Purdue University
Purdue e-Pubs

Proceedings of the IATUL Conferences | 2007 IATUL Proceedings

Jun 12th, 12:00 AM

Facilitating the deposit of experimental chemistry data in institutional repositories: Project SPECTRa

Peter Morgan
Cambridge University Library

Jim Downing
Cambridge University Library

Peter Murray-Rust
Cambridge University Library

Alan Tonge
Cambridge University Library

Fiona Cotterill
Cambridge University Library

See next page for additional authors

Peter Morgan, Jim Downing, Peter Murray-Rust, Alan Tonge, Fiona Cotterill, Henry Rzepa, and Lorraine Windsor, "Facilitating the deposit of experimental chemistry data in institutional repositories: Project SPECTRa." *Proceedings of the IATUL Conferences.* Paper 19.
https://docs.lib.purdue.edu/iatul/2007/papers/19

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Presenter Information

Peter Morgan, Jim Downing, Peter Murray-Rust, Alan Tonge, Fiona Cotterill, Henry Rzepa, and Lorraine Windsor

This event is available at Purdue e-Pubs: https://docs.lib.purdue.edu/iatul/2007/papers/19

ENABLING INTERNATIONAL ACCESS TO SCIENTIFIC DATA SETS: CREATION OF THE DISTRIBUTED DATA CURATION CENTER (D2C2)

James L. Mullins

Purdue University
United States of America

jmullins@purdue.edu

Abstract:

The challenge of accessing, maintaining, sharing and preserving massive datasets, generally referred to as data curation, has been a direct result of computational e-science. Although scientists and engineers recognized the problem, the solution was not apparent. The principles that underlay library science are not widely understood or appreciated by those outside librarianship. The theory and principles behind librarianship are obscured by historical application primarily to print materials – books and journals – however, the same principles that apply to organization, retrieval, and preservation of print materials apply to the digital realm as well.

The National Science Foundation (NSF) in the United States is concerned that much of its funding was committed to creating datasets, used for a specific research project, and then discarded. The question was: couldn't a dataset be "mined" for more than one research project? The NSF has begun to assess and research the issues associated with "archiving" datasets for present and future research use.

Purdue University Libraries, after having observed the need that domain researchers have, determined that the creation of a center to focus on and research these issues while fostering collaboration between librarians and domain researchers was needed. The Distributed Data Curation Center (D2C2) was created at the end of 2006.

Keywords: e-science; curation of scientific data sets; international scholarly communication.

ENABLING INTERNATIONAL ACCESS TO SCIENTIFIC DATA SETS: CREATION OF THE DISTRIBUTED DATA CURATION CENTER (D2C2)

James L. Mullins

Purdue University
United States of America

jmullins@purdue.edu

Introduction:

The research role of university librarians is changing. With the need to manage massive digital datasets by domain researchers, opportunities for new collaboration are developing between the disciplinary researchers in science, engineering, and technology with librarians. Challenges faced by researchers in organizing and accessing massive datasets have created the need, and a new appreciation for the training and knowledge of librarians. The opportunity for librarians to participate as co-investigators in sponsored/funded research is increasing.

The recent report from the US National Science Board (NSB*), Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* confirmed this crisis in data management and called for the creation of new research positions (pp. 26-27) that closely replicate the professional knowledge of librarians [US National Science Board 2006]. The positions of Data Manager and Data Scientist have important elements that the education and experience of librarians fulfill.

At Purdue University, the Libraries created the position of Associate Dean for Research (ADR) in 2005. The ADR position grew out of a need for the libraries to explore collaborative research projects with academic departments in the sciences, engineering and technology. The decision to create the ADR was, in part, to replicate the structure and responsibilities within the academic departments. The dean of libraries met with deans and department heads and described ways in which librarians could add strength to their research proposals. The ADR developed this further through meetings with interested faculty and researchers and was invited to serve on the University Research Council comprised of all associate deans for research from the colleges and schools, as well as central university research administrators. The results have been quite successful, making it difficult for the Libraries to keep up with the requests from academic departments to discuss and explore collaborative research proposals. To facilitate this collaborative research, Purdue University Libraries created the Distributed Data Curation Center (D2C2) to serve as a mechanism to bring researchers together to investigate ways in which optimal dataset management can be achieved at Purdue and throughout the research world.

