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)

Hó országa by Kiss György - Rippl-Rónai Municipal Museum with County Scope - Kaposvár, Hungary - CC BY.

Enrichments for the common data space for Cultural Heritage

Donkey Rides on the Beach - Rijksmuseum, Netherlands - Public Domain.

europeana

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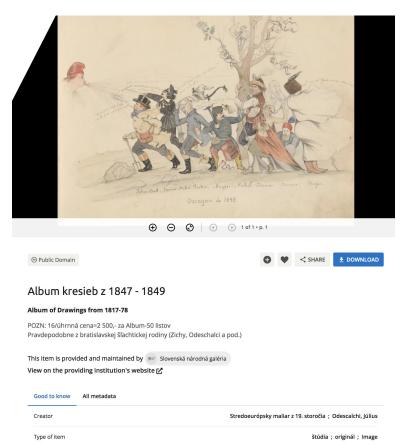
Enrichment effort: an activity or workflow [...] that either leads to new enrichments [...]. It may reflect manual, crowdsourced, user-assisted/semi-automated or fully automated processes.

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

…

Public Domain

Album kresieb z 1847 - 1849

Album of Drawings from 1817-78

POZN: 16/úhrnná cena=2 500,- za Album-50 listov
Pravdepodobne z bratislavskej šľachtickej rodiny (Zichy, Odeschalci a pod.)

This item is provided and maintained by 📇 Slovenská národná galéria
View on the providing institution's website 📎

Good to know	All metadata

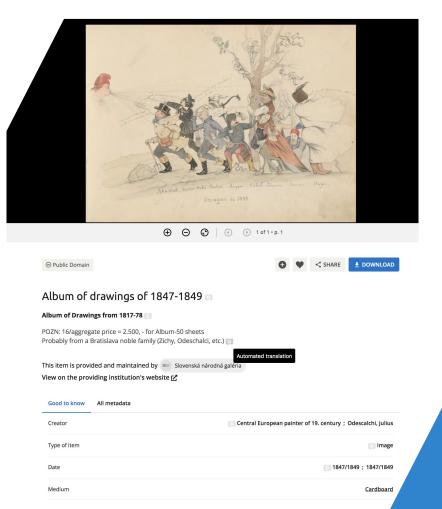| Creator	Stredoeurópsky maliar z 19. storočia ; Odescalchi, Július
Type of item	štúdia ; originál ; Image
Date	1847/1849 ; 1847/1849
Medium	cardboard ; Cardboard

Enrichment examples

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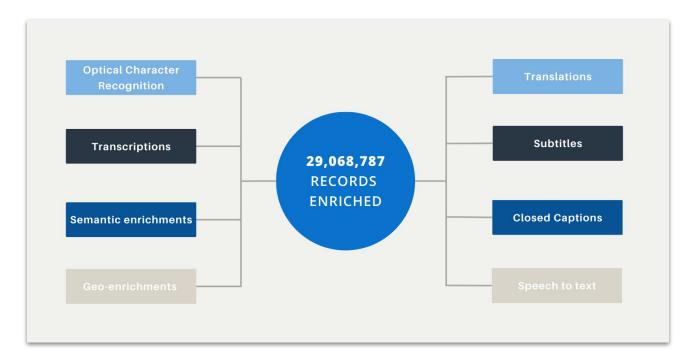- suggesting corrections to the name of an author
- translating metadata fields
- generating text transcription of a recorded speech
- generating a subtitle for a video

...

⊕ Public Domain   SHARE   ⬇ DOWNLOAD

Album of drawings of 1847-1849

Album of Drawings from 1817-78

POZN: 16/aggregate price = 2.500, - for Album-50 sheets
Probably from a Bratislava noble family (Zichy, Odeschalci, etc.)

This item is provided and maintained by   Slovenská národná galéria

Automated translation

View on the providing institution's website ☑

Good to know	All metadata
Creator	Central European painter of 19. century ; Odescalchi, Julius
Type of item	Image
Date	1847/1849 ; 1847/1849
Medium	Cardboard

https://www.europeana.eu/en/item/07101/K_6079_27

We want more of it!

