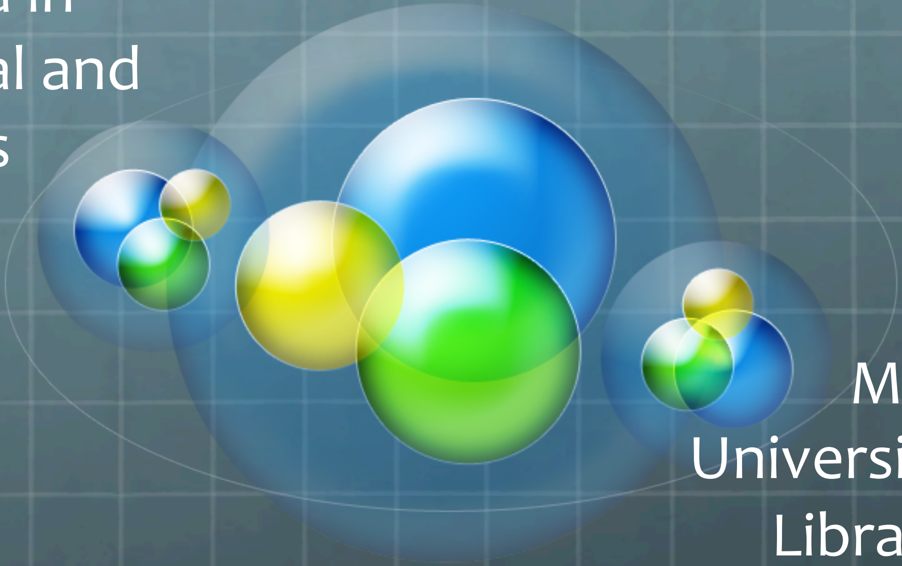
Towards Open Access to Research Data in the Mathematical and Physical Sciences



