Data Management Plan (DMP)
for
Princeton Plasma Physics Laboratory (PPPL)
February 2015, Revision 0

Reviewed by: **SIGNATURES ON FILE**

R. Hawryluk (ITER & Tokamaks)

M. Ono (NSTX-U)

A. Bhattacharjee (Theory)

H. Neilson (Advanced Projects)

M. Williams (Engineering)

P. Efthimion (PS&T)

Approved: _____
M. Zarnstorff (Ex Officio and Deputy Director for Research)

Approved: _____
Deputy Director for Operations (Adam Cohen)

Approved: _____
Director (Stewart Prager)

Approved: _____
Princeton University Vice President for PPPL (A.J. Stewart Smith)

TABLE OF CONTENTS

INTRODUCTION AND DATA MANAGEMENT POLICY

As of October 1, 2014, the Office of Science and Technology Policy (OSTP) requires the implementation of a Digital Data Management Plan for all federally funded scientific research. That plan, which is to be included in all grant proposals, aims at facilitating the sharing of digital data resulting from experiments or simulations. The main requirement is to make available at the time of publication all the data displayed in charts, figures, and images included in published articles. This can be accomplished by submitting supplemental material along with the article to to the publisher. Links to the supplemental material appear on the same web page as the article and an "Archival Resource Key" (ARK) [http://en.wikipedia.org/wiki/Archival_Resource_Key] or "Digital Object Identifier (DOI)" [http://en.wikipedia.org/wiki/Digital_object_identifier] is added for easy access (see, for example, rules for Supplemental Material for the American Institute of Physics journals [http://publishing.aip.org/authors/supporting-data]. This is the best way to archive the data along with the article.

Since most of the research carried out at PPPL is also published in publicly available technical reports, supplemental data linked to the reports will be stored as follows:

(1) Princeton University capabilities in this area, which this plan proposes to adopt, are quite similar to that provided at similar Universities and fusion research centers, and deploy similar open source software -- i.e., "Dataspace Software" [http://dataspace.princeton.edu/jspui/] using Archival Resource Keys "ARKs" instead of the equivalent "Digital Object Identifiers" (DOIs). In sharing digital data resulting from experiments and simulations, the DMP for PPPL fulfills the key requirement of making available at the time of publication all data displayed in charts, figures, and images included in published articles. This is accomplished by links to supplemental material appearing on the same web page as the article -- together with an ARK for easy access.

(2) Since most of the R&D carried out at PPPL is published in the form of publicly available technical reports, supplemental data linked to such reports will be stored at a centralized repository developed in collaboration with Princeton University. Reports and articles will include the ARK/DOI needed to access the data files and/or the instructions on how to access them in cases where much larger datasets might be made available.

(3) For much larger amounts of data (i.e., approaching the PB range or greater), the Princeton University's DataSpace services cannot provide associated resources – they are limited to 2GB unless special arrangements are made. For the larger volume of data, a repository will be located at PPPL following existing practices in, for example, the Engineering Department. We note here that the NSTX GPI data (in the 10 to 12 Gbytes per file range) will also continue to be located at PPPL.

Non-U.S. collaborators working on PPPL facilities (such as NSTX-U and PS&T devices) and/or with theoretical and design R&D led by PPPL are subject to all of the same DMP rules as U.S. researchers with respect to, for example, sharing published data. They are not governed by their home institution guidelines if they are working on U.S. facilities and/or with theoretical and design R&D projects that are led by scientists within our country.

PART A. PRIMARY DATA MANAGEMENT PLAN – NSTX-U (PPPL'S FLAGSHIP FACILITY)

The NSTX-U Data Management Plan (DMP) describes the elements of data from measured to analyzed and also describes the resources available for the data management and preservation during the course of research operations. In addition, this page describes the resources available for sharing of data and provides a link to user requirements for data access. Finally, web links to the NSTX-U and PPPL computing and analysis resources are provided. Any NSTX-U data management plan questions should be directed to the NSTX-U Head of Physics Analysis: HeadPhysicsAnalysis@pppl.gov.

I. Data Categories

Data from NSTX-U discharges will be obtained from a suite of diagnostics measuring a broad range of plasma characteristics. The three main categories of NSTX-U data are raw, reduced, and analyzed.

A. Raw

Raw (measured) data may take the form of voltages, emissivities, etc., and are not directly useable as input to higher level analysis routines. The raw data will be:

(1) OD - temporally and spatially constant information during the course of a plasma discharge such as fixed operational settings, device/facility conditions, etc.