A National and International Challenge:

From where has this need come? It is an integral factor in the process of e-science. What exactly qualifies as e-science? For that we need to review an accepted definition as provided by the UK e-Science Programme:

What is meant by e-Science? In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprise is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists [Research Councils 2007].

What is the scale that is being proposed and discussed here? It can range from a few gigabytes to the creation of thirty terabytes of data every day, as is being projected for the Large Synoptic Survey Telescope once it is operational. The 8.4m Large Synoptic Survey Telescope (LSST), to be located in Peru, is a wide-field telescope facility that will provide new capabilities in astronomy. The LSST will provide time-lapse digital imaging of faint astronomical objects across the entire sky. The LSST has been identified as a national scientific priority in reports by diverse national panels, including several National Academy of Sciences and federal agency advisory committees. This judgment is based upon the LSST's ability to address some of the most pressing open questions in astronomy and fundamental physics, while driving advances in data- intensive science and computing. [LSST 2007]

The current estimates for the LSST data generation is for 36 gigabytes (GB) of data every 30 seconds, thereby, over a ten-hour night will generate approximately 30 terabytes of data. Within ten years it is estimated that it will require computational power at the rate of nearly 100 teraflops.

The plan for managing the creation of the data from the LSST is illustrated in Figure 1 below.

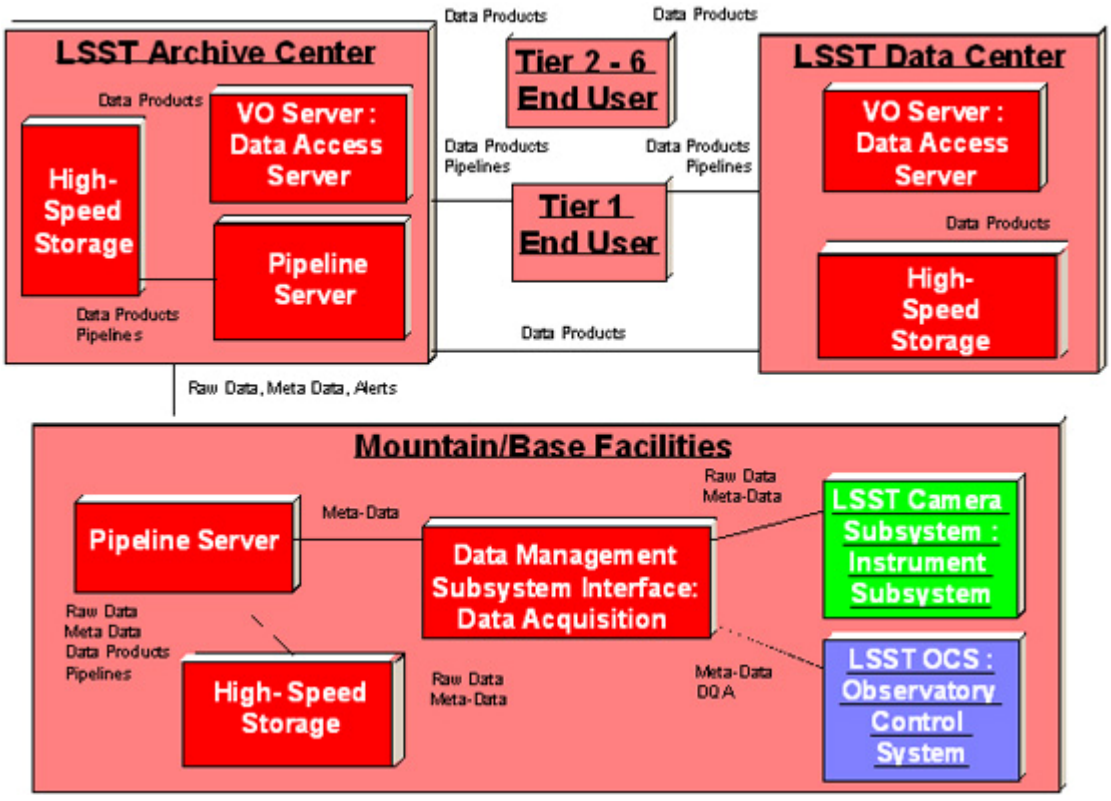


Figure 1: LSST Data Management Facilities [LSST 2007]

Does the LSST project represent the largest computation projected? Only time will tell, whether the mapping of the universe, the earth, the atmosphere, plant and animal genomes, all will generate massive amounts of data, requiring computational power, storage, and curation.