Mike Hildreth
University of Notre Dame

Library of Congress
September 19, 2016

mathematical & physical sciences
OPEN MPS DATA
preservation

mpsopendata.crc.nd.edu

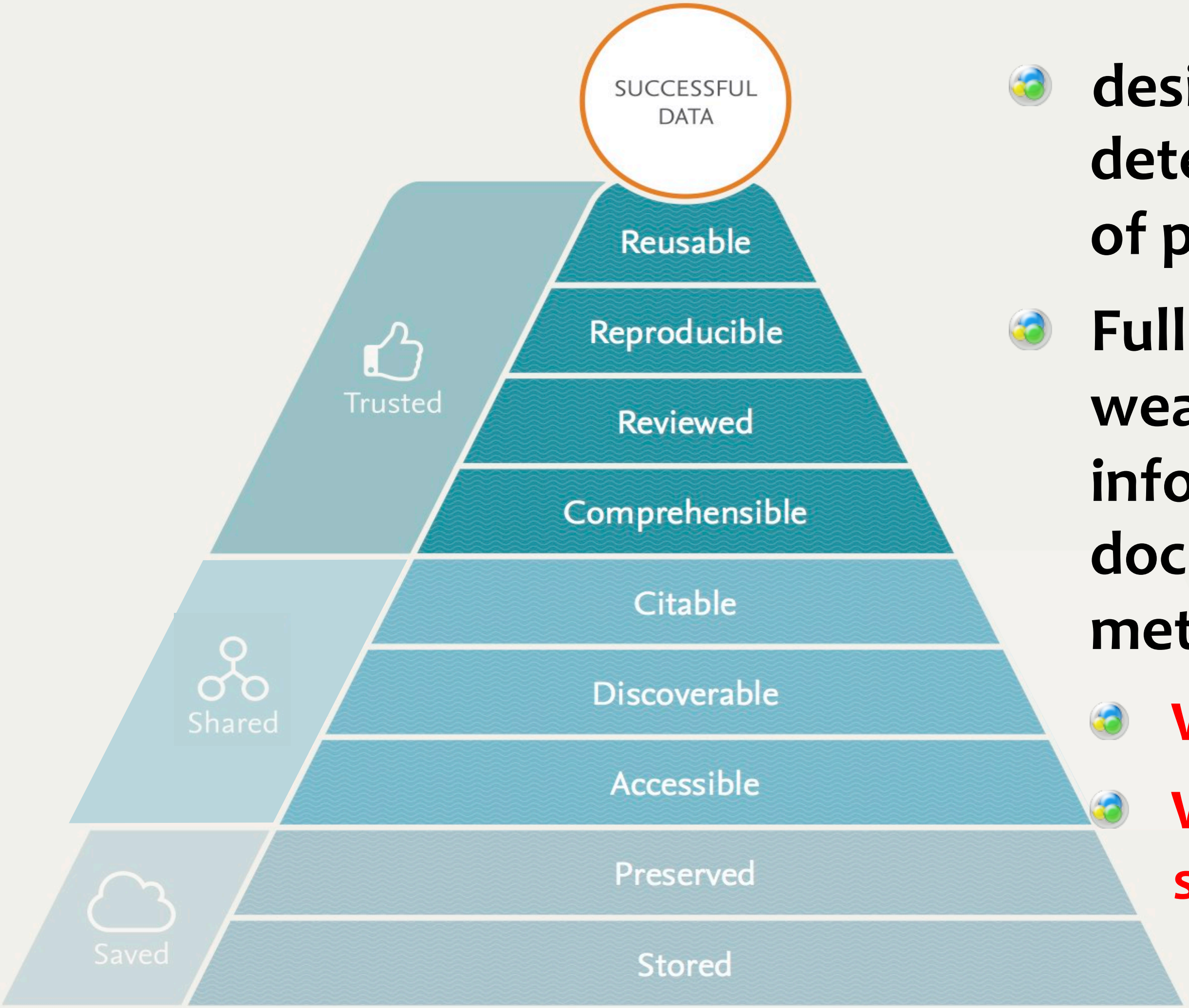
The Landscape

- OSTP Directive, February 2013
  - Research results and supporting research data acquired with public funds must be available to the public
  - Agencies must put forth their plans on how to comply
- NSF Open Data Policy, Report 15-52 (NSF Reply)
  - Mandates (from 2016) deposit of published articles in public archive
  - lays out future directions NSF will explore to make research data more available
    - builds upon existing requirements of DMPs
  - promises to consult community for implementation

NSF Open Data Workshops

- Purpose: “take the pulse of the research community on public access to research data” in the MPS directorate
- Goals:
  - feedback to NSF on current best practices with regard to research data curation and access
  - suggestions for areas of improvement and investment to facilitate broader access to research data in the future
- First workshop Nov 2015, produced Draft Report
  - Researchers, funders, agencies, librarians, publishers
- Second workshop Fall 2016: December 1-2, Arlington, VA
- Final Report will be submitted to NSF
- “Meta” NSF-wide workshop in planning stages

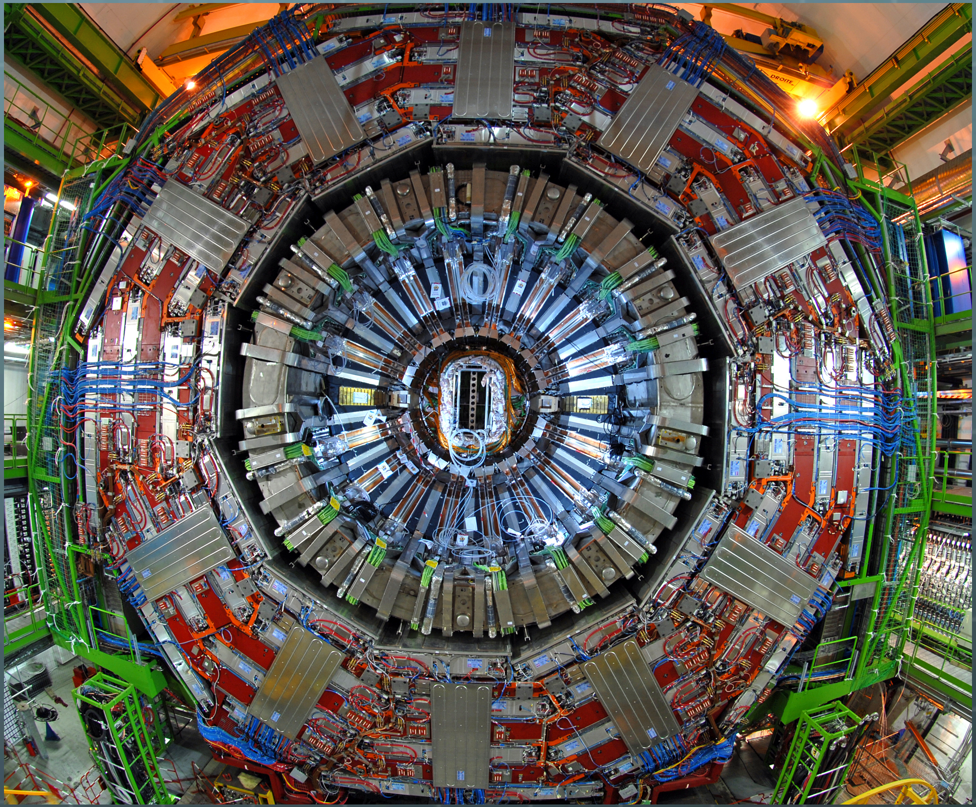
What is Research Data?



