REPRODUCIBLE RESEARCH AND ELECTRONIC NOTEBOOKS

Daureen Nesdill, MS, MLIS
Research Data Management Librarian
J. Willard Marriott Library, University of Utah
daureen.nesdill@Utah.edu

Reproducibility of research is an increasing concern as researchers move from print to a hybrid print/electronic to a totally electronic research project. In addition, research in many disciplines rely on large datasets, i.e. Big Data. Funding agencies have responded to this concern by addressing the 2013 White House OSTP mandate that ensures publications and data resulting from research projects they fund are freely available to other researchers and to the public. The funding agencies are also requesting data management plans. They realize that research projects must be adequately managed so, in addition to being freely available, the research is reproducible and the data can be repurposed.

THE ISSUES: Managing research data has changed over time and has become more complex

Paul Ehrlich at Rockefeller Institute

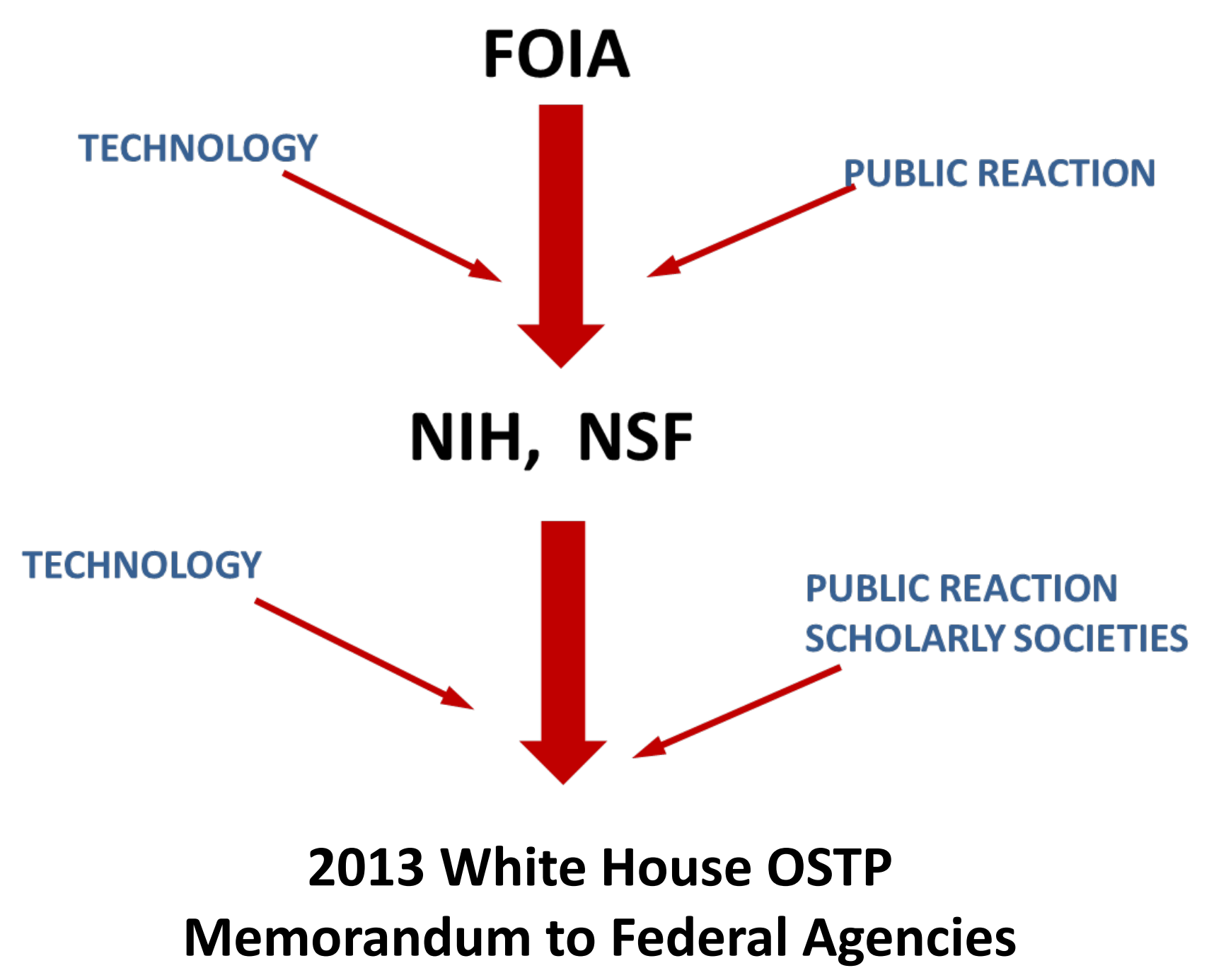
The amount of data accumulated over a career can be large and unmanageable. If in print, how do you find the information needed? If digital, where are all the servers to store all that data? Today's datasets can be 100s of Terabytes.

