Hunting for Semantic Clusters

Shenghui Wang\textsuperscript{1} \hspace{1em} Antoine Isaac\textsuperscript{2} \hspace{1em} Valentine Charles\textsuperscript{2} \\
Rob Koopman\textsuperscript{1} \hspace{1em} Anthi Agoropoulou\textsuperscript{2} \hspace{1em} Titia van der Werf\textsuperscript{1}

\textsuperscript{1}OCLC Research, Leiden, The Netherlands \\
\textsuperscript{2}Europeana Foundation, The Hague, The Netherlands

KnowEscape2013
Outline

1. Introduction

2. Hierarchically structuring Cultural Heritage objects

3. Results and evaluation
1. Introduction

2. Hierarchically structuring Cultural Heritage objects

3. Results and evaluation
Online Computer Library Center (OCLC), founded in 1967, is a nonprofit, membership, computer library service and research organisation dedicated to the public purposes of furthering access to the world's information and reducing library costs.

OCLC and its member libraries cooperatively produce and maintain WorldCat (www.worldcat.org)
WorldCat statistics (as of June 30 2012)

- WorldCat encompasses records for books, serials, sound recordings, musical scores, maps, visual materials, mixed materials and computer files.
- 273 million records and 1.85 billion holdings (reached 2 billion on 4 May 2013)
- Items from more than 74,000 libraries, 170 countries and 485 languages
- Date ranging from 4800 B.C. to the present.

http://www.oclc.org/content/dam/oclc/publications/AnnualReports/2012/2012.pdf
Linked open data

Diagram showing interconnected data sources such as WorldCat, VIAF, FAST, DDC, etc.
Large-scale aggregators

Domain Aggregators

- Libraries
  - e.g. The European Library
- Archives
  - e.g. APEX

National initiatives

- National Aggregators
  - e.g. Culture Grid, Culture.fr
- Regional Aggregators
  - e.g. Musées Lausannois
- Thematic collections
  - e.g. Judaica Europeana, Europeana Fashion
Pair of silver rimmonim

Description: A pair of late eighteenth century Italian silver rimmonim.
Repository/Location: Jewish Museum London
Geographic coverage: Italy [Place] [Production]
Type: rimmonim
Format: 406 mm; silver
Identifier: [local JML 58095809] [Metadata]
Rights: Digital image: Jewish Museum London
Source: Jewish Museum London

View item at Jewish Museum London
Duplicates?

- Same objects, different providers

Shipping In The Open Sea

**Description:** This is a fine seascape by the North East artist John Wilson Carmichael. The picture uses one of his favourite types of composition. A small boat filled with fishermen in the foreground points to the most important ship, which is highlighted by a patch of rough, white-capped waves. This ship itself points to another

Shipping in the Open Sea

**Description:** This is a fine seascape by North-East artist John Wilson Carmichael. The picture uses one of his favourite types of composition. A small boat filled with fishermen in the foreground points to the most important ship, which is also highlighted by a patch of rough, white-capped waves. This ship itself points to another ship further away on the left. The zig-zag line of the ships creates lively movement in the picture. The
Duplicates?

- Same objects, different providers
- Same page digitised three times
Challenges in large-scale aggregators

- Aggregation of metadata from heterogeneous collections leads to data quality issues (e.g., duplicates)
Challenges in large-scale aggregators

- Aggregation of metadata from heterogeneous collections leads to data quality issues (e.g., duplicates)
- Mapping from different formats and vocabularies to a shared data model may cause information missing (e.g., internal and external links between objects)
Challenges in large-scale aggregators

- Aggregation of metadata from heterogeneous collections leads to data quality issues (e.g., duplicates)
- Mapping from different formats and vocabularies to a shared data model may cause information missing (e.g., internal and external links between objects)
- Cultural Heritage objects could be linked differently (e.g., duplication, depiction/representation, derivation, succession, etc.)
Challenges in large-scale aggregators

- Aggregation of metadata from heterogeneous collections leads to data quality issues (e.g., duplicates)
- Mapping from different formats and vocabularies to a shared data model may cause information missing (e.g., internal and external links between objects)
- Cultural Heritage objects could be linked differently (e.g., duplication, depiction/representation, derivation, succession, etc.)
- Keyword-based search does not provide end users a global overview of what is available.
Outline

1. Introduction

2. Hierarchically structuring Cultural Heritage objects

3. Results and evaluation
Hierarchical structuring based on levels of similarity

Our method contains three parts:

- Fast clustering algorithm based on minhashes and compression similarity
- Field selection for focal semantic clusters
- Hierarchically structuring records based on similarity
Fast clustering based on minhashes and compression similarity

