Enabling International Access to Scientific Data-sets: creation of the Distributed Data Curation Center (D2C2)

IATUL
Stockholm

June 12, 2007

James L. Mullins, PhD
Dean of Libraries &
Professor of Library Science
Purdue University Libraries
USA

e-Science

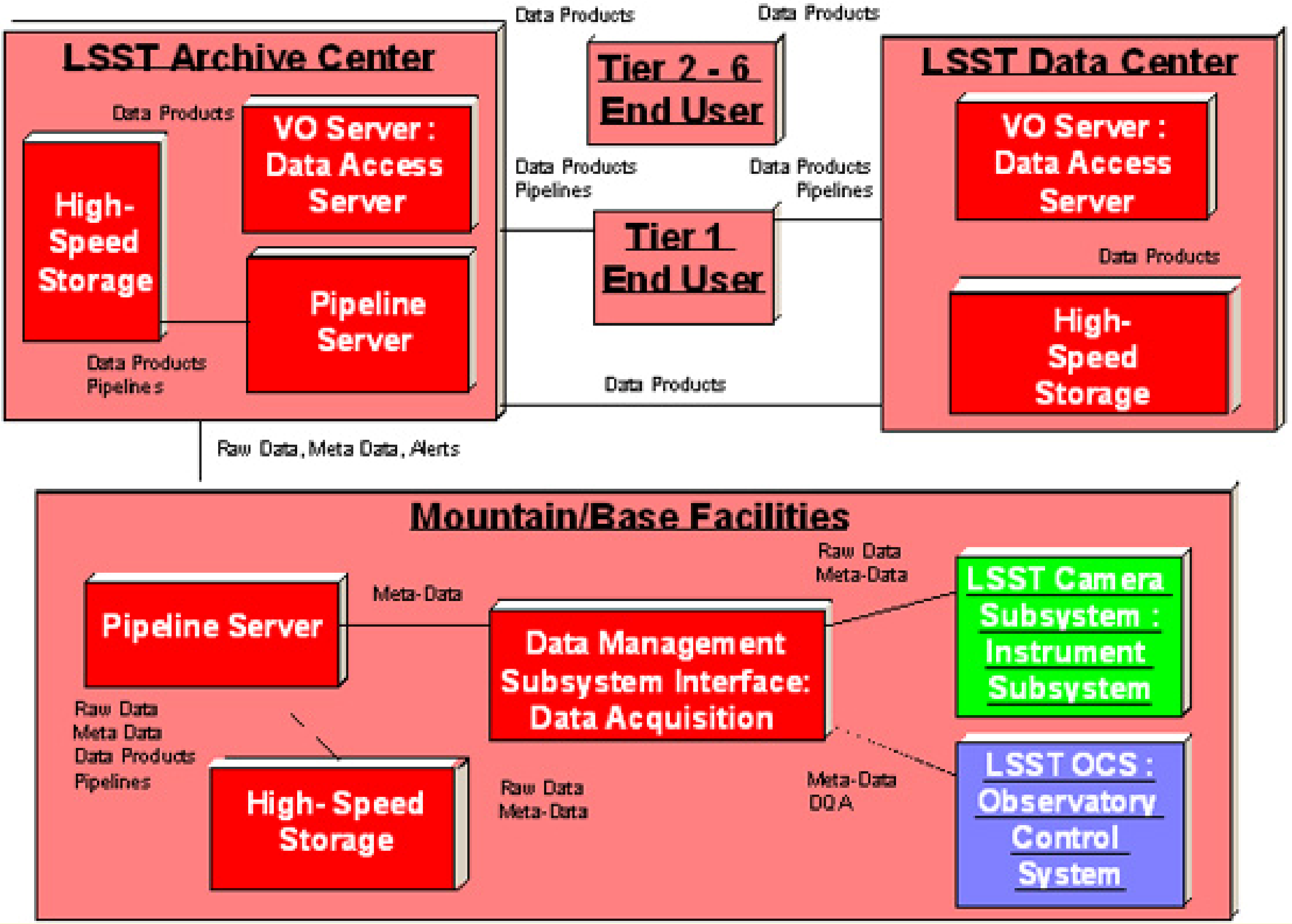
What is meant by e-Science?

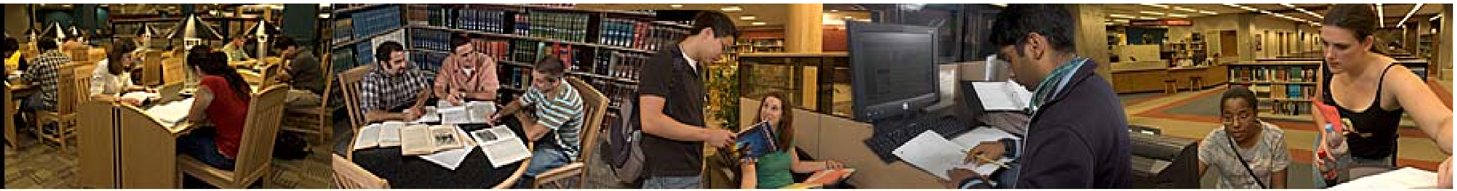
- Large scale science increasingly carried out through distributed global collaborations enabled by the Internet

- Such collaborative scientific enterprise will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists

