DIGITAL COLLECTION CONTEXTS:
INTELLECTUAL AND ORGANIZATIONAL
FUNCTIONS AT SCALE

Carole L. Palmer & Megan Senseney
Center for Informatics Research in Science & Scholarship
GSLIS - University of Illinois at Urbana-Champaign

Karen Wickett
School of Information, University of Texas at Austin
IMLS Digital Collections and Content, 2002-2012

Fourfold growth from 2006 to 2012:

1,700 Collections
1.2 Million Items
1,492 Institutions

From 46 states and 2 territories

Collection registry and item-level repository providing a single point of access for IMLS-digitized content and US history collections.
IMLS DCC Research and Development

Research and development:
- metadata quality,
- aggregation workflows,
- collection representation,
- subject analysis and visualization,
- collection evaluation,
- metasearch,
- faceted search refinement, and
- usability.

Retained vital **institutional identities** and **collection contexts** lost in many large-scale aggregations.

Disseminated 60+ publications and presentations; supported 11 students’ research, including completion of three doctoral dissertations.
Extending access through Flickr and a Blog

Mining and Mother Jones in Mount Olive
Springfield Aviation Company Archives
Charles W. Cushman Collection
Charles Overstreet Collection

Sowing Culture
Selections from the IMLS Digital Collections & Content Project
DLF/DCC: DPLA Beta Sprint, 2011

Effort to leverage DCC advances to develop new modes of interaction and contribute lessons learned from building a nationally scoped aggregation to Digital Public Library of America initiative.

http://www.diglib.org/community/collaborations/dpla-beta-sprint/
Early Engagement with Europeana

- Technical exchange – peer aggregation in terms of many aspects of approach
- Many shared processes and areas of advancement
- Very different governance
- Important interactions around the Europeana Data Model and future of linked open data
Technical report proposing a formal extension of the Europeana Data Model that explicitly accommodates representation of collections and collection/item relationships.

- Characterizes concept of collection
- Discusses roles of collections in aggregations
- Outlines requirements for collection representation and description
- Provides recommendations for extending EDM
From “Digital Collections and Content”…

…to “Digital Collection Contexts”
On Digital Collection Contexts

“Collections” - prevalent and important

- Conceptual constructs
- Analog and digital constructions

Problems and potentials in digital environment

- Loss of presence
- New international collecting opportunities
Full-day workshop at iConference 2014

Around two core themes
  • Conceptual Foundations of Collections
  • Practical Implications of Collections Implementation

Organized by
  • Carole L. Palmer & Megan Senseney, CIRSS
  • Antoine Isaac, Europeana Foundation
  • Karen Wickett, University of Texas at Austin

Workshop Goals

• Broaden the conversation across an international community

• Further the research and development agenda for digital aggregations

• Relate conceptual advances to implementation goals

• Identify realistic approaches for collection representation, contextualization, and interoperability at scale
Panelists: Conceptual Foundation of Collections

**Hur-Li Lee**, School of Information Studies, University of Wisconsin-Milwaukee

**Karen M. Wickett**, School of Information, University of Texas at Austin

**Martin Doerr**, Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas

**Carlo Meghini**, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche

Panelists: Practical Implications of Collections Implementation

Amy Rudersdorf, Digital Public Library of America

Sheila Anderson, Centre for e-Research, King’s College London

Shenghui Wang, OCLC Research

Paul Clough, Department of Computer Science, University of Sheffield

Workshop Attendees

- **38** Registrants
- From **15** different countries on **4** different continents
CONCEPTUAL FOUNDATIONS OF COLLECTIONS
On the Notion of Collections

A retrospective overview presented by Hur-Li Lee.

Provides 15-year overview of scholarly literature on collections, beginning with her 2000 article “What is a Collection?”

Identifies three basic dimensions of collections:

- Characteristics and elements
- Objectives and functions
- Representation and organization

Is Collection Modeling Contextual Modeling?