Collaborations introduce:

- Multicultural differences
- Multidisciplinary differences
- Differences in lab culture
- Geographically dispersed research group

Technology has allowed researchers to collaborate across campus and the world and also with teams in other disciplines. The management of documents is more complex. Language becomes an important issue. We obtain our terminology from our ethnic group, parents, area in the US where we grew up, our graduate advisor and our discipline/sub-discipline. When research teams collaborate language is something that needs to be addressed.

The mandate for researchers to retain data for 3 years is a result of FOIA, Freedom of Information Act. Over time technology advanced and public reaction to the cost of research led to the mandates by first NIH in 2003 and then NSF in 2011 to improve the management of research data and to share research outputs when possible. The OSTP issued a memorandum in 2013 stating that the direct results of federally funded scientific research are to be made available to and useful for the public, industry, and the scientific community.



FOIA

TECHNOLOGY

PUBLIC REACTION

NIH, NSF

TECHNOLOGY

PUBLIC REACTION
SCHOLARLY SOCIETIES

2013 White House OSTP
Memorandum to Federal Agencies

Federal agencies have responded to the mandate with:

Peer-Reviewed Publications
Agencies are requiring that Principal Investigators (PI) make peer-reviewed publications publicly accessible after a 12-month embargo from the original date of publication. In order to be compliant, PIs will need to understand the specific requirements of the agency's programs.

CHORUS (www.chorusaccess.org/) and FundRef (www.crossref.org/fundingdata/index.html) have developed systems to assist with depositing published articles. Submit your article to the publisher along with the name of the funding agency and in a year your article will be in the funder-designated repository.

Digital Data
Agencies are requiring that PIs develop data management plans. In order to be compliant, PIs will need to:

- understand the specific requirements of the agency program
- share the data publicly unless the data management plan justifies not sharing

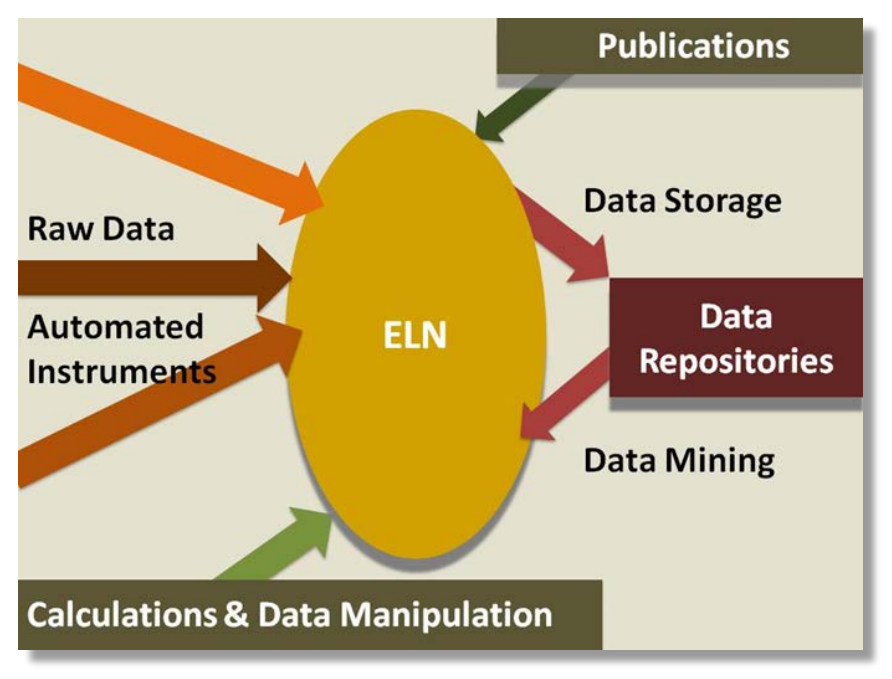
NOTE: Librarians can assist with writing the data management plan and determining the best data repository for your subject and file format.