(2) 1D - temporally varying measurements (magnetic fluxes, neutron emissivity, etc.), or spatially varying data taken only at one time

(3) 2D - measurements that vary both in time and space (kinetic profiles, etc.)

(4) 3D - temporally varying 2D images (visible camera, gas puff imaging, etc.)

B. Reduced

Raw data will be converted to reduced data through diagnostic-specific analysis software. Reduced data will be in real physics units (e.g., temperatures, densities, etc.), and once validated by the responsible diagnostician, can be used as input to high level analysis codes. A listing of NSTX-U diagnostics, units for the measurements, and the person responsible for the diagnostic is provided at: http://nstx-u.pppl.gov/diagnostics.

C. Analyzed

Validated reduced data that has been synthesized through direct analysis or through higher level analysis codes. Analyzed data, along with some validated reduced data, form the basis for figures and physics conclusions presented in publications.

II. Data Management Resources, Storage, and Archival

A. Resources

On-site data management resources include real-time and post-experiment data reduction, standardized data acquisition architecture and storage (MDSPlus), on-call help for software and hardware issues (software and hardware engineers), coordinated hardware maintenance, upgrades and compatibility affecting computers owned by PPPL as well as those owned by

collaborators, shared CPU resources with some CPUs dedicated to specific data acquisition and reduction tasks, and web-based visualization tools. Code for generating plots via web-based plotting tools are maintained on the local PPPL cluster. Outside resources include Google mail, sites, and docs – with NSTX-U web pages managed by Google. Information on downloading and documentation for

MDSPlus – the standard architecture for data management within the magnetic fusion community – is provided at:
http://mdsplus.org/index.php/Introduction

B. Storage

On-site data is stored in MDSPlus, the standard architecture for data management within the magnetic fusion community. All data are stored within this architecture except for certain exceptions (such as fast camera videos, which is stored in its own repository, CAMDATA). Data storage is centrally managed and is contained in a dedicated project space. There is no standard format required for the video data, but the data format for this has evolved into a *de facto* standard. We tend to keep the data in the vendor's file format; e.g., for Vision Research cameras these are called CINE files. Data contributed to international databases are stored on off-site servers but are accessible through the Web.

C. Archival

Data is archived using the EPICS archiver (engineering operations data repository) and using VERITAS for data backup for end users with a self-help archiving system for long-term storage. Procedure ITD-003 (Nov. 2010) governs the PPPL backup policy (available on request) and it includes both on- and off-site storage, data formats include those specified by MDSPlus, NETCDF, SQL Server databases and Plasma State files. Assistance on storage and archival is obtained from the Helpdesk, and from MDSPlus and Unix backup system administrators.

III. Data Access and Sharing

A. Resources

Data sharing is facilitated through Web-based visualization tools accessible to public, common MDSPlus architecture/tools including shared analysis code, NTCC module library, FTP services, common login cluster (ability to access main computer cluster from on- or off-site), trusted data movement mechanisms among PPPL, ORNL, GA, NERSC, MIT and ITER, common output file formats (e.g., Plasma State file from TRANSP runs, NETCDF files), 10 Gigabyte ESNET connection to all National Labs, GLOBUS on-line for transferring data over the internet. Data provenance is limited to maintaining histories of data calibrations, etc. through MDSPlus and keeping track of data smoothing, averaging, etc. in UFILES (for TRANSP runs).

B. Access and Sharing

All research data displayed in publications will be made digitally accessible to the public at the time of publication. This will include data displayed in charts, figures, images, etc., and they will be identified uniquely by Archival Resource Keys (ARKs). The ARKs and/or URLs for accessing the data files will be given in the publication. The data files will be

stored in the Princeton University Data Repository. The underlying digital research data used to generate the displayed data will be made available through the establishment of a collaboration, whose requirements are given below (Sec. III.C.)

C. Requirements

The establishment of a collaboration is contingent on both identifying a point of contact with an NSTX-U researcher and reading and signing the NSTX-U Data Usage and Publication agreement. The use of data and the publication policy are governed by the NSTX-U Data Usage and Publication Form.

IV. Links to NSTX-U and PPPL Data Management Resources

The following links provide additional information for the management and analysis of NSTX-U data.

PPPL Research Computing

PPPL Information Technology

NSTX-U Software Page

NSTX-U Diagnostics

TRANSP

NSTX-U EPICs

V. Digital Data

Digital data in support of publications will be provided in accordance with DOE policy, when developed.

