Fishs Seller, Národní muzeum - Náprstkovo muzeum, Czech Republic, CC BY

europeana
network
association

INTEROPERABILITY OF ANNOTATIONS AND USER SETS

Report and recommendations

Co-financed by the Connecting Europe Facility of the European Union

Table of Content

Table of Content 2

Task Force members 3

1. Introduction 4

2. Case studies 4

2.1. FindSampo - Finnish Archaeological Finds Recording Linked Open Data Service and Semantic Portal 4

2.1.1. Data requirements 5

2.1.2. Relation to EDM profile for annotations 6

2.2. EnrichEuropeana - Document transcription, translation and semantic enrichments for historical manuscripts 6

2.2.1. Data requirements 11

2.2.2. Relation to EDM profile for annotations 13

2.3. CrowdHeritage - CrowdSourcing Platform for Enriching Europeana Metadata 14

2.3.1. Data requirements 15

2.3.2. Relation to EDM profile for annotations 17

2.4. SOCH - Semantic Annotations 17

2.4.1. Data requirements 17

2.4.2. Relation to EDM profile for annotations 19

2.5. Burckhardt Source correspondence text analysis 19

2.5.1. Data requirements 21

2.5.2. Relation to EDM profile for annotations 24

2.6. 50s in Europe Kaleidoscope - Dealing with multiple layers of annotations 25

2.6.1. Data requirements 27

2.6.2. Relation to EDM profile for annotations 28

2.7. The Europeana Media Generic Services project - Transcriptions, subtitles and captions for A/V content 28

2.7.1. Data requirements 29

2.7.2. Relation to EDM profile for annotations 31

3. Conclusion 32

Task Force members

Editor(s)	Antoine Isaac Nuno Freire
Contributor(s)	Eero Hyvönen Sergiu Gordea Maria Ralli Albin Larson Giulio Andreini Luca De Santis Hugo Manguinhas
Other Task Force members	Vladimir Alexiev Valentina Bachi Erik Buelinckx Bernadine Brocker Wieder Rossana Damiano Mar Dixon Antonella Fresa Paul Jongsma Marijn Koolen Pieter van Leeuwen Marian Lefferts Tom Miles Vahur Puik Marco Rendina Mia Ridge Robert Sanderson Alexandru Stan Frederik Truyen Uldis Bojars Jordi Cabeza Olav Kwakman

1. Introduction

The Task Force on Interoperability of annotations and user sets has been launched by the EuropeanaTech community in order to mature specifications for profile(s) for the Europeana Data Model (EDM) that enhance interoperability of applications producing and consuming data for annotations and other forms of user-contributed (meta)data, such as transcriptions, translations, tags, reviews and groupings of objects like virtual collections, personal galleries, "stories"....

The task force has inventorized use cases from its participants, in order to clarify with concrete examples what are the annotations that can be relevant for Europeana. In a later phase, the data requirements raised from the use cases were analysed, and related to the relevant part of the EDM profile for Annotations[1], which was in parallel specified by Europeana Foundation and (partly) implemented in the Annotations API. This document frequently references sections of the EDM profile for Annotations, which, at the time of writing, is in version 1.2.

2. Case studies

In this report, use cases, data requirements and links to the EDM profile for Annotations are grouped into case studies that present in one go each one's use cases, requirements and their support by the EDM specification.

Editor's note: in the remainder of the document, the future tense may be used in some project's description, while the project has ended. We have preferred to keep the submitted descriptions for the projects as such, to better reflect their goals with respect to annotations and their interoperability.

2.1. FindSampo - Finnish Archaeological Finds Recording Linked Open Data Service and Semantic Portal

The Finnish Archaeological Finds Recording Linked Open Database[2] (SuALT) is a multidisciplinary project developing innovative solutions to respond to metal detecting and other non-professional encounters with archaeological material, applying semantic computing to "citizen science".

The growing flow of uncovered archaeological material poses challenges to researchers and collections finds data managers. We must support finders with legislative and also archaeological information. Easy to use tools to report finds and provide structured metadata are needed. Leaving finds cataloguing to curators is increasingly unfeasible given the increase in metal detecting. To make use of new data, cultural heritage managers, researchers and the public need search and analysis tools. Since finds are

[1] The EDM profile for annotations is available at https://pro.europeana.eu/page/edm-profiles#annotations-profile.