In the past year:

Optical Character Recognition

Transcriptions

Semantic enrichments

Geo-enrichments

29,068,787 RECORDS ENRICHED

Translations

Subtitles

Closed Captions

Speech to text

europeana

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

Enrichments Policy
for the common European data space for cultural heritage

Revision	Version 1
Last updated	August 2023
Author(s)	Hugo Manguinhas, Ariadna Matas, Maša Škrinjar
Contributor(s)	Europeana Foundation staff Europeana Data governance stakeholder working group
Dissemination Level	Public

How to cite this:

CNECT/LUX/2021/OP/0070 - Deployment of a common European data space for cultural heritage: Enrichments Policy (V1). Europeana Initiative. August 2023. URL: https://pro.europeana.eu/post/enrichments-policy-for-the-common-european-data-space-for-cultural-heritage. CC BY-SA

https://pro.europeana.eu/post/enrichments-policy-for-the-common-european-data-space-for-cultural-heritage

Quality?

Antoine Antoine Antoine Antoine Antoine Antoine Antoine.

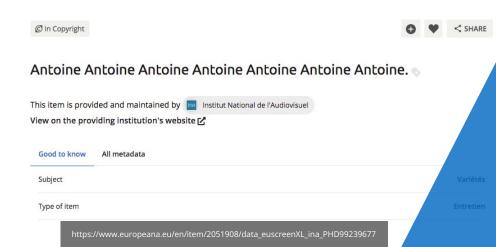This item is provided and maintained by Institut National de l'Audiovisuel

View on the providing institution's website

Good to know   All metadata

Subject

Type of item

Variétés

Entretien

https://www.europeana.eu/en/item/2051908/data_euscreenXL_ina_PHD99239677

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 […]

In Copyright

Antoine Antoine Antoine Antoine Antoine Antoine Antoine.

This item is provided and maintained by [ina] Institut National de l'Audiovisuel
View on the providing institution's website

Good to know   All metadata

Subject   Variétés

Type of item   Entretien

https://www.europeana.eu/en/item/2051908/data_euscreenXL_ina_PHD99239677

Towards a methodology for validating enrichments

europeana

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

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

Donkey Rides on the Beach - Rijksmuseum, Netherlands - Public Domain.

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
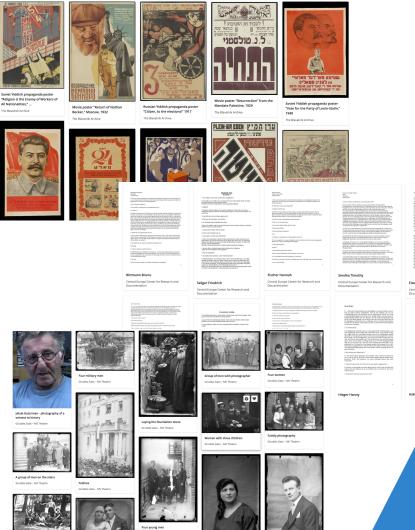

Walking **tour** 9 months ago

Lublin. From Jewish Piaski to Brama Krakowska. Jewish History Tour

⭐ (0) 🏛 Ośrodek "Brama Grodzka - Teatr NN" in Lublin

▶ Play   🌐 EN ⌄

Overview   Stories 15   Reviews 0

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

https://j-story.org/experiences/from-jewish-piaski-to-brama-krakowska

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...)

europeana

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

europeana

Manual evaluation

Setting

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

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

Rating system

9 - 10 stars: complete and precise enrichment.

dc:title	Entity name	Entity URI
Plac Litewski w Lublinie	Plac Litewski w Lublinie	http://www.wikidata.org/entity/Q11819103

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.

dc:title	Entity name	Entity URI
Korowód Festiwalu "Zaczarowany Lublin" na ul. Archidiakońskiej	Lublin	http://www.wikidata.org/entity/Q37333

Manual evaluation

Setting

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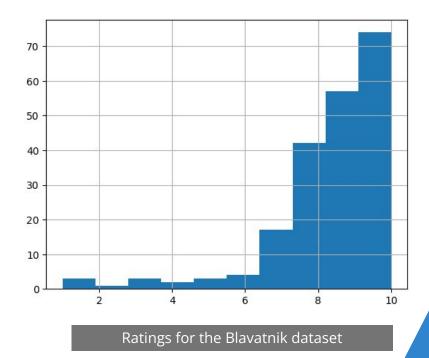
Manual evaluation

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

Ratings for the Blavatnik dataset

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

Type of object	🏷️ Maps
Date	1941 to 1945
Provider	Jewish Heritage Network
Rights statement for the media in this item (unless otherwise specified)	http://rightsstatements.org/vocab/InC/1.0/
User generated content	false
Places	🏷️ Budapest, Hungary
Identifier	https://api.blavatnikarch... /app/blava......

Enrichment by partner projects

Final points

europeana

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

europeana

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?

europeana

View of Delft by Johannes Vermeer - Mauritshuis, Netherlands - Public Domain