OCLC Research and Europeana

Shenghui Wang
Research Scientist
OCLC

Valentine Charles
Interoperability Specialist
Europeana
OCLC Research is one of the world’s leading centers devoted exclusively to the challenges facing libraries and archives in a rapidly changing information technology environment. Our mission is to expand knowledge that advances OCLC’s public purposes of furthering access to the world’s information and reducing library costs. Since 1978, we have carried out research and made technological advances that enhance the value of library services and improve the productivity of librarians and library users.
OCLC Research: Three roles

1. To act as a community resource for shared Research and Development (R&D)
2. To provide advanced development and technical support within OCLC itself
3. To enhance OCLC’s engagement with members and to mobilize the community around shared concerns.

http://www.oclc.org/research.html
OCLC Research – 3 constituencies

OCLC Membership

OCLC Services

OCLC Research
Library Partnership
OCLC Research work agenda

1. Research Information Management
   - Opportunities for libraries in support of research process and outputs

2. Mobilizing Unique Materials
   - Describe, disclose, discover, deliver effectively

3. Metadata Support and Management
   - New models, workflows for network level services

4. Infrastructure and Standards Support
   - Support new architectures and their adoption

5. System-wide Organization
   - Cooperative models of acquiring and managing collections

6. User behavior studies & Synthesis

---

DEFINE FUTURE RESEARCH LIBRARY SERVICES – REVITALIZE OUR VALUE PROPOSITION

TRANSFORM OUR CURRENT OPERATING PRACTICES AND PROCESSES – IMPLEMENT SYSTEMIC CHANGE
OCLC Research Library Partnership

156 Partners at January 2012

50% of ARL
63% of RLUK

25 of top 30 in the World University Rankings
Strength/weakness

OCLC Research in Europe

• Strength:
  – 50 experts dedicated to innovation for the library community globally
  – Applied research, hands-on
  – Little overhead
  – No political/commercial agenda
  – Results are shared and in the open

• Weaknesses:
  – European partners in the minority, cultural/language differences
  – ORLP partnership weak on the continent; little awareness
  – Image problem (OCLC as vendor; strong association with metadata)
  – OCLC IPR regime with metadata needs clarification
Positioning OCLC Research in Europe

Develop a strategy

• ORLP: too few members in Europe => no impactful cooperation opportunities yet
• Choose for strategic cooperation with influential consortia: The European Library, Europeana, Open Planets Foundation (OPF)
• Make use of the networking strength of existing associations in Europe: LIBER
Positioning OCLC Research in Europe

Develop a strategy

- Encourage European partners to participate in ongoing OCLC Research activities
- Engage with existing networks in areas where OCLC Research can help make a difference
Outline of an European Research Programme

Three collaboration areas:

1. with Europeana: Innovation pilots
2. with OPF: Preservation Health Check pilot
3. with national libraries: Develop strategies for the scalable and sustainable management of digital collections.
Leading to:

1. **Metadata quality services** (dedup, enrichment, intelligent clustering, NER and automatic tagging)
2. **Health check services** (quality assessment, risk assessments)
3. **Good practices** for the scalable and sustainable management of digital collections and infrastructures
4. **Usage data analysis** (web site traffic, added value of aggregations, hard data on real user behaviour)
Explore Europe's cultural collections

Playlist: Jean-Phillipe Rameau
Featured Search: The Museum of Architecture, Berlin
Exhibition: European Sport Heritage
Featured search: The Saratov State Art Museum collection

From the blog

Caravaggio: a Murderous Artist
The personalities featured on the blog this week do nothing to debunk the idea of artists having fiery temperaments. On Tuesday, we told the tale of Baroque...

Borromini: rivalry, illusion and tragedy
Greatness, illusion, bitter rivalry, and a tragic end. It's a life story that could easily have been penned for a Hollywood epic drama. Today we are ce...