A comparison of Cal Lee’s framework for contextual modeling to DCC/EDM recommendations for modeling collections by Karen Wickett.

Reviews six relevant classes of contextual entities:

- Object
- Relationship
- Purpose
- Agent
- Time
- Place

Based on analysis, collection modeling can act as contextual modeling for items in digital aggregation systems.

Unity Criteria: Why are Items Together?

*Martin Doerr on determining why items are together and whether they belong in a given collection.*

The reasons why items are together in a collection, are a key to better understand both the **processes of collecting** and the **information value** the membership in a collection adds to the item.

Four general categories for determining an item’s suitability for collection membership:

- Nature of Object
- Example Function
- Witness Function
- Aboutness
On the Logical Foundations of Digital Collections

An explanation of collection intension by Carlo Meghini.

Collections are not just containers. Two collections with the same content are not necessarily one and the same collection.

The condition of identity for collections is more complex than sameness of extension in a specific situation: it amounts to sameness of content in every possible situation.

A collection is therefore similar to a predicate symbol in logic:
  - it has an extension, that is a set of objects that belong in it
  - it has an intension, that is a property that defines its meaning
Morning Breakout Sessions

1) Unity criteria
   Facilitator: Martin Doerr

2) Formalizing collection structures
   Facilitator: Karen Wickett and Carlo Meghini

3) Scholarly use of collections
   Facilitator: Carole L. Palmer and Hur-Li Lee

4) Non-scholarly use of collection
   Facilitator: Antoine Isaac
Unity Criteria

“Start with material facts, from materials facts you can infer psychological relationships [motivations], from this you can draw conclusions about collective behavior and categorical principles”
~Martin Doerr

Alternate argument against modeling intension:
Intension can only be expressed as a historical fact.

Arrangement according to different institutional contexts.
Arrangement for potential personal research.

Physical objects vs. digital proxies.
Formalizing Collection Structures

“What can be said can be said clearly, and this is our manifesto” ~Carlo Meghini

Core questions:
• Is this kind of approach sufficient?
• Should we treat collections as objects?

Disagreement among the group on both accounts.
Scholarly Use of Collections

• Challenges
  • Retaining meaningful collection relationships at scale
  • Collections as dynamic, non-static research objects (for purposes of retrieval and re-use)

• What is the minimum criteria for…
  • Establishing a coherent collection?
  • Establishing authenticity?
  • Allowing variability?

• Open Questions
  • Can we categorize users as collections? Collections of users?
  • Can we integrate interpretive structures in scholarly collections? Which versions of interpreted data should be shared?
Non-Scholarly Use of Collections

Emphasizing users’ perspective over managers’ perspective.

Is it realistic or cost-effective to create collection-level data in DPLA and Europeana?

General agreement that collections are valuable
But what is the cost-benefit ratio of representation?
PRACTICAL IMPLICATIONS OF COLLECTIONS IMPLEMENTATION
Implementing Collection Contexts, and Metadata Issues Related to Normalization and Shareability

An account of data remediation and collection representation challenges at the DPLA by Amy Rudersdorf.

DPLA metadata application profile includes `dpla:Collection` class, but use is inconsistent and varies widely.

• Collection information may be found in other values of DC records or at the OAI-PMH set level.
• Collections may be conceived differently by different institutions.

DPLA plans to present collection information as data variations are addressed and remediated.

http://dp.la
The CENDARI Knowledge Framework

The Collaborative Digital Archive Research Infrastructure (CENDARI) was presented by Sheila Anderson as a dynamic research environment for humanities scholars.

Goal: facilitate and enhance research by increasing access to and use of records of historic importance across Europe

Challenges discovered through user studies:

- Collections and archives are often fragmented and dispersed
- Epistemological differences in approaches to knowledge making by historians and librarians/archivists
- Greater need for structured documents that leverage semantic linking

CENDARI addresses challenges by combining

- the development of an integrated metadata strategy with
- the development of dynamic domain ontologies.

http://www.cendari.eu/
Hunting for Semantic Clusters in Aggregations

An initiative to develop object clustering for use in the case of heterogeneous datasets presented by Shenghui Wang.