PART B. CONTRIBUTIONS TO PPPL DATA MANAGEMENT PLAN FROM OTHER PPPL DEPARTMENTS

I. Theory Department

PPPL authors can leverage the services offered by the DOE National Energy Research Scientific Computing Center (NERSC). Data from numerical simulations can be stored in NERSC's data archive in a web-accessible location, which is linked to a particular person or project. NERSC already has a data management strategy in place (http://www.nersc.gov/users/data-and-file-systems/policies/) to help users comply with the new OSTP rules. URLs will be provided in publications.

The computational codes used for conducting the research activities in the PPPL Theory department are made available to external Users under the conditions stipulated in the "Theory Code License Release Form", which must be filled out and signed by the User before access to the code can be granted. The form, along with information about the codes, can be found on the following public website:

http://theorycodes.pppl.wikispaces.net/Theory+Department+Codes

External users are encouraged to enter into an official collaboration with the PPPL researcher in charge of the development of the particular code of interest.

II. Advanced Projects

The Data Management Plan for the Advanced Projects Department differs from that described for the NSTX-U with respect to how data generated in the course of R&D will be shared and preserved.

The department conducts its experimental research on overseas facilities. In particular, this department's fusion system studies and socioeconomic studies generate data from analyses using both commercial and research codes. The department's DMP addresses both components:

Experimental Data

The Advanced Projects Department generates experimental data from diagnostics that we build and deploy on overseas facilities, and we rely on host systems for data storage and archiving. We also provide analysis that are applied to data from both U.S. and non-U.S. diagnostics. We cannot impose U.S. requirements on data generated by our overseas colleagues, or analysis results that rely all or in part on host or non-U.S. partner resources.

- Raw and reduced experimental data (as defined by NSTX-U) generated by PPPL researchers in the course of collaboration with non-U.S. facilities will be archived by the hosts and made accessible to researchers involved in the collaboration. Access for others will be considered on a case-by-case basis.
- Analyzed data (as defined by NSTX-U), which relies on raw and reduced data from U.S., host, and non-U.S. partner resources, is maintained by individual researchers. As a rule it is made available only to researchers involved in the collaboration, upon request to the originating individual. Access for others is generally not provided.

Fusion System Studies and Socioeconomic Studies

Design data and analysis results are developed in the course of our DOE-funded studies. For example if we are evaluating a candidate fusion facility, we would typically generate a CAD model of the system and perform analyses using a wide range of physics and engineering analysis tools. We apply codes and generate data files in various formats, but we do not develop codes. Input and output data files and code version information are maintained by the individual researcher.

Publications

- For publications for which a PPPL researcher is the first author, the author will provide a publically accessible supplementary information package as described by the Office of Science and Technology Policy (OSTP), above. The content will include information produced entirely or primarily with PPPL resources. Content produced primarily with the resources of the host or non-U.S. partners will be included on a case-by-case basis, and only with the consent of those parties.
- For publications for which the first author is from a U.S. institution other than PPPL, the DMP of that institution will apply.
- For co-authors on a paper that is first-authored by a member of a host institution outside of the U.S., the U.S. DMP rules do not apply.

- Regarding the mechanics for making supplementary information available, we will follow the guidelines prescribed by this PPPL DMP, including those of the Engineering Department for large amounts of data -- in petabyte (PB) range and above -- created during system and socioeconomic studies, etc. and follow existing practices at PPPL for local data storage.

III. Engineering Department

The Engineering Department does not have a formal data management and storage procedure in place due, in part, to the high volume of data involved with engineering analysis and the time and resources required to process this data. For modest volume data in publications, including information in published graphs, the Princeton University repository will be used. For the archiving and maintaining of engineering data, including large petabyte scale range of data, our plan is to follow existing practices at PPPL for local data storage.

Virtually all engineering analysis work is performed by commercial analysis codes. These proprietary codes have been well documented and tested for accuracy and serve as a common platform, which can be used by engineers worldwide to independently check and verify results. Input and output data from these codes are routinely shared. For projects like ITER, specific codes are mandated as the standard for engineering calculations. While data files are routinely shared, the codes themselves are not, due to their proprietary nature. It is the responsibility of each institution to install these codes and acquire licenses to run them. Typical commercial codes used for engineering analysis and computer aided design include ANSYS, CATIA, OPERA, NASTRAN, Pro Engineer, and MAXWELL. In recent years, individually developed codes (unless they are in support of one of the commercial codes) have been frowned upon as not having been benchmarked and rigorously tested for wider use.

