Collections in the Age of E-Research; Realizing Potential through Curation and Aggregation

Carole L. Palmer

Center for Informatics Research in Science & Scholarship
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign

Digital Library Federation Fall Forum
2 November 2010,
Palo Alto, CA

The iSchool at ILLINOIS

CIRSS

What were we worried about 50+ years ago?

1948 - Royal Society Scientific Information Conference

1952 - Chicago School Symposium on Special Information

1958 - International Conference on Scientific Information

- Growth of the literature
- Complexity of formats
- Interdisciplinarity
- Connectivity among systems
- Faster circulation
- Access to pre-publications
- Document components of documents

What is fundamentally different now?

- Involvement of scientists and scholars in information system R&D
- Access to scholarly content by general public
- Modes of engagement with the universe of information
  - convenience, speed
  - inking
  - small window of access
- Higher stakes …

In the contemporary context of e-science, where the aim is to

re-shape scientific endeavours

and provide new infrastructures to support them, [the] goal of studying the detail of actual practice takes on a new significance.

(Hine, 2005)

"Reshaping" suggests professional imperatives

- Provide access to the broad landscape of information

  across disciplines - LIS as a metascience (Bates, 1999)

  across generations, to preserve the scholarly record.

- Foster sharing across institutions.

- Manage scatter of information resources.

- Coordinate in alignment with complex social structures and practices. (Shera, 1972)

- Add value to improve current use and potential for future use. (Taylor, 1986)

Value of digital collections is in the collective

- Tremendous investment in generating content.
- Material exposed but growing base of interrelated content not exploited or broadcast.
- Scholarly and scientific potential and economies of scale obscured in the scatter.

Curated collections essential units of high-value information.

Aggregation of collections essential organizational layer for access, interpretation, and use.

Two initiatives

IMLS Digital Collections & Content (DCC)

Data Conservancy

Two large-scale digital collection efforts

Different kinds of LIS / research library R & D contributions to the emerging network of digital collections and services.

IMLS DCC - CIRSS & University Library collaboration

Carole Palmer (2007-2012) Tim Cole (2002-2007)	PIs
Allen Renear, Miles Efron, Mike Twidale Sarah Shreeves, Bill Mischo, Tom Habing	CIRSS co-PIs Library co-PIs
Katrina Fenlon Jacob Jett, Piotr Organisciak, Sunah Suh, Richard Urban, Karen Wickett, Oksana Zavalina	Project Coordinator Research Assistants
Martin Doerr - Centre for Cultural Informatics, Greece Jeremy Frumkin - Arizona Jonathan Furner - UCLA Jean Godby - OCLC Josh Greenberg - NYPL Pete Johnston - Eduserve. UK Bill Landis - Yale Jenn Riley - Indiana Jim Scheppke - Oregon State Library Guenter Waibel - OCLC	Advisory Committee

IMLS DCC - Phase 1 (2003-2007)

- Single point of access to IMLS-funded content (NLG, some LSTA).

  collection registry
  item-level metadata repository using OAI-PMH

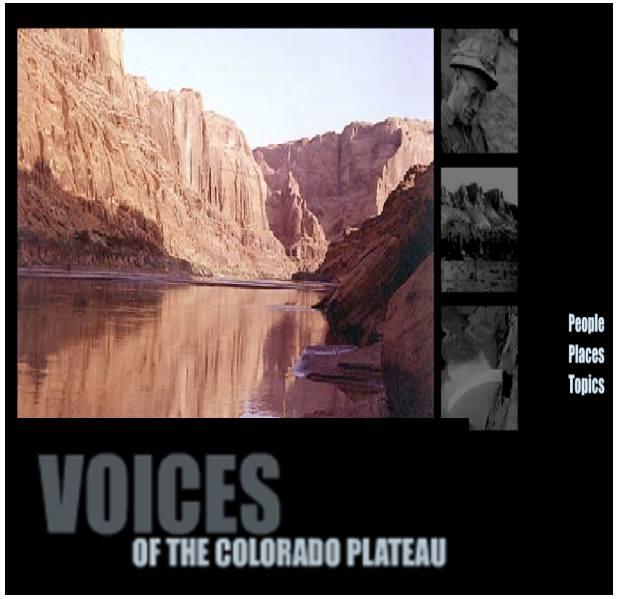
- Focus on interoperability issues and trends.

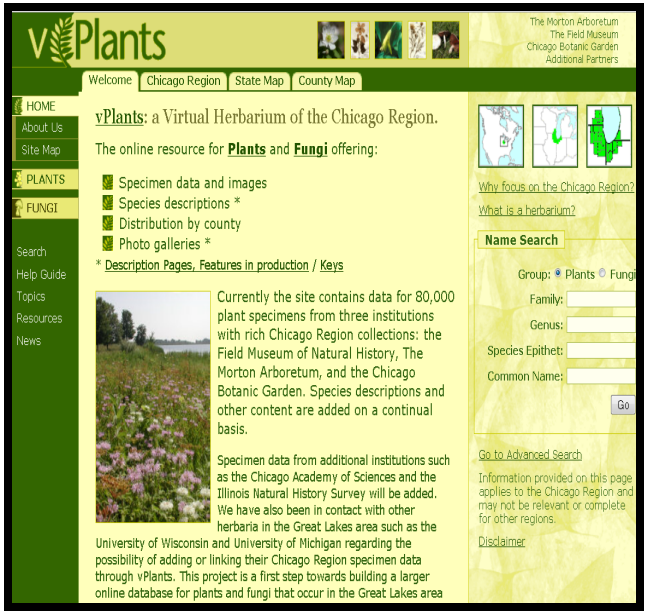
  metadata sharing, harvesting, quality, best practices

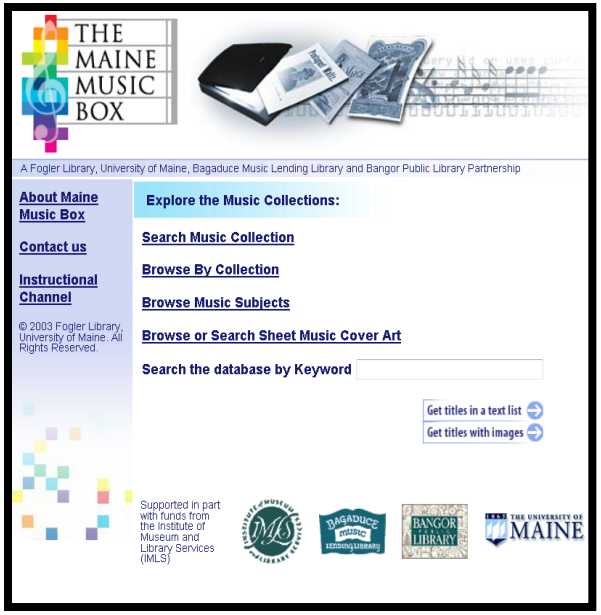
- Collection representation and identity.

  conceptualizations of collections: exhibits, displays, tours …
  cultures of description in libraries, museums, archives

DCC in 2007 – 202 diverse collections, 300,000 items







Questions driving Phase 2

What's in it, really?