[2] Project homepages: https://blogs.helsinki.fi/sualt-project/ and https://seco.cs.aalto.fi/projects/sualt/

connected to existing collections, we also address cross-collection data interoperability. The methods and Open Source tools developed are also applicable to other cultural heritage citizen science fields.

(Wessman et al, 2019)[3] describe the workflow for the SuALT system: the end user is facilitated with a web interface that helps her in 1) analysing the find, 2) creating the find report with high quality metadata at the find site (via a mobile phone), and 3) later on in managing her personal finds data via a PC with a larger screen. The user can interact with the metal-detecting and expert community through the web interface to get feedback and help the identification of the find. The systems also connect with existing finds data from the Finnish Heritage Agency as seed data, and other archaeological information (e.g., museum item collections data). The knowledge base is also linked to other European Finds databases, providing users access to contextual information. The Linked Data Finland platform is used for hosting the data as a service

Stakeholders:

- Finnish Heritage Agency (FHA). Collecting finds data from metal detectorists and supporting their citizen science activities.
- University of Helsinki (UH). Research on archaeological data, metal detection, and citizen science.
- Aalto University. Designing and development of the technical solutions and archaeological infrastructure.
- AriadnePlus EU Project. The data will be included in AriadnePlus services on an EU level and beyond.

Related projects in other countries:

- PAS (UK)
- DIME (Denmark)
- PAN (The Netherlands)

2.1.1. Data requirements

In the Finds database for archaeological finds, a find has the following properties:

- find_number
- municipality
- province
- area
- specification
- coordinate_system
- height_system
- x_coordinate
- sequence_number
- find_name
- material
- type
- part
- main_material
- other_materials
- sub_category
- diameter
- period
- start_year
- extension_one
- end_year
- dating_reason
- extension_two

[3] Anna Wessman, Suzie Thomas, Ville Rohiola, Jutta Kuitunen, Esko Ikkala, Jouni Tuominen, Mikko Koho and Eero Hyvönen: A Citizen Science Approach to Archaeology: Finnish Archaeological Finds Recording Linked Open Database (SuALT). DHN 2019 Digital Humanities in Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, pp. 469-478, CEUR Workshop Proceedings, Vol-2364, Copenhagen, Denmark, March, 2019. Available online: http://ceur-ws.org/Vol-2364/43_paper.pdf

- y_coordinate
- z_coordinate
- x_square
- y_square
- layer
- depth
- unit
- structure
- typological_extension
- description
- amount
- weight
- length
- width
- thickness
- thickness_two
- original_image_name
- image_name
- photographer
- date_photo_taken
- image_description
- image_url
- archaeological_site_url
- project_id

The data model records finds based on new Finnish ontologies to be interlinked with international ones in use in AriadnePlus, especially AAT (Getty) for object types, Periodo (for times), Geonames, and Nomisma (coins). Also vocabularies in use in UK, Denmark, the Netherlands, and Belgium in related projects are considered.

2.1.2. Relation to EDM profile for annotations

The workflow for the SuALT system supports commenting on the finding aids by the experts community (Wessman et al, 2019). The detail of the data requirements, however, makes one wonder whether this is an annotation case. The data fields do not hint at an essential difference, for a find, between what could come from the 'seed data' (from an institution) and the data entered by the user of the web interface. In particular, there is no indication of who creates which field. It looks like one record with data that could come from several sources.

This is certainly right for the case at hand, of course. However, that would position the use case a bit further away from the other annotation cases in the task force, where the data maps more directly to the W3C Web Annotation Model.

2.2. EnrichEuropeana - Document transcription, translation and semantic enrichments for historical manuscripts

The EnrichEuropeana project aims at developing a crowdsourcing platform that will enable citizens to transcribe and enrich cultural heritage material from Europeana Collections and national aggregator portals.

The Transcribathon platform is used to transcribe historical manuscripts. These manuscripts contain handwritten text which in most cases is very hard to read. For several languages, like German for example, the script standard has been considerably changed since these documents were created.