Curation - the Role of Librarians:

In September 2005, the National Science Board published a report titled, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* [US National Science Board 2005]. Beginning on page 25 in the third chapter of the *Report*, the roles and responsibilities of individuals and institutions are described. The individual roles that are most relevant to this paper are the definitions of data authors, data managers, and data scientists. In simple terms these three roles break down as follows: data authors – domain scientists, educators and students who have a vested interest involved in the research generated from the data; data managers – information technologists, computer scientists, and information scientists responsible for the computing, storage and access of the data for analysis; and data scientists – curators, expert annotators, librarians and archivists, among others. The Data Scientists have responsibility to undertake creative inquiry and analysis to enhance the undertaking of research by the data authors, and to apply a consistent methodology and best practices to the curation of data.

As a follow-up to the *Long-lived Digital Data Collections* report, the Association of Research Libraries (ARL) with a grant from the National Science Foundation (NSF) brought together around thirty individuals to discuss the issues of long-term data stewardship. The workshop was held September 26-27, 2006, in Arlington, VA, USA, bringing together domain scientists, engineers, computer scientists, information technologists, librarians, and NSF and ARL staff. The culmination of the workshop was a report issued Fall, 2007 with the title: *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships – the Role of Academic Libraries in the Digital Data Universe* [Association of Research Libraries 2006].

What is curation of data? As with many words, the word curation means different things within different fields and professions. Curation to a scientist or engineer could be defined as "the process of examining, testing and selecting information to be deposited into a database. Typically this adds considerably to the value of a database."[7] However, to a librarian/archivist to speak of "curation of data," the intent is to store, provide access, preserve, and carry forward into the future with assurance that the data will be accessible and retrievable for future verification or use. Although the definitions are not overly different, it does require an awareness of this difference in definition to foster conversations between a scientist/engineer and a data scientist (librarian/archivist).

So how does the specialized knowledge of a librarian/archivist apply itself to the needs of scientific inquiry by the curation of datasets? For hundreds (possibly thousands) of years there have been people designated to organize collections of objects, whether they were papyrus scrolls, cuneiform tablets, vellum manuscripts, or early codex volumes. In order to locate a particular piece of information, a person or group of people worked together to devise a logical order and structure, a retrieval system methodology, that would assign and locate all materials within the "library." The organizational scheme might serve only one repository (library/archives) or a group for an entire region, generating a consistency and ease of retrieval among distributed researchers. However, it wasn't until the late 19th century with the proliferation of publications – books, periodicals, magazines, newspapers – that a standardized approach was necessary. Professional library education developed in the mid-19th century.

Soon thereafter professional codes for organizing materials, first within a library, such as the Library of Congress classification scheme, and more universally such as the Dewey Decimal System. In order to reflect a more international approach (away from United States), in the 20th century the Colon Classification scheme was developed by Shiyali Ramamrita Ranganathan in 1931 in India.

Cataloging rules were created by the American Library Association in collaboration with the Library Association in Britain in the late 19th century to describe the format of books and periodicals as well as create the subject retrieval. The cataloging rules (equate today to metadata) and subject headings (equate today to taxonomies) organized materials and made accessible within Libraries, while providing a consistent framework allowing scholars to move from one library to another, knowing that materials would be organized in a consistent manner.

Toward the end of the 19th century, researchers realized that order must come to the proliferation of professional journals that were being published within the subject domains: chemistry, physics, philosophy, geology, etc. Indexing/abstracting services were created linking authors to articles and to subject content, thereby simplifying researcher access.

