# Realizing the Potential of Research Data

Carole L. Palmer

Information School
University of Washington

Coalition for Networked Information
14 April 2015

- Are we the experts we need to be?
- What are the exemplar for data resources and services?
- Can we learn and lead at the scale and pace needed?

Fall Forum

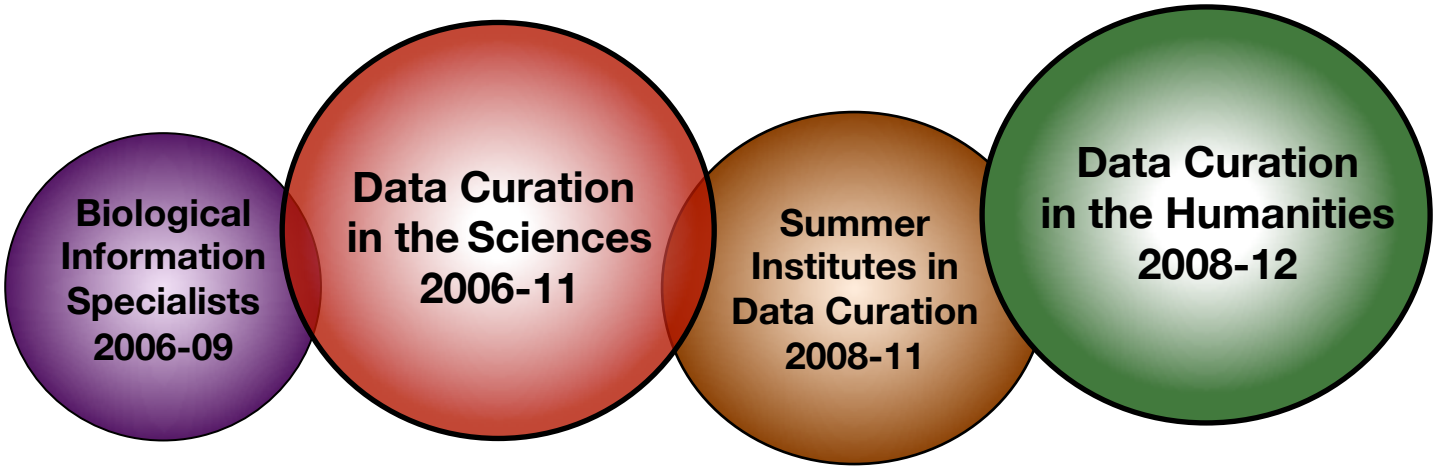
Reinventing Science Librarianship: Models for the Future

# ARL-CNI Forum 2008

October 16-17, 2008
Arlington, Virginia 22202

*Sponsored by ARL & CNI*

## Preparing e-Science Information Specialists: New Programs and Professionals




Well positioned—institutional and human infrastructure, expertise, commitment

*Going forward,* ***must not underestimate challenge.***

# Deluge of discourse and directives

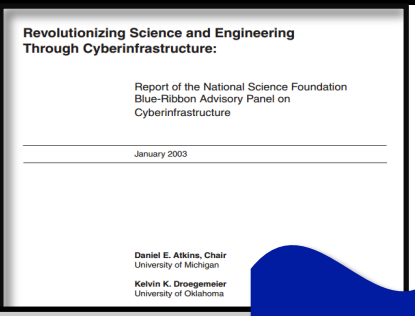
2003 at least 11 reports from NSB, NRC, NSF