Two-step approach:
- Grouping records which could potentially be further clustered
  - Transform metadata into a set of minhashes
  - Group records with similar minhashes
- Iterative parallel clustering records based on compression similarity
  - Select cluster heads which are far apart
  - Greedily assign records to the closest cluster head
  - Divide clusters if the clusters are not “compact” enough

By varying the similarity level, clusters with different compactness can be produced.
Fast clustering based on minhashes and compression similarity

Two-step approach:

- Grouping records which could potentially be further clustered
  - Transform metadata into a set of minhashes
  - Group records with similar minhashes

- Iterative parallel clustering records based on compression similarity
  - Select cluster heads which are far apart
  - Greedily assign records to the closest cluster head
  - Divide clusters if the clusters are not "compact" enough

- By varying the similarity level, clusters with different compactness can be produced.
Focal semantic clusters

- At level 100, many clusters contain duplicates, or records with almost identical metadata.

---

Focal semantic clusters

- At level 100, many clusters contain duplicates, or records with almost identical metadata.
- At level 80, many clusters are of specific interests, e.g., pages of the same book, pictures of the same building, etc.
Focal semantic clusters

- At level 100, many clusters contain duplicates, or records with almost identical metadata.
- At level 80, many clusters are of specific interests, e.g., pages of the same book, pictures of the same building, etc.
- These *focal semantic clusters* often represent small cultural entities, which can be connected to other entities.
Field selection for focal semantic clusters

- However, different data providers do not apply same standards in the same way.
  - Same information could be put into different metadata fields
  - The extent to which an object is described varies a lot provider by provider.
Field selection for focal semantic clusters

- However, different data providers do not apply same standards in the same way.
  - Same information could be put into different metadata fields
  - The extent to which an object is described varies a lot provider by provider.
- Not all metadata fields should be used for clustering.
  - Otherwise, the pages of one book are scattered in multiple clusters.
However, different data providers do not apply same standards in the same way.

- Same information could be put into different metadata fields
- The extent to which an object is described varies a lot provider by provider.

Not all metadata fields should be used for clustering.

- Otherwise, the pages of one book are scattered in multiple clusters.

We applied a standard Genetic Algorithm to automatically select the important fields which give the best focal clusters.
Hierarchical structuring based on levels of similarity

Level 100
Original
G.A.
Repr.
Field selection 1
Provider 1
Field selection 2
Provider 2
Field selection 3
Provider 3
Level 80
Level 60
Level 40
level 20
Shenhui Wang, Antoine Isaac, Valentine Ch
Hunting for Semantic Clusters
KnowEscape2013 17 / 27
Hierarchy example
1 Introduction

2 Hierarchically structuring Cultural Heritage objects

3 Results and evaluation
Experiments with a small dataset

- We applied the method on 1.1 million records from the UK.
- Manually check randomly chosen clusters and try to understand what made these records clustered together, i.e., identify the semantic links between records.
Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings
Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings

Bowburn, boiler house
Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings
Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings

Letter from Capt. John Livingston
RAMC
Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings

Categories of clusters

- Same objects/duplicate records
- Views of the same object
- Derivative works
- Parts of the same object
- Collections
- Thematic groupings

"Rural life"
Working with the full Europeana dataset

- 23.6M records from 2428 data providers across Europe (a data dump on Feb 2013)
Working with the full Europeana dataset

- 23.6M records from 2428 data providers across Europe (a data dump on Feb 2013)
- At level 100, we found more than 200K clusters which contain highly similar records
Working with the full Europeana dataset

- 23.6M records from 2428 data providers across Europe (a data dump on Feb 2013)
- At level 100, we found more than 200K clusters which contain highly similar records
- At level 80, we found nearly 1.5 million focal clusters from all individual data providers.

<table>
<thead>
<tr>
<th>Similarity level</th>
<th>#Records to be clustered</th>
<th>#Clusters</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>23,595,555</td>
<td>200,245</td>
<td>6m2.82s</td>
</tr>
<tr>
<td>80</td>
<td>23,595,555</td>
<td>1,476,089</td>
<td>*</td>
</tr>
<tr>
<td>60</td>
<td>6,407,615</td>
<td>382,268</td>
<td>3m35.26s</td>
</tr>
<tr>
<td>40</td>
<td>2,431,753</td>
<td>212,389</td>
<td>2m28.79s</td>
</tr>
<tr>
<td>20</td>
<td>1,068,188</td>
<td>84,554</td>
<td>1m20.99s</td>
</tr>
</tbody>
</table>

Working with the full Europeana dataset

- 23.6M records from 2428 data providers across Europe (a data dump on Feb 2013)
- At level 100, we found more than 200K clusters which contain highly similar records
- At level 80, we found nearly 1.5 million focal clusters from all individual data providers.