Who's it for?

Limitations to opportunistic, critical-mass approaches to growth.

Need to understand whole, develop for user communities, leverage the collective.

Opening History (2007 -)

Build on extensive base of historical materials

Policy driven collection

Emphasis on research threads – metadata relationships, emergent collections for scholarly use

Opening History - 900+ collections, over 1 million items

Opening History Activities:

- Sowing Culture
- IMLS DCC
- Metadata Roundtable
- Research
- Working Groups

CIRSS

Search For Items

Search

Advanced Search for Items

Browse Collections By:

Subject	Object	Place
Social Studies (823) Arts (209) Science (69) Religion (37) Language Arts (24) Vocational Education (16) View all subjects...	image (645) text (477) physical object (119) sound (72) moving image (22) Interactive Resource (17) View all objects...	Midwest U.S. (general region) (209) Southwest U.S. (general region) (193) Southern U.S. (general region) (173) Pacific Coast U.S. (general region) (120) Mountain Region U.S. (general region) (71) Northeastern U.S. (general region) (55) View all places...

also browse by: Title | Hosting Institution

Search Collections Only

Suggest a collection

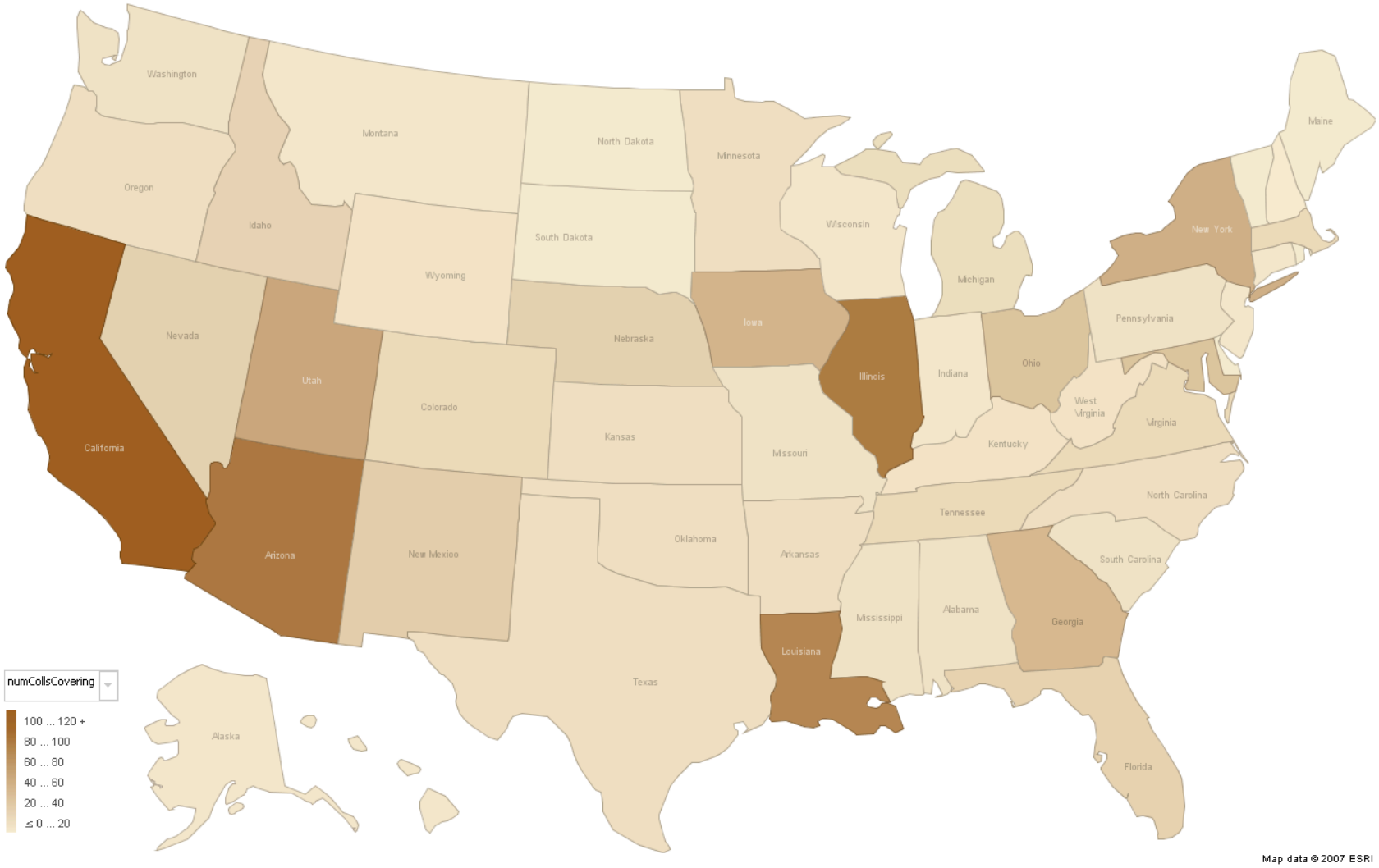
Currently 901 collection records. Last record updated On 10/20/2010.

Collection Development Policy (pdf)

(http://imlsdcc.grainger.illinois.edu/history)

Broad geographic coverage

Accelerated growth through engagement with state libraries



Moving beyond size and search

- Can we improve how we collect collections by applying knowledge of:
  - how scholars build and engage with research collections.
  - research library collection development practices - conspectus-style subject evaluation.

