Entity Management at Europeana

Antoine Isaac
Europeana Foundation
with slides from Hugo Manguinhas and Dasha Moskalenko

DCMI Virtual 2021, 14 October 2021
Europeana.eu

- An initiative from the European Union
- Publishes 52M digitized objects
- Using metadata from 4,000 libraries, archives and museums in 44 countries
Enriching metadata

The Europeana Data Model (EDM) gives a base for (linking to) multilingual metadata

We encourage data providers to contribute their own links/data to local or external vocabularies
Currently recognized by Europeana in providers' metadata:

- The Getty - Art & Architecture Thesaurus (AAT) [http://vocab.getty.edu/aat/](http://vocab.getty.edu/aat/)
- The Getty - Union List of Artist Names (ULAN) [http://vocab.getty.edu/ulan/](http://vocab.getty.edu/ulan/)
- Getty Thesaurus of Geographic Names (TGN) [http://vocab.getty.edu/tgn/](http://vocab.getty.edu/tgn/)
- Virtual International Authority File (VIAF) [http://viaf.org/viaf/](http://viaf.org/viaf/)
- Geonames [https://sws.geonames.org/](https://sws.geonames.org/)
- Gemeinsame Normdatei (GND) [http://d-nb.info/gnd](http://d-nb.info/gnd)
- Library of Congress Subject Headings (LCSH) [http://id.loc.gov/authorities/subjects/](http://id.loc.gov/authorities/subjects/)
- UDC [http://udcdata.info/rdf/](http://udcdata.info/rdf/)
- UNESCO Thesaurus [http://vocabularies.unesco.org/thesaurus/](http://vocabularies.unesco.org/thesaurus/)
- Wikidata [http://www.wikidata.org/entity/](http://www.wikidata.org/entity/)
Enriching metadata

EDM gives a base for (linking to) multilingual metadata

We encourage data providers to contribute their own links/data to local or external vocabularies

In parallel, we apply automatic enrichment to link object metadata to reference datasets for places, persons, concepts
We are building an "Entity Collection":

- Centralized point of reference and access to data about contextual entities: places, agents, concepts...
- Caching and curating data from the wider Linked Open Data
- A sort of Europeana knowledge graph

Available via our Entity API: [https://pro.europeana.eu/page/entity](https://pro.europeana.eu/page/entity)
Data currently in the Entity Collection

216,302 Places
- a subset of Geonames, corresponding to places which are part of European countries and of some specific feature classes

165,005 Agents
- a subset of DBpedia corresponding to most of the instances of dbp:Artist with some exceptions, and integrated from 49 DBpedia language editions

1,573 Concepts
- a subset of DBpedia corresponding to a selection of concepts matching the needs from Europeana Collections (e.g., WWI battles).
- Europeana Sounds music genres (obtained from Wikidata)
- Photo Consortium’s photography vocabulary

3,304 Organizations
- Extracted from Europeana’s CRM and aligned to Wikidata when possible

21 Timespans
- Centuries, assembled manually
Entity data example

```json
{
  "@context": "http://data.europeana.eu/entity/context.json",
  "id": "http://data.europeana.eu/agent/base/146351",
  "type": "Agent",
  "preferredLabel": {
    "en": "Wolfgang Amadeus Mozart",
    "cs": "Wolfgang Amadeus Mozart",
    "de": "Wolfgang Amadeus Mozart",
    "fr": "Wolfgang Amadeus Mozart",
    "it": "Wolfgang Amadeus Mozart",
    "ja": "ワーグン・アマデウス・モーツァルト",
    "ko": "모차르트"
  },
  "allLabel": {
    "en": "Wolfgang Amadeus Mozart",
    "cs": "Wolfgang Amadeus Mozart",
    "de": "Wolfgang Amadeus Mozart",
    "fr": "Wolfgang Amadeus Mozart",
    "it": "Wolfgang Amadeus Mozart",
    "ja": "モーゼル・アマデウス",
    "ko": "모차르트"
  },
  "biographicalInformation": {...},
  "dateOfBirth": [
    "1756-01-27"
  ],
  "dateOfDeath": [
    "1791-12-05"
  ],
  "end": "1791-12-05",
  "identifier": ["32197206"],
  "placeOfBirth": [
    "Salzburg"
  ],
  "placeOfDeath": [" liners"]
}
```

- **Inter-linking information**
- **Preferred labels for many languages**
- **Coreference links to other datasets (e.g. DBpedia, Wikidata)**
Use cases for the Entity Collection

Improve user experience on Europeana services

- **Findability**: search with and for people, places and subjects. In many more languages, and with less ambiguity
- **Contextualization**: users see context about objects. Entity Pages group objects and link to other entities

Search for **verme**

Search for **Johannes Vermeer**

Search for **Jan Cornelisz Vermeyen**

Search for **Frederik Vermeiren**

Search for **Jan Vermeer van Utrecht**

Semantic auto-completion

Entity Pages
Multilinguality

PERSON

Johannes Vermeer

COLLECTIONS YOU MIGHT LIKE

Watercolor painting  Drawing  Engraving  Calligraphy

Type of media  Can I use this?  Providing country  More filters

Results: 43

Images of drawings and engravings by Johannes Vermeer.
Johannes Vermeer

Multilinguinity
Browsing entities
Organizing and promoting organizations *(coming Oct 21)*
Entity Management & Curation
Basic Entity Collection process
Example: integrating DBpedia resources

Dumps were carefully selected and downloaded for all languages that Europeana supports...

