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
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.
|--------------------------|----------------------|-----|
| Allen Renear, Miles Efron, Mike Twidale  
Sarah Shreeves, Bill Mischo, Tom Habling | 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
Broad geographic coverage

Accelerated growth through engagement with state libraries
• 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
- known repositories
- long term personal collections
- exploration and gathering
- deep study over time

- intentionally created wholes
- texts and images
- evidence for inquiry
- exploration and gathering
- 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

Conspectus-style collection evaluation (manual)
- Analysis of collection metadata
- Supplementary log analysis & interviews
- Subject concentrations
- Thematic strengths
- Collection development policy
- Collection-centric interface features
### Subject evaluations – concentrations & growth

<table>
<thead>
<tr>
<th>Subject concentration</th>
<th>Collections with subject, April 2009</th>
<th>Collections with subject, April 2010</th>
<th>% of OH Collections with subject, April 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Military history</td>
<td>53</td>
<td>83</td>
<td>10.3%</td>
</tr>
<tr>
<td>Native American history</td>
<td>32</td>
<td>78</td>
<td>9.7%</td>
</tr>
<tr>
<td>Transportation history</td>
<td>33</td>
<td>48</td>
<td>6.0%</td>
</tr>
<tr>
<td>Asian American history</td>
<td>41</td>
<td>44</td>
<td>5.5%</td>
</tr>
<tr>
<td>African American history</td>
<td>29</td>
<td>33</td>
<td>4.1%</td>
</tr>
<tr>
<td>Mining history</td>
<td>17</td>
<td>26</td>
<td>3.2%</td>
</tr>
<tr>
<td>Exploration &amp; travel history</td>
<td>19</td>
<td>23</td>
<td>2.9%</td>
</tr>
<tr>
<td>Military History Strength Factors</td>
<td>Frequencies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject-focused collections</td>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Items in subject-focused collections</td>
<td>19,388</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject-inclusive collections</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Items in subject-inclusive collections</td>
<td>48,085</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item types represented</td>
<td>39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U.S. states covered</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Events covered – subject specific factor</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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
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**

PI, Sayeed Choudhury,

Tim DiLauro et al., Sheridan Libraries

<table>
<thead>
<tr>
<th>CoPIs and Partners</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Carl Lagoze</strong></td>
<td>Cornell University</td>
</tr>
<tr>
<td><strong>Mary Marlino</strong></td>
<td>National Center for Atmospheric Research (NCAR)</td>
</tr>
<tr>
<td><strong>Carole Palmer</strong></td>
<td>CIRSS, GSLIS, University of Illinois at U-C</td>
</tr>
<tr>
<td><strong>Paddy Patterson</strong></td>
<td>Marine Biological Laboratory</td>
</tr>
<tr>
<td><strong>Chris Borgman</strong></td>
<td>University of California Los Angeles</td>
</tr>
<tr>
<td><strong>Ruth Duerr</strong></td>
<td>National Snow and Ice Data Center</td>
</tr>
<tr>
<td><strong>Mark Evans</strong></td>
<td>Tessella, Inc.</td>
</tr>
<tr>
<td><strong>Eileen Fenton</strong></td>
<td>Portico</td>
</tr>
<tr>
<td><strong>Sandy Payette</strong></td>
<td>DuraSpace / Fedora Commons</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th></th>
<th>Astronomy</th>
<th>Life Sciences</th>
<th>Earth Sciences</th>
<th>Social Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NCAR</strong></td>
<td></td>
<td>Task-based design</td>
<td></td>
<td>Use cases, system requirements</td>
</tr>
<tr>
<td><strong>UCLA</strong></td>
<td>Ethnography, deep case study</td>
<td>Disciplinary differences</td>
<td>Curatorial processes</td>
<td>ILLINOIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMALL SCIENCE</td>
<td></td>
<td>- reuse potentials</td>
</tr>
</tbody>
</table>
Small science is big, and poorly curated

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

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Grants</td>
<td>2405</td>
<td>9621</td>
</tr>
<tr>
<td>Total Dollars</td>
<td>$1,747,957,451</td>
<td>$1,117,431,154</td>
</tr>
<tr>
<td>Range</td>
<td>$300,000 - $38,131,952</td>
<td>$579 - $300,000</td>
</tr>
</tbody>
</table>

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
Doctoral research awards
2011 -

Masters in bioinformatics
2006 -

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

Data Curation in Research Centers

Curation in the Sciences

Curation In the Humanities

Biological Information Specialist

Summer Institutes

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 ...
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


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

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

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.
### 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