- Can we begin to realize the collective value of the DCC aggregation by:
  - uncovering new, cohesive areas of content as they build up.
  - systematically building cohesive units of value to users.

Scholarly work with collections

special collections

unpublished, primary materials

intentionally created wholes

texts and images

dense units of content

known repositories

browsing

long term personal collections

evidence for inquiry

exploration and gathering

deep study over time

records in context

. . . Brockman et al. · Brogan · Buchanan et al. · Buckland
Case · Currall, Moss & Stuart · Duff & Johnson
Ellis & Oldman · Lee · Palmer · Tibbo . . .

Guiding concepts

Contextual mass – inquiry & collecting by humanities scholars

(Brockman, Palmer, Neumann, & Tidline, 2001)

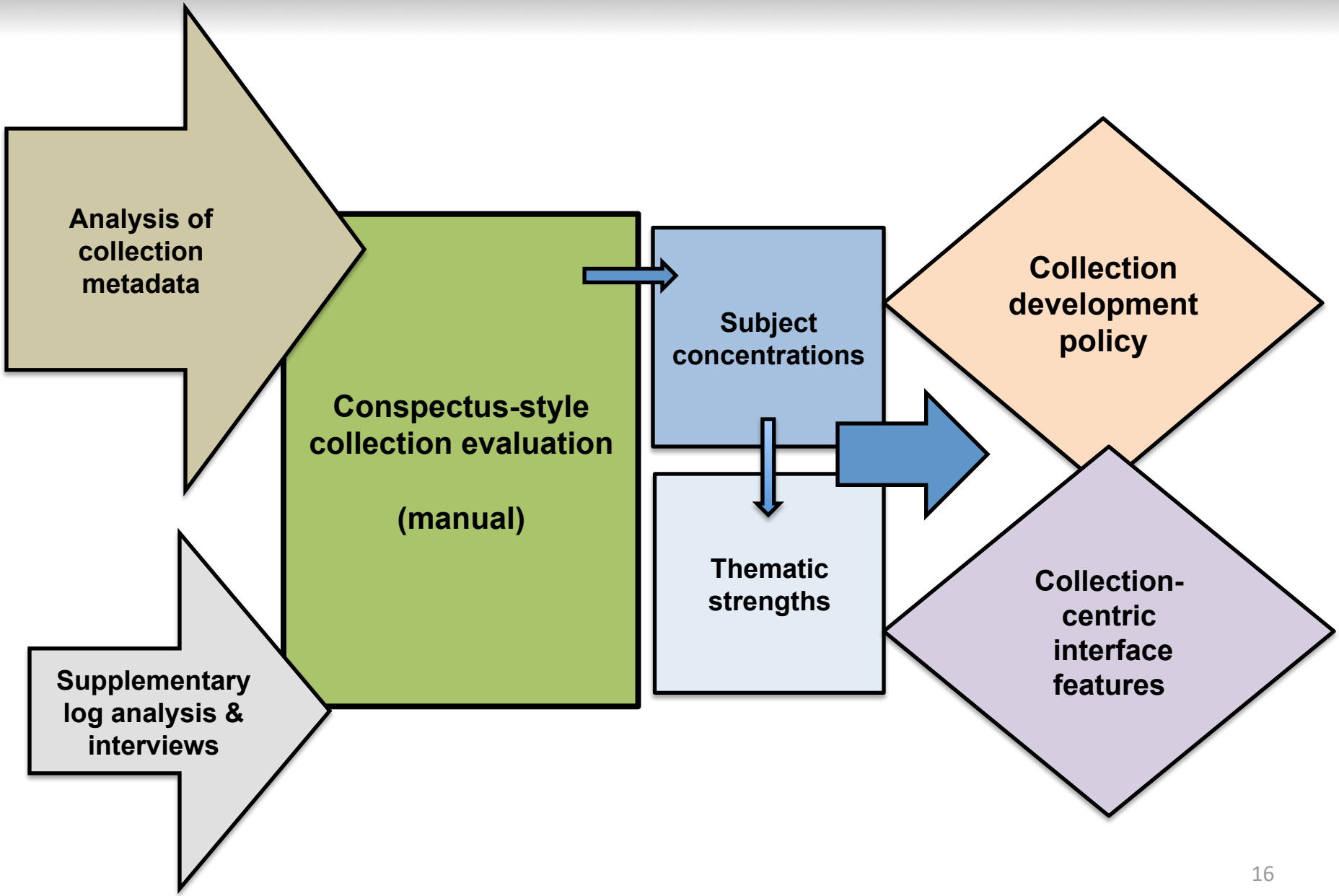
- size not a criteria for inclusion (bigger is not better)
- focus on meaningful interrelationships among distributed materials
- strive for dense, rich, cohesive groups of sources to support research

Thematic research collections – created by digital humanities scholars

(Palmer, 2004)

- primary sources and related materials sought and collected by scholars and special collections curators
- latent in the subject strengths emerging in OH

Aggregation strategy



Subject evaluations – concentrations & growth

Subject concentration	Collections with subject, April 2009	Collections with subject, April 2010	% of OH Collections with subject, April 2010
Military history	53	83	10.3%
Native American history	32	78	9.7%
Transportation history	33	48	6.0%
Asian American history	41	44	5.5%
African American history	29	33	4.1%
Mining history	17	26	3.2%
Exploration & travel history	19	23	2.9%

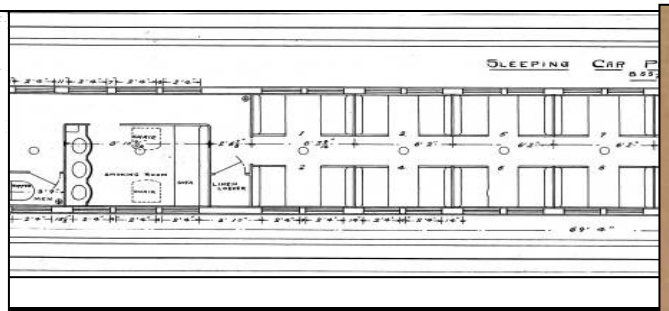
Subject concentration - contextual mass factors