ADDRESSING THE ISSUES:

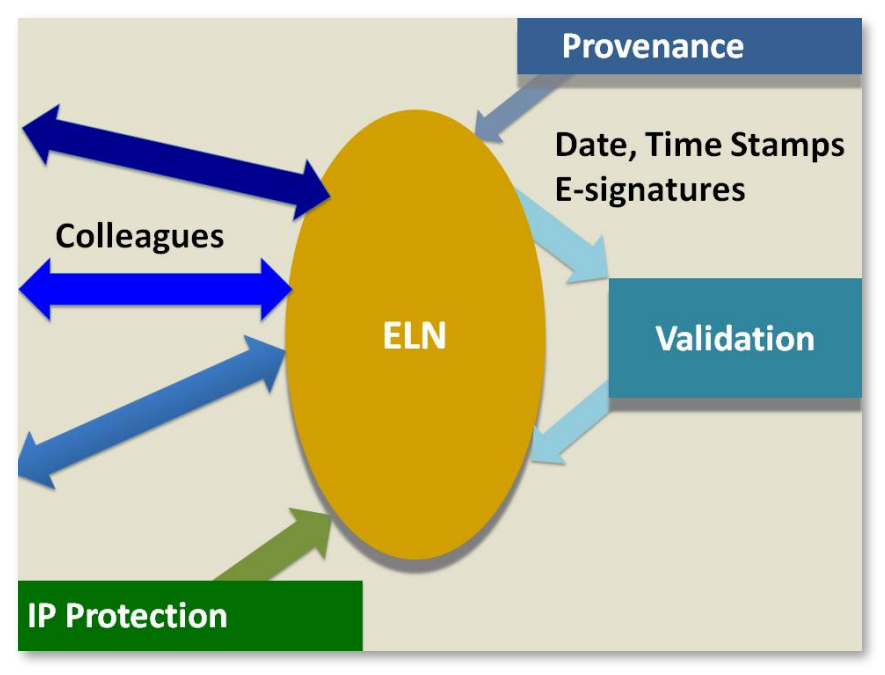
Electronic Lab Notebooks, ELNs

An ELN is not an input device, but a research management tool. It is not just a place to store files, but replaces the print notebook.

There are two layers to an ELN. The **data layer** is where researchers access and work with the raw data and/or data that has been manipulated, refers to information in publications, mines additional data in repositories and eventually will store the data in a repository. This is the lab bench.



The **people layer** includes the colleagues in the research group – across the lab bench or across the ocean. Permissions are set as to who has access to the various areas of the ELN. Here provenance (audit trail) is recorded, that is, how the data was collected, what was done with the data, when and by whom. This validates that the different steps of the experiments were performed, by whom and when. The PI can monitor progress of each team member and also of the entire project from his or her computer.



ELNs presently being used at the University of Utah:

VisTrails was developed at the University's Scientific Computing and Imaging Institute by Prof Juliana Friere. "It is an open source scientific workflow and provenance management system that supports data exploration and visualization".

