Microsoft® Research
Faculty Summit 2010

Opportunities for Libraries: Bringing Data to the fore in Scholarly Communication and potential implications for promotion and tenure.

James L. Mullins, PhD
Dean of Libraries and Professor
Purdue University

Data

- Prolific growth – large and small science
- Lifeblood – scientific and engineering research
- Modeling – demands massive amounts of data
- Funding Agency Expectation – public accessibility

National Science Foundation
WHERE DISCOVERIES BEGIN

Press Release 10-077

Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans

Government-wide emphasis on community access to data supports substantive push toward more open sharing of research data

May 10, 2010

During the May 5th meeting of the National Science Board, National Science Foundation (NSF) officials announced a change in the implementation of the existing policy on sharing research data. In particular, on or around October, 2010, NSF is planning to require that all proposals include a data management plan in the form of a two-page supplementary document. The research community will be informed of the specifics of the anticipated changes and the agency's expectations for the data management plans. The changes are designed to address trends and needs in the modern era of data-driven science.

http://www.nsf.gov/index.jsp

Data - Problem Statement

How could a dataset be identified to allow discovery, access/retrieval, provenance, citation, and accreditation/impact?

Data Challenges (and not)

- Discovery and Retrieval
- Authentication/Curation
- Attribution
- Provenance/Citation (impact of dataset to research)
- Archiving (and not)
- Storage – not a major technical/funding challenge (for the most part)

Data - The Problem

"I like to think of data in three categories, using a mining metaphor: 'raw ore', 'concentrate', and 'virgin metal.' The question is which data are worth saving and which throwing away?"

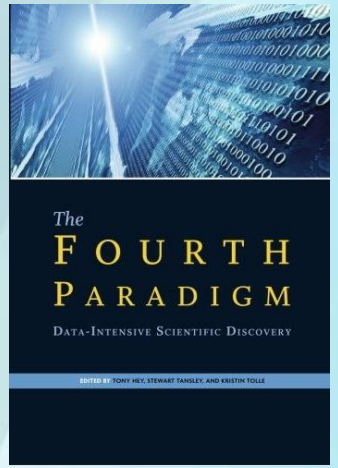
Arden Bement, former director of National Science Foundation.

Arden Bement, *Remarks before IATUL*, June 21st, 2010, Purdue University, West Lafayette, Indiana. http://docs.lib.purdue.edu/iatul2010/conf/day1/7/

Data Management - Recent Research & Publications

- ***Accessibility, and Stewardship of Research Data in the Digital Age***, **summarizes the data management challenges facing the scientific research community. The report was issued by the Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, National Academy of Sciences, Fall, 2009. http://www.nap.edu/catalog.php?record_id=12615**

- ***The Fourth Paradigm: Data-Intensive Scientific Discovery*.**
- **Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, Inc., 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm/**



- ***Keeping Research Data Safe 2: a cost model and guidance for UK Universities.*** **Neil Beagrie, Brian Lavoie, and Matthew Woollard. Final Report, - April 2010. Prepared by Charles Beagrie Limited. Studied funded by JISC. Copyright, 2010. http://www.beagrie.com/KRDS2_selectioncriteria.pdf**

The Digital Object Identifier DOI®) System identifies content objects in the digital environment.

- CrossRef assigns a DOI to articles submitted by publishers

- An efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article

Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004): *Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.* Marine Geology, 207(1-4), 209-224, doi:10.1016/j.margeo.2004.03.017

http://www.datacite.org/

Crossref Indicators (July 01, 2010)

Total no. participating publishers & societies	3108
% of non-profit publishers	57%
Total no. participating libraries	1,597
No. journals covered	22,282
No. DOIs registered to date	41,913,980
No. DOIs deposited in previous month	375,951
No. DOIs retrieved (matched references) in previous month	15,247,161
DOI resolutions (end-user clicks) in previous month	n/a

http://www.crossref.org/

So What about Data?

Data: Sharing

"Asking to see a researcher's data is like asking to see their underwear!"

Dr. Sylvia Brouder, Professor of Agronomy, Purdue University, at the IATUL Conference, June 21, 2010, Purdue University.