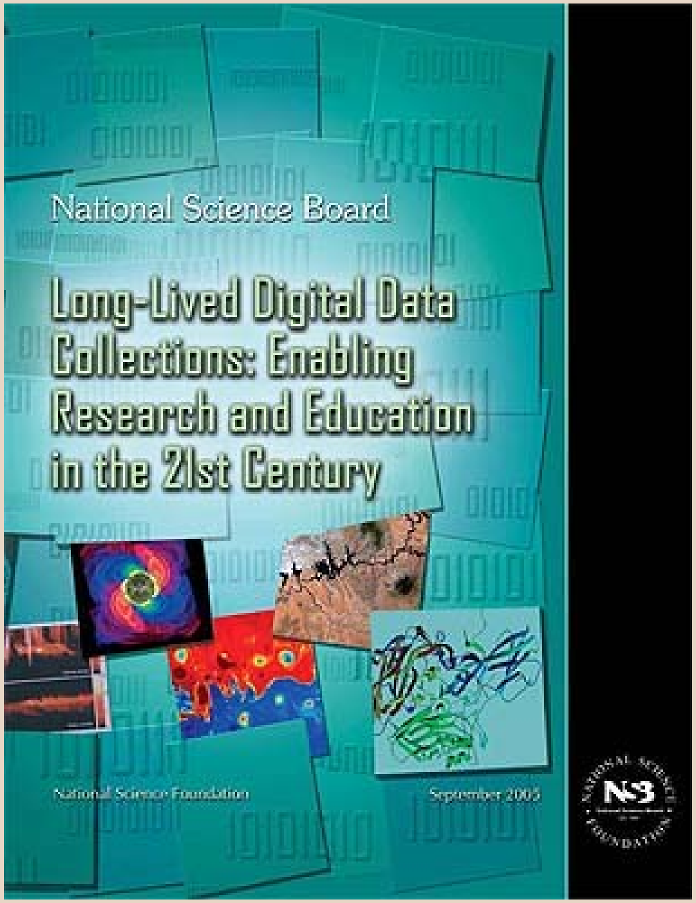
- Requiring large scale storage, retrieval and transfer

Large Synoptic Survey Telescope (LSST) – e-Science example





Innovative Research Concepts



National Science Board

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

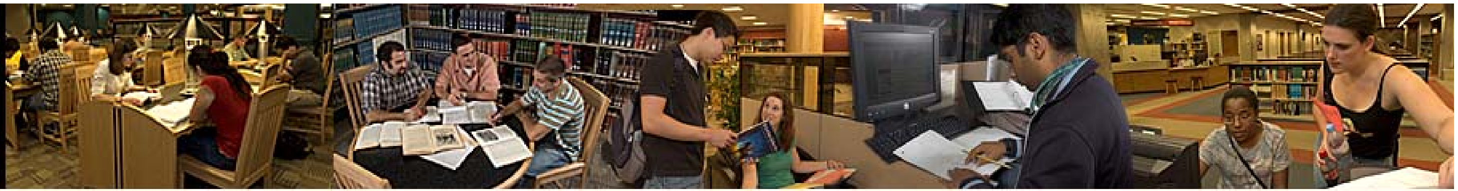
National Science Foundation

September 2005

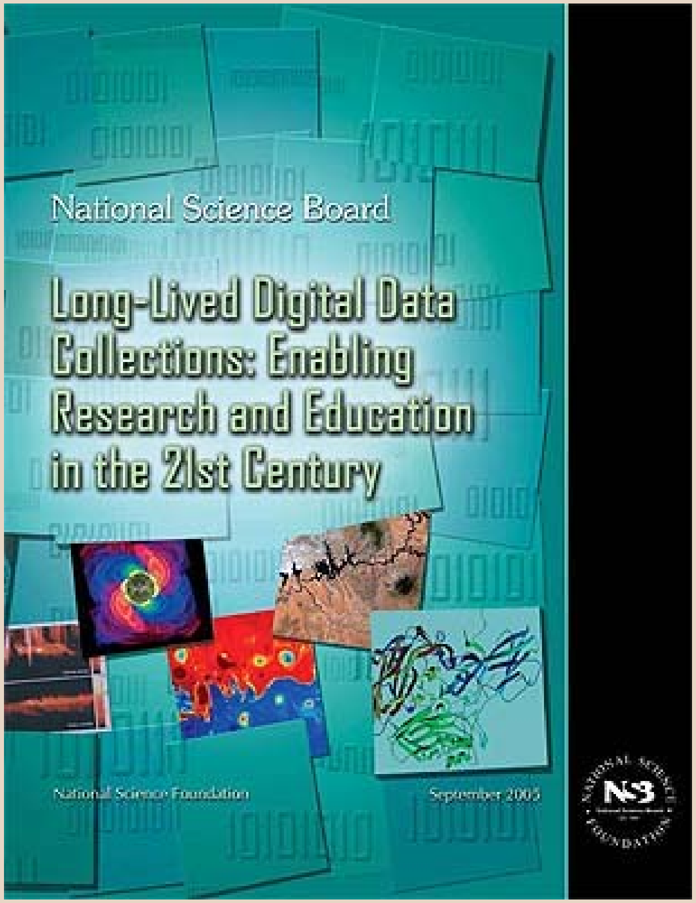
"Sound policy development and implementation rest on the recognition of the roles and responsibilities of those who play an active part in the digital data collection universe: authors; managers, and funding agencies."