The main goal of transcribing is to enable better access to the textual information available in these documents, by converting it into human and machine readable text. Additionally the textual information is enriched by linking named entities with web resources (i.e. Wikidata, DBpedia, Europeana Entity Collection). This can use translation of the transcriptions.

The Champion users on the Transcribathon platform are empowered with curation privileges. The fully curated enrichments and (transcription) documents will be submitted to the Europeana core service platform (e.g. via the Annotation API).

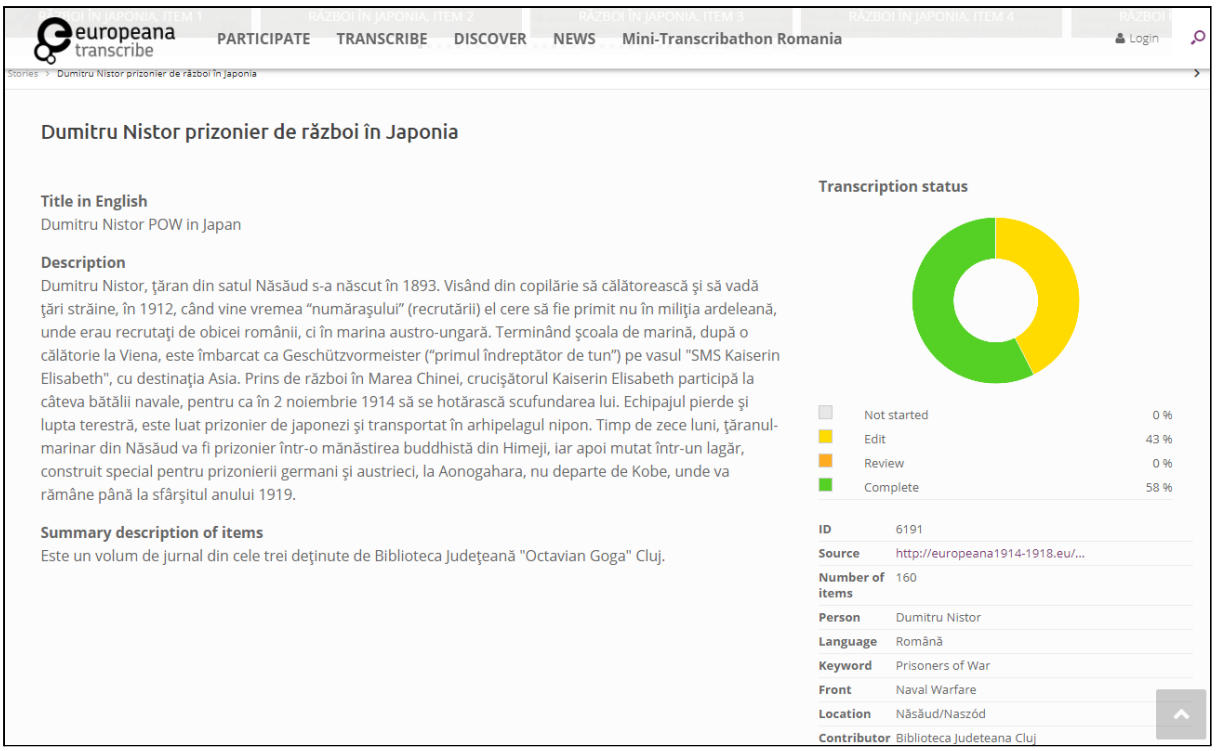
For manual enrichment the Transcribathon platform supports users to manually enrich items with.:

- A description of what the item is about, including details on images and objects that might appear in the item (e.g. description of the image on a postal stamp, or individual images in a photo album)
- A date that reflects the content (text) of the item
- A classification of the item (either Letter, Diary, Postcard or Picture)
- A person somewhat associated to or mentioned by the item (e.g. creator, subject) further detailing it with its name, place and date of birth and death
- A precise location (ie. place) indicated by names and/or the geo-coordinates indicated by using a pinpoint on the map and/or a street address
- A free-text keyword pertaining to the topic and content of the item

The enrichments can be added both for the full document (using the information from the title and the description of the document) or for individual pages of the document (using the information from the transcribed text).