In the late 20th century the advent and application of computer technology provided remote access initially to automated catalogs and subsequently to digital indexes and abstracts. Now, much of the research generated today is published in both print and digital format and, more and more, it is being disseminated without even appearing in print. One result of this digital transformation has been that researchers want to be able to "work backwards" from the reported research findings to the underlying dataset used in the research. The published research article itself becomes "metadata par excellence" to the underlying dataset. The illustration below created by D. Scott Brandt, Purdue University, illustrates the new cycle underlying the need to provide access presently and in the future to data sets:

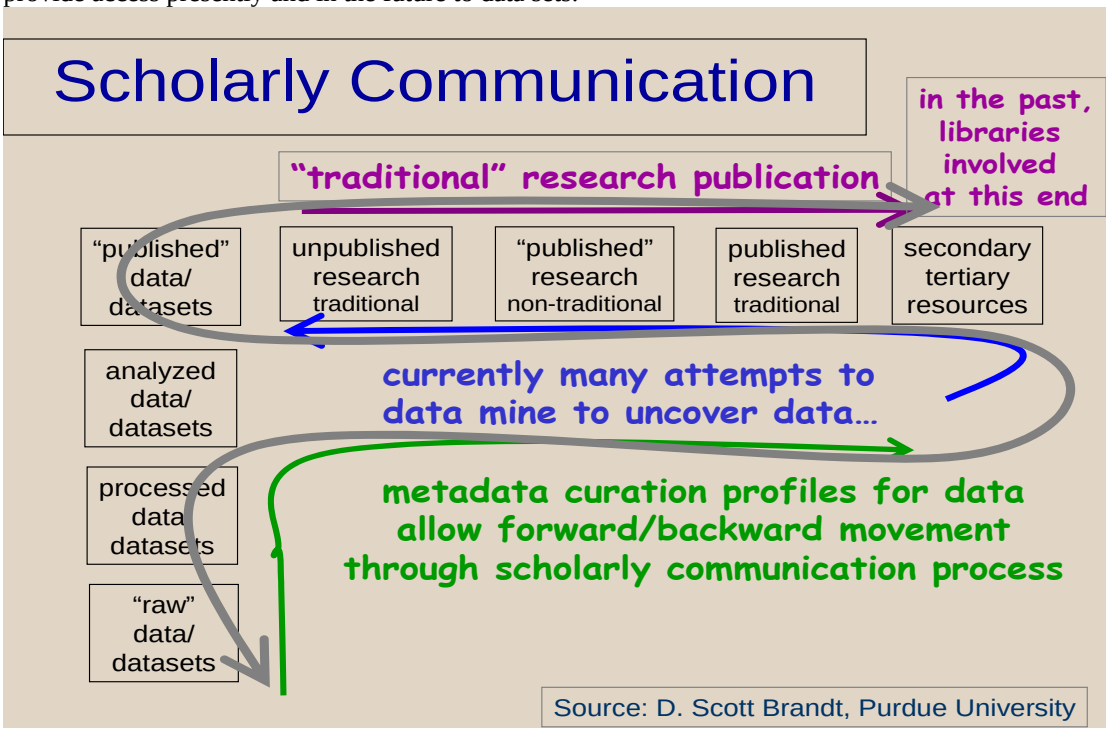


Figure 2: Scholarly Communication, D. Scott Brandt [Association of Research Libraries 2006, p.39]

Dr. Christopher Greer, program director of the Office of Cyberinfrastructure, National Science Foundation, USA, has developed a pentagram that represents the role of the library in the future of data curation within the larger field of e-Science. Referring to it as the I-Center, in order to break down perceptions or stereotypes of the role of a library, he places the I-Center in the middle of the pentagram with the other five points identified as: domain science, computer science, library/ information science, archival science, and cyber infrastructure. His contention is that it will require all five of these specialties to collaborate to develop a model that will be practical and workable to curate the massive datasets that are now being generated.

