November 2007

Preserving Undergraduate Research Data

Jake R. Carlson
Purdue University, jakecar@umich.edu

Jeremy R. Garritano
Purdue University, jgarrita@umd.edu

Preserving Undergraduate Research Data

Jake Carlson, Data Research Scientist, Purdue University Libraries/Distributed Data Curation Center
Jeremy Garritano, Chemical Information Specialist, Purdue University Libraries

About the Center for Applied Science Practice in Education (CASPiE)

The Center for Authentic Science Practice in Education (CASPiE) is a multi-institutional collaborative effort designed to address major barriers to providing research experiences to younger undergraduate science students.

The objectives of the CASPiE program are to:

1. Provide first and second year students with access to research experiences as part of the mainstream curriculum.

2. Create a collaborative, "research group" environment for students in the laboratory.

3. Provide access to advanced instrumentation for all members of the collaborative to be used for undergraduate research experiences.

4. Help PUI faculty develop research projects so that their own research capacity is enhanced and the students at these institutions can participate in this research.

5. Create a research experience that is engaging for women and ethnic minorities and appropriate for use at various types of institutions, including those with diverse populations.

This text was taken from the CASPiE website (http://www.caspie.org) on 11/22/07

About the Distributed Data Curation Center (D2C2)

The practice of science has been undergoing a dramatic change as advances in technology are enabling experiments that were previously considered impossible. The lifeblood of computational science, data, are rapidly becoming essential sources of information in scientific practice, even beyond their initial research purpose. Technological advances have made the collection and storage of terabytes or petabytes of data achievable, however using and repurposing data requires more than just its immediate storage. Data has to be curated to have value; that is it has to be structured and described in ways that can be easily understood and used by humans and machines, it has to be readily discoverable and accessible, and it has to be preserved for long term access. Data that is not curated will lose its utility as an information resource.

The Distributed Data Curation Center is a research center sponsored by the Purdue University Libraries whose aim to address curation issues and work on problems related to unorganized, disparate, heterogeneous and distributed data, data workflows and environments. The D2C2 works closely with the efforts of other agencies, centers, and groups which are doing related work so that practices and standards can be shared, reviewed and evaluated. As it is not likely that one "fits all" solution in this area will ever be agreed upon, it is the aim of the D2C2 to research insights, applications and systems to facilitate the distributed nature of curation.

More information about the D2C2 is available at our website: http://d2c2.lib.purdue.edu/