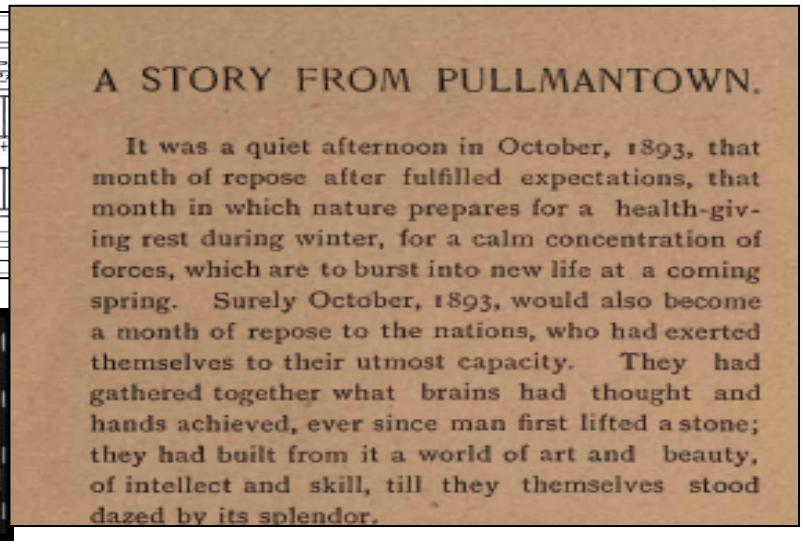
Military History Strength Factors	Frequencies
Subject-focused collections	56
Items in subject-focused collections	19,388
Subject-inclusive collections	27
Items in subject-inclusive collections	48,085
Item types represented	39
U.S. states covered	24
Events covered – subject specific factor	12

Transportation thematic strength – Pullman

- Demonstrates value of small collections
  - 1500 blueprints - Newberry's Pullman Car Co. Collection
  - Scattered photos of Pullman porters from Pittsburgh, CT, WA
  - 40,000 glass negatives and prints, relevant to Pullman and era - Bain LOC,
  - Videos and songs referencing Pullman in pop culture - other LOC

1 rare book satirizing the Pullman Co from 17-item
Illinois Art and Literature digitized books collection





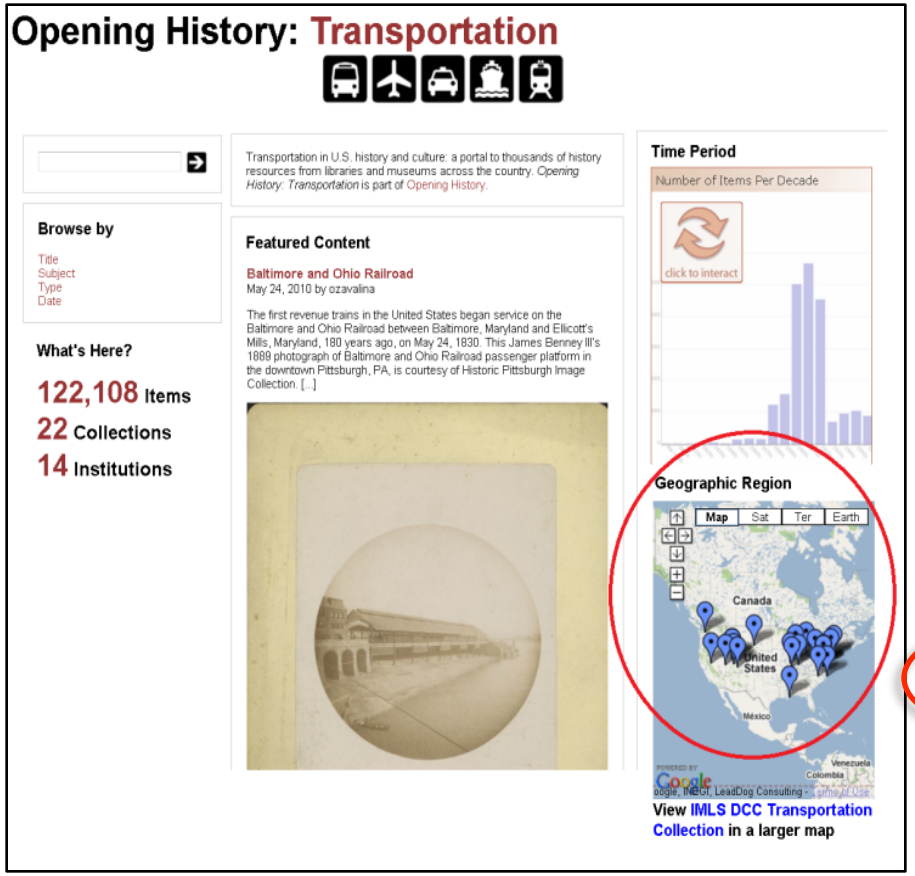
A STORY FROM PULLMANTOWN.

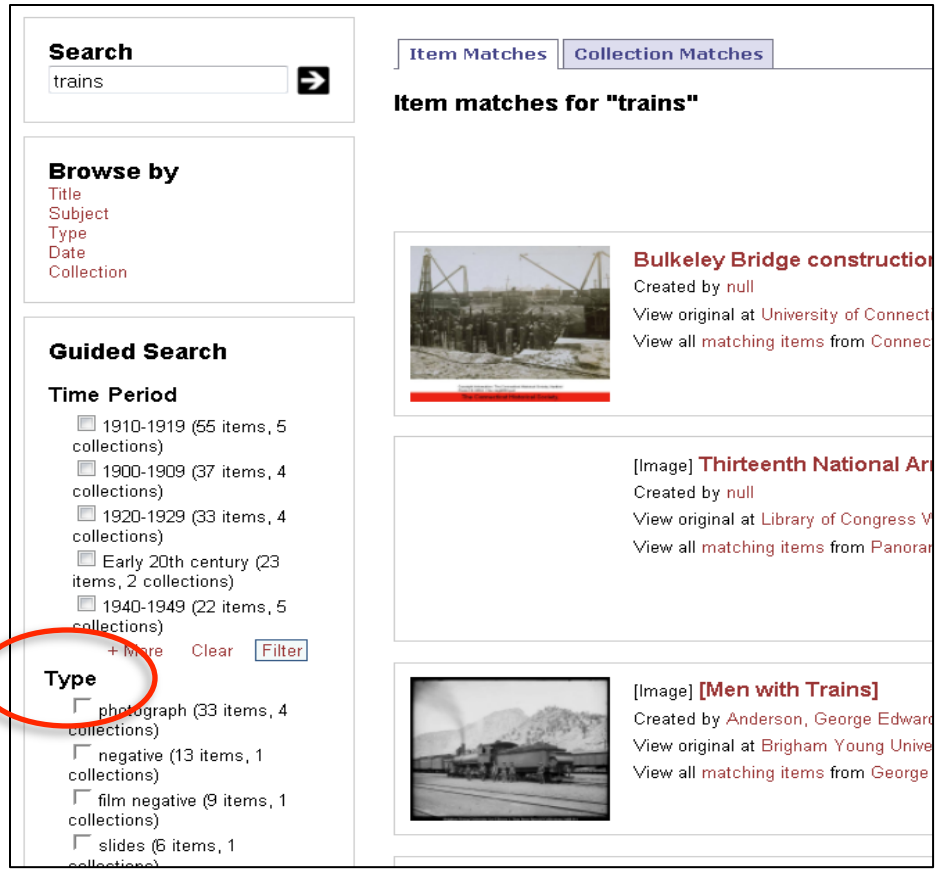
It was a quiet afternoon in October, 1893, that month of repose after fulfilled expectations, that month in which nature prepares for a health-giving rest during winter, for a calm concentration of forces, which are to burst into new life at a coming spring. Surely October, 1893, would also become a month of repose to the nations, who had exerted themselves to their utmost capacity. They had gathered together what brains had thought and hands achieved, ever since man first lifted a stone; they had built from it a world of art and beauty, of intellect and skill, till they themselves stood dazed by its splendor.