National Science Board,
Long-lived digital data collections: Enabling research and education in the 21st century, p. 25.

http://www.nsf.gov/pubs/2005/nsb0540/



Innovative Research Concepts

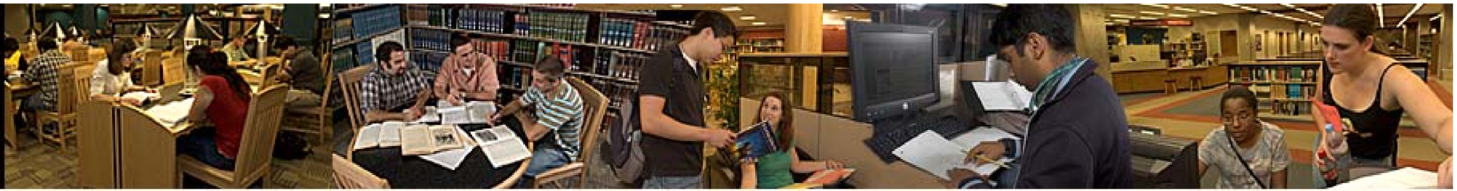

National Science Board
Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century
National Science Foundation
September 2005

- ***Data Authors – benefit from their own work, broadly disseminated, safely archived.***

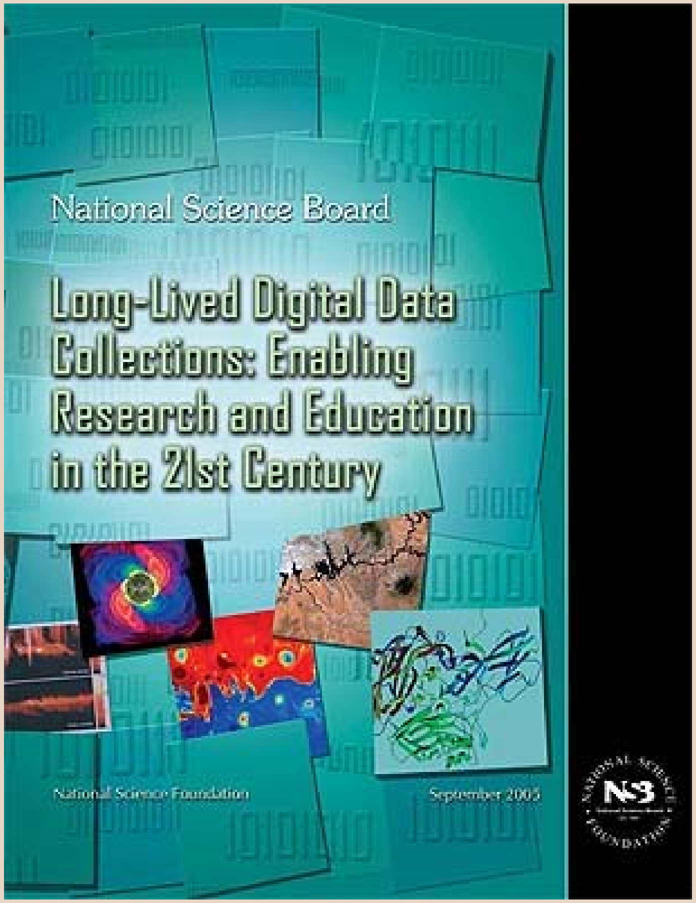
- ***Data Managers -- collaborates by insuring successful retention and dissemination through technical infrastructure***

- ***Data Scientists – conduct creative inquiry and analysis, enhance the research of data authors***

National Science Board, *Long-lived digital data collections: Enabling research and education in the 21st century*, p. 27.



Innovative Research Concepts



National Science Board

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

National Science Foundation

September 2005

Data Scientists:

… crucial to the successful management of a digital data collection – lie in having their contributions fully recognized

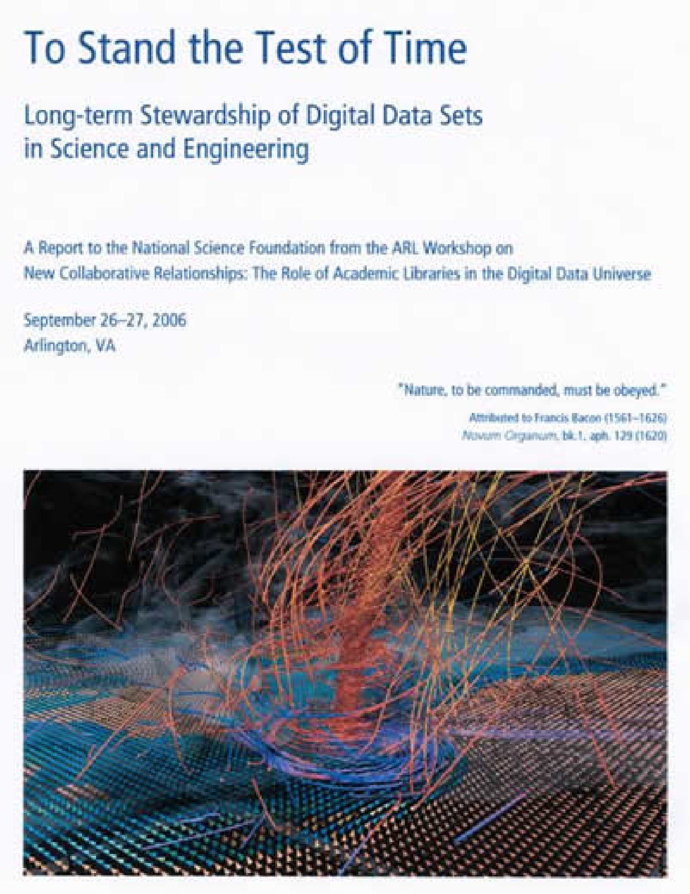
National Science Board,
Long-lived digital data collections: Enabling research and education in the 21st century, p. 27.

