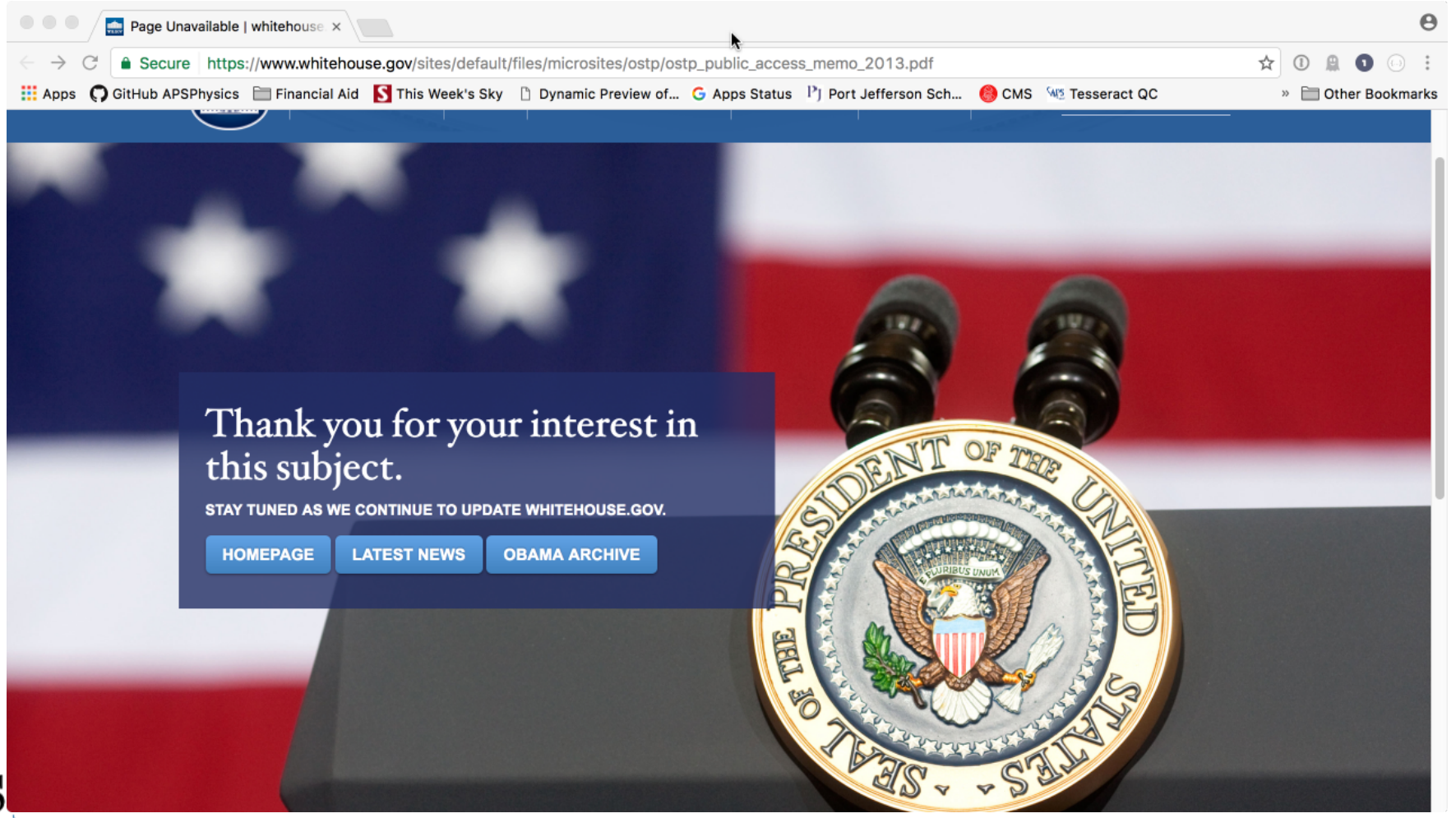
APS and Open Data

Mark Doyle
Chief Information Officer
Unit Convocation – January 27, 2017

Why are we talking about Open Data?

- **Feb. 2013 OSTP memorandum**: "The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. **Such results include peer-reviewed publications and digital data**."