Making subject concentrations more explicit

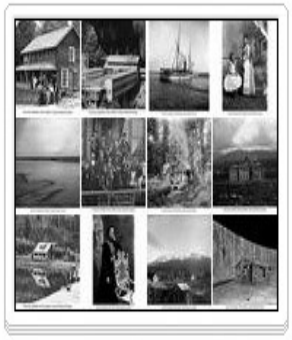
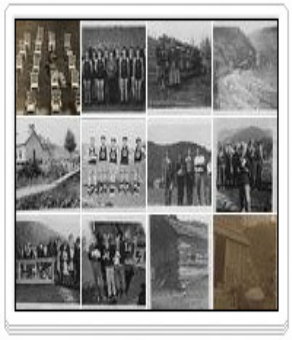
Transportation History experimental interface

Enhanced access to a concentration with relatively low investment





Extending access via Flickr

Flickr Feasibility Study

- Developing metadata, workflows, policy, technical routines
- Differences in content and representation based on type of collection and institution
- Weekly batch uploads of 100-200 (distributed across collections)
- Assessing public interaction with content
- Evaluating value of service and participation

Thematic Flickr groups using IMLS DCC images

- Cologne – Black and White
- Emmett Kelly Sr or Jr - Weary Willie
- Vintage Photos from World War II
- Vintage Monocle
- Legendary Adventures Guild
- Ford Farlaine/Torino
- Worldwide Ship Spotters (WSS)
- Hairygit's Nature Group
- Lighthouses of the World
- Clan Ross
- Too Tuff To Die! Hop Picking
- Hop Picking

- Bitterlake Seattle
- I, Magnin, & Co.
- Ticket Booth
- Seattle Dwellings
- Mail Pouch
- Bomarc Missile
- King County Creative Commons
- Roosevelt and Ravenna (Seattle)
- Snow Plows and Snow Removal Equipment
- Diamond T Trucks

For cultural heritage . . .

Principled collecting of collections valuable alternative to search and retrieval of scattered items or opportunistic aggregations.

- context of original curated collections
- links to complementary collections

Conspectus-style evaluation provides systematic process for evaluating and enhancing aggregations of research collections.

- Identify coherent bodies of emergent content
- contributions of small collections not obscured

Nationally scoped research collections can be developed out of subject concentrations through targeted recruitment.

In the scientific data arena

- Complex set of activities, dependencies, constituencies
  - Domain researchers and informaticists often leading development
  - Data and research centers with deep down-stream knowledge
  - Funding agencies and universities invested in data assets - producing competitive science; data management plans

New demands for data services requires libraries to collaborate with range of data experts and initiatives in domains, IT, data centers, etc.

We have much to contribute, and oh so much to learn.

Data Conservancy

JOHNS HOPKINS UNIVERSITY

PI, Sayeed Choudhury,

Tim DiLauro et al., Sheridan Libraries

CoPIs and Partners	
Carl Lagoze	Cornell University
Mary Marlino	National Center for Atmospheric Research (NCAR)
Carole Palmer	CIRSS, GSLIS, University of Illinois at U-C
Paddy Patterson	Marine Biological Laboratory
Chris Borgman	University of California Los Angeles
Ruth Duerr	National Snow and Ice Data Center
Mark Evans	Tessella, Inc.
Eileen Fenton	Portico
Sandy Payette	DuraSpace / Fedora Commons

Aims and approach

Integrated and comprehensive data curation strategy

- to collect, organize, validate, and preserve data
- to address grand research challenges that face society

Network of expertise: domain and data scientists, information and computer scientists, librarians, engineers, enterprise experts.

Infrastructure builds on and connects existing exemplar initiatives

- deep engagement with scientists
- extensive experience with large-scale, distributed systems.

Astronomy as an exemplar community

Success in data standards, practices, documentation, and associated services

Ingest astronomy data into preservation archive,
connect data to existing services used by astronomers.

Demonstrate utility of hosting data in environment that supports existing scientific capabilities in a sustainable manner.