<table>
<thead>
<tr>
<th>Similarity level</th>
<th>#Records to be clustered</th>
<th>#Clusters</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>23,595,555</td>
<td>200,245</td>
<td>6m2.82s</td>
</tr>
<tr>
<td>80</td>
<td>23,595,555</td>
<td>1,476,089</td>
<td>*</td>
</tr>
<tr>
<td>60</td>
<td>6,407,615</td>
<td>382,268</td>
<td>3m35.26s</td>
</tr>
<tr>
<td>40</td>
<td>2,431,753</td>
<td>212,389</td>
<td>2m28.79s</td>
</tr>
<tr>
<td>20</td>
<td>1,068,188</td>
<td>84,554</td>
<td>1m20.99s</td>
</tr>
</tbody>
</table>

Field selection for focal semantic clusters

- For the 10 providers with most records (covering 35% of the whole Europeana dataset), it took 161 minutes on average.
- Datasets with 200-250 records cost 21 minutes on average.
Field selection for focal semantic clusters

- For the 10 providers with most records (covering 35% of the whole Europeana dataset), it took 161 minutes on average.
- Datasets with 200-250 records cost 21 minutes on average.

<table>
<thead>
<tr>
<th>#Providers</th>
<th>metadata field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>dc:title</td>
</tr>
<tr>
<td>2</td>
<td>dc:type</td>
</tr>
<tr>
<td>3</td>
<td>dc:language</td>
</tr>
<tr>
<td>4</td>
<td>dc:rights</td>
</tr>
<tr>
<td>5</td>
<td>dc:subject</td>
</tr>
</tbody>
</table>

(a) Top 10 most selected fields

<table>
<thead>
<tr>
<th>#Providers</th>
<th>field combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>dc:title</td>
</tr>
<tr>
<td>2</td>
<td>dc:title dc:type</td>
</tr>
<tr>
<td>3</td>
<td>dc:title dc:creator</td>
</tr>
<tr>
<td>4</td>
<td>dc:title dc:identifier</td>
</tr>
<tr>
<td>5</td>
<td>dc:description</td>
</tr>
</tbody>
</table>

(b) Top 5 most selected field combinations
Randomly select 100 clusters at each level
7 evaluators categorised these clusters, based on the categories found in the first round
Manual evaluation

- Randomly select 100 clusters at each level
- 7 evaluators categorised these clusters, based on the categories found in the first round

<table>
<thead>
<tr>
<th>Cluster Category</th>
<th>100</th>
<th>80</th>
<th>60</th>
<th>40</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same objects/duplicate records</td>
<td>11</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Views of the same object</td>
<td>61</td>
<td>33</td>
<td>6</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Parts of an object</td>
<td>10</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Derivative works</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Collections</td>
<td>1</td>
<td>4</td>
<td>27</td>
<td>13</td>
<td>43</td>
</tr>
<tr>
<td>Thematic grouping</td>
<td>9</td>
<td>34</td>
<td>36</td>
<td>29</td>
<td>22</td>
</tr>
<tr>
<td>Nonsense</td>
<td>2</td>
<td>3</td>
<td>30</td>
<td>57</td>
<td>28</td>
</tr>
</tbody>
</table>
Conclusions

- Finding similar CH objects is the first step towards identifying semantic links and groups of objects within large-scale aggregations.
- CH objects can be linked via different types of relations which can be used in different applications.
- Hierarchical structuring invites new ways of exploring CH collections.
Sir James Eyre (1734-1799), Chief Justice of the Common Pleas (Government Art Collection)
Multidimensional similarities

Sir James Eyre (1734-1799), Chief Justice of the Common Pleas (Government Art Collection)

Sir John Eardley Wilmot (1709-1792) Chief Justice of the Common Pleas (Government Art Collection)

Eyre, James (Austrian National Library)

Shenghui Wang, Antoine Isaac, Valentine Ch
Hunting for Semantic Clusters
KnowEscape2013
Continue hunting for semantic clusters

Thank you!

Shenghui Wang (shenghui.wang@oclc.org)
Antoine Isaac (aisaac@few.vu.nl)
Valentine Charles (Valentine.Charles@kb.nl)
Rob Koopman (Rob.Koopman@oclc.org)
Anthi Agoropoulou (anthi.agoropoulou@gmail.com)
Titia van der Werf (titia.vanderwerf@oclc.org)