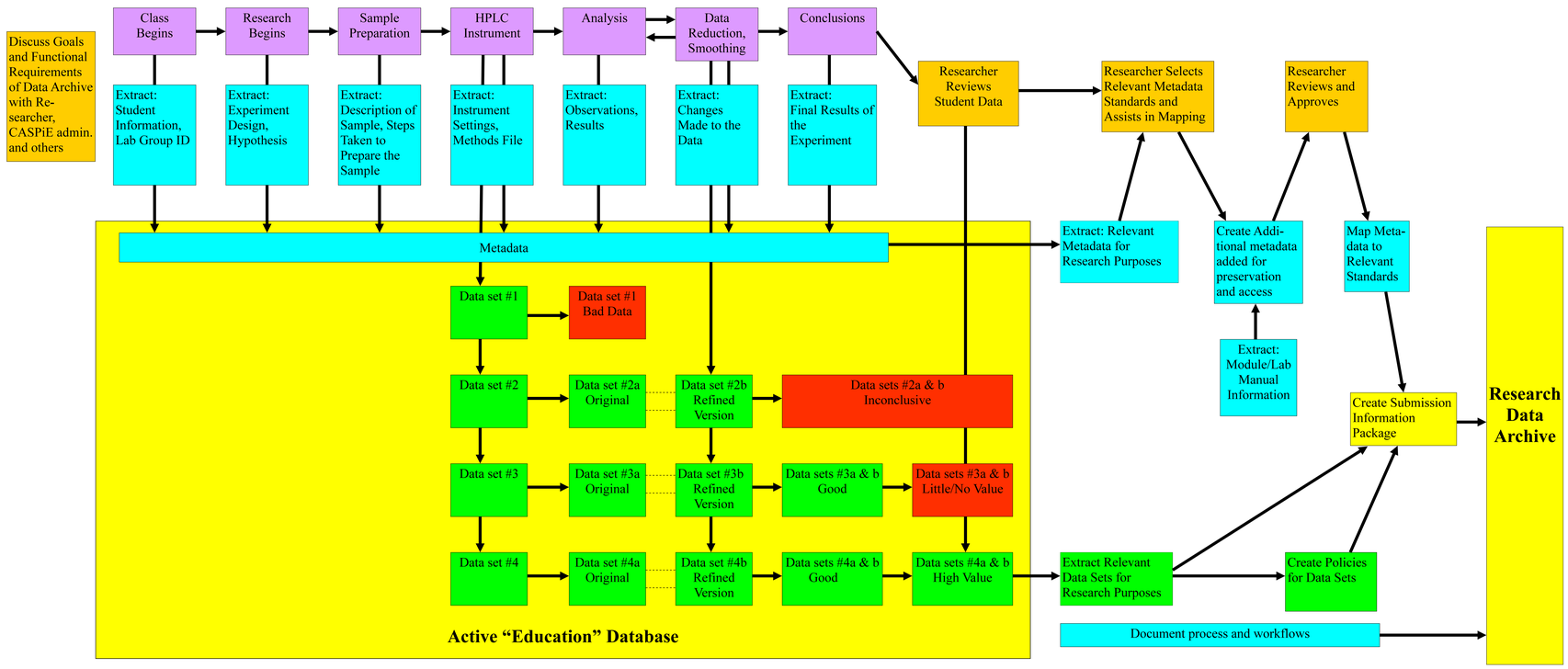
About the Purdue University Libraries

Purdue University librarians believe that librarians and related information professionals are well positioned to help address some of the issues and challenges of research data curation and preservation:

- At its core, Library Science is about describing and organizing information resources, and making them available to the people who need them. Librarians have extensive knowledge, expertise and training in these areas.
- Librarians approach their work with a long term perspective in mind. In many ways, academic libraries represent an institutional commitment to capturing and preserving the legacy of the work done at the university.
- Subject librarians have built strong relationships with their faculty and understand the research, and scholarly communication processes in their disciplines. Libraries are generally trusted organizations by faculty.

Several librarians at Purdue University have participated or are currently participating in collaborative or sponsored interdisciplinary research projects with Purdue faculty. Librarians have worked on data curation and/or preservation projects with faculty from Agronomy, Biology, Civil Engineering, and multiple units within Discovery Park.

A Proposed Model for a CASPiE Data Archive



Why Archive CASPiE Data?

- The research being generated through CASPiE projects is "authentic" research data and therefore should be treated as such. It has value beyond the classroom setting as it is meant to address issues and solve problems in research.
- CASPiE researchers need an efficient method to access this data to do their own research. Dumping data into an Excel spreadsheet or onto a CD limits the value of the data, while limited or no metadata effectively curtails future utility of the data and subsequent research.
- Access to data in an archive can be controlled according to the wishes of the researcher, CASPiE administration and by the status of the user. In the early stages of analyzing the data, the researcher may wish to limit access to him/herself or others in CASPiE. When the initial analysis winds down access to the data can be expanded to others outside of CASPiE.
- Data in an archive can be structured to be machine-readable and interoperable with other data sets. This enables the metadata to be harvested by search engines or subject repositories, making the data more discoverable by others. It also enables the data to be repurposed beyond its original intent through such activities as combining several data sets into a new data set, commonly known as a "mash-up".
- It is good scientific practice to document, archive, and share your data with others. Data in an archive has a permanent identifier, which enables it to be cited when the results are presented or published.
- Storing data in an archive makes good financial sense. It can help prevent duplication of previous experiments, and archived data can serve as the basis of future research, collaborations or other endeavors. Data management plans which include provisions for archiving data are increasingly becoming a requirement to receive funding from government agencies.

Issues and Challenges

- Faculty teaching CASPiE courses are charged with teaching first and second year students actual science practices using a complex system of networked instrumentation in a relatively short time period. Therefore the process of archiving the data cannot present too much of an additional burden upon the faculty, researcher or students.
  - Response: We will work with the researcher, faculty, CASPiE administration, and ITaP to automate as much of the data collection and archiving process as possible, and to closely align the data archiving workflows with the workflows of the experiment itself.
- As CASPiE continues to grow, data will be generated by hundreds of students across scores of class sections at multiple institutions. Of these datasets, what data will need to be preserved in the Research Archive Database, how long should it be preserved, and who should be able to access it?
  - Response: We will develop collection development, access and other related policies for the CASPiE data archive in conjunction with the researcher(s) and CASPiE administration to decide what should be kept and for how long. These decisions will be incorporated and, to the extent possible, automated in the data archive. Usage statistics will be tracked and policies will be revisited on a regular basis.
- Each analytical technique will require unique metadata elements to adequately describe and preserve the experiment. Currently there is a lack of metadata standards for some analytical methods while others standards are still in development and may not yet be widely accepted.
  - Response: If we are unable to discover appropriate metadata standards for a precise analytical method we will employ broader metadata standards to capture the appropriate amount of information for each technique. We will also identify and consult with professional societies / experts in the field who are involved with creating and adopting metadata standards. It is imagined that our work may lead to our involvement in the creation of new or the enhancement of current metadata standards.