Anita De Waard

- desired level of re-use determines complexity of preserved data
- Full re-use implies a wealth of ancillary information, documentation, metadata, software, etc.
  - Where to draw the line?
  - What “level” of data to save?

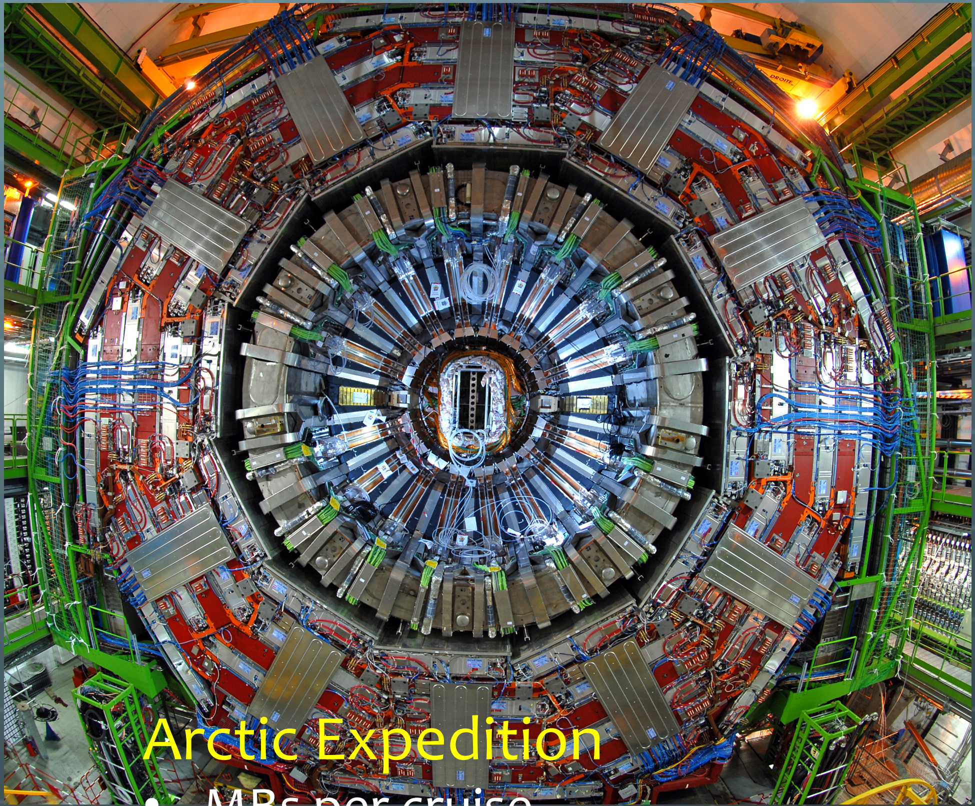
Broad Spectrum of Data



Large Hadron Collider

- 10's PB/year
- 10,000 scientists
- published results require many processing steps
- ~ 500k computers in world-wide computing grid
- huge resources required