Early leadership from Information Schools

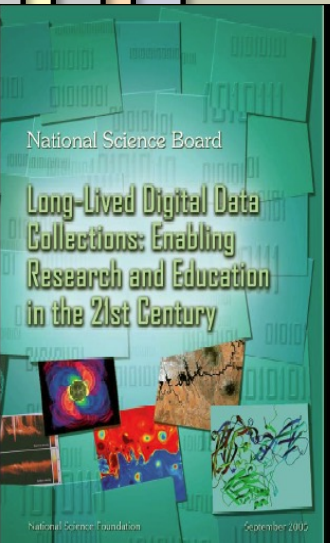
Atkins

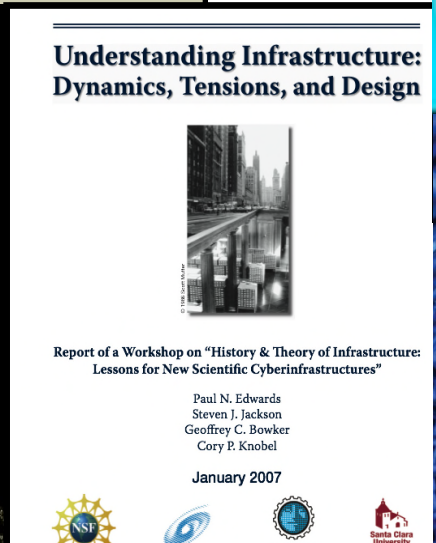
Larsen

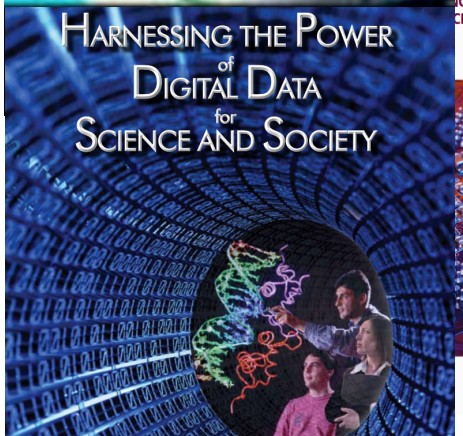
Unsworth

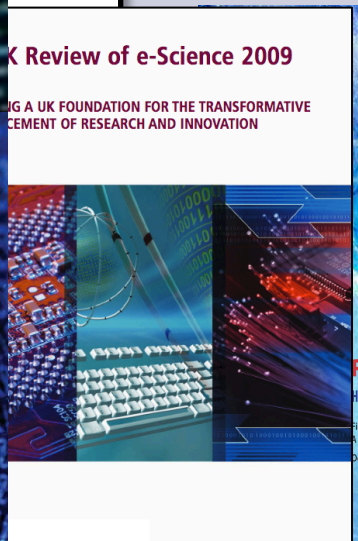





# Deluge of repositories and standards

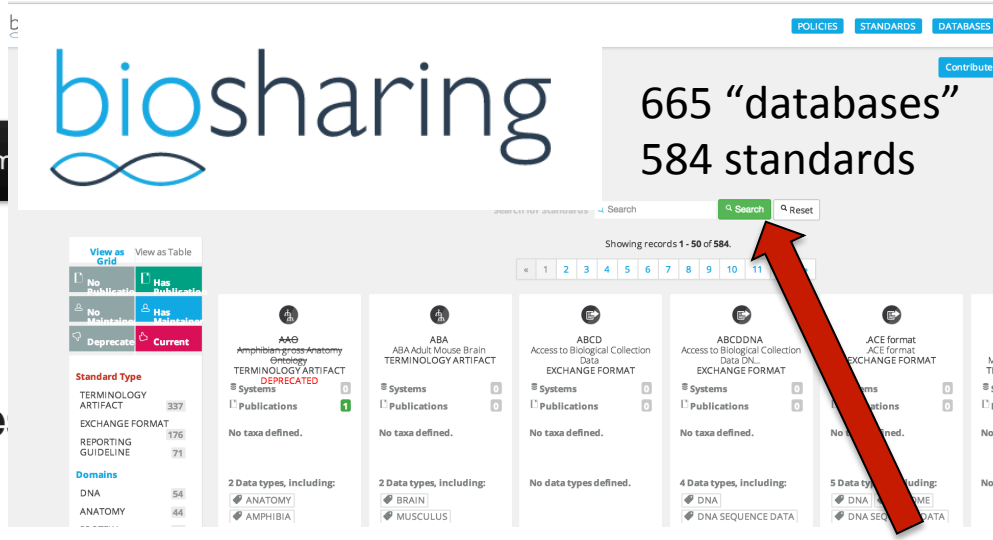
re3data.org

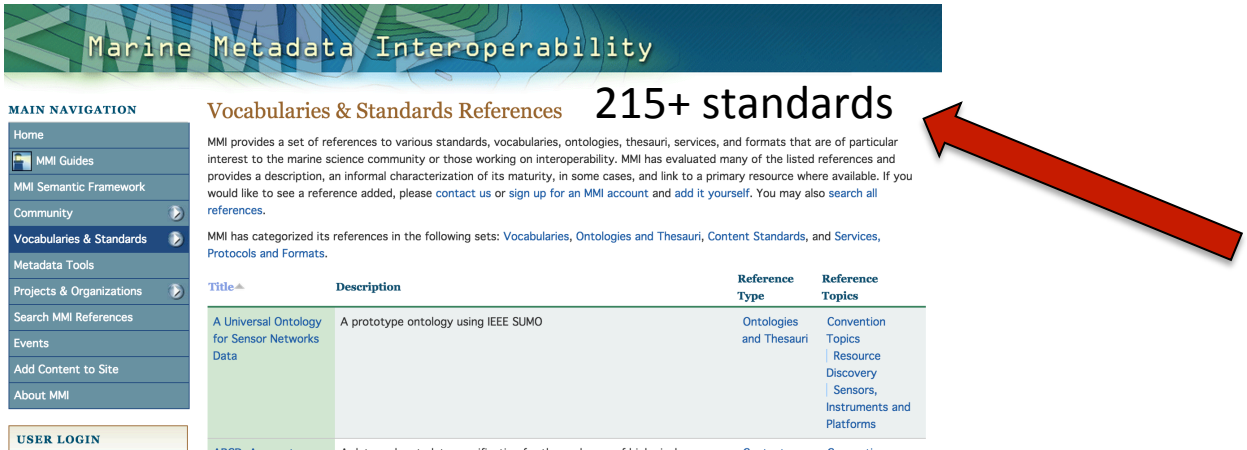
REGISTRY OF RESEARCH DATA REPOSITORIES

Home Search Browse Suggest FAQ About Schem

TAG ARCHIVES: DATABIB

**Over 1,000 research data repositorie**



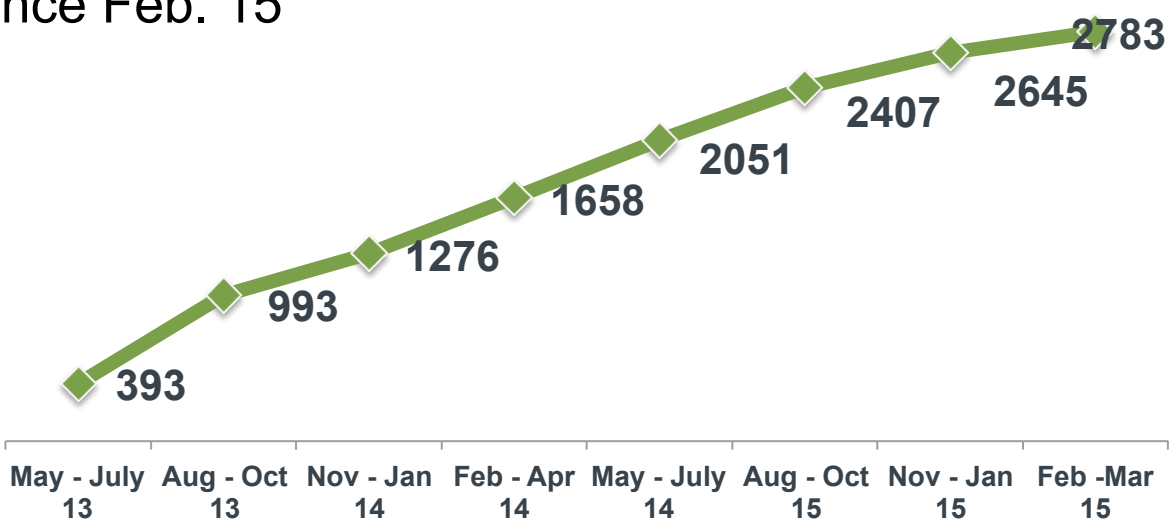
RESEARCH DATA ALLIANCE

Reduce barriers to data sharing

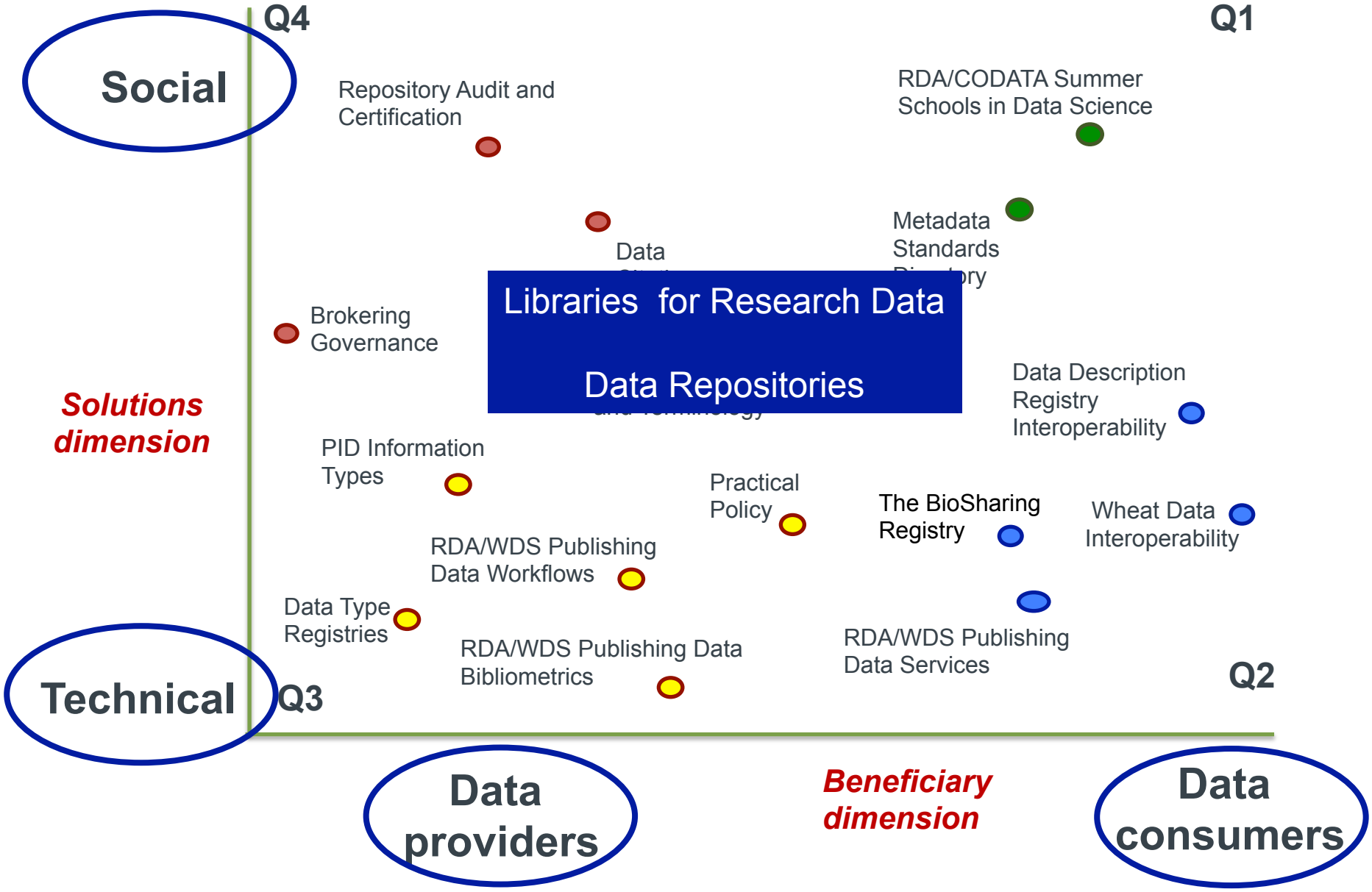
Accelerate coordinated global data infrastructure

