Towards a methodology for validation of metadata enrichments in Europeana

Antoine Isaac, Valentine Charles, José Eduardo Grano de Oro, Nuno Freire, Lianne Heslinga, Hugo Manguinhas, Mónica Marrero, Paolo Scalia, Eleftheria Tsoupra (Europeana Foundation) Alexander Raginsky, Vadim Shestopalov (Pangeanic)
Enrichments for the common data space for Cultural Heritage
Data enrichment - Europeana definitions

**Enrichment**: data [...] that augments, contextualises or rectifies the authoritative data made available by cultural heritage institutions for an object.

- Enrichments of metadata aim to improve it by adding new, or refining or rectifying information about the object
- Content enrichments aim to produce alternative representations of the object in various media types.

**Enrichment effort**: an activity or workflow [...] that either leads to new enrichments [...]. It may reflect manual, crowdsourced, user-assisted/semi-automated or fully automated processes.

https://pro.europeana.eu/page/glossary
Enrichment examples

- connecting subject terms with vocabularies
- suggesting corrections to the name of an author
- translating metadata fields
- generating text transcription of a recorded speech
- generating a subtitle for a video

...
Enrichment examples

- connecting subject terms with vocabularies
- suggesting corrections to the name of an author
- translating metadata fields
- generating text transcription of a recorded speech
- generating a subtitle for a video
...
We want more of it!

In the past year:
Trying to bring direction and consistency

Via a policy for enrichments

Vision and principles for enrichment efforts

- **Quality**
- Transparency of provenance
- Interoperability, reusability and copyright
- Collaboration and capacity building
- Participation, diversity and inclusion
- Environmental sustainability
Quality?

**Principles in the Enrichments Policy**

- Enrichment efforts lead to improved quality and increase the value of data for users.
- Quality improvements [...] are measured on the basis of transparent and standardised criteria [...] 
- A process for validating enrichments is designed prior to the execution of enrichment efforts as an instance of a validation methodology [...]
Towards a methodology for validating enrichments
High-level steps for validation

- Analysis and documentation of enrichment effort
- Iterative acceptance validation workflow
  - Sampling enrichments for validation
  - Assessing sample of enrichments
  - Validate against quality requirements
  - Using assessment to identify reliable confidence score and thresholds
  - Improve enrichments

https://pro.europeana.eu/post/methodology-for-validating-enrichments
Documenting enrichments and evaluation

*Enrichment questionnaire to be filled by partners*

Enrichment performed and outcomes
- Goal of the enrichment
- Source of enrichments
- Type and target of enrichments
- Dataset(s) selection
- Overall enrichments produced and delivered to Europeana, enrichments not accepted and motive

Quality Assessment
- Framework (tools, methodology followed)
- Coverage
- Reviewers / Annotators
- Results
- Source and/or rationale for the confidence scores (or any other quality metric)
- Recommended threshold for confidence scores
Application in the Jewish History Tours project
Jewish History Tours

2021-2023

Create online tours using Jewish heritage data on pan-European topics, such as multiculturalism, persecution and migration

https://pro.europeana.eu/project/jewish-history-tours
JHT Partners

- Pangeanic
- Jewish Heritage Network
- Association Européenne pour la Préservation et valorisation de la culture et du patrimoine Juifs
- Clio Muse Tours
- Moses Mendelssohn Akademie
- CEC- Zentrum zur Erforschung und Dokumentation jüdischen Lebens in Ost- und Mitteleuropa
- Ośrodek “Brama Grodzka – Teatr NN
- Europeana Foundation

https://pro.europeana.eu/project/jewish-history-tours
Automatic geoenrichment pipeline

*Using metadata to link items to places with coordinates*

- Record Parser extracts texts from the record's title
- Named Entity Extraction service extracts geo-entities labels (names of streets, squares, landmarks, monuments and buildings)
- Multilingual Generative Entity Retrieval (mGENRE) service retrieves (Wikidata) entities
- Wikidata API and an (optional) Custom Places DB are used to get info about entity (coordinates, multilingual labels and descriptions...)*
Geoenrichment integration into Europeana

**Process**

- Establish relevant documentation for the enrichment process
- First round of enrichments
- First manual analysis: identifying issues
  - E.g., enrichment with "USSR" but with lat 18.2223 and long 109.5156 (in China)
- Second version of enrichment process
- Manual evaluation
- Integration into Europeana
Manual evaluation

Setting

- Samples of enriched collections (not yet manually reviewed)
- Annotation guidelines

**Rating system**

- **9 - 10 stars**: complete and precise enrichment.
- **7 - 8 stars**: enrichment is correct, but too general or incomplete. I.e. out of several possible locations only one was found, or found city only while street or place were in the original text.

### Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Records</th>
<th>Records enriched</th>
<th>Records in sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNN</td>
<td>62,172</td>
<td>23,838</td>
<td>304 (1.3%)</td>
</tr>
<tr>
<td>Centropa</td>
<td>1787</td>
<td>1638</td>
<td>120 (7.3%)</td>
</tr>
<tr>
<td>Blavatnik_MSYT</td>
<td>578</td>
<td>553</td>
<td>125 (22.6%)</td>
</tr>
<tr>
<td>Blavatnik_postcards</td>
<td>1327</td>
<td>366</td>
<td>65 (17.8%)</td>
</tr>
<tr>
<td>Blavatnik_posters</td>
<td>74</td>
<td>41</td>
<td>16 (39.0%)</td>
</tr>
</tbody>
</table>

**dc:title**

<table>
<thead>
<tr>
<th>Entity name</th>
<th>Entity URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plac Litewski w Lublinie</td>
<td><a href="http://www.wikidata.org/entity/Q11819103">http://www.wikidata.org/entity/Q11819103</a></td>
</tr>
</tbody>
</table>

Korowód Festiwalu "Zaczarowany Lublin" na ul. Archidiakońskiej

<table>
<thead>
<tr>
<th>Entity name</th>
<th>Entity URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lublin</td>
<td><a href="http://www.wikidata.org/entity/Q37333">http://www.wikidata.org/entity/Q37333</a></td>
</tr>
</tbody>
</table>
Manual evaluation

Setting

- Samples of enriched collections
- Annotation guidelines
- As annotation tasks on LabelStudio
- By TNN (content partner), Pangeanic (technical partner), Europeana
Manual evaluation

Results

Average ratings
- TNN: 8.76
- Centropa: 8.58
- Blavatnik: 8.59

Conclusion: the enrichment can be applied to the JHT datasets or very similar ones

Applying to different types of dataset would require more evaluation
Integration into Europeana

26,436 enriched records ingested using an extension of the Europeana Data Model for enrichments and their provenance
Integration into Europeana

26,436 enriched records ingested using an extension of the Europeana Data Model for enrichments and their provenance.

<table>
<thead>
<tr>
<th>Type of object</th>
<th>Maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>1941 to 1945</td>
</tr>
<tr>
<td>Provider</td>
<td>Jewish Heritage Network</td>
</tr>
<tr>
<td>Rights statement for the media in this item (unless otherwise specified)</td>
<td><a href="http://rightsstatements.org/vocab/InC/1.0/">http://rightsstatements.org/vocab/InC/1.0/</a></td>
</tr>
<tr>
<td>User generated content</td>
<td>false</td>
</tr>
<tr>
<td>Places</td>
<td>Budapest, Hungary</td>
</tr>
<tr>
<td>Identifier</td>
<td><a href="https://api.blavatnikarchive.org/app/blavatnikarchive/search/">https://api.blavatnikarchive.org/app/blavatnikarchive/search/</a></td>
</tr>
</tbody>
</table>
Final points
Work in progress

- Though it builds on experience from the past years, our methodology is a first version

- We want to work on it with the Europeana community

- There is still a lot to do, and we anticipate some things will be hard
  - Identifying evaluation criteria and quality requirements that apply across projects and types of enrichment (as opposed to agree on them at the beginning of every enrichment effort)
  - Having reliable confidence scores
All this feels like the proper thing to do, right?

Really? We were doing these evaluations already, no?
- Our experience is slightly different...
- We often face a perfect storm:
  - lack of resources
  - perception that it's ok to rely a more-or-less informed gut feeling that the value of an enrichment effort outweighs its issues

Show of hands for the audience!
- Who does such validation for all of their enrichment efforts?
- Should we try to make it mandatory for every enrichment effort in the common European data space for cultural heritage?