Broad Spectrum of Data



Large Hadron Collider

- 10’s PB/year
- 10,000 scientists
- published results require many processing steps
- ~ 500k computers in world-wide computing grid
- huge resources required

Arctic Expedition

- MBs per cruise
- 10 scientists
- very diverse data
- value comes from linking many different datasets
- huge number of small datasets

Needs for an Open Access Future

Expanded concept of “Repository”:

- Infrastructure for software/environment preservation
  - interfaces to computational resources
- Means for revision/correction and versioning; embargo
- Data quality assurance infrastructure
- actionable links between publications and research data/ software
- Federated storage infrastructure
  - globally accessible and interoperable archives
- Global search capabilities
  - automatic metadata generation, appropriate discovery tools

Needs for an Open Access Future

A "Repository" must provide (and researchers must have)

Means/tools to preserve and discover/access/re-use:

- **Software:** the software used to create, process, and analyze the data
- **Workflows:** instructions, frameworks, or scripts use to run the software
- **Software environment:** a specification or a instantiation of the requisite operating system, architecture, libraries, etc., that are necessary to run the software/workflows
- **Simulation capabilities:** the capability to run the software with different parameters than used to generate the original data
- **Documentation:** a description of the software, workflows, and other information describing how the data were derived, processed, and analyzed.
- **Data characterization:** documentation of data (formats, content, etc.) and the metadata that describes it and makes it discoverable.

Needs for an Open Access Future

- **Normative and Policy Considerations:**
  - Establishment of best practices in data management & experimental reproducibility
    - Through what review process are these criteria established?
  - Establishment of ways to quantify the usefulness of data
    - metrics for support of reward structure
  - Establishment of a culture of data citation
  - Establishment of a de-accession policy
  - Establishment of a policy for preserving data for non-published experiments
  - Establishment of a communication structure for published data
  - Establishment of training/workforce development programs

Pilot Projects (Stepping Stones)

- **Certified repositories:**
  - Support creation of "advanced" repository systems that can ingest the broad spectrum of data associated with knowledge preservation
  - Curate lists of certified archives and their uses
  - Inreach to the scientific communities in order to
    - Publicize the capabilities and uses of new repositories, such as embargo capabilities, cross-platform data sharing and computation, etc.
    - Initiate discussion of standards
  - develop guidelines for trusted repositories
    - minimum requirements for due diligence
    - data security, licensing, bit-level integrity checking

Pilot Projects (Stepping Stones)

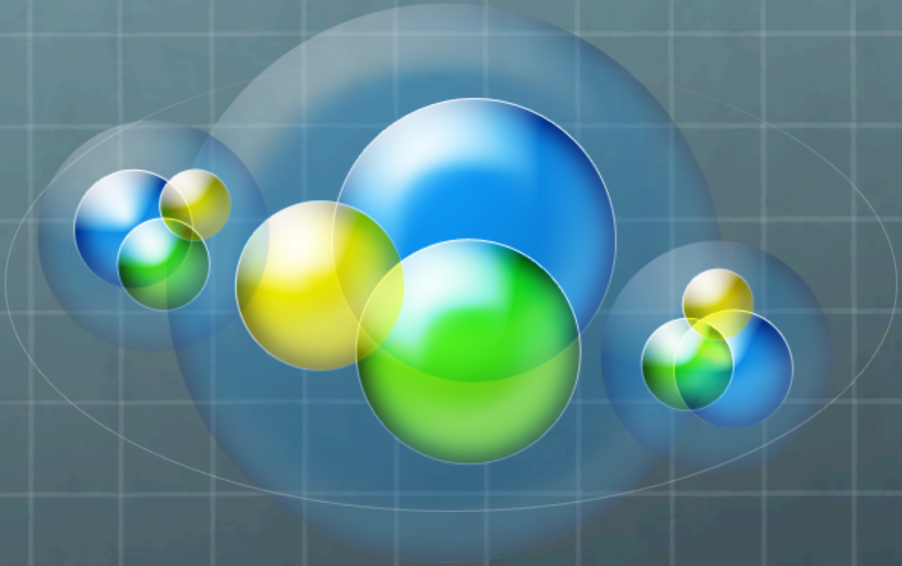
- **Establish prototype federated archival systems:**
    - Create interoperable links between disparate domain-specific resources
- **Attach additional funding or new RFPS for new modes of work in terms of data/knowledge preservation**
- **Projects to demonstrate benefits of workflow preservation, use of data management tools, etc.**
- **Tools for automatic metadata generation**
- **Metadata development:**
    - Develop searchable and computable ontologies for knowledge preservation, including workflows, multiple data sources, etc.
- **Development of training materials for data and workflow preservation tools**