95 countries - 50% Europe, 37% US

2783 RDA Community Members
+138 since Feb. 15



# 56 Working and Interest Groups



# National data services



The National DATA SERVICE






CSC - IT CENTER FOR SCIENCE

*Data Archiving and Networked Services*

DANS

# Abundance of data science initiatives


Stanford Data Science Initiative

UNIVERSITY OF ROCHESTER
INSTITUTE FOR DATA SCIENCE


Data Sciences Initiative
UCDAVIS
University of California, Davis

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK
Data Science Institute

UNIVERSITY of VIRGINIA
Data Science Institute


UCI Data Science Initiative

# 2014 - UW eScience Celebration of data intensive research


Data Science Kickoff Session:
137 posters from 30+ departments and units

**http://escience.washington.edu/**

# Problem: dynamics of systems and networks

from homogeneous, centralized, local

to heterogeneous, distributed, coordinated

- consolidation
- gateways for interoperation


(Edwards, et al., 2007)

"make-or-break" phase (Parsons & Berman, 2013)

Early choices constrain options

# Institutions as intellectual habitat

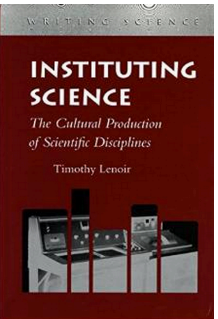
Research programs of researchers

- extend and legitimate products of work
- dominate cycles of credit and resources

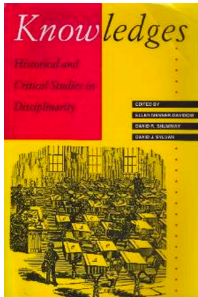
Institutions support

**new routines long enough for distinctive types of work to emerge**

- establish service roles
- facilitate links with other disciplines
- enable transmission of techniques and information




Lenoir, Timothy. 1993. "The Discipline of Nature and the Nature of Disciplines." In *Knowledges: Historical and Critical Studies in Disciplinarity*, edited by Ellen Messer-Davidow, David R. Shumway, and David J. Sylvan. Charlottesville: University Press of Virginia.



# 5-year, $37.8 million cross-institutional collaboration to create a *data science environment*

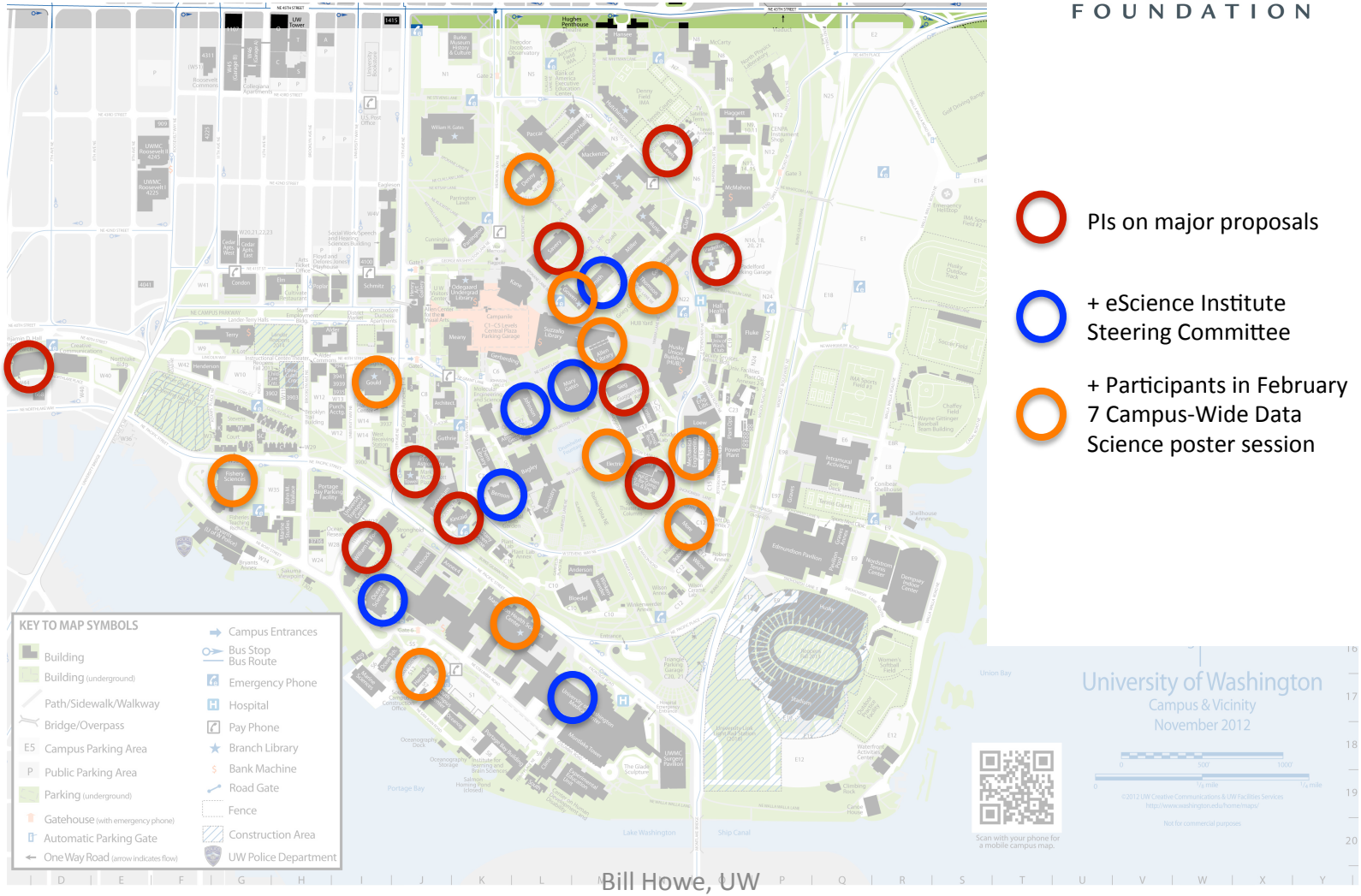
NYU

Berkeley
UNIVERSITY OF CALIFORNIA

UNIVERSITY of WASHINGTON

ALFRED P. SLOAN FOUNDATION

GORDON AND BETTY MOORE FOUNDATION




PIs on major proposals

+ eScience Institute Steering Committee

+ Participants in February 7 Campus-Wide Data Science poster session

Washington Research FOUNDATION

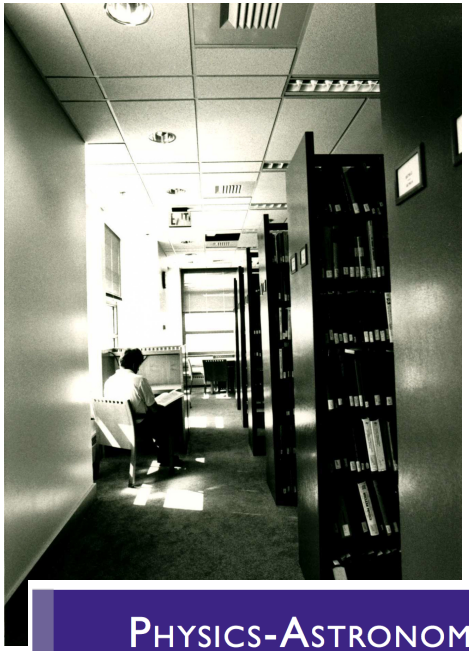
# Data Science Studio

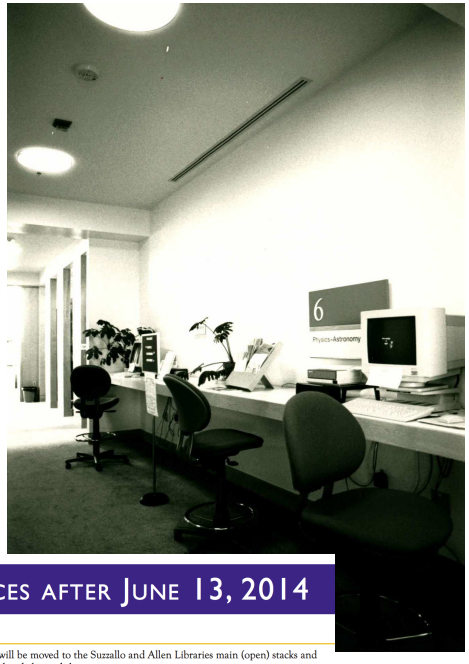
6th floor Physics Astronomy Building

Partnership among:

- Provost
- UW Libraries
- Physics, Astronomy, Arts & Sciences
- eScience Institute

# Revisioned the library
# focus on working spaces and culture




## Physics-Astronomy Reading Room Transition to Data Science Studio

### Timeline

**June 13, 2014, 5pm**

The Physics-Astronomy Reading Room (the entire 6th floor of the Physics/Astronomy Building) and book drop will close permanently.