Scope to include: life sciences
earth sciences
social sciences

Science and library based hubs

Marine Biological Laboratory

- Encyclopedia of Life - taxonomic organization, ontology indexing
- species identification queries for climate change analyses

National Snow & Ice Data Center

- extensive sensor network, fieldwork, aircraft and satellite data
- access node on the DC network, test bed for distributed services

National Center for Atmospheric Research

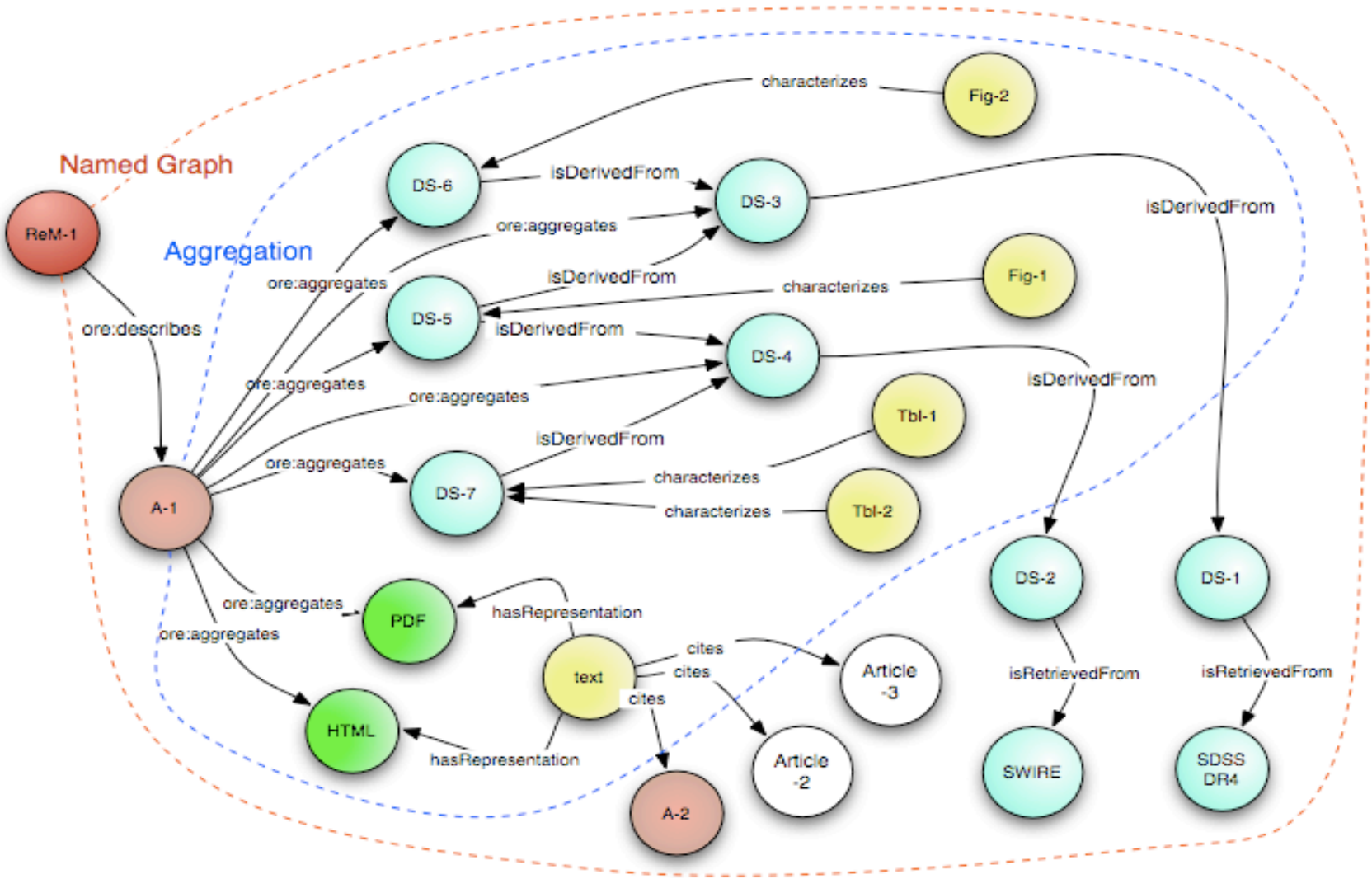
- civic decision making and climate science in megacities

Cornell University Library

- DataStar - promotes archiving to disciplinary data centers
- arXiv eprints service capability for linking research data with publications

Resource map for astronomy compound object

OAI-ORE (object exchange & reuse) protocol



Use cases, data practices & curation processes

	Astronomy	Life Sciences	Earth Sciences	Social Sciences	
NCAR	Task-based design Use cases, system requirements				
UCLA	Ethnography, deep case study	Disciplinary differences Curatorial processes SMALL SCIENCE - reuse potentials			**ILLINOIS**

Small science is big, and poorly curated

12,025 NSF grants awarded in 2007 = $2,865,388,605

	20%	80%
Number of Grants	2405	9621
Total Dollars	$1,747,957,451	$1,117,431,154
Range	$300,000 - $38,131,952	$579 - $300,000

Top 254 grants received 20% of the total awarded

(Heidorn, 2009)