APS physics


Page Unavailable | whitehouse
Secure
https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
Apps
GitHub APSPhysics
Financial Aid
This Week's Sky
Dynamic Preview of...
Apps Status
Port Jefferson Sch...
CMS
Tesseract QC
Other Bookmarks
Thank you for your interest in this subject.
STAY TUNED AS WE CONTINUE TO UPDATE WHITEHOUSE.GOV.
HOMEPAGE
LATEST NEWS
OBAMA ARCHIVE
SEAL OF THE PRESIDENT OF THE UNITED STATES
E PLURIBUS UNUM

APS physics

DOE Data Management Plans

- … researchers must consider the data needed to **validate** research findings
- … making all research data displayed in publications resulting from the proposed research **open**, **machine-readable**, and **digitally accessible** to the public at the time of publication
  - data that are displayed in charts, figures, images, etc.
  - underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the principles stated above.
  - This requirement could be met by including the data as supplementary information to the published article, or through other means.
  - The published article should indicate how these data can be accessed.

APS physics

Sharing and Preservation (from DOE)

- ***Data sharing*** means making data available to people other than those who have generated them.
- ***Data preservation*** means providing for the usability of data beyond the lifetime of the research activity that generated them.
- In the context of this statement, ***validate*** means to support, corroborate, verify, or otherwise determine the legitimacy of the research findings. Validation of research findings could be accomplished by reproducing the original experiment or analyses; comparing and contrasting the results against those of a new experiment or analyses; or by some other means.

APS physics

NSF

- "What constitutes such data will be **determined by the community of interest** through the process of peer review and program management. This may include, but is not limited to: **data, publications, samples, physical collections, software and models**."
- "What constitutes reasonable data management and access will be **determined by the community of interest** through the process of peer review and program management. In many cases, these standards already exist, but are **likely to evolve** as new technologies and resources become available."

https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp

OPEN MPS DATA

Search MPS Open Data

MPSOpenData | Draft Report | Workshop 1 | Workshop 2 | Reference Reading Materials

WORKSHOP 2

Gauging the Impact of Requirements for Public Access to Data

Dec 1-2, 2016

Arlington, VA

MPS Open Data Workshop Series

Taking the pulse of the research community on open data issues

Funded by the National Science Foundation, this workshop series will generate discipline-specific responses from the **Mathematical and Physical Sciences** research communities to the federal policy requiring open data and the recently-released NSF policy statement on open data.

In order to decide how and what to preserve for public consumption, and in what manner the data will be stored and accessed, a series of dialogues is required. Discussions within individual disciplines must reach a consensus on data preservation procedures and data access guidelines consistent with discipline-specific expectations for data re-use, access policies, and the level of burden implied by conservation that is placed on the individual investigator.

These workshops are designed to "*take the pulse*" of the research community on these issues. A final report containing suggestions for best practice and implementation will be submitted to the NSF upon completion of the workshop series.

The MPS Open Data workshop series is supported by the National Science Foundation.

APS physics

What is data?

- *Raw data*: data captured from an experimental apparatus or initial data generated for a theoretical study
- *Structured database*: A curated database that stores data for querying and analysis
- *Processed data*: Data that has been calibrated, organized, and selected (through binning or cuts, for example) from raw data for further analysis
- *Figure/Plot/Table data*: Machine-readable data files that contain the data displayed in a figure, plot, or table in a publication
- *Software*: The software used to create, process, and analyze the data
- *Workflow*: Instructions, frameworks, or scripts use to run the software
- *Software environment*: A specification or an instantiation of the requisite operating system, architecture, libraries, machine state, etc., that are necessary to run the software/workflows

APS physics

Gathering community feedback

- Mike Lubell, Matthew Salter, and I met briefly with Michael Hildreth (University of Notre Dame) at last year's April meeting
- Office of Public Affairs (Allen Hu) solicited feedback from unit leaders in April
  - Do members of your unit typically have the infrastructure required to carry out the mandate contained in the recommendation?
  - As digital archiving methods change over time, do most members of your unit have the capability to maintain access to data for the foreseeable future?
  - Assuming there are no additional federal funds available, would your institution be able to provide the necessary resources to implement the recommendation?
  - Do you have other concerns about the ability of your unit members and their institutions to implement the recommendation?
- Prepared draft report (Hu) and discussions with APS Physics Policy Committee

APS physics

Summary of Concerns Expressed in Unit Leadership Feedback

- **Lack of clarity** in what and how much data are expected to be included under the mandate.
- Placing **additional significant requirements on researchers** to store and prepare files for open data without a concomitant increase in resources.
- The challenge of widely varying data intensity and **disparate levels of effort** in implementing open data, across a **vast array of scientific fields** and Federal agencies
- Potential for misunderstanding data sets without **significant additional context and effort**.
- Potential difficulties in maintaining access to data as digital **archiving methods change over time**.
- A **lack of accurate estimates of all the costs involved with setting up and maintaining an open data system**, including the time and effort of researchers and the administrative and technology burdens on research institutions.

