by which the policy is currently upheld.***
  Many Federal agencies, Laboratories, Universities, scientific collaborations, and user facilities have their own policies regarding the dissemination of research results including digital data. There may also be established practices that are not formally enforced by any institution but are broadly followed. For example, research communities may have dissemination practices that are followed, independent of agency/institutional requirements.
- ***Whether, in addition to dissemination, long-term stewardship is accounted for by the existing policy or practice.***
  For digital data, the report could mention whether associated software for accessing data is also available and maintained.

In the case of digital data, these descriptions will likely depend on the type, size, and structure of the data sets under consideration. It would be useful, therefore, to include in your discussions, a brief survey of the kinds of data that are generated, the size of the data sets, and how they are stored.

As part of this report, I welcome the Committee's perspective on which dissemination models, if any, successfully maximize the potential benefit of research results in a way that is sustainable within the research community. I also invite you to include any observations regarding opportunities where public access policies or practices could enhance the discovery potential of Office of Science research results.

Sincerely,

W. F. Brinkman
Director, Office of Science

represented on the Committee, to identify and reduce regulatory, logistical, and fiscal barriers within the Federal government and State governments that inhibit United States manufacturing;

(4) facilitate the transfer of intellectual property and technology based on federally supported university research into commercialization and manufacturing;

(5) identify technological, market, or business challenges that may best be addressed by public-private partnerships, and are likely to attract both participation and primary funding from industry;

(6) encourage the formation of public-private partnerships to respond to those challenges for transition to United States manufacturing; and

Strategic plan. Deadline.

(7) develop, and update every 5 years, a strategic plan to guide Federal programs and activities in support of advanced manufacturing research and development, which shall—

(A) specify and prioritize near-term and long-term research and development objectives, the anticipated time frame for achieving the objectives, and the metrics for use in assessing progress toward the objectives;

(B) specify the role of each Federal agency in carrying out or sponsoring research and development to meet the objectives of the strategic plan;

(C) describe how the Federal agencies and Federally Funded Research and Development Centers supporting advanced manufacturing research and development will foster the transfer of research and development results into new manufacturing technologies and United States based manufacturing of new products and processes for the benefit of society to ensure national, energy, and economic security;

(D) describe how Federal agencies and Federally Funded Research and Development Centers supporting advanced manufacturing research and development will strengthen all levels of manufacturing education and training programs to ensure an adequate, well-trained workforce;

(E) describe how the Federal agencies and Federally Funded Research and Development Centers supporting advanced manufacturing research and development will assist small- and medium-sized manufacturers in developing and implementing new products and processes; and

(F) take into consideration the recommendations of a wide range of stakeholders, including representatives from diverse manufacturing companies, academia, and other relevant organizations and institutions.

(c) REPORT.—Not later than 1 year after the date of enactment of this Act, the Director shall transmit the strategic plan developed under subsection (b)(7) to the Senate Committee on Commerce, Science, and Transportation, and the House of Representatives Committee on Science and Technology, and shall transmit subsequent updates to those committees as appropriate.

42 USC 6623.

SEC. 103. INTERAGENCY PUBLIC ACCESS COMMITTEE.

(a) ESTABLISHMENT.—The Director shall establish a working group under the National Science and Technology Council with

the responsibility to coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, including digital data and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies.

(b) RESPONSIBILITIES.—The working group shall—

(1) identify the specific objectives and public interests that need to be addressed by any policies coordinated under (a);

(2) take into account inherent variability among Federal science agencies and scientific disciplines in the nature of research, types of data, and dissemination models;

(3) coordinate the development or designation of standards for research data, the structure of full text and metadata, navigation tools, and other applications to maximize interoperability across Federal science agencies, across science and engineering disciplines, and between research data and scholarly publications, taking into account existing consensus standards, including international standards;

(4) coordinate Federal science agency programs and activities that support research and education on tools and systems required to ensure preservation and stewardship of all forms of digital research data, including scholarly publications;