Access will be limited to construction activities beginning at 5pm.

**beginning June 16 through Summer 2014**

Library materials will be moved out of the space and relocated to other libraries. *

Interior construction and remodeling will take place. Access for construction only.

## Libraries Resources after June 13, 2014

### Books

- The books from the Physics-Astronomy Reading Room will be moved to the Suzzallo and Allen Libraries main (open) stacks and shelved with the astronomy and physics research books already located there.
- A few dozen books will be moving to the Mathematics Research Library (Padelford Hall), Engineering Library, and Odegaard Undergraduate Library, where the content of those books will be closely associated with the subject coverage and clientele of those units.

### Journals:

- Journals available online will be moved to the Libraries Auxiliary Storage, and can be requested through the Library catalog
- Journals unavailable online will be moved to the Suzzallo and Allen Library main stacks, and will be shelved alongside related material

### Reference

- The core Reference collection, i.e., encyclopedias, dictionaries, handbooks, and directories will be moved to the Suzzallo Library first floor reference collection. This open browsing area near the elevators has room to read and study, and is conveniently located near several bookscan stations.
- The remainder of the Reference collection will move to the Suzzallo and Allen Library main stacks and will be available for checkout
- A few basic and introductory encyclopedias will be moving to Odegaard Undergraduate Library.

### Course Reserves

- Odegaard will be the default location for physics and astronomy course reserves after June 13, 2014.
- Starting Summer Quarter 2014, instructors will have the option to have their graduate course reserves at Odegaard Undergraduate Library, Built Environments Library, Health Sciences Library, Engineering Library, or Mathematics Research Library
- Additionally, instructors may choose to have their course materials in the Suzzallo Library first floor reference collection, which cannot be checked out, but would be available for browsing or scanning in the library.

### Science Fiction Browsing Collection

- The browsing collection of science fiction books from the former Chemistry Library will be moving to Odegaard Undergraduate Library and the Suzzallo and Allen Libraries.

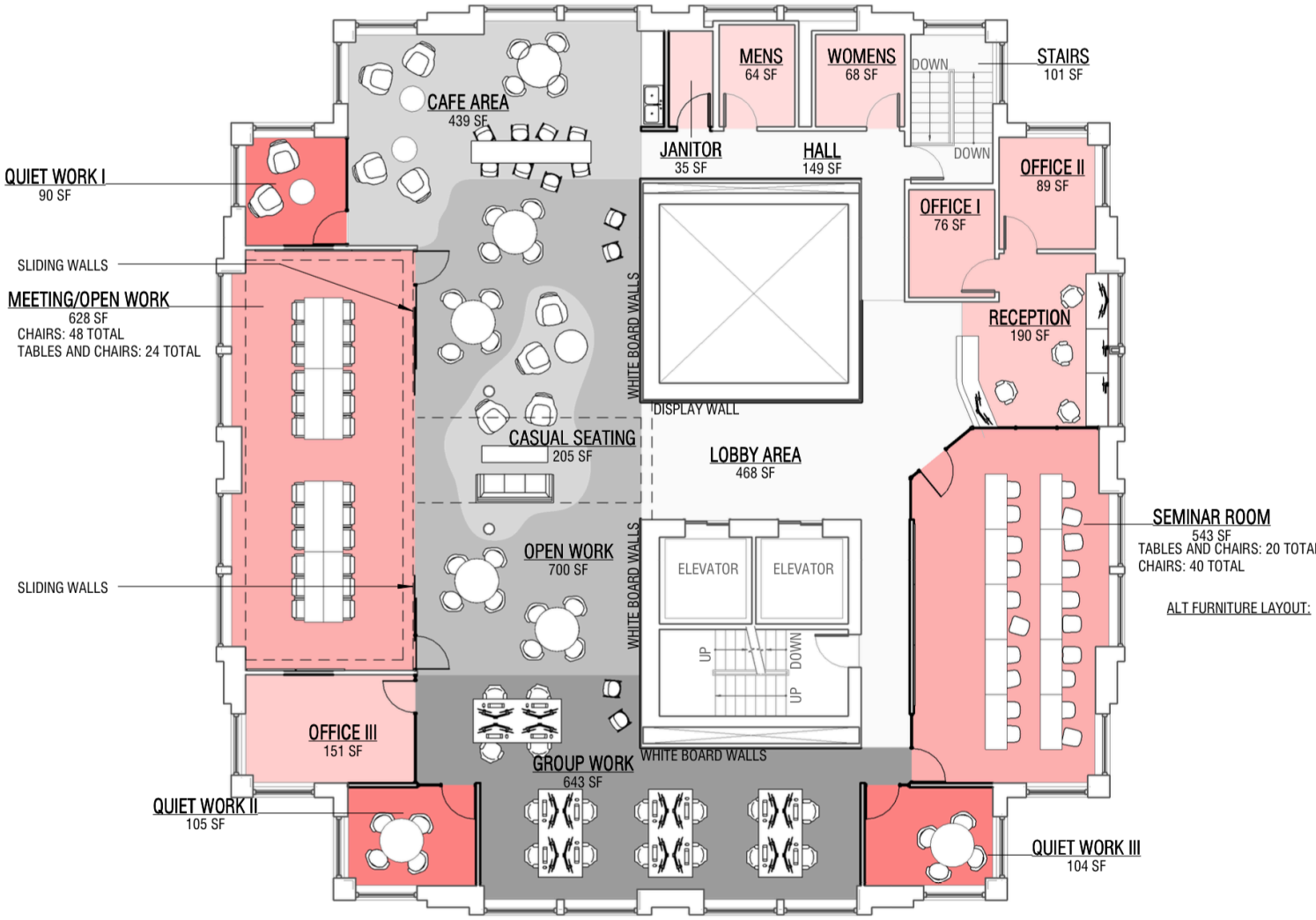
### Holds and Requests

Bill Howe, UW

**Casual & Open work**

**Seminar & Group work**

**Quiet work**

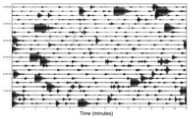



1 DATA SCIENCE STUDIO FLOOR PLAN
3/32" = 1'-0"


WRF Data Science Studio


A case for teaching
SQL to scientists
Daniel Halperin
#w2btac @SESYNC 2013-07-09

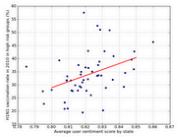
# Spring 2014 Incubation Projects



## Automated Detection and Analysis of Repeating Earthquakes

*Alicia Hotovec-Ellis, Kate Allstadt, Jon Connolly, and John Vidale — Earth and Space Sciences*

***eScience Contact: Jake Vanderplas***



## Using social media data to identify geographic clustering of anti-vaccination sentiments

*Benjamin Brooks, Abraham Flaxman — Institute for Health Metrics and Evaluation*

***eScience Contact: Andrew Whitaker***

## Analysis of Kenya's Routine Health Infor

*Gregoire Lurton, Abraham Flaxman, Emmanuela Gakidou — I*

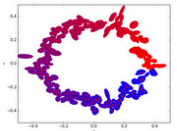
***eScience Contact: Daniel Halperin***

Project leads must physically co-locate with the incubator staff.