RDF dumps
48 Language Editions (~3.6GB/GZ)

SPARQL queries select DBpedia resources for each EDM Contextual Class

Each DBpedia resource is converted to EDM using XSLT and further filtered

http://data.dws.informatik.unimannheim.de/dbpedia/2014/
Example: organizations

To make it possible, we have:

- Extracted organization data from our CRM
- Translated all organization names into English
- Augmented our metadata with Wikidata
- Linked Europeana items to organizations
A new service for
  ● managing entities and data sources
  ● editing entity metadata

So far adding new entities was very time consuming
With the new service, new entities can now be added using external linked data sources (e.g. Wikidata, VIAF, AAT)

Some entities have descriptions that are not a great fit for Europeana
Selected users will be able to edit the description for entities
Showcasing the best items on an Entity Page

- Partners will be able to curate the top items appearing on an Entity Page
- Developed as a “collaborative” user gallery associated with an Entity
- Suggesting new items using a recommendation engine
### Entity statistics

- **Overview per language and per type of entity**
- **Allows us to track the multilingual coverage and general curation efforts**

#### Entities per Language

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Agent</th>
<th>Concept</th>
<th>Place</th>
<th>Timespan</th>
<th>Organization</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>100</td>
<td>93.8</td>
<td>5.79</td>
<td>100</td>
<td>78.5</td>
<td>47</td>
</tr>
<tr>
<td>de</td>
<td>27.5</td>
<td>56.3</td>
<td>2.77</td>
<td>100</td>
<td>11.3</td>
<td>13.6</td>
</tr>
<tr>
<td>fr</td>
<td>26.7</td>
<td>69.3</td>
<td>2.48</td>
<td>100</td>
<td>9.53</td>
<td>13.2</td>
</tr>
<tr>
<td>fi</td>
<td>8.77</td>
<td>36.2</td>
<td>15.8</td>
<td>100</td>
<td>3.63</td>
<td>12.8</td>
</tr>
<tr>
<td>it</td>
<td>17.2</td>
<td>55.1</td>
<td>5.64</td>
<td>100</td>
<td>5.99</td>
<td>10.8</td>
</tr>
<tr>
<td>es</td>
<td>19</td>
<td>60.7</td>
<td>1.68</td>
<td>100</td>
<td>6.63</td>
<td>9.38</td>
</tr>
<tr>
<td>sv</td>
<td>12.5</td>
<td>40.2</td>
<td>5.33</td>
<td>100</td>
<td>5.05</td>
<td>8.54</td>
</tr>
<tr>
<td>nl</td>
<td>12.2</td>
<td>60.9</td>
<td>2.25</td>
<td>100</td>
<td>13.1</td>
<td>6.84</td>
</tr>
<tr>
<td>pl</td>
<td>11.9</td>
<td>46.2</td>
<td>2.08</td>
<td>100</td>
<td>4.78</td>
<td>6.46</td>
</tr>
<tr>
<td>pt</td>
<td>11.7</td>
<td>46.9</td>
<td>2</td>
<td>100</td>
<td>5.21</td>
<td>6.36</td>
</tr>
<tr>
<td>bg</td>
<td>2.93</td>
<td>26.8</td>
<td>3.77</td>
<td>100</td>
<td>1.63</td>
<td>3.49</td>
</tr>
<tr>
<td>cs</td>
<td>6.08</td>
<td>37.3</td>
<td>0.74</td>
<td>100</td>
<td>3.12</td>
<td>3.2</td>
</tr>
<tr>
<td>da</td>
<td>5.05</td>
<td>30.8</td>
<td>0.55</td>
<td>100</td>
<td>3.3</td>
<td>2.63</td>
</tr>
</tbody>
</table>

#### Entity per Type over Time

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>386.2k</td>
</tr>
<tr>
<td>Place</td>
<td>216.3k</td>
</tr>
<tr>
<td>Agent</td>
<td>165k</td>
</tr>
<tr>
<td>Organization</td>
<td>3,304</td>
</tr>
<tr>
<td>Concept</td>
<td>1,573</td>
</tr>
<tr>
<td>Timespan</td>
<td>21</td>
</tr>
</tbody>
</table>
Challenges
Challenges

Technical

- Modeling
- Data integration
- API design and infrastructure
  - serve but also update data, search (and rank) entities
- Related platforms design & implementation
  - Europeana portal, curation environment, monitoring
Challenges

Organizational

● Getting everyone buy the idea
● Getting people involved in curation:
  ○ selection of entities
  ○ edition of metadata
  ○ links between entities and items
● Planning all this!

May requires shifting priorities

*Compare this with DCMI 2018 presentation*