Work to date:
Automated clustering process used to generate hierarchies of records based on similarity.

Open research:
• combine current clustering algorithm with user-defined constraints to generate more meaningful topical clusters
• Provide greater semantic enrichment to item-level metadata to experiment with multi-dimensional clustering
Supporting Users’ Navigation and Exploration of Large Digital Collections

A project using metadata enrichment and object linking for personalized exploration of aggregations presented by Paul Clough.

PATHS (Personalised Access To cultural Heritage Spaces):

• Enriched data from Europeana using techniques from natural language processing and information retrieval.
• Designed system incorporating various navigational and exploratory search aids.
• Allowed users to create narrative trails/paths through the data, including:
  • workspace feature to store items during exploration
  • editing feature for arranging and structuring gathered items
  • functions for sharing the paths

http://www.paths-project.eu/
Afternoon Breakout Sessions

1) Digital library aggregation and interoperability
   Facilitator: Amy Rudersdorf and Antoine Isaac

2) Data enrichment for collections
   Facilitator: Paul Clough and Shenghui Wang

3) Use scenarios for collections
   Facilitator: Carole L. Palmer and Sheila Anderson

4) Collection visualization
   Facilitator: Marian Dörk and Karen Wickett
Digital Library Aggregation and Interoperability

Relationship between Aggregator and Provider:

- Data transformation between provider and aggregator is continuous
- Clear need to identify business relationship, what are the processes between the provider and aggregator
- Must document mappings more clearly
- Aggregator and provider terminologies may not match

Recommendation:

Develop a reference model that outlines a given set of processes and assigns responsibility for implementation
Data Enrichment for Collections

Improving data quality through enrichment
  • Policies and standards
  • Who are the stakeholders?

Open questions:
  • Levels of enrichment
  • Maintaining clarity of provenance for enrichment (and including confidence ratings for enriched data)
  • Reliability of non-authoritative sources (e.g., wikipedia)
Use Scenarios for Collections

How do you leverage what people are actually doing with content in ways that enrich your system?

Three emerging themes:
1) Journey/narrative
2) Disciplinary differences
3) Clustering
Collection Visualization

The most effective visualizations speak to the special characteristics of the collection, but different communities might find different aspects significant…

Know communities:

- Aggregators/Administrators
- Scholars
- General Public

Visualization can be used to provide views that are specialized for different audiences as a complementary feature that doesn’t compromise authoritative practices of information organization.
OUTCOMES AND NEXT STEPS
Outcomes

• Clear community of interest in digital collection
  • Attendees requested to form a group for ongoing discussions
  • Suggested formats include listserv or wiki

• Indication that collections issues extend beyond scholarly community to general public

• Many open questions for ongoing research and development…
Emerging Research Agendas

Participants identified the following areas as next steps for their research:

- Reconstructing lost or previously non-existent collection descriptions
- Collections visualization
- Developing processes and business plans for aggregations
- Relating collection description with overarching content strategies
- User-defined collections in relation to semantic clusters
- Minimum criteria for establishing new, user-generated collections
- Offering alternative representations for different user groups
- Convergence of collections from memory institutions with broader cultural dialogues
CIRSS Next Steps

• Workshop report

• Proposal development
  • recovering lost collection contexts
  • supporting scholarly creation of new, referential collections

• Synergies
  • Collection contexts in science data repositories
  • HTRC and Cultural Heritage Aggregations
Open for discussion

• Relating Wang’s semantic clustering work and Clough’s PATHS project to collection description and representation.

• Addressing needs of the multiple user communities for large-scale aggregations
  • Epistemological differences across scholarly disciplines
  • Scholarly vs. non-scholarly use of collections
  • Etc.
Acknowledgements