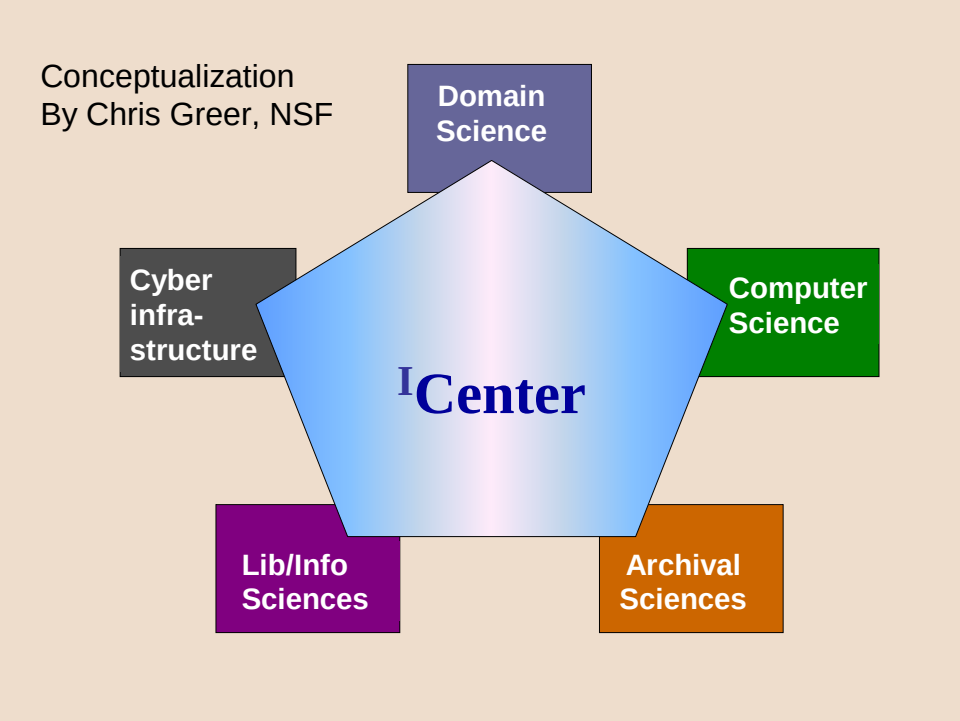


Figure 3: The I-Center, Dr. Christopher Greer, Office of Cyberinfrastructure, National Science Foundation [Greer 2006].

Several of the points of the pentagram are logical: domain science, computer science, and cyber infrastructure, it would be doubtful that anyone would disagree with these as necessary or critical components. However, how do library/information science and archival science figure in, and why are they separate?

What is library science? The typical definition describes library science as the practice of administering a library, however, there are more descriptive definitions that define library science for example: *the systematic study and analysis of the sources, development, collection, organization, dissemination, evaluation, use, and management of information in all its forms, including the channels (formal and informal) and technology used in its communication* [Reitz 2007]. This definition allows a conceptualization of the theory and principles behind librarianship that the more prosaic definition does not. Therefore, using the definition above it is apparent that the library as the I-Center is consistent with the principles and theory of library science. Using the principles of library science to evaluate the sources (domain scientists), evaluate the collection and create an organization structure (metadata/database creation),

develop dissemination avenues using taxonomies/ontologies (subject classification) and management (store and preserve). Librarians are logical as collaborators in this aspect of data curation and management. However, there are two necessary aspects of the curation of data that library science is not prepared to deal with particularly well, and in fact, fly in the face of the principles of librarianship (information for all, any time and any place). This is where archival science comes into play.

As defined by the Society of American Archivists, *archival science a systematic body of theory that supports the practice of appraising, acquiring, authenticating, preserving, and providing access to recorded materials* [Pearce-Moses 2007]. Anyone who works in archives becomes familiar with the "gift terms" of a donor. Often a donor of personal papers will restrict for a period of time who can be provided access to the papers, and the terms are agreed upon and followed. Confidential material can be withheld upon the request of the donor for a specified time, or access could be restricted to a particular group. Additionally, archivists are trained to review massive amounts of paper files, and evaluate and retain "examples" rather than the entire collection for the sake of space and timely access. Although archivists rarely de-accession materials there are times when collections may be weeded or lent to another institution for "contributory collection building." However, the major difference between a librarian's responsibility and that of an archivist is that most often the archivist is handling one of a kind original materials. Archives could be defined as a repository of raw data, historically in print, that until a researcher (most often a humanist or social scientist) comes to "mine" is nothing but bits of raw data, not that dissimilar from massive digital datasets that the scientist will mine.