EuroPEACEana: What does peace mean to you?
Today, 21 September, is the International Day of Peace. Established by a United Nations resolution, the first event of its kind took place in

Featured item

Malden
Vladimir Dimitrov (1882-1960) was a Bulgarian painter, considered to be one of the countries.

Follow us on Pinterest
Ceramic section from the front of an alter in Lisbon, Portugal. 1670 - Unknown creator. Museu Nacional de Machado de Castro (Public Domain)
A short introduction on Europeana

- Europeana is a service that aggregates data from the cultural heritage sector in Europe.
  - libraries, museums, archives and audio-visual archives
  - [http://www.europeana.eu/](http://www.europeana.eu/)
- Provides a portal for users to access that data
  - Metadata, previews and links to source
- Will make the metadata freely available for anyone to re-use
  - under Creative Commons Zero (CC0) -public domain dedication
- Enriches data, provides tools
  - Link to data from other sites, embed on wikipedia, API
- Makes data available as Linked Open Data
  - [http://data.europeana.eu/](http://data.europeana.eu/)
Context of collaboration between OCLC & Europeana

• In Europeana:
  – R&D is driven by funded EU projects
  – Aggregation of metadata from heterogeneous collections leads to data quality challenges

• OCLC Research has extensive experience and provides expertise in metadata quality management.

• The collaboration serves research objectives which are open-ended.
Innovation pilot 1

– Connect as many Europeana objects (books, paintings, etc) to resources of the Virtual International Authority file.

• Europeana is currently enriching resources that represent places, time periods, concept and persons with selected vocabularies and datasets.

The Raven
Creator: Poe, Edgar Allan (1809-1849)
Date: 1875
Type: TEXT
Subject: Lithographie -- France
Description: Allan Poe (1809-1849), "spiritualist" and author of "The Raven" in Lemerre, who was one of the founders of "Parisian Neoclassicism." http://viaf.org/viaf/60351476
Innovation pilot 1

– The Europeana case is quite different from many library-focused ones
  • Persons are referred to in the simple ESE (Europeana Semantic Element) metadata
  • There is no indirect linking, for example, via a reference to an authority number used at a national library.

– The project would allow an improvement of the enrichment process.
Innovation pilot 2

• Connect related Europeana records
  – Detect duplicates or near-duplicates
  – Identify and create semantic links between objects that are related
    • translated copies of the same publication
    • a painting and a photograph of that painting
    • different editions of one book, or
    • a collection of letters that belong to the same person.
Current situation in Europeana

- A current related items feature already exists
  - based on the enrichment fields what, who, where, when and the similarities in the metadata fields such as dc:title and dc:description.
  - But an improvement of the enrichment process would be needed to make the relations more explicit.

Schoemaker Atlas: Utrecht, Maartensdijk

Explore further!

Similar content
1. Rough clustering millions of records into small clusters
   - Clustering 1 million records costs less than one minute
     • Using min-hashes, compression-based similarity measures, parallel computing
   - Using different similarity thresholds for a hierarchical view of objects

2. Categorising clusters and identifying specific semantic links within clusters.
Analysis of the results

– A selection of clusters have been analysed.
  • Selection of examples
  • Formulation of hypothesis of the cluster generation
  • Comparison of the clusters with the similar items found in the Europeana portal

– Clusters have been categorised
Clusters overview

Europeana relevant clusters

- Parts of one CHO
- Duplicates objects
- Derivative works
- Views of the same CHO
- Thematic clusters
  - CHO #s related to one location
  - CHO #s related to an event

Collections

Other complete collections
Personal collections
Categories of clusters

- **Same objects/duplicates**
  - clusters with same objects that have been either:
    - provided more than once to Europeana within the same dataset or via two different channels.
    - duplicated during the Europeana ingestion process (quality issue)
Categories of clusters

- Parts of one Cultural Heritage Object (CHO)
  - clusters of objects that are structurally composed of other objects/parts.