APS physics

Survey

- Went on to do a survey sent to over 5,000 APS journal authors to assess more quantitatively current practices and attitudes. About 550 responses.
- Our goal is to try to characterize what, if any, extra expense and effort would be needed to make various types of data publicly accessible.

Q8: In the past three years, have you or your research group privately shared data and/or software with other researchers or research groups?

Answer Choices	Responses	
Yes	**78.92%**	438
No	**14.41%**	80
Unsure	**6.67%**	37
Total		**555**

Q9: In the past three years, has another researcher or research group privately shared their data and/or software with you or your research group?

Answer Choices	Responses	
Yes	**72.76%**	406
No	**19.00%**	106
Unsure	**8.24%**	46
Total		**558**

Q7: If the necessary infrastructure and funding were available, how inclined would you be to make your data, software, and documentation necessary for its interpretation publicly available?

Answer Choices	Responses	
I already make my data publicly available.	**13.25%**	73
I would endeavor to make more of my data publicly available	**47.19%**	260
I'd like to make my data publicly available, but it would be too time consuming or expensive to prepare it for use by others	**30.67%**	169
I would not be inclined to make my data publicly available.	**8.89%**	49
Total		**551**

Q5: Please indicate the level of staffing/funding that would be required for you or your research group to make publicly accessible the following items on a sustained basis.

	Already doing it	Could be done with existing staff/funding	Would need additional staff/funding	Not practicable	N/A	Total	Weighted Average
Raw data	**7.94%** 44	**21.48%** 119	**35.92%** 199	**18.41%** 102	**16.25%** 90	554	3.14
Structured databases	**4.93%** 27	**14.96%** 82	**31.57%** 173	**8.58%** 47	**39.96%** 219	548	3.64
Processed data	**16.97%** 94	**31.77%** 176	**33.03%** 183	**5.23%** 29	**13.00%** 72	554	2.66
Figure/Plot/Table data	**32.43%** 179	**38.77%** 214	**19.57%** 108	**2.54%** 14	**6.70%** 37	552	2.12
Software	**21.80%** 121	**28.11%** 156	**27.21%** 151	**6.13%** 34	**16.76%** 93	555	2.68

Survey Takeaways

- Significant majority of responders have recently shared table/plot data, processed data, and/or software
  - relationship to and impact on publications less clear
- Majority have established archival policies for data & software
  - are already doing the preservation necessary for open access
  - less than half say they could do preservation and sharing with current resources if more than figure/table data is required
  - ~3/5 of respondents would make their data publicly-accessible with appropriate resources and infrastructure (without a mandate)
- No one is sure who has the infrastructure that could function as a true open data archive
  - local repositories seem to be held in best esteem
- Broad spectrum of respondents; predominantly small research groups
  - is this representative of the "long tail"?

APS physics

Courtesy of M. Hildreth

Comments from Respondents

More than 100 people wrote comments. Some themes:

- raw data is worthless without further processing
- in some disciplines, data sharing is the norm and works fine
  - In some disciplines it doesn't work and no one shares, even when asked
  - subfields are definitely distinct and should be treated that way
- intellectual capital invested in software can be substantial, so mandating sharing or open-source is problematic
- excessive cost (time and money) of mandate to go beyond simple figures/tables data
- data/software sharing is important and should be supported
- need for infrastructure to support open access and archiving
- importance of embargos