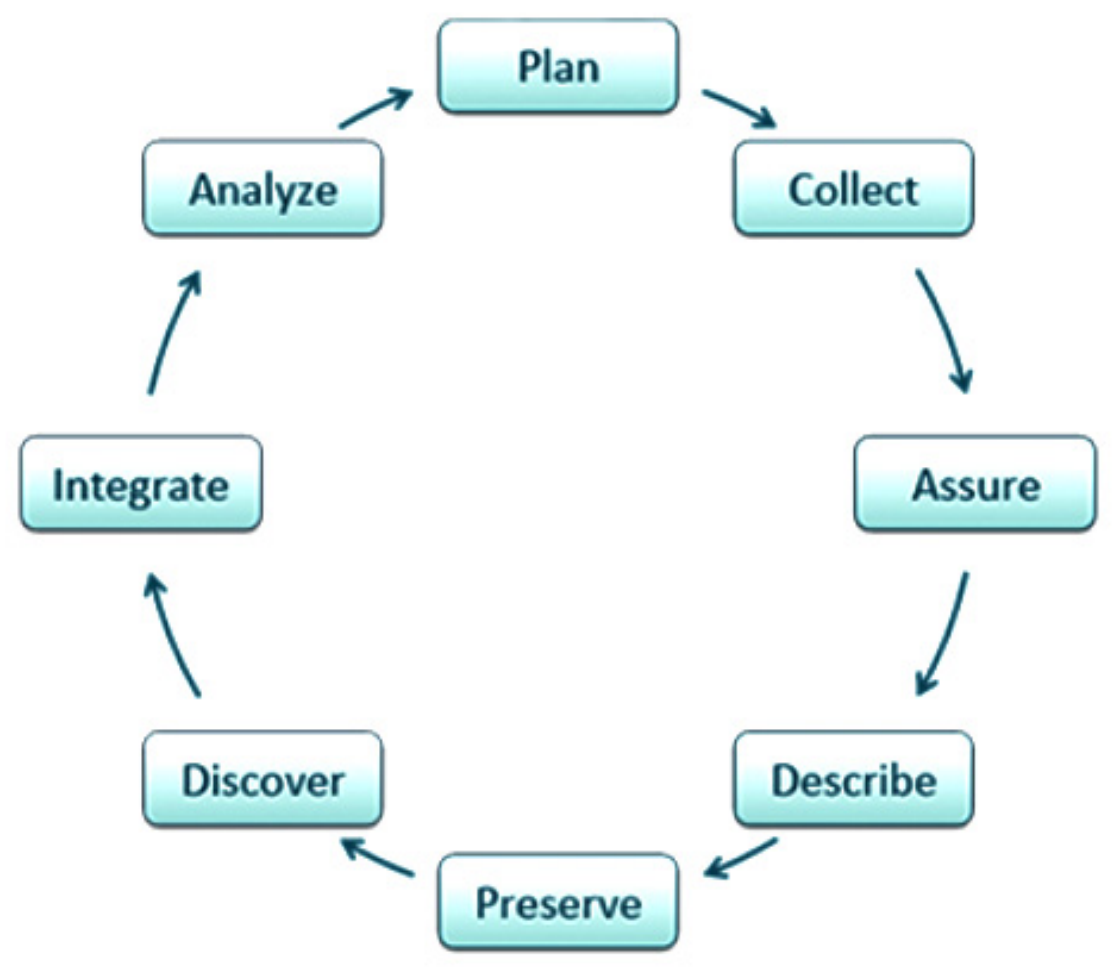
Emulab is a testbed for research on networked computer systems. Emulab was developed by the Flux Research Group, School of Computing. It provides researchers with computing, storage, and network resources on which to run a wide variety of experiments. Emulab is being used by over 4,500 researchers in computer science worldwide.