Shared metadata

- dcterms_spatial : City Of London
- dcterms_spatial : 531500, 181500
- dcterms_isPartOf : Victorian popular music. Collect Britain
- dc_date : [1873]
- dc_subject : 780
- dc_subject : Sheet music -- Great Britain -- History -- 19th century
- dcterms_created : 1873
- dc_creator : Composer : Dallas, John

- 20240196 : The Oil Shop part 03
- 20240197 : The Oil Shop part 04
- 20240198 : The Oil Shop part 05
- 20240199 : The Oil Shop part 06
- 20240200 : The Oil Shop part 07
- 20240201 : The Oil Shop part 08
Categories of clusters

– **Views of the same CHO**
  - clusters of objects which have multiple representations. Each representation offers a different view of the CHO.
    - In most of the case metadata is the same. It would be possible to attach all these views to the same record.

– **Derivatives works**
Categories of clusters

– **Thematic clusters**
  
  • These clusters are often too small to be considered as a complete collection. They have in common some metadata that relate them to a similar topic, location, event…
  
  • Depending of the focus, the way we define the CHO they could be considered as different views of the same CHO.

– **Collections**

  • 21991936 : Letter from Dr Evans to Capt. Livingston (2)
  • 21991937 : Letter from Capt. John Livingston RAMC (1)
  • 21991938 : Letter from Capt. John Livingston RAMC (2)
  • 21991939 : Letter from Capt. John Livingston RAMC (3)
  • 21991940 : Letter from Capt. John Livingston RAMC (4)
  • 21991941 : Letter from Capt. John Livingston RAMC (5)
  • 21991942 : Letter from Capt. John Livingston RAMC (6)
  • 21991961 : Two letters from Capt. John Livingston RAMC (1)
  • 21991964 : Two letters from Capt. John Livingston RAMC (2)
  • 21991965 : Two letters from Capt. John Livingston RAMC (3)
  • 21991966 : Two letters from Capt. John Livingston RAMC (4)
  • 21991967 : Two letters from Capt. John Livingston RAMC (5)
  • 21991968 : Two letters from Capt. John Livingston RAMC (6)
  • 21991969 : Two letters from Capt. John Livingston RAMC (7)
Findings

– On the clusters
  • Clusters are generally good but are limited to close relationships

– On the data use for the research
  • Quality issues in the data
    – Standard are interpreted differently by providers despite the presence of guidelines
    – Creation of digital object is not always in line with the creation of descriptive metadata
  • Logical structure of cultural heritage object is not always reflected in the metadata.
Next steps (1)

- Re-use the categories to find ways of automatizing the finding of such categories.
  - some cluster categories may be deduced from common metadata values in given fields
  - Patterns might exist for each type of categories.

- Categorise the clusters in terms of FRBR entities and relation (like a manifestation of an expression).

- Experiment with visualization methods.
Next steps (2)

- Applying the types of relations available in EDM to the types of clusters found during the experiment.
  - `dc:subject`, `edm:isRepresentationOf` for "aboutness" links (Mona Lisa and a historical picture of Mona Lisa)
  - `edm:realizes`, which is quite FRBR-related (An item of the Gutenberg’s edition realizes the Bible)
  - `edm:isSimilarTo` (covering true and cases of derivation) and its sub-properties `edm:isDerivativeOf` (for real derivation cases like re-working, extension), `edm:incorporated` (for inclusion / re-use) and `edm:isSuccessorOf` (for "sequels")
  - more general links (`dc:relation`), general part-whole relation (`dcterms:hasPart`), citation (`dcterms:references`), direct versioning links (`dcterms:hasVersion`).

- Findings from the pilot could feed into best practice guides for content providers and thereby improve the quality of the whole Europeana dataset.
Everyone is happy

Mutual benefits

Methods

Clustering and enrichment innovation

Results

Europeana data model

New browsing experiences

OCLC internal data (digital gateway, worldcat, etc)

Data services for third parties
Thank you!

Valentine Charles at valentine.charles@kb.nl
Shenghui Wang at shenghui.wang@oclc.org