## Efficient Computation on Large Spatiotemporal Network Data

*Ian Kelley, Josh Blumenstock — Information School*

***eScience Contact: Andrew Whitaker***



## Scalable Manifold Learning for Large Astronomical Survey Data

*Marina Meilă — Statistics*

***eScience Contact: Jake Vanderplas***

## ASPASIA: Adult Service Providers and Some Incidental Addenda

*Sam Henly — Economics*

***eScience Contact: Andrew Whitaker***

# Resident data science team

- Permanent staff of ~5 *data scientists* – applied research and development
- Drop-in open workspace
- Studio "Office Hours"
- Incubation Program

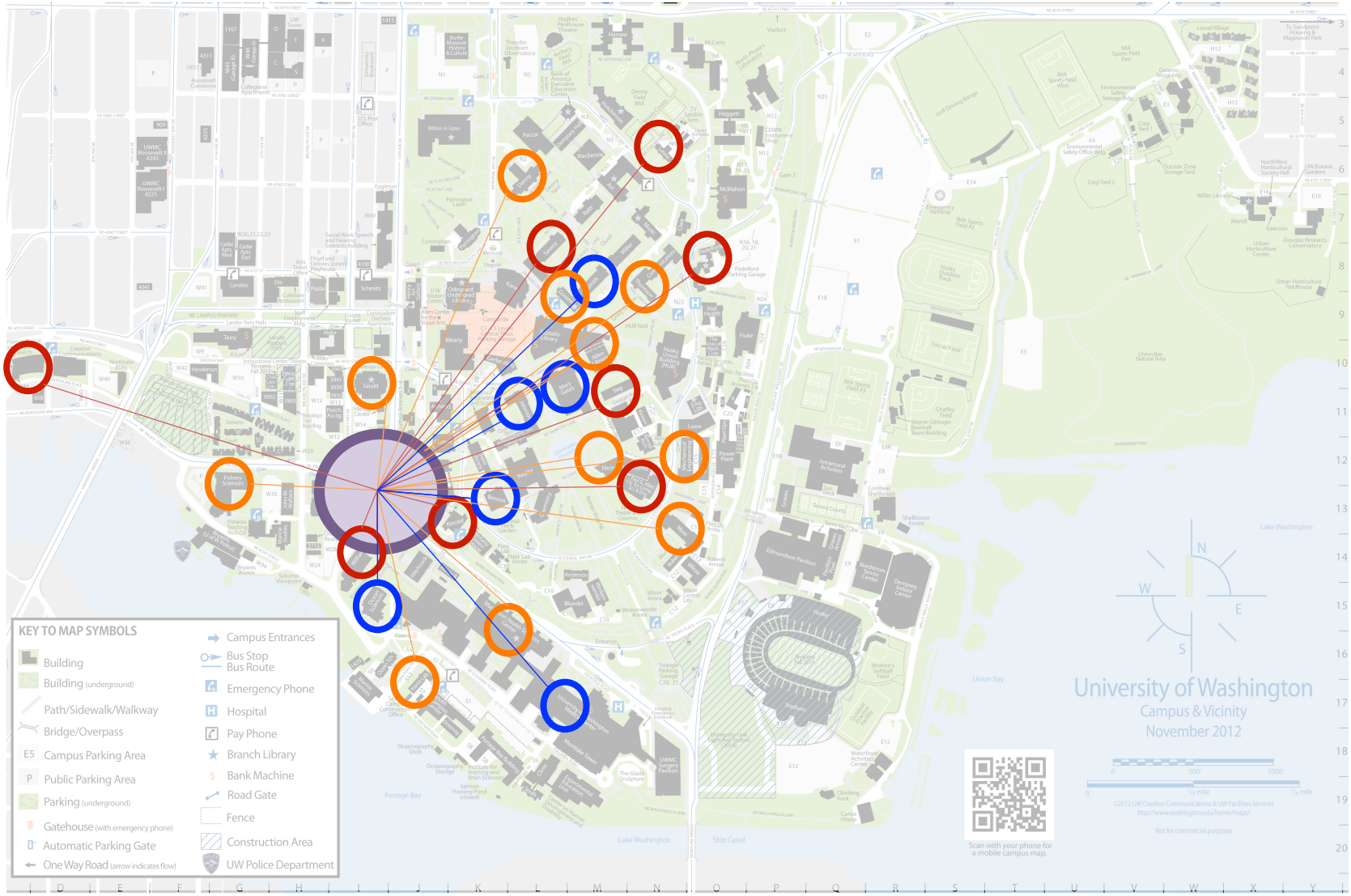
"Don't see how you do it without the library."

…plus seminars, sponsored lunches, workshops, bootcamps...

Library in the mix:

- Office hours
- Reproducibility and Open Science Group
- Site visits

# Data Science: the rising tide that lifts all boats



# Problems - eScience vs. open, curated data

*How much time do you spend "handling data"*
*as opposed to "doing science"?*

Mode answer: *90%* *(Bill Howe, 2015)*

What qualifies as releasable data?

Open data constrained by evidential cultures -

Individualism vs. Collectivism (Collins, 1998)

Who takes responsibility for validity and meaning?

# Why do we invest in data?

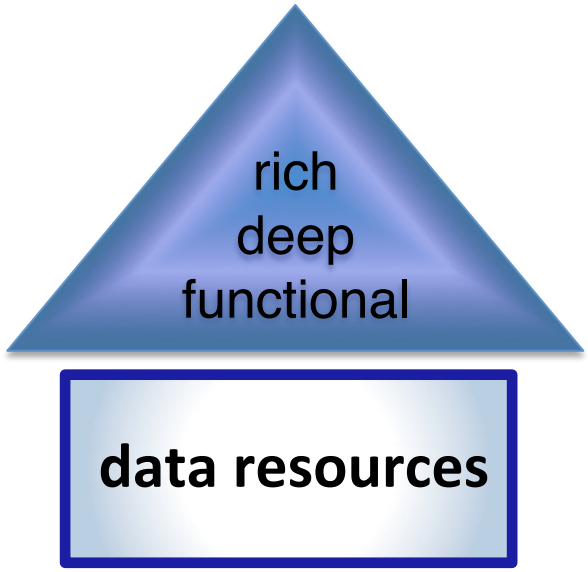
- Open data requirements and expectations
- Reproducibility, replication, and other "Rs"
- Stewards of the common good / scholarly record
- Competitive, innovative research
  - exemplars of "open" research
  - **centers of excellence, research prominence**

# Optimizing data for reuse

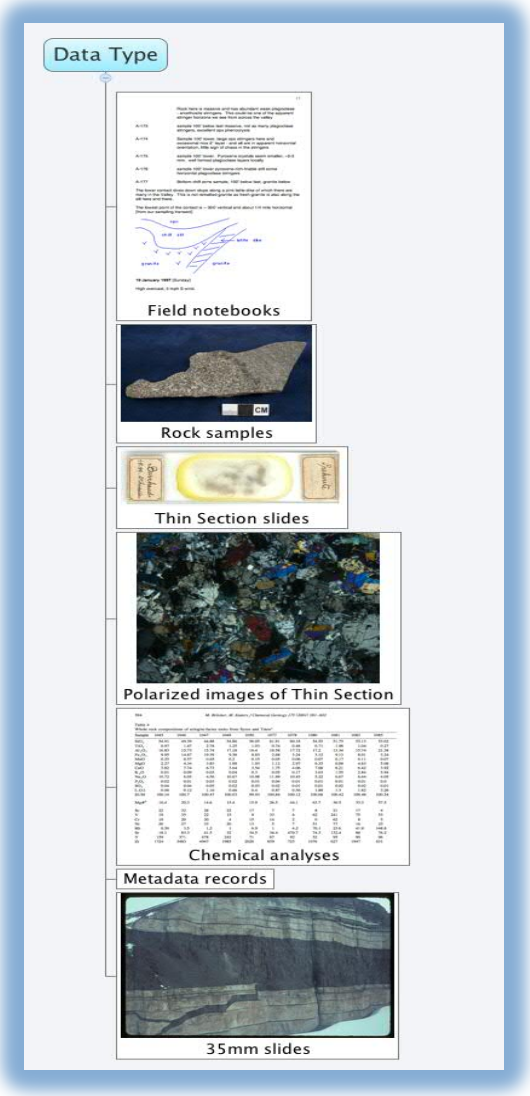
Different objectives and expertise than:

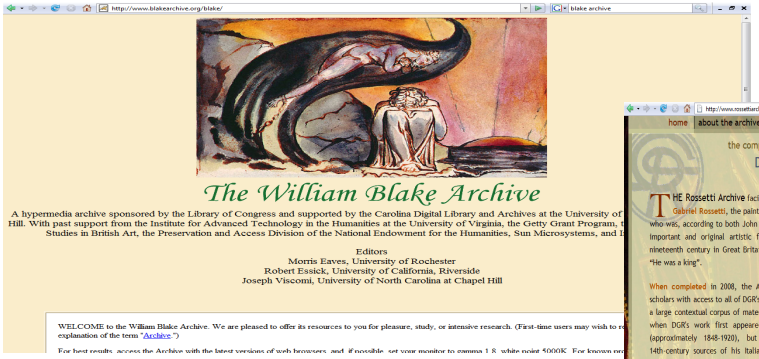
- preserving a record of research
- providing access and transparency

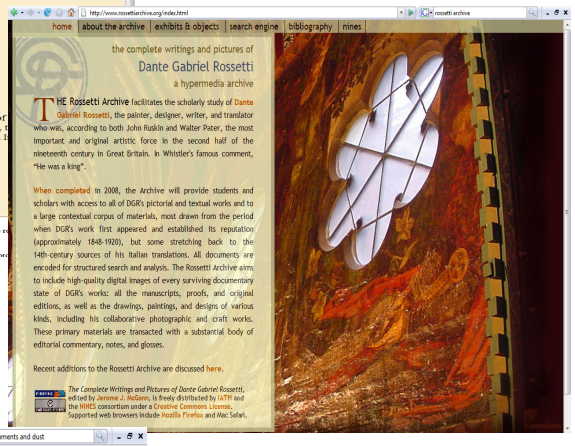
and much more resource intensive.

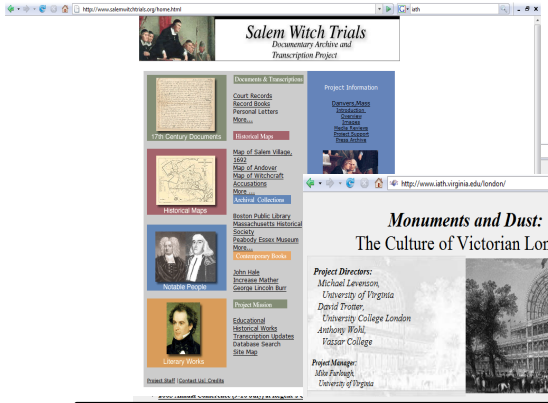


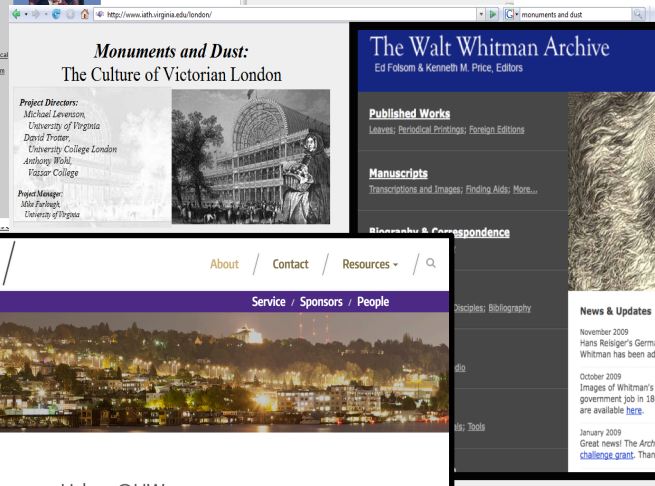
# Promoting our deep, rich, functional data



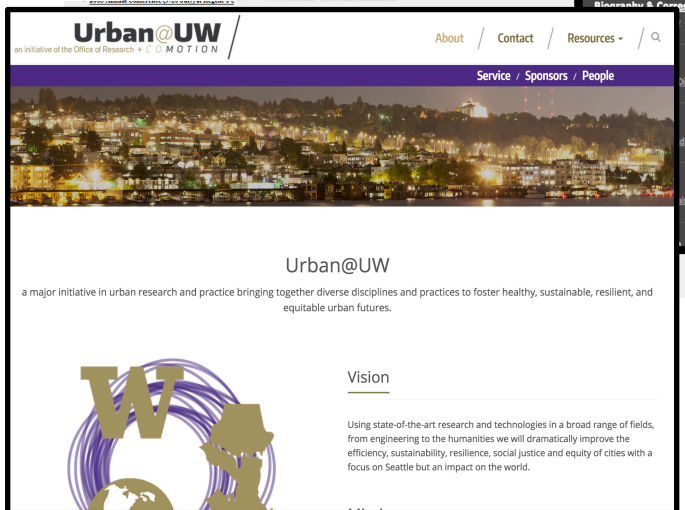









# Empirically derived reuse principles

Data Curation Profiles Project

Data Conservancy

Site-Based Data Curation at YNP

- *Releasable ≠ reusable*
- *Producer sets / consumer subsets*
- *Indicators of reuse value*
- *Primacy of method*

# True reusability for site-based data

Retain value and promote reuse of data from scientifically significant sites.

Geobiology data from Yellowstone National Park

**Reuse dependent on**

Sampling procedures

**within**

Field campaign context:

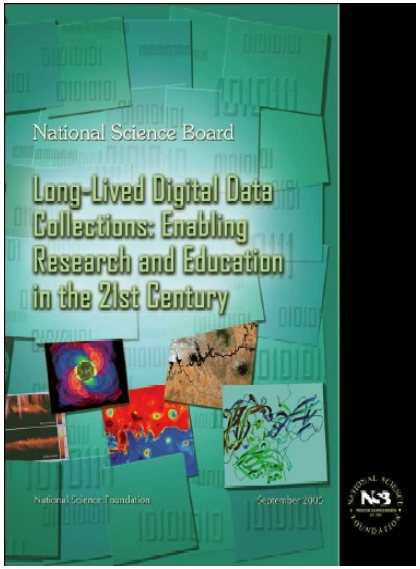
geological feature

*** **new measurements**

vent location, etc.

Used with permission from B. Fouke

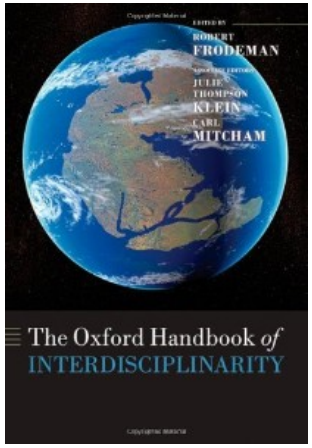
# Crisis in resource collections




**NSB 2005:**…ever increasing investment in creating and maintaining collections, and the rapid multiplication of collections, with a potential for decades of curation.