http://www.nsf.gov/pubs/2005/nsb0540

Curation - the Role of Librarians:

- Definition of Data Curation (librarian)
  - Store
  - Provide Access
  - Preserve
  - Carry Forward
- Definition of Data Curation (scientist)
  - Validate
  - Authenticate
  - Maintain

National Science Foundation Recognition of the Challenge for Data Curation



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

"Nature, to be commanded, must be obeyed."

Attributed to Francis Bacon (1561–1626)
Novum Organum, bk.1, aph. 129 (1620)

Dr. Christopher Greer
Program Director
Office of Cyberinfrastructure, NSF, USA

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
- Supported by NSF, September 26-27, 2006
- Attendees: NSF program directors; disciplinary researchers; information technologists; computer scientists; and librarians
- http://www.arl.org/bm~doc/digdatarpt.pdf

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering - Findings

- Ecology - digital data reflects array of stakeholders – institutions with variety of policies and practices
- Scale requires shared stewardship and partnerships
- Historically universities and their libraries have preserved the record – now expanded role and partnerships required
- Stewardship should be distributed by federated access
- Stakeholders have different expertise and need – requiring new partnerships

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering – Findings (continued)

- Stewardship of digital requires both preservation and curation
- Infrastructure is a shared common good as the data generated from federal funding is
- Requires sustainable economic models
- Linkages between data archives, publications, and associated communication – libraries
- Requires change in federal funding by agencies for data stewardship
- NSF & other funding agencies must raise awareness of issues

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering – Overarching Recommendation

- NSF should facilitate the establishment of a sustainable framework for the long-term stewardship of data. This framework should involve multiple stakeholders by supporting:
  - Research to understand, model, & prototype data stewardship
  - Training and educational programs to develop new workforce
  - Efforts to effect change in the research enterprise regarding the importance of the stewardship of digital data produced

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering – Specific Recommendations

- Fund projects that address issues concerning ingest, archiving, and reuse of data by multiple communities.
- Foster the training and development of a new workforce in data science
- Support the development of usable and useful tools, including:
  - Automated manipulation of data
  - Data registration
  - Definition of commonly used terms and concepts
  - Rights management and other access control considerations

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering – Targeted Recommendations

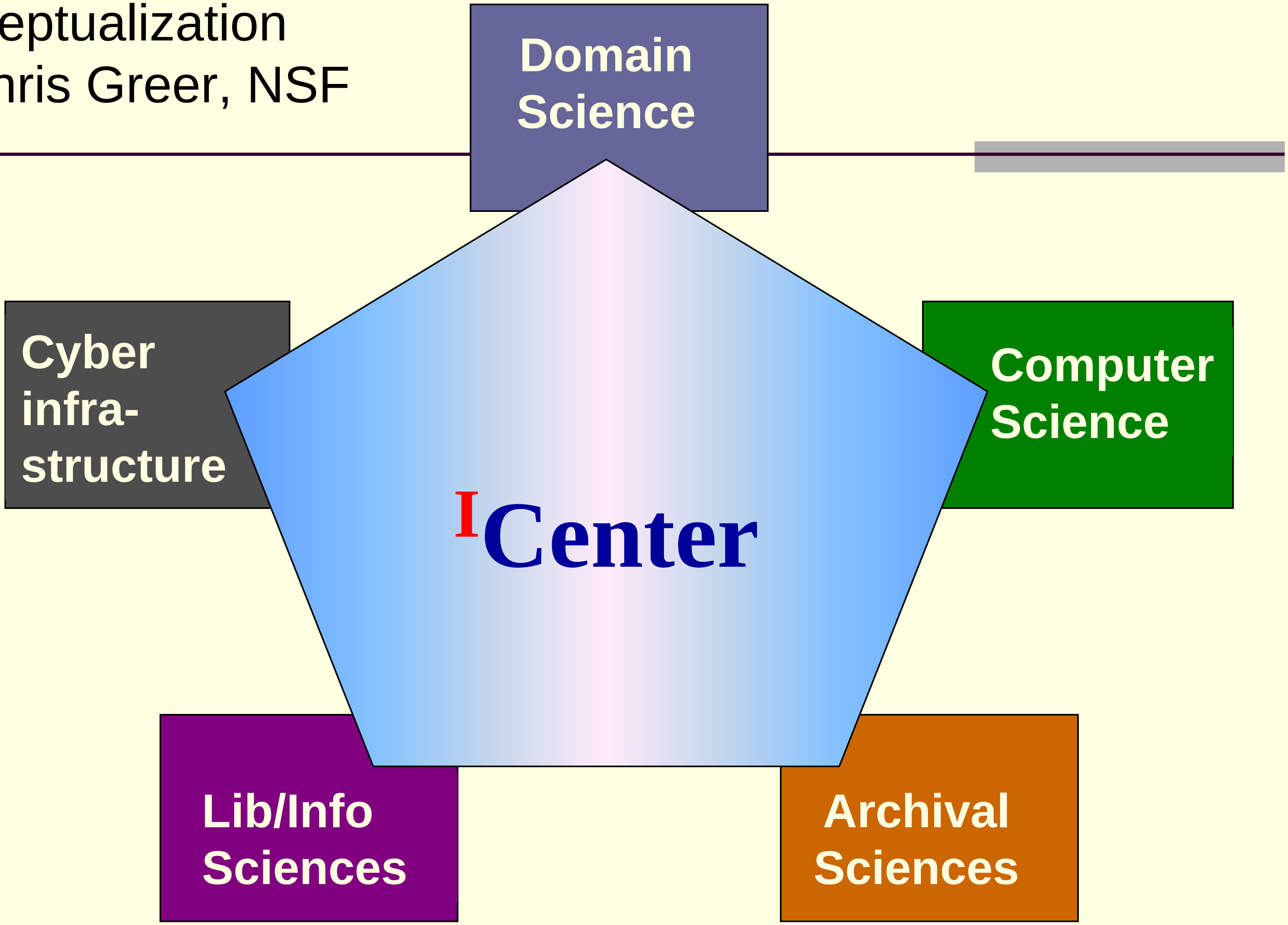
- NSF should develop a program to fund projects/case studies for digital data stewardship and preservation in science and engineering.
- NSF, along with IMLS, LIS schools, support training and education initiatives
- NSF should support development of usable and useful tools and automated services (e.g, metadata)
- Economic and Social Science experts should be employed to investigate sustainable models
- NSF should require data management plans in all NSF grant applications, and emphasis on sustainability of such plans
- NSF should encourage data sharing among communities of scholars and researchers