The Research Electronic Data Capture (REDCap) application is a browser based tool that allows investigators and/or their staff to create web based case report forms (CRFs) and surveys to collect data from a variety of clinical research study types. This ELN is also good for mouse colony management. Since the data are restricted. It is on HIPAA compliant servers at the CHPC- Research Computing Support for the University. Biomedical Informatics provides the support and training.

LabArchives Is the newest ELN on campus. LabArchives is multidisciplinary, and supports collaboration, provenance, access rights, sensitive data and 27/7 access worldwide. It also protects intellectual property. Students, visiting researchers, postdocs and staff can leave with a copy of the research, but not the original.

Using an ELN will go along way towards your research being reproducible, but following best practices, understanding provenance and metadata, and implementing them into the workflow can ensure the reproducibility of your research and the ability to repurpose the data.

Best Practices for Conducting Research



Plan

Collect

Assure

Describe

Preserve

Discover

Integrate

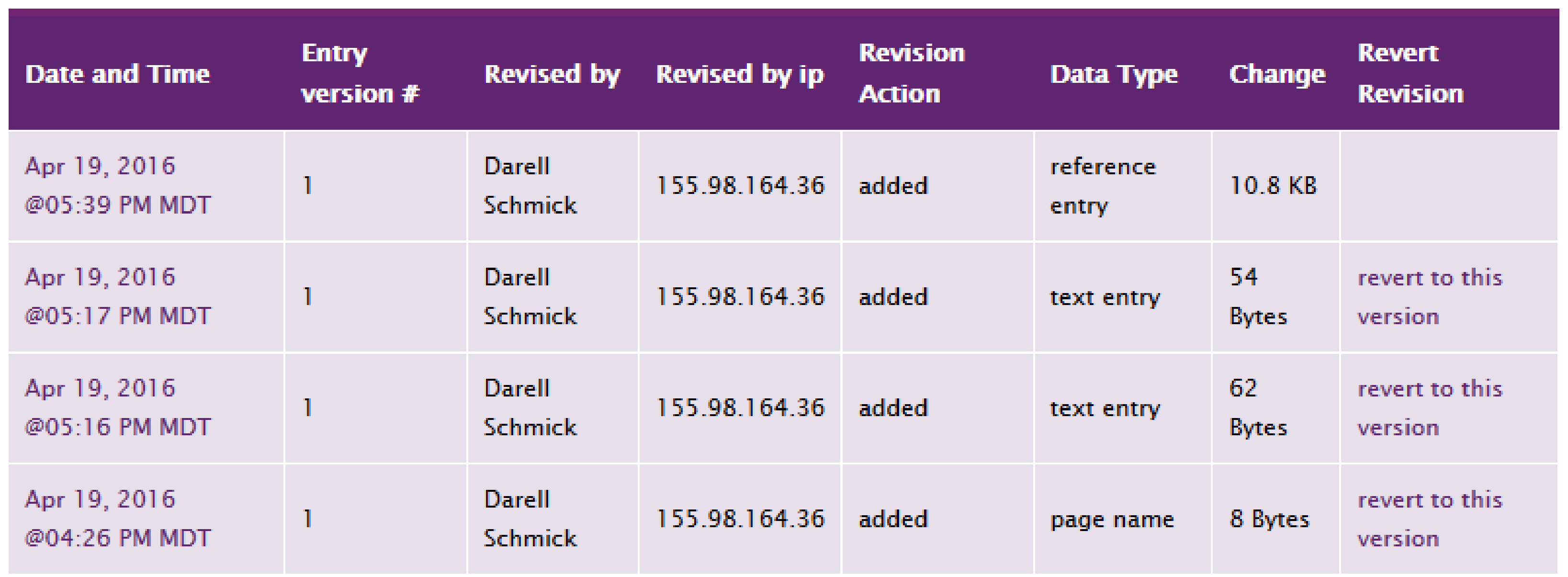
Analyze

Disciplinary best practices and federal agency best practices have been developed to assist researchers with managing their data. DataOne has produced a website and published a Best Practices Primer on data management. In addition, disciplines have developed best practices. For example:

- The Royal Society of Chemistry has an online resource for teaching best practices, Learn Chemistry.
- Engineering Research Centers (sponsored by NSF) has also published an online Best Practices Manual.
- The FDA has developed many best practices including Good Laboratory Practices (GLP) for Non-clinical Laboratory. GLP is defined as a "set of principles intended to assure the quality and integrity of non-clinical laboratory studies that are intended to support research or marketing permits for products regulated by government agencies."