The five points on the I-Center Pentagram come together this way: domain scientists develop a problem, collect data into a dataset to observe, experiment, and develop a conclusion; the dataset is housed and accessible through software written by a computer scientist; the analysis is undertaken and the results shared through the technology infrastructure; general database access and taxonomic/ontological structure is developed by library scientists through use of metadata and classification designation; access control and review is managed through archival science. Through this collaboration, the investment of funding agencies and the time committed by the domain scientists, there is greater assurance that the data will be more generally available today and in the future.

Case Study: Distributed Data Curation Center (D2C2) Purdue University, USA

How can the five specialties on the I-Center Pentagram come together? Is it possible? Is it likely? Will librarians and archivists step forward and fill this void? Or, will it be undertaken by a new professional group that will emerge to meet this need while librarians/archivists continue to be more concerned about the end product, the published research?

Since 2004 the Purdue University Libraries have been undertaking research into the methods by which domain scientists create, use, and share datasets. At first it was assumed (from looking in and not being a part of it) that dataset creation and management was for the most part the same from discipline to discipline. However, after about six months of extensive conversations with researchers in biology, earth and atmospheric science, astronomy, chemistry, chemical engineering, plant science, ecological sciences, it is apparent that there is no single method or process that would describe or meet the disparate needs.

For some researchers, the dataset is created outside of their department or laboratory (e.g., astronomy, atmospheric and earth sciences, human and plant genomics) where the database is a public domain entity and it is only once the researcher begins to apply their question and create algorithms to query the data do they become "proprietary." At this middle point, the dataset

becomes held by the researcher and/or laboratory, and is only shared once the research is published (and then most likely only a subset of the dataset).

Other domain scientists (e.g., chemists, chemical engineers, medicinal/pharmacological scientists) pose their problem and collect the data; therefore from the very beginning it is proprietary. These are simplistic examples and are only used to illustrate a few instances and are not to be considered exhaustive or definitive.

Once it was known at Purdue that there was this need to foster collaboration between the domain scientists, librarians/archivists, computer scientists, infrastructure technologists; it was decided that a center that would foster this collaboration was necessary. At Purdue University to foster interdisciplinary collaboration a research center is created. During 2006 plans were laid for the new center through discussion at the University with stakeholders. Before the center could be approved by the University criteria had to be met, including: a charge and mission; advisory board; administration and budget, and a projection on the likelihood of outside sponsored funding at a minimum of $1,000,000 per year, ramping up after the first year. Under the leadership of D. Scott Brandt, professor of library science and associate dean for research of the Purdue Libraries, the proposal was created, reviewed and accepted by the vice-president for research, the provost and the president of the University. The name agreed upon was the Digital Data Curation Center (D2C2) with the following charge:

> *The vision for the Distributed Data Curation Center (D2C2) is based on the premise that the framework of scholarly communication has been impacted drastically by the increasingly data intensive and highly networked nature of academic research in the 21st century. Research has the capability of creating and sharing outputs faster and earlier than in previous eras, and omnipresent sensors, high performance computing and network grids will facilitate the proliferation of data and capabilities to generate, manipulate and disseminate it at ever increasing rates. Reports of petabyte units of data and megaflop units of processor cycles have become commonplace, and the realization of such systems adds fuel to the fire. And where there is data there must be curation.*



Figure 4: Distributed Data Curation Center, Purdue University Libraries, 2007 [Brandt 2006].

> *Curation facilitates deposition and preservation of data through community-based policies and application of metadata standards and ontologies to ensure integrity of long term access. It encompasses a set of essential protocols and systems which facilitate access, dissemination and archiving of e-research. These protocols include management policies and tools that provide descriptive analysis of digital collections and objects to augment discovery, administration, use, reuse, and preservation. Research takes the form of investigating policies to guide and facilitate the work of data generators in submitting and exposing data, applying enhanced metadata schemas to describe digital preservation and access, designing systems to*

facilitate long term discovery of collections of objects, and developing middleware to address problems of interaction between systems and applications.