To Stand the Test of Time: Long Term Stewardship of Digital Data Sets ion Science and Engineering – Specific Recommendations

How can we respond?

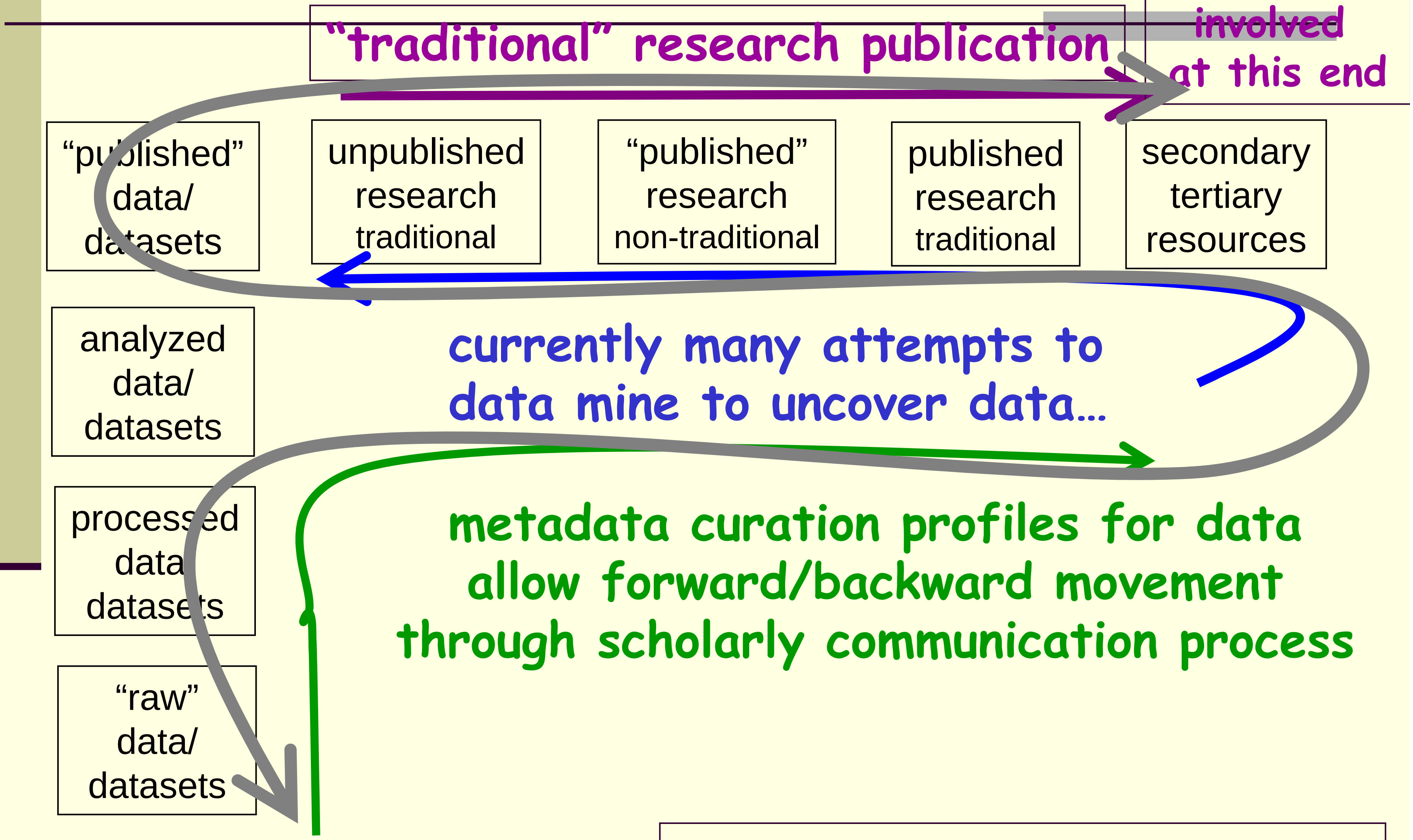
How can we prepare?