Submit

Data

DataCite
A global registration agency for research data.

http://www.datacite.org/

Data

Our [DataCite] long term vision is to support researchers by providing methods for them to locate, identify, and cite research datasets with confidence.

http://www.datacite.org/

Data: Attribution

- Descriptive Metadata - author (person or corporate), research variables, etc.
- Subject descriptors - disciplinary taxonomy
- Digital Object Identifier - persistent identifier

The dataset with DOI: Kuhlmann, H et al. (2009): *Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.* *doi:10.1594/PANGAEA.727522,*

http://www.datacite.org/

Linking of Dataset to Article

The DOI system offers an easy way to connect the article with the underlying data:

The dataset:

Kuhlmann, H et al. (2009): *Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.*

doi:10.1594/PANGAEA.727522,

Is supplement to the article:

Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004): *Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.* Marine Geology, 207(1-4), 209-224,

doi:10.1016/j.margeo.2004.03.017

http://www.datacite.org/

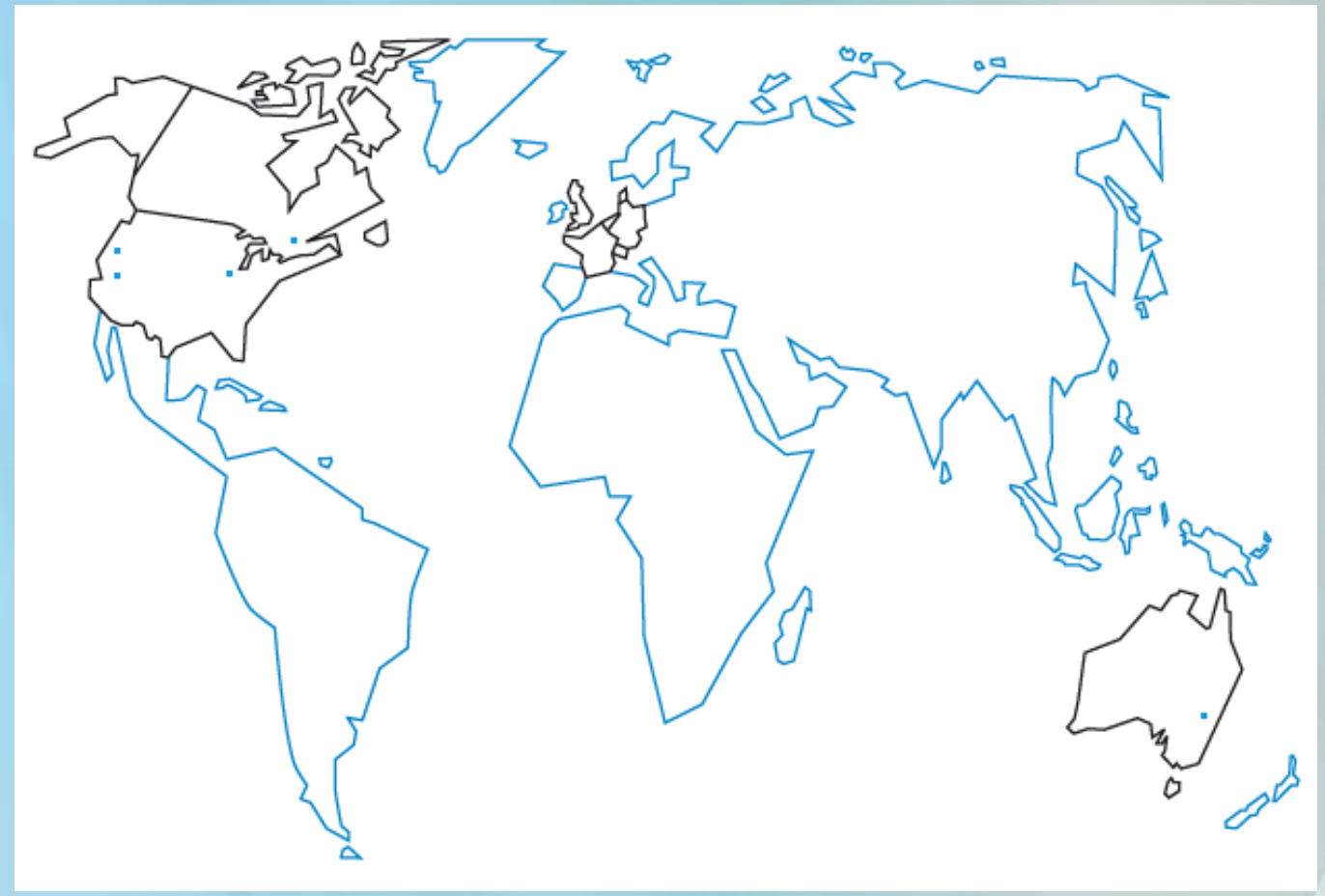
The Digital Object Identifier (DOI®) System is for identifying content objects in the digital environment.

The use of DOI names for the citing of data sets makes their provenance trackable and citable and therefore allows interoperability with existing reference services.