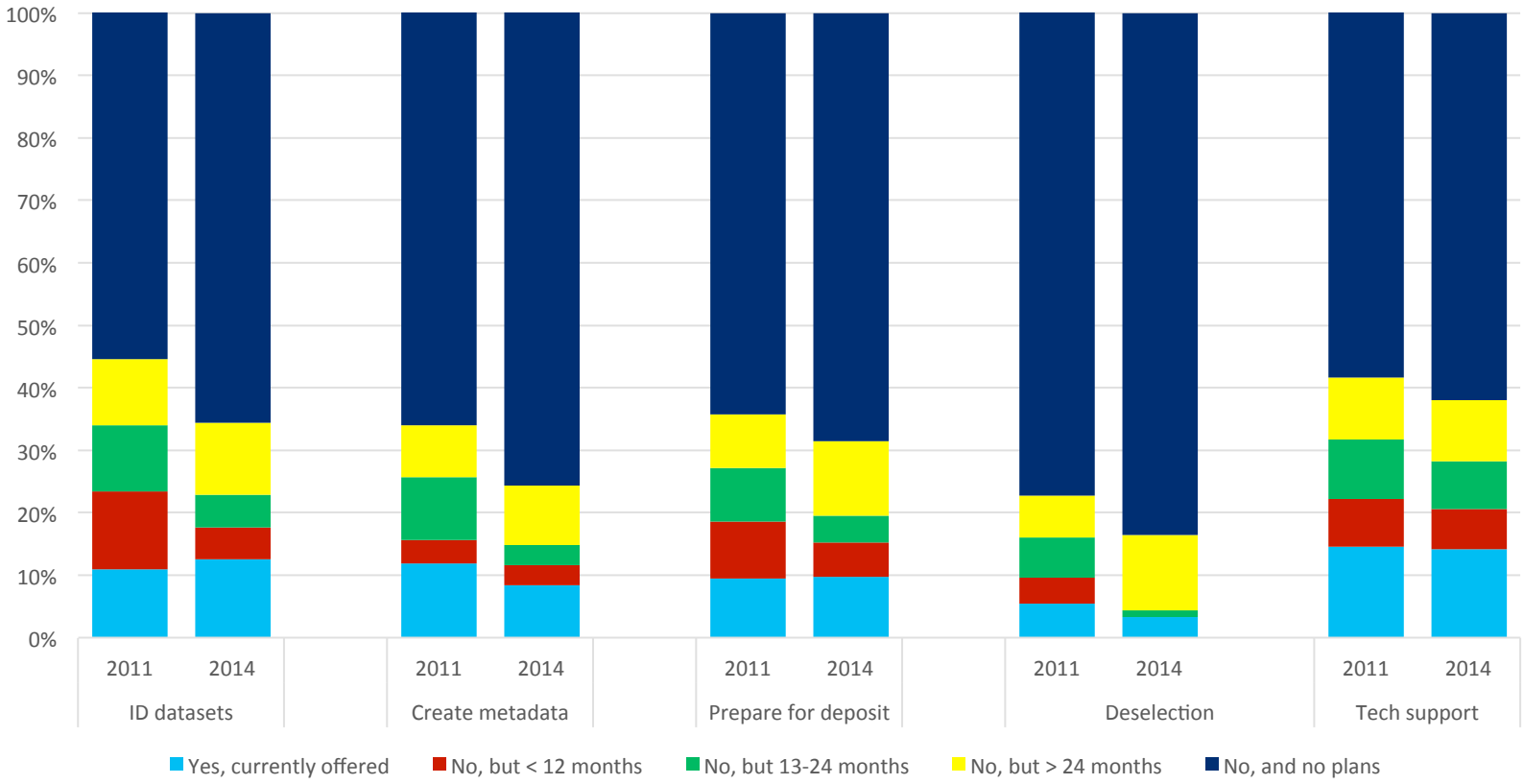
**Atkins and Unsworth**: Value-added … widely shared … collections…enabling …interdisciplinary research …

Greatest challenges not ability to move across disciplinary boundaries but in maintaining the increasingly long and mutable intellectual paths to our disciplinary past. (Palmer, 2010)



# Where are we with the workforce?

## Research Data Services Offered or Planned in ACRL Libraries

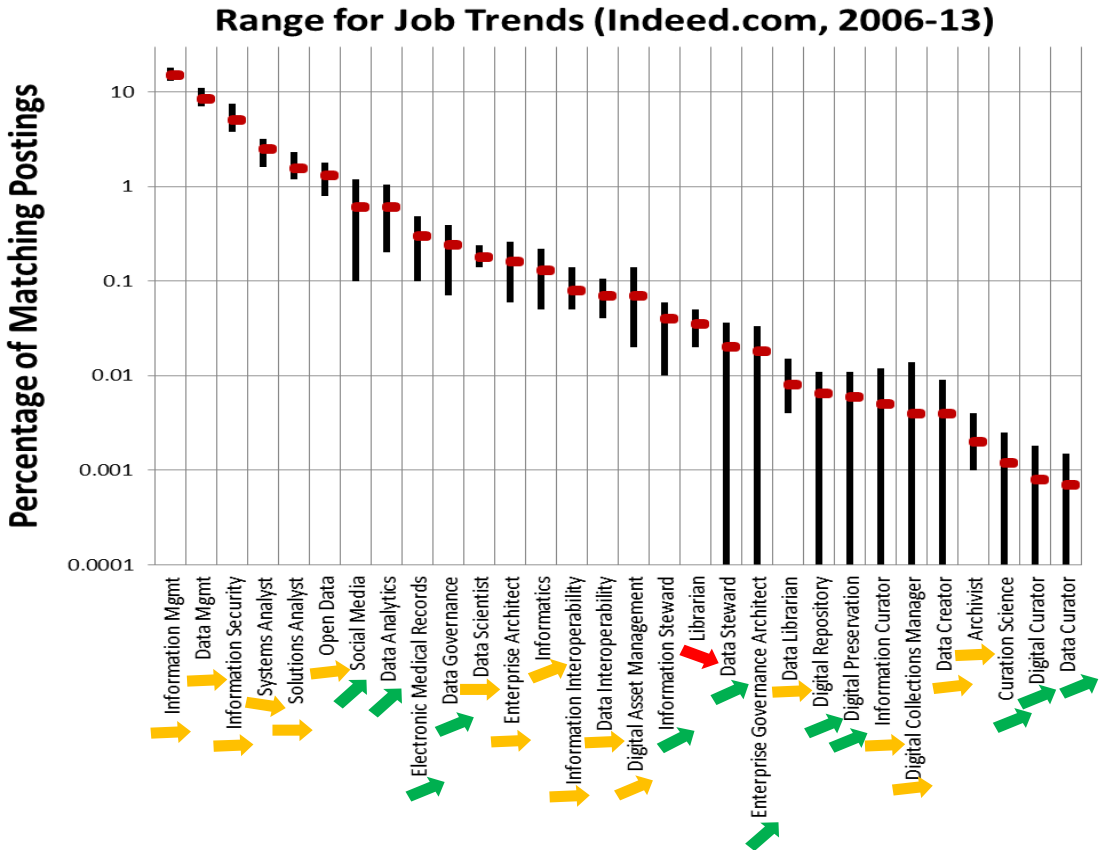



(Tenopir, ASIST, 2014)

THE UNIVERSITY of TENNESSEE

Center for Information and Communication Studies

# Forthcoming - Preparing the Workforce for Digital Curation




(R. Larsen, IDCC, 2014)

Trending up:

- Information Steward
- Data Steward
- Digital repository
- Digital preservation
- Curation Science
- Digital Curator
- Data Curator

Trending down:

Librarian

# Illinois data curation placements

## Academic

- 40% of placements,
  ¼ of those outside library
- Many focused on metadata and technology

## Positions that (probably) didn't exist 5 years ago

- Research Data Management Service Design Analyst
- Data Management Consultant
- Data Science & Informatics Librarian
- Data Curator
- Assistant Dean, Digital Humanities Research

## Non-academic positions

- Data Steward Consultant
- Solutions Analyst
- Senior General Engineer
- GIS Specialist
- Director of Archive Technology
- Digital Asset Manager
- Information Architect
- Information Systems Associate
- Digital Project Coordinator
- Media Content Specialist

# More expertise

Classroom experiences with multiple experts

- Earth science data center services
- Cyberinfrastructure R & D
- International data sharing coordination
- Funding & policy perspectives

Field experiences
with multiple mentors

Data / Science / Peer mentors

(See DCERC video at: https://www.youtube.com/watch?v=mbX5bvgT1ME)