Conceptualization
By Chris Greer, NSF



Scholarly Communication

in the past, libraries involved at this end



"traditional" research publication

"published" data/ datasets

unpublished research traditional

"published" research non-traditional

published research traditional

secondary tertiary resources

analyzed data/ datasets

processed data datasets

"raw" data/ datasets

currently many attempts to data mine to uncover data…

metadata curation profiles for data allow forward/backward movement through scholarly communication process

Source: D. Scott Brandt, Purdue University

Case Study:

Distributed Data Curation Center (D2C2) -

Purdue University

Purdue University

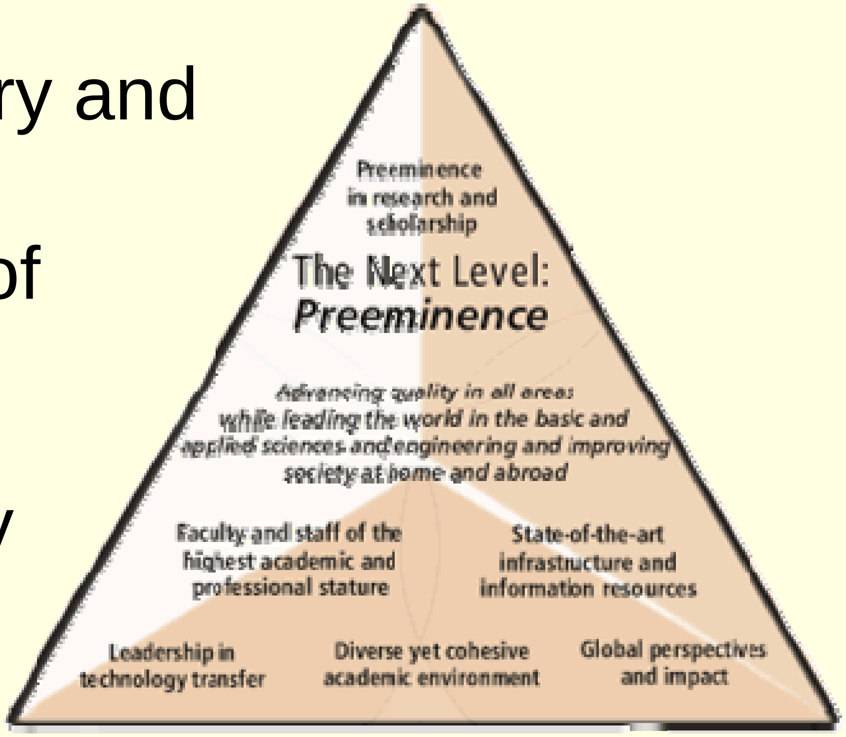
- Founded 1869 by gift from John Purdue
- 39,228 students – 31,290 undergrads; 7,938 graduate and professional.
- Premier programs: engineering (astronautics: alumnus Neil Armstrong); agriculture; hospitality and tourism; business; computer science; communications.
- Third largest international student enrollment in U.S. – 4,824 for 2006/07 (over 2,000 from India, China and Korea combined).

Purdue University

- **Nine** Colleges: Agriculture, Consumer & Family Sciences, Education, Engineering, Liberal Arts, Management, Pharmacy/ Nursing/Health Sciences, Technology, Vet Medicine
- **73** Departments, several cross-disciplinary: e.g. Agricultural & Biological Engineering

Strategic directions

- **University**: “interdisciplinary and collaborative endeavors grounded in the strengths of academic disciplines”
- **Libraries**: “Libraries faculty are better integrated into campus research agenda”


Preeminence in research and scholarship
The Next Level: Preeminence
Advancing quality in all areas while leading the world in the basic and applied sciences and engineering and improving society at home and abroad
Faculty and staff of the highest academic and professional stature
State-of-the-art infrastructure and information resources
Leadership in technology transfer
Diverse yet cohesive academic environment
Global perspectives and impact