IV. ITER and Tokamaks Department

External Collaborations: Open access to full-text

All publications resulting from external collaborations, either domestic or international, will report a DOI number that links to the full text document. Where the Journal provides open access to the full, published, text, the DOI will be associated with the Journal archive. Where the Journal does not provide open access to the full text, the DOI will be associated to a separate archive. For domestic collaborations (DIII-D and C-Mod) the archive can be either the GA or the MIT DSpace archive. For international collaborations this will be either the archive designated by the various facilities, or the Princeton University archive, when the international facility does not have a designated storage archive. In all cases where the open access is not provided by the Journal, the publication of the published article on a local server must comply with the Journal policy, which might require the layout of the archived document to be not the final, published one.

Data storage and archival

Raw data storage and archival is managed by the individual facilities, both domestic and international. Dedicated servers will be setup at GA and MIT for storage of analyzed data that are included in a publication. It is desirable that all these data, including those that have been reduced and analyzed on the PPPL cluster by a PPPL author, are centralized on the dedicated

servers. For TRANSP simulations that use experimental data from domestic facilities, a dedicated server already exists.

Guidelines for Data Management and Open Access of fusion research in Europe (the new Eurofusion consortium) are available at:

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Each European member is in the process of drafting a Data Management Plan, in compliance with the general guidelines. A complete and finalized plan is not available yet.

Storage of analyzed data from international collaborations must be coordinated between the collaborating institutions. If the Data Management Plan in place does not explicitly require that analyzed data reside on specific server, or if a Data Management Plan does not exist, there is no obligation for PPPL to transfer data to the remote server, and analyzed data can reside locally.

Collaborations with ITER are regulated by contracts between PPPL and the ITER Organization (IO) and are funded directly by the IO. Under these Task Agreements (TA), the ITER and Tokamaks Department typically provides simulations. As part of the TA, the results of these simulations are stored in the Data repository at the IO under the form of Tables, which contain time slices profiles and equilibria (eqdsk format files) and time traces of waveforms. Collaborations with the IO that are not funded directly by the IO, but by DOE thru the Office of Science are still subject to the Data Sharing Agreement in place at the IO.

Data access and sharing

Data access and sharing is subject to the specific Publication Agreement in place at domestic and international facilities.

For DIII-D, copy of the "Data usage & publication agreement" can be found at:

https://diii-d.gat.com/ssl_form/datausage

According to the guidelines published by Eurofusion, open access to published and analyzed data is voluntary: "In the context of research funding, open access requirements in no way imply an obligation to publish results. The decision on whether or not to publish lies entirely with the fundees. Open access becomes an issue only *if* publication is elected as a means of dissemination." There is no obligation for PPPL to provide open access to data that are published as part of an international collaboration regulated by the Eurofusion consortium.

The existing Publication Agreement in place at this time at CCFE and IPP Garching, which regulates the policy for publication of scientific data and collaborations with JET and ASDEX-U, is intended to be valid until otherwise publicly stated. PPPL employees who are lead authors of scientific publications that use experimental data from international facilities must comply with the Publication Agreement in place and are not required to provide a DOI for published data.

Access to Metadata that appear in publications and that are generated as part of ITER collaborations under a TA or under funding by the DOE Office of Science are regulated by the Data sharing agreement in place at the IO. There is no obligation for PPPL Authors to provide open access to these metadata.

When a Data Management Plan does not exist, open access to published data is regulated by the Publication Agreement in place.

V. Plasma Science and Technology (PS&T)

The data management plan for the Plasma Science and Technology Department will be similar to that for NSTX-U. All data shown in figures that appear in publications will be made digitally accessible to the public at the time of publication. The data will be identified uniquely by ARKs/DOIs. The ARKs/DOIs and/or URLs for accessing the data files will be given in the publication. The data files will be stored in the Princeton University Data Repository. In cases where a member of the public wishes to access the data used to generate processed or derived data presented in a publication, the underlying data used to generate the presented data will be made available through formal establishment of a collaboration between the person wishing to access the data and the appropriate person in the PS&T Department. A requirement for establishment of such a collaboration will be that the person wishing to access the data agree to the PS&T Department's Data Use and Publication Policy. With regard to collaboration agreements, the PS&T Department will provide a separate document to which the PPPL Data Management Plan will provide a link.

Data produced by the PS&T Department experiments is stored locally on the experimental data acquisition computers. These are regularly backed up to ensure that the data are not lost due to computer hardware failure.

REFERENCES

DOE Office of Science

"Statement on Digital Data Management"

"Suggested Elements for a Data Management Plan"

"Frequently Asked Questions"

PPPL Procedure GEN-032, "Sharing PPPL Theory and Computation Department Codes with Researchers at Other Institutions"

PPPL Procedure GEN-034 "Sharing PPPL Engineering, Experimental and Analysis Codes with Researchers at Other Not-for-Profit Institutions"