Integrating best practices into LabArchives from the beginning of any project and communicating them to your entire research team will lead to the results of your research being reproducible and repurposed.

Provenance



Date and Time	Entry version #	Revised by	Revised by ip	Revision Action	Data Type	Change	Revert Revision
Apr 19, 2016 @05:39 PM MDT	1	Darell Schmick	155.98.164.36	added	reference entry	10.8 KB	
Apr 19, 2016 @05:17 PM MDT	1	Darell Schmick	155.98.164.36	added	text entry	54 Bytes	revert to this version
Apr 19, 2016 @05:16 PM MDT	1	Darell Schmick	155.98.164.36	added	text entry	62 Bytes	revert to this version
Apr 19, 2016 @04:26 PM MDT	1	Darell Schmick	155.98.164.36	added	page name	8 Bytes	revert to this version

As stated above, provenance is how the data was collected, what was done with the data, when and by whom. From a computer science perspective, Simmhan et al. define data provenance as "information that helps determine the derivation history of a data product, starting from its original sources. The term data product or dataset to refer to data in any form, such as files, tables, and virtual collections."

Labarchives has a method for capturing provenance – referred to as an audit trail. LabArchives not only documents provenance, but researchers can revert to a previous entry and start again from there.

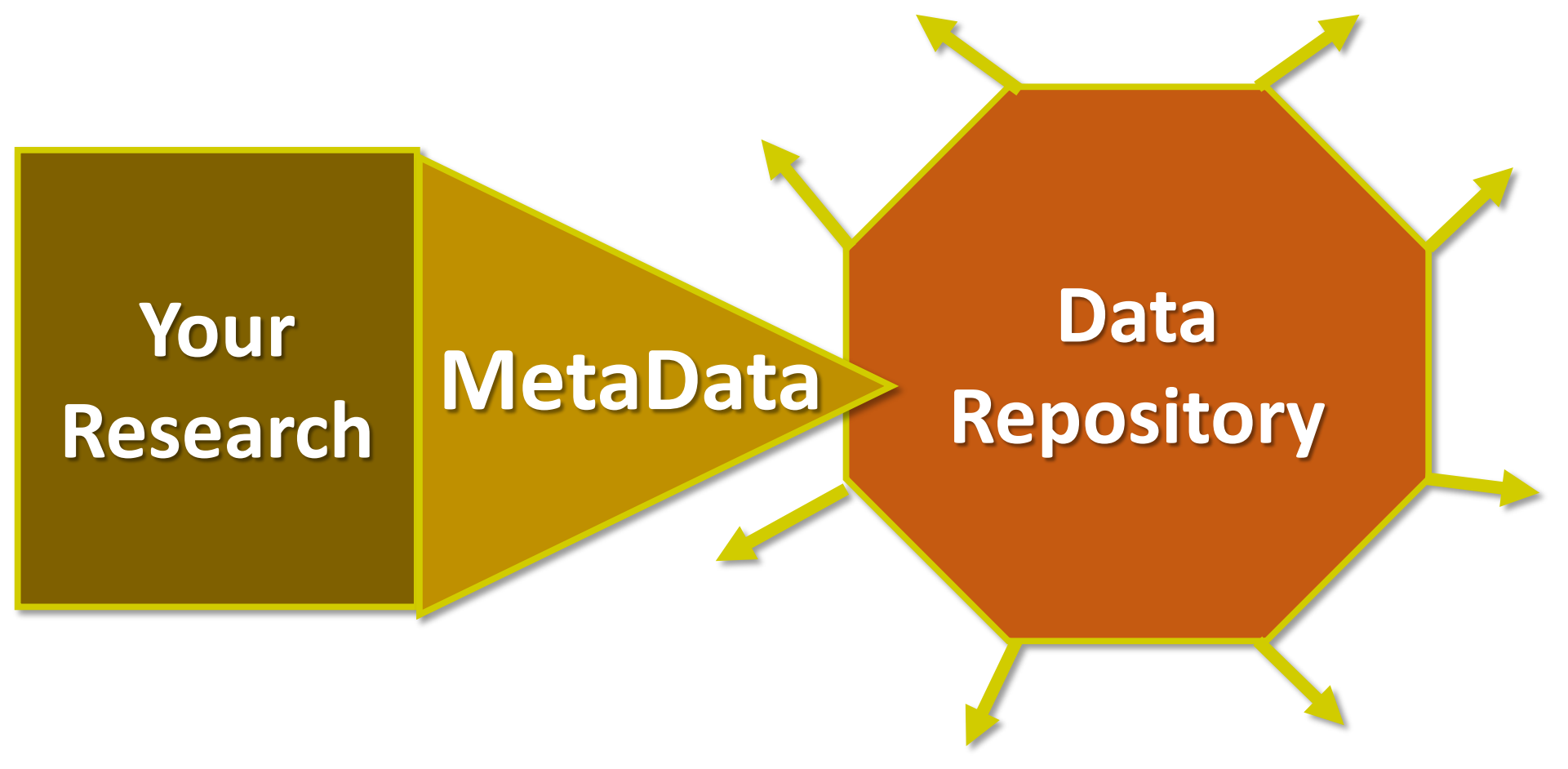
Provenance is important for FDA studies and patents. It is also important as proof the research was completed and followed the specifications of the submitted grant – including the data management plan.