Over 800,000 datasets assigned DOIs!

http://www.datacite.org/

Data - Problem Statement



http://www.datacite.org/

Members

AUS

Australian National Data Service (ANDS)

CAN

Canada Institute for Scientific and Technical Information

CH

Library of the ETH Zurich

DK

Technical Information Center of Denmark

FR

Institute for Scientific and Technical Information

GER

German National Library of Science and Technology (TIB)

German National Library of Medicine (ZB MED)

GESIS - Leibniz Institute for Social Science

NL

TU Delft Library

UK

The British Library

USA

California Digital Library (CDL)

Purdue University Libraries

Associated Members

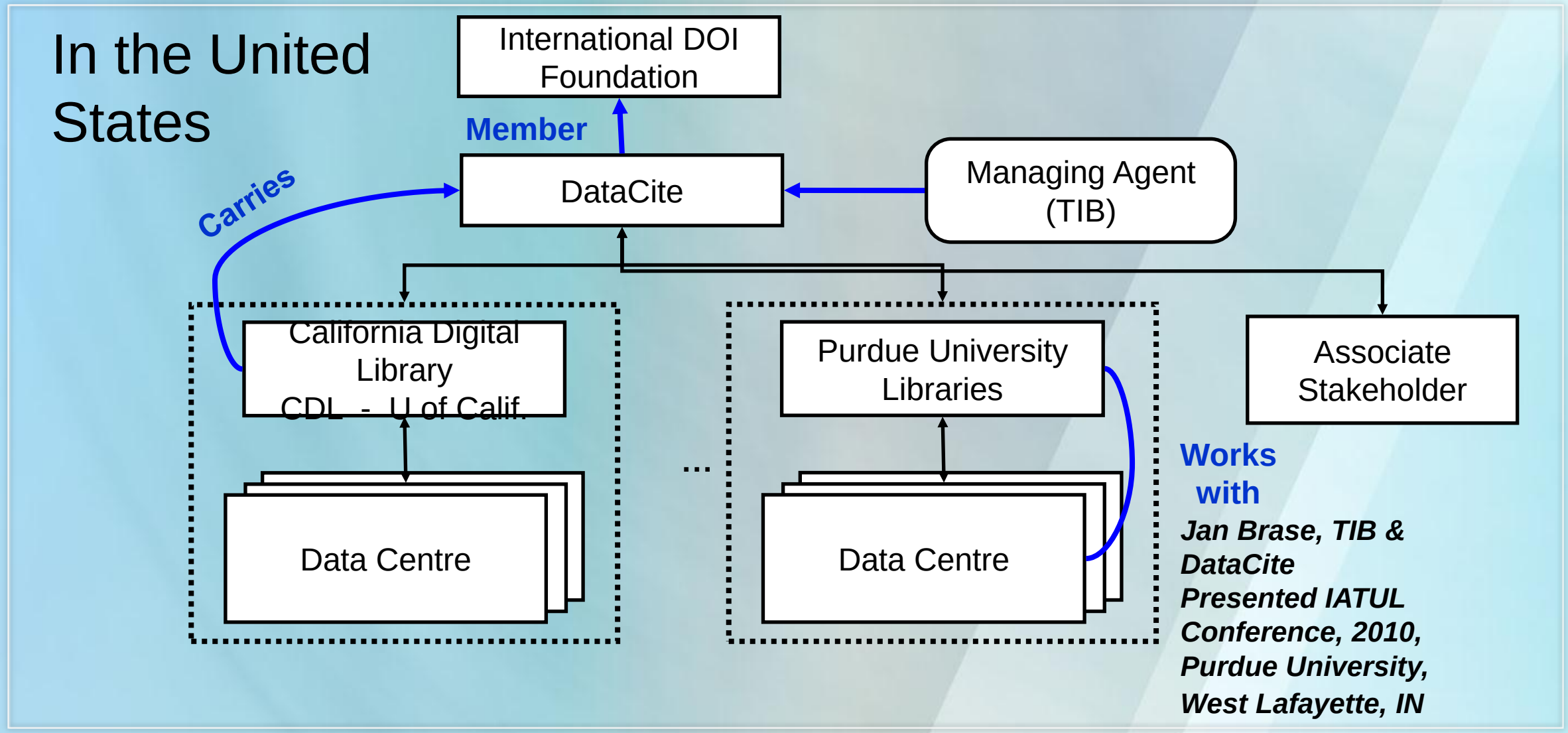
UK

Digital Curation Centre

USA

Microsoft Research

DataCite Structure



Data: Acceptance, Attribution, Citation, and Impact

- Acceptance by research community of the importance of data in research
- Assignment of authorship/creation – metadata & DOI
- Citation by research undertaken and reported through DataCite
- Establishment of h-index for impact

Dataset Creation: A Criterion for Promotion & Tenure

Once the creation of a dataset has been accepted by the investigator's disciplinary field and academic institution as a valid contribution to research, and once the h-index can be calculated for the impact of a dataset (through DataCite), the creation of an important dataset could be a criterion for promotion and tenure.

Microsoft® Research
Faculty Summit 2010

Opportunities for Libraries: Bringing Data to the fore in Scholarly Communication and potential implications for promotion and tenure.

James L. Mullins, PhD
Dean of Libraries and Professor
Purdue University

Microsoft® Research

Faculty Summit 2010