# NCAR internships

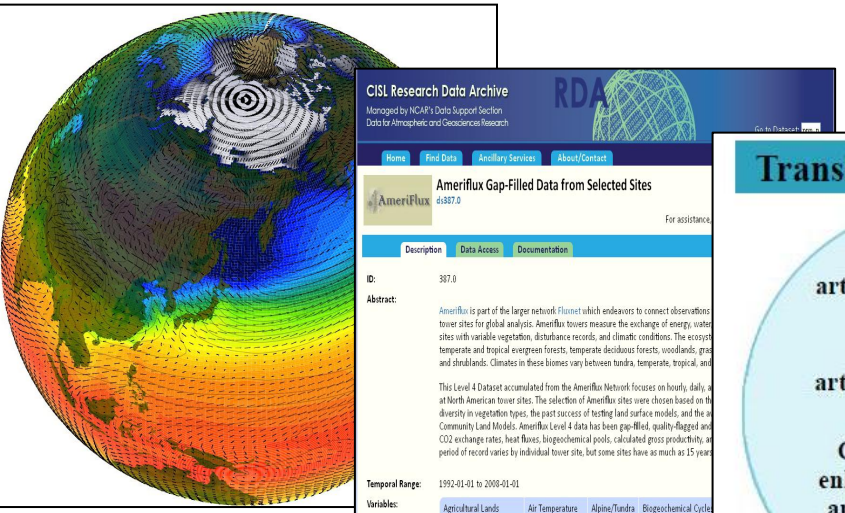
Climate model metadata

Sensor data archiving

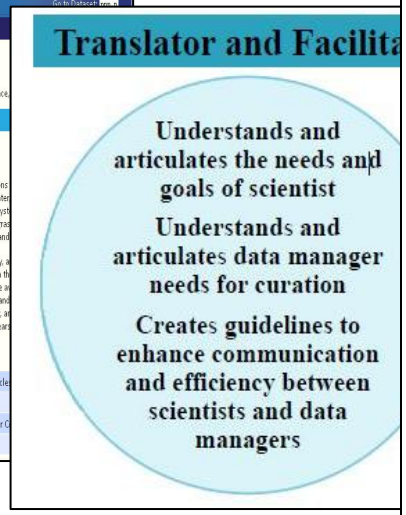
Social science data organization

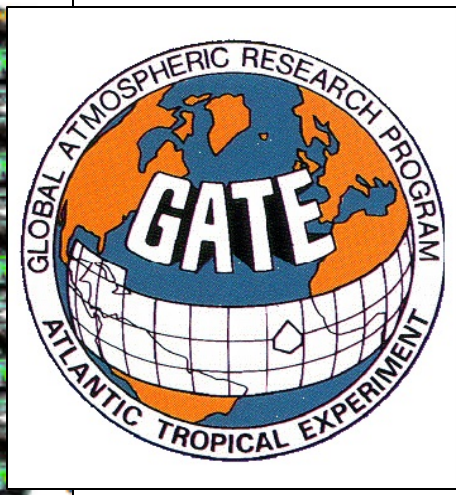
Time-series temporal spatial

Analog data for digital access






metadata harvesting, standards compliance, quality

processing & file migration

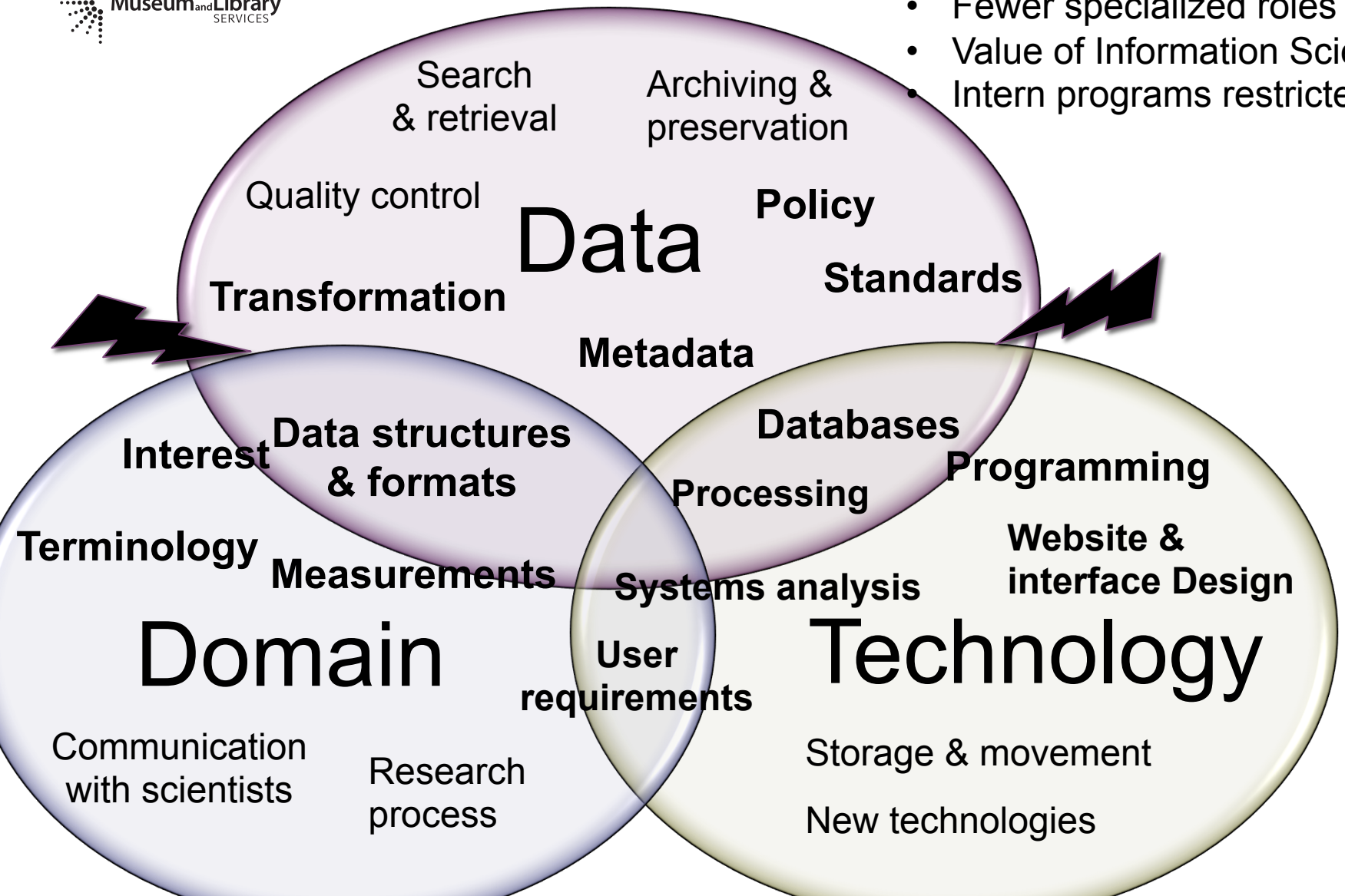
cross-disciplinary data curation; subsetting

high resolution, provenance, NetCDF

50 international collections, OAIS, DOIs

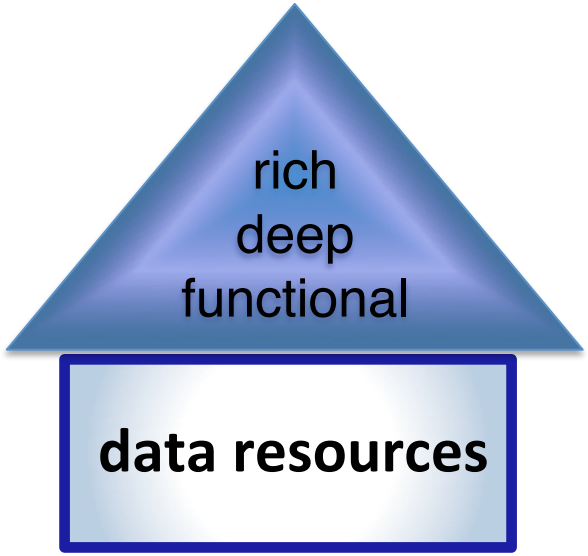
# Trends in national data facilities

- Some growth in positions
- Fewer specialized roles
- Value of Information Science
- Intern programs restricted



# Too much to lose, if we don't get it right.

"Your analytics are only as good as your curation."




- marshal our strengths in LIS
- leverage progress across disciplines
- build a new LIS foundation in the science of data

*Thank you for your attention.*

Information School
UNIVERSITY of WASHINGTON
we make information work