Metadata

In NISO: Understanding Metadata the word metadata is defined as:
"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information."

Research data repositories require appropriate metadata to store and preserve the data appropriately and to insure accessibility. There are three types of metadata, Administrative, Descriptive and Structural. Repositories are concerned with Administrative and Structural. Researchers are concerned with Descriptive. Many disciplines have developed metadata schemas and vocabularies for their research – others have not.

If researchers have a standard vocabulary to work with, and it has been incorporated into LabArchives, then everyone in the research group will be using the same terminology. If a standard vocabulary does not exist, then a vocabulary for the research project and group can be developed.



Your Research

MetaData

Data Repository

Incorporating a standard vocabulary into the research process ensures improved communication among the research team and collaborators. It also can reduce the workload of the librarians helping researchers in submitting data to a discipline data repository.

References

Bird. C.L; Willoughby, C; Frey, J.G Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences Chem. Soc. Rev., 2013, 42, 8157-8175 http://pubs.rsc.org/en/content/articlehtml/2013/cs/c3cs60122f

DataONE https://www.dataone.org/best-practices

Emulab https://www.emulab.net/

ERC Best Practices Manual http://erc-assoc.org/best_practices/best-practices-manual

FDA-GLP https://www.certara.com/2013/12/09/what-is-glp-good-laboratory-practice/

LabArchives http://campusguides.lib.utah.edu/labarchives

Learn Chemistry http://www.rsc.org/learn-chemistry/resource/res00001418/laboratory-best-practices?cmpid=CMP00002777

NISO: Understanding Metadata http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Paul Erlich http://centennial.rucares.org/index.php?page=Chemotherapy

REDCap https://redcap01.brisc.utah.edu/ccts/redcap/

Simmhan, Y.L., Plale, B., Gannon, D. A survey of data provenance in e-science. (2005) SIGMOD Record, 34 (3), pp. 31-36. doi: 10.1145/1084805.1084812 http://www.sigmod.org/sigmod/record/issues/0509/p31-special-sw-section-5.pdf

VisTrails https://www.vistrails.org/index.php/Main_Page