APS physics

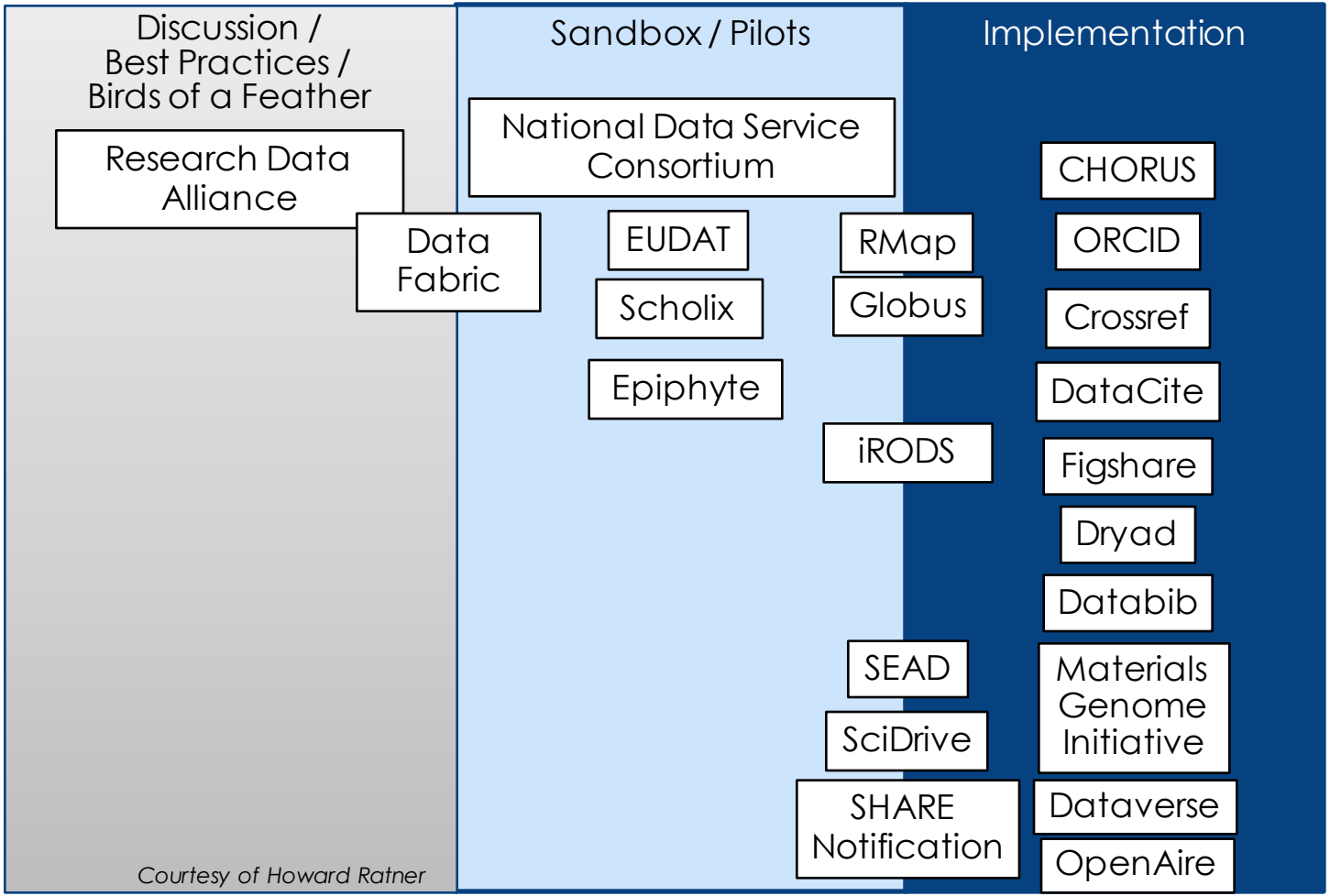
Courtesy of M. Hildreth

Next Steps

- Gather more guidance from you, the APS Unit Leadership (smart phone, tablet, or laptop)
- Continue to work with OPA to formulate APS recommendations
- APS journals enhancing our editorial policies, submissions server, and journal platform to better handle data and the associations among publications and data sets.

Additional Slides

The Data Landscape



APS physics

APS Collaborative Involvement (Journals)

- **National Data Service Consortium**
  - Open Linked Data Repository for Article-Data Associations (OLDRADA) – proposal from Elsevier with short timeline – now called Scholix
  - Materials Genome Initiative – White House, $250 million effort
- **Force11/Data Citation Implementation Group**
- **NISO Journal Article Tagging Suite (JATS) Standing Committee**
- **Research Data Alliance** – standards setting group
- **AAHEP Information Provider Summits** – Collaboration among high energy physics, astronomy, and astrophysics information providers (publishers, INSPIRE, PDG, JaCoW, ADS, arXiv, HepData, etc.)
- **CHORUS**

APS physics

Repositories and Services

- **figshare** – Digital Science (part of Macmillan) – rich set of services for disseminating and viewing data files; long-term preservation prospects unclear

- **Dryad Digital Repository** - DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.

- **Zenodo** - A CERN service, Zenodo is an open dependable home for the long-tail of science, enabling researchers to share and preserve any research outputs in any size, any format and from any science.

- **HepData** – University of Durham - The Durham HepData Project has for more than 30 years compiled the Reactions Database containing what can be loosely described as cross sections from HEP scattering experiments. They are all made publically available in the Reaction Data Database.

- **DataCite** - a DOI registration authority for data sets; akin to CrossRef for journal articles; managed by TIB, the German National Library of Science and Technology, and other member organization (mostly libraries)

APS physics

Data Journals – Another Approach

- ***GigaScience*** (BioMedCentral/Springer) aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish 'big-data' studies from the entire spectrum of life and biomedical sciences. To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources.

- ***Scientific Data*** (Nature Publishing Group) is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the Data Descriptor, is designed to make your data more discoverable, interpretable and reusable.

APS physics