Conclusions

- Much work ahead if we are to provide "open access" to all results/data from federally-funded research
  - clearly won't happen overnight
- The concept of "Repository" is rapidly evolving
  - encompass requirements for reproducibility, re-computablility, "knowledge" preservation
  - oh, and massively heterogeneous data, too.
- "Global" access and storage will require federated architecture of thousands of small repositories
  - linking domain-specific and institutional archives
  - discovery and visualization tools

mpsopendata.crc.nd.edu

OPEN MPS DATA



OPEN MPS DATA

Collective Suggestions

- Baseline recommendation:
  - Data that appear in publications should be available in machine-readable digital format, and persistently linked to those publications
  - simple starting point, but one that is not common to all MPS disciplines
    - would be a major step forward
  - Will require partnership with publishers
- Discipline-specific policy discussion will be required in order to decide an appropriate level of preservation and re-use

Needs for an Open Access Future

- Normative and Policy Considerations: (Social?)
  - for broad adoption, tools enabling preservation for open access must make doing science easier
    - "economic incentive"
  - Modifiication of Incentive Structure
    - Data citation
    - Software citation
    - Change metrics for promotion and tenure
    - Institutional recognition
    - Recognition by funders

Needs for an Open Access Future

- Normative and Policy Considerations: (Policy)
  - Establishment of best practices in data management & experimental reproducibility
    - Through what review process are these criteria established?
  - Establishment of ways to quantify the usefulness of data
    - metrics for support of reward structure
  - Establishment of a culture of data citation
  - Establishment of a de-accession policy
  - Establishment of a policy for preserving data for non-published experiments
  - Establishment of a communication structure for published data
  - Establishment of training/workforce development programs

With open access to data, I could…

- Discover what's available
- Find data that does not support the investigators' expectations, but could be useful in another context
- Make better decisions regarding experiment planning and laboratory safety
- Train students in data analysis, data quality assessment, experiment design

What things would help in research?

- Long-term access to trusted data
- Tools that help to automate metadata annotation, e.g., ELNs (not necessarily commercial products)
- Agreed-upon formats and metadata standards
- Get government agencies to insist on non-proprietary formats for instruments procured with federal funds
- Incentives (i.e., budget) for implementing good data management practice
- Flexibility in generating outputs, e.g., for reporting out to funders

Reproducibility

- Not all research is reproducible (e.g., correlations between natural events)
- Important to document the entirety of the experimental process
  - Allows repurposing of data for new research questions

Reviewing and sharing code

- Peer review of code is impractical
- "Software as Data": code should be shared and described
- Describing code is analogous to describing instrumentation, experimental configuration, etc.
- Software citation is important for credit, establishing precedence

Incentives

- Data citation
- Software citation
- Change metrics for promotion and tenure
- Institutional recognition
- Recognition by funders

Minimum requirements for data associated with publications

- Data needed to support the conclusions drawn in the paper, but what does that mean?
- Data behind the figures
- But how far back do you need to go?
- Can peer review answer this question? Add instruction to reviewer "Is the supplemental information provided sufficient to support the conclusions?"
- Trust and reputation of data provider
- How long to keep? indefinitely

What needs to be done to make open access data useful?

- Share raw data, processed data, derived data and processing steps/tools
- Or trusted, science-ready data
- Data and context
- Some authors are reluctant to have journal host data because they are transferring copyright to the journal ◊ data need home that retains full public access

OPEN MPS DATA