Interdisciplinary collaboration

Cyber Infrastructure

Energy

Oncology

Entrepreneurship

Manufacturing

e-Enterprise

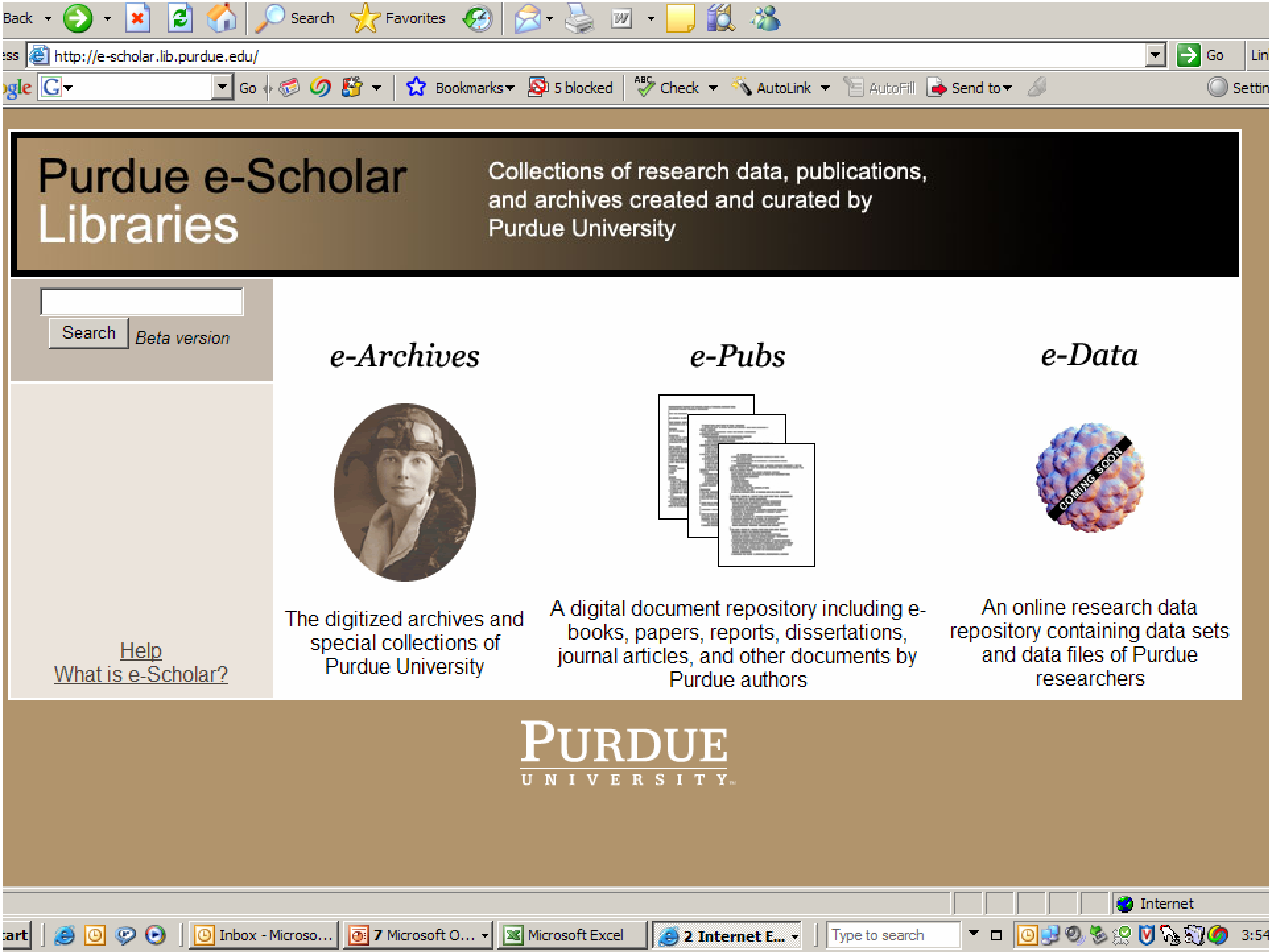
Nanotechnology

Environment

Bioscience

Learning Center

Discovery Park: Ten recent interdisciplinary centers which are designed to facilitate and promote leading edge research



Purdue e-Scholar
Libraries

Collections of research data, publications, and archives created and curated by Purdue University

Search *Beta version*

e-Archives

The digitized archives and special collections of Purdue University

e-Pubs

A digital document repository including e-books, papers, reports, dissertations, journal articles, and other documents by Purdue authors

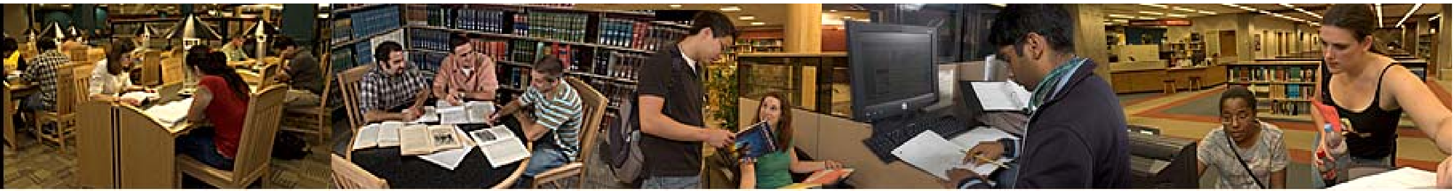
e-Data

COMING SOON

An online research data repository containing data sets and data files of Purdue researchers

Help
What is e-Scholar?

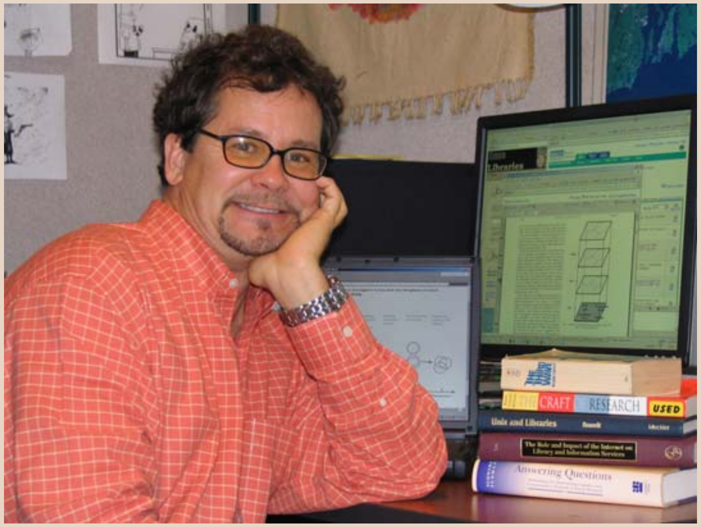
PURDUE
UNIVERSITY



Envisioning New Interdisciplinary Collaborations

Associate Dean for Research,
D. Scott Brandt,
Professor of Library Science

Facilitates individual and interdisciplinary research efforts of the fifty-two Libraries faculty