As curation involves both essential activities and critical systems to facilitate access, dissemination and archiving of e-research, it is useful to look at the technical issues and the framework in which we look at them. Curation can be described as protocols and tools that provide descriptive analysis of digital collections and objects to augment discovery, management, use, reuse, and preservation. These protocols can take the form of schemas to describe digital objects, systems to facilitate discovery of collections of objects, and middleware to resolve problems of interaction between systems and applications. At the same time, curation also provides policies and consultation to facilitate the work of data providers in submitting and exposing data, as well as enhance capabilities to navigate to data. Storage and handling of data are important and necessary components of local, regional and national solutions to the "data deluge" problem—it is critical to ensure that various data can reside somewhere and that it can be reached. But where storage and handling usually seek to remove the human element by automating processes, curation seeks to account for the human element in discovery and access [Brandt 2007].

Conclusion:

As D2C2 establishes itself as a player at Purdue University, nationally and internationally, it is actively engaging in discussions and preparing to respond to requests for proposals (RFP) from various funding agencies (National Institutes for Health, National Science Foundation, Department of Energy). Already D2C2 has collaborated successfully in the following projects:

Investigating Data Curation Profiles Across Multiple Research Disciplines

(Institute for Museum and Library Services, USA) D. Scott Brandt (PI), Michael Witt (co-PI), Jake Carlson (co-PI) - Libraries, Purdue University. Melissa Cragin (co-PI), P. Bryan Heidorn (co-PI), Carole L. Palmer (co-PI), Sarah Shreeves (co-PI) - University of Illinois at Urbana-Champaign, USA.

Mountain Store, Acquisition of a High-speed Petascale Storage System for Data Intensive Science

(National Science Foundation MRI) Matt Huber (PI), Earth and Atmospheric Science, Purdue University. D. Scott Brandt (co-PI), Libraries, Purdue University.

DataSpan, an Extensible Mapping Toolkit for Data Workflows and Repositories

(National Science Foundation SDCI) Tom Hacker (PI), The Cyber Center, Purdue University. Michael Witt (co-PI), Libraries, Purdue University.

iShare, New Middleware Development for Internet Sharing

(National Science Foundation SDCI) Rudolf Eigenmann (PI), Electrical and Computer Engineering, Purdue University. D. Scott Brandt (senior personnel), Libraries, Purdue University. [Brandt 2007]

Although D2C2 has blazed a trail at Purdue University for the advancement of data curation, it is only a blip on the horizon that foretells the growing need for data curation in the e-science research world. The challenges are immense and avenues of successful approaches can only be

achieved through collaboration of institutions throughout the world. Purdue's D2C2 may be one of the first, but soon it will be one among many. Tackling the massive datasets to be generated by projects such as LSST will be a challenge to be accepted and explored.

References:

Association of Research Libraries. 2006. "To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering: A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26-27, 2006. Arlington, VA., USA. http://www.arl.org/bm~doc/digdatarpt.pdf downloaded April 21, 2007.

Brandt, D. Scott. 2007. "D2C2, Distributed Date Curation Center." Purdue University Libraries, West Lafayette, IN, USA. *http://d2c2.lib.purdue.edu/d2c2about.html downloaded April 21, 2007*

Greer, Christopher. 2006. "*A* Vision for the Digital Data Universe*.*" Purdue University, West Lafayette, IN, USA, presentation, December 19, 2006.

LSST. 2007. Large Scale Survey Telescope, home page: http://www.lsst.org/About/lsst_about.shtml downloaded April 21, 2007.

Pearce-Moses, Richard. 2007. *A Glossary of Archival and Records Terminology* http://www.archivists.org/glossary/term_details.asp?DefinitionKey=1814 downloaded April 21, 2007.

Reitz, Joan M. 2007. *ODLIS — Online Dictionary for Library and Information Science*. **http://lu.com/odlis/search.cfm**

Research Councils UK. 2006. "About the UK e-Science Programme, Arrangements for the Future of the UK e-Science Programme. URL: http://www.rcuk.ac.uk/escience/default.htm downloaded April 21, 2007.

University of Edinburgh, School of Information Science. 2007. "Glossary of Library and Information Science." http://www.axiope.org/glossary.html Downloaded April 21, 2007.

US National Science Board (NSB). 2005. "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century." NSB-05-40, September 2005, http://www.nsf.gov/pubs/2005/nsb0540/ downloaded April 21, 2007.