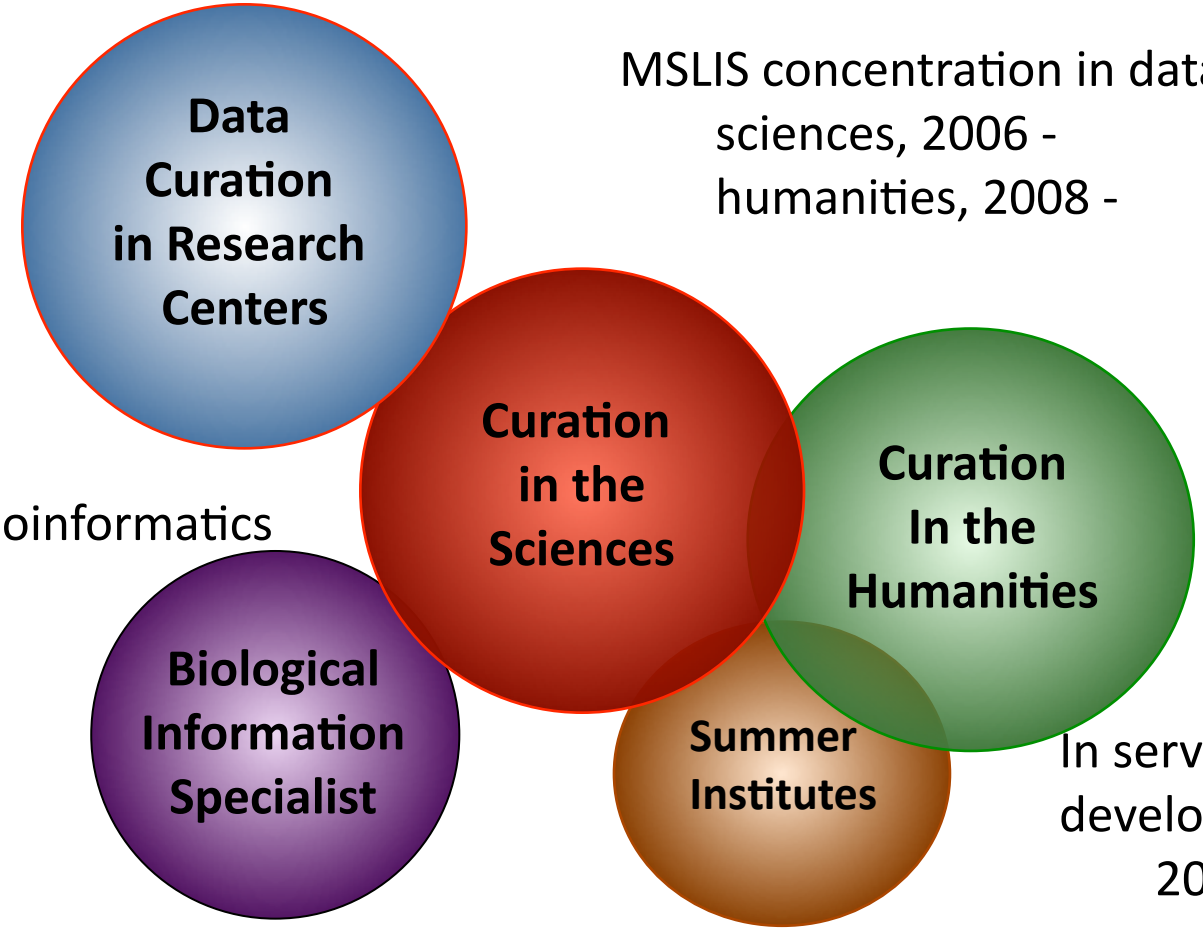
Formalizing data concepts

- Conceptual precision and accuracy necessary for the reliable data curation and integration across disciplines
  - dataset, version of a dataset, part of a dataset, collection of datasets, data, granule, datum, etc.
  - representation levels (data, representation, encoding, format, etc.
  - related problems in measurement

Research informing education

Doctoral research awards
2011 -

MSLIS concentration in data curation
sciences, 2006 -
humanities, 2008 -



Masters in bioinformatics
2006 -

In service professional development
2008 -

Research libraries are mission-ready

- *Support research, scholarship, teaching*
- *Advance development & transmission of knowledge*
- *Preserve our intellectual heritage*
- *Contribute to the common good*
- *Promote the development of models, standards and infrastructure*
- *Develop distinctive collections, services, staff and facilities*
- *Leaders, collaborators, and innovators, to enable our communities—locally and beyond*

Many models and processes for remix

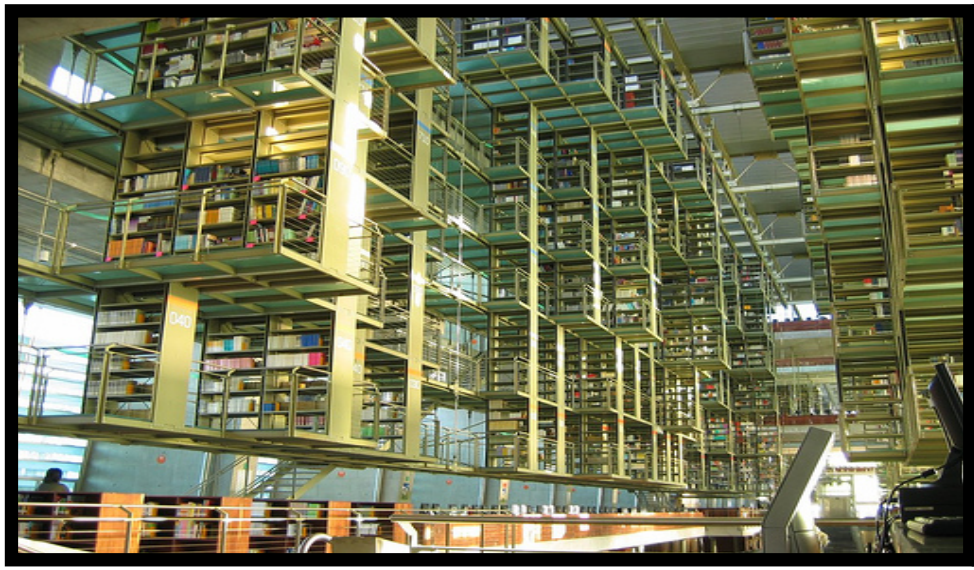
research and national libraries, consortia, CRL …

union catalogs, bibliographies of bibliographies

collocation, just-in-time collecting …

The stacks

Flickr users: stancia, rh creative commons



Special collections

flickr.com/photos/001fj/2907653323/

Profession is well positioned, but field is in its infancy

Compatible institutional and human infrastructure,

expertise, principles, and commitment.

Can leverage deep knowledge of the bibliographic universe, digital libraries, and scholarly user communities.

But that professional knowledge began generations ago.

Studies of use of scientific literature emerged in 1960s, began in earnest in 1980s.

Just beginning to investigate, document, and respond to the current complex e-research collection and service environment.

Questions, comments, discussion

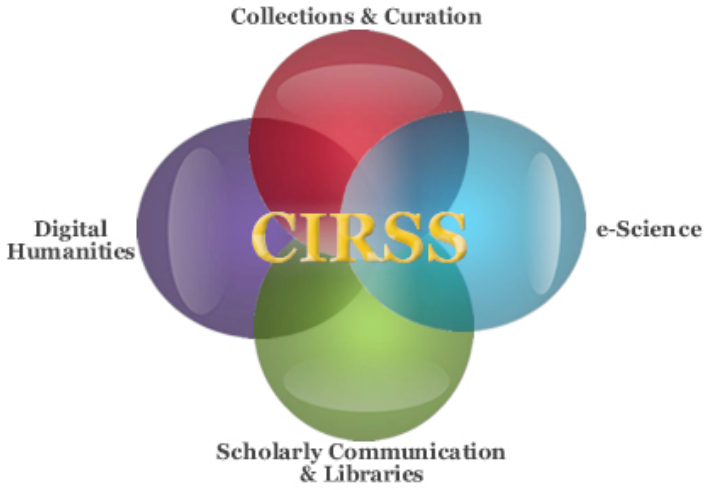
clpalmer@illinois.edu

IMLS DCC: http://imlsdcc.grainger.Illinois.edu

Data Conservancy: http://DataConservancy.org

http://cirss.lis.uiuc.edu/

Center for Informatics Research in Science and Scholarship



References

Palmer, Carole L., Zavalina, Oksana, & Fenlon, Katrina (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. Proceedings of the ASIS&T Annual Meeting (Pittsburgh, PA, Oct. 22-27).