Purdue University Libraries

Since 2004, initiative for Libraries faculty to collaborate with other faculty across campus—apply library science knowledge and expertise to research problems:

collect, organize, describe, curate, archive, disseminate data/information

Determine need for collaboration

- Hypothesized that researchers have data management needs and that librarians can help meet them
- Employed top-down and bottom-up investigation for data collection
- Verified: PU researchers said they need help in collecting, organizing and providing access to their data

Outside of the library

- Attended research seminars, call-outs, etc., to identify collaboration and funding opportunities
- Built relationships - found researchers who understood that collecting, organizing and providing access to data and information are not only important, but critical, and thus, librarians need to be involved
- Found problems to solve, then collaborated on solutions
- Talked about what we know—organizing data and information (different meanings to different groups)
- Brought something to the table. Had to be prepared to demonstrate something tangible (initially a proof-of-concept or a prototype).

Current areas of collaboration

- Agronomy
- Biology
- Cancer Center
- Center for the Environment
- Chemical Engineering
- Chemistry
- Civil Engineering
- Cyber Center
- Discovery Learning Center
- Earth & Atmospheric Sciences
- English
- IT at Purdue
- Mechanical Engineering Technology
- Regenstrief Center
- Graduate School
- Oncological Sciences

Motivation (participants)

- Directly related to work, and makes something difficult easier
- It’s an extension of “our everyday job”
- Something new and exciting to do
- Breaking new ground, want to contribute to interdisciplinary initiative
- Force the issue of how it gets done (i.e., more people added to help out)

Motivation (non-participants)

- Articulation of what is expected by the Dean
- Partly determined on a case-by-case basis
- Has to be “interesting to me”
- Something that uses “the skills I can bring to it”
- Need to get credit for it (recognition, reward)
- Important to allow individual to define what interdisciplinary research is
- Should be opportunities to "stick your toe in the water" before making big commitment
- Need time to do it, and to do the “things I want to do”

Within the library

- Identify research agendas (“hot spots”) and prioritize.
- Apply library science to interdisciplinary problems.
- Recognize a percentage of time that is dedicated to research.
- Carry out research interest discussions, talking to individuals about their research.
- Discuss published research (brown bag sessions)—identify journal articles which fit with current or future research endeavors.
- Consult on research grant applications, both brainstorming and draft proposals.
- Send out research updates of progress and successes.

Distributed Data Curation Center – D2C2

- Sustainability for data curation repositories Ontological and taxonomic organization of disciplinary datasets
- Metadata to facilitate access to data collections
- Data curation profiles for archiving and preserving datasets http://d2c2.lib.purdue.edu/

Distributed Data Curation Center – D2C2

D2C2

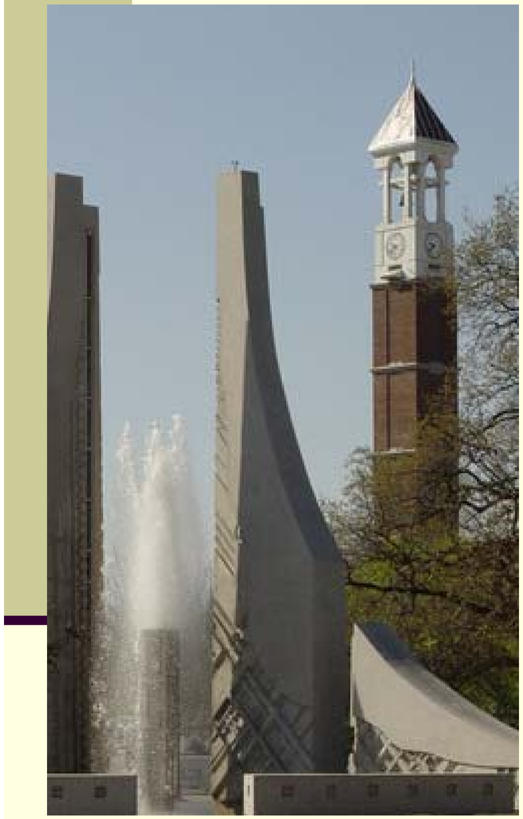
Distributed Data Curation Center

Necessary to create a Logical Structure to foster Collaboration with Research Faculty

Distributed Data Curation Center – D2C2

- Reviewed and Approved by University Process for Authorized Centers
- Charged within First Year to Generate $1,000,000 in New Revenue
- Advisory Board
  - Deans of Agriculture, Engineering, Libraries, Science, and Technology
  - Vice President for Information Technology
  - Director, Cyber Center
  - Acting Director, associate dean for research, Libraries

Recap…



- 11 Libraries faculty involved in 9 grants since April of last year
- New positions: data research scientists to support research
- Researchers are starting to come to Libraries for collaborations
- "*100 conversations, lead to 20 discussions, lead to 5 grants, lead to 1 award*…"

Thank you!

Questions and Answers?

James L. Mullins – Purdue University, USA

jmullins@purdue.edu