(5) work with international science and technology counterparts to maximize interoperability between United States based unclassified research databases and international databases and repositories;

(6) solicit input and recommendations from, and collaborate with, non-Federal stakeholders, including the public, universities, nonprofit and for-profit publishers, libraries, federally funded and non federally funded research scientists, and other organizations and institutions with a stake in long term preservation and access to the results of federally funded research;

(7) establish priorities for coordinating the development of any Federal science agency policies related to public access to the results of federally funded research to maximize the benefits of such policies with respect to their potential economic or other impact on the science and engineering enterprise and the stakeholders thereof;

(8) take into consideration the distinction between scholarly publications and digital data;

(9) take into consideration the role that scientific publishers play in the peer review process in ensuring the integrity of the record of scientific research, including the investments and added value that they make; and

(10) examine Federal agency practices and procedures for providing research reports to the agencies charged with locating and preserving unclassified research.

(c) PATENT OR COPYRIGHT LAW.—Nothing in this section shall be construed to undermine any right under the provisions of title 17 or 35, United States Code.

(d) APPLICATION WITH EXISTING LAW.—Nothing defined in section (b) shall be construed to affect existing law with respect to Federal science agencies' policies related to public access.

(e) REPORT TO CONGRESS.—Not later than 1 year after the date of enactment of this Act, the Director shall transmit a report to Congress describing—

(1) the specific objectives and public interest identified under (b)(1);

(2) any priorities established under subsection (b)(7);

(3) the impact the policies described under (a) have had on the science and engineering enterprise and the stakeholders, including the financial impact on research budgets;

(4) the status of any Federal science agency policies related to public access to the results of federally funded research; and

(5) how any policies developed or being developed by Federal science agencies, as described in subsection (a), incorporate input from the non-Federal stakeholders described in subsection (b)(6).

(f) FEDERAL SCIENCE AGENCY DEFINED.—For the purposes of this section, the term "Federal science agency" means any Federal agency with an annual extramural research expenditure of over $100,000,000.

42 USC 6624.

SEC. 104. FEDERAL SCIENTIFIC COLLECTIONS.

(a) MANAGEMENT OF SCIENTIFIC COLLECTIONS.—The Office of Science and Technology Policy shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, access, including online access, and long-term preservation of such collections for the benefit of the scientific enterprise. In developing those policies the Office of Science and Technology Policy shall consult, as appropriate, with—

Consultation.

(1) Federal agencies with such collections; and

(2) representatives of other organizations, institutions, and other entities not a part of the Federal Government that have a stake in the preservation, maintenance, and accessibility of such collections, including State and local government agencies, institutions of higher education, museums, and other entities engaged in the acquisition, holding, management, or use of scientific collections.

Web site.

(b) CLEARINGHOUSE.—The Office of Science and Technology Policy, in consultation with relevant Federal agencies, shall ensure the development of an online clearinghouse for information on the contents of and access to Federal scientific collections.

(c) DISPOSAL OF COLLECTIONS.—The policies developed under subsection (a) shall—

(1) require that, before disposing of a scientific collection, a Federal agency shall—

(A) conduct a review of the research value of the collection; and

Consultation.

(B) consult with researchers who have used the collection, and other potentially interested parties, concerning—

(i) the collection's value for research purposes; and

(ii) possible additional educational uses for the collection; and

(2) include procedures for Federal agencies to transfer scientific collections they no longer need to researchers at institutions or other entities qualified to manage the collections.

(d) COST PROJECTIONS.—The Office of Science and Technology Policy, in consultation with relevant Federal agencies, shall develop a common set of methodologies to be used by Federal agencies for the assessment and projection of costs associated with the management and preservation of their scientific collections.

Appendix B: Sub-Committee Membership

Marina Artuso, Syracuse University
Andrew Cohen, Boston University
Lance Dixon, SLAC
Douglas Glenzinski (Chair), Fermilab
Patricia McBride, Fermilab
Ian Shipsey, Purdue University
Mel Shochet (ex-Officio), University of Chicago