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Washington, DC: Digital Library Federation/Council on Library and Information Resources.

Brogan, M. (2006). *Contexts and Contributions: Building the Distributed Library*. Washington, DC: Digital Library Federation/Council on Library and Information Resources.

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J., & Warwick, C. (2005). Information seeking by humanities scholars. In *Proceedings of the European Conference on Digital Libraries* (Vienna, Austria, Sept. 18-23, 2005), 218-229.

Buckland, M. K. (1999). *Library Services in Theory and Context*. (2nd ed.). New York: Pergamon Press.

Case, D. O. (1991). The collection and use of information by some American historians: A study of motives and methods. *Library Quarterly, 61*, 61-82.

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria, 58*, 131-146.

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behaviors of historians in archives. *Library Quarterly, 72*(4), 472-496.

Ellis, D., & Oldman, H. (2005). The English literature researcher in the age of the Internet. *Journal of Information Science*, *31*(1), 29-36.

Lee, H. L. (2000). What is a collection? *Journal of the American Society for Information Science, 51 (12),* 1106-1113.

Palmer, C. L. (2004). Thematic research collections. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 348-365). Oxford: Blackwell.

Tibbo, H. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. *The American Archivist, 66*(1), 9-50.

Image attributions

- Philips, G. (1884). The Porters on the Pullman palace cars. In ***Music for the Nation: American Sheet Music, 1820-1860, 1870-1885*** digital collection. Retrieved from: http://hdl.loc.gov/loc.music/sm1884.16714.

- Pullman Company (1898). Sleeping Car. Drawing. In ***Pullman Company Archives, Newberry Library*** digital collection. Retrieved from: http://collections.carli.illinois.edu/cdm4/results.php?CISOOP1=any&CISOFIELD1=CISOSEARCHALL&CISOROOT=/nby_pullman&CISOBOX1=Company.

- Bech-Meyer, N. (1894). A Story from Pullmantown. C.H. Kerr. In ***Illinois Art an Literature*** digital collection. Retrieved from: http://hdl.handle.net/10111/UIUCOCA:storyfrompullman00bech.

- Edison, T. A. (1903). A romance of the rail. Film. In ***Early Motion Pictures, 1897-1920*** digital collection. Retrieved from: http://hdl.loc.gov/loc.mbrsmi/edmp.2174.

6th International Digital Curation Conference

Chicago, IL
Dec. 6-8, 2010

hosted by
CIRSS / GSLIS

in partnership with
Digital Curation Centre, UK

- pre-conference Research Data Workforce Summit
- post-conference LIS Research Summit

Data curation is . . .

The active and on-going management of (research) data
through its lifecycle of interest and usefulness
to scholarship, science, and education.

Functions

- enable discovery and retrieval
- maintain data quality
- add value
- provide for re-use over time
- archiving
- preservation

Tasks

- appraisal and selection
- representation
- authentication
- data integrity
- maintaining links
- format conversions

Convergent user data

Transaction Logs

User interactions	Page views
Viewing collection metadata records	1,760
Viewing item metadata records	368
Collection browse:	2,939
Subject browse	953
Geographic browse	533
Project browse	502
Object type browse	487
Institution browse	311
Collection title browse	153
Item browse	4,388
Collection search	880
Item search	1,860

Sessions with historians

- Collection information useful for navigation, interpretation, and assessment of content.
- OH navigated as central hub,
  - context of collection,
  - exploration of related collections at host institutions.
- Provenance information at collection level is critical for assessment of value.

Core curation content

Foundations of Data Curation

- Digital Data
- Scholarly Communication
- Lifecycles
- Collections
- Infrastructures & Repositories
- Selection and Appraisal
- Metadata
- Standards & Protocols
- Archiving & Preservation
- Intellectual Property & Legal Issues
- Workflows; Data Re-use & Value
- Policy & Cooperative Alignments
- Scientific Information Work

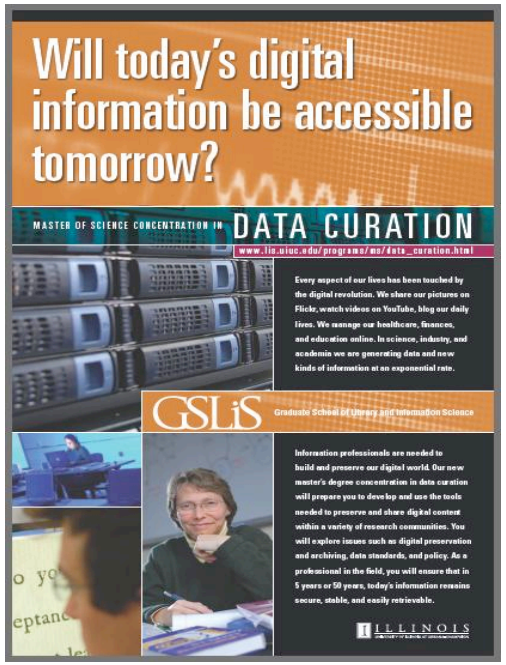
Assignments:
20 cases developed this semester
Critiques of data management plans

Digital Preservation

- Archival Theory & Diplomatics
- OAIS Reference Model
- Data Formats
- Digital Archival Objects
- Preservation Strategies:
- Emulation vs. Migration
- Authenticity, Integrity & Trust
- Evaluation & Value
- Digital Preservation & The Law

Assignments:
Planning Grant Application
Trusted Repository Assessment

Basic curation curriculum



Required Core Courses

Foundations of Data Curation
Digital Preservation
Systems Analysis & Management

Selected Electives – require 2, recommend 4

Metadata in Theory & Practice
Information Modeling
Ontologies in Natural Science
Foundations of Information Processing
Digital Libraries: Research & Practice
Representing & Organizing Info Resources

Internships: JHU, Smithsonian, NLM, USDA, Purdue, NSIDC . . .

MASTER OF SCIENCE CONCENTRATION IN DATA CURATION