Automatic Semantic Enrichments with Persons and Locations are computed within the scope of the EnrichEuropeana project. This functionality will be integrated in the new version of the Transcribathon platform. The enrichments are computed by using DBpedia Spotlight and Stanford NER. DBpedia Spotlight is linking the entities with DBpedia resources, while named entities identified by Stanford NER are resolved against Wikidata and/or Europeana Entity Collection.

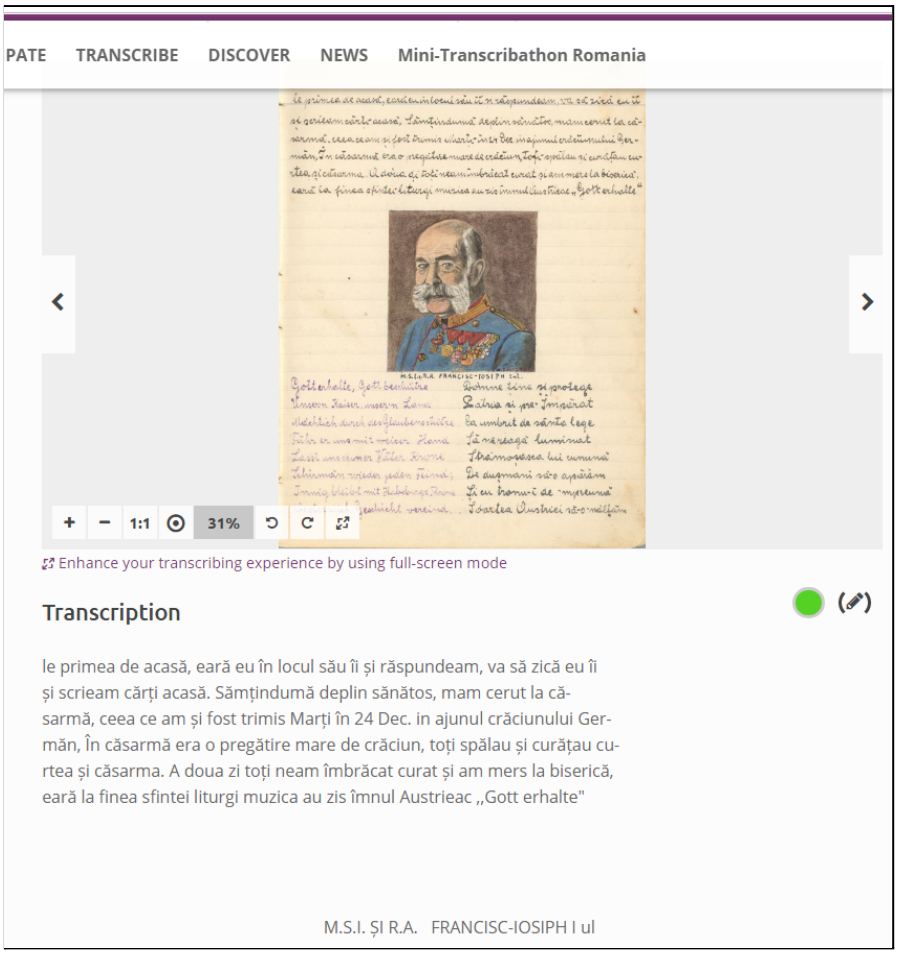
The following figure presents an example of a document in Transcribathon:



Document on the Transcribathon platform
(source: https://transcribathon.com/en/documents/id-6191/[4])

[4] Translation of the document description:
Dumitru Nistor, a peasant from Năsăud village, was born in 1893. Dreaming from childhood to travel and see foreign lands, in 1912, when the time of the "counting" (recruiting) came, he asked to be received not in the Transylvanian militia, where usually the Romanians were recruited, but in the Austro-Hungarian marina. By completing the naval school, after a trip to Vienna, he is boarded as Geschützvormeister ("the first cannons setter") on the "SMS Kaiserin Elisabeth" ship to Asia. Caught as capture of war in the Sea of China, cruiser Kaiserin Elisabeth participates in some naval battles, so that on 2 November 1914 his sinking will be decided. The crew loses ground battle as well and is captured by [...].

The following figure presents an example of a transcription of a document page in Transcribathon:



Transcription of a document page
(source: https://transcribathon.com/en/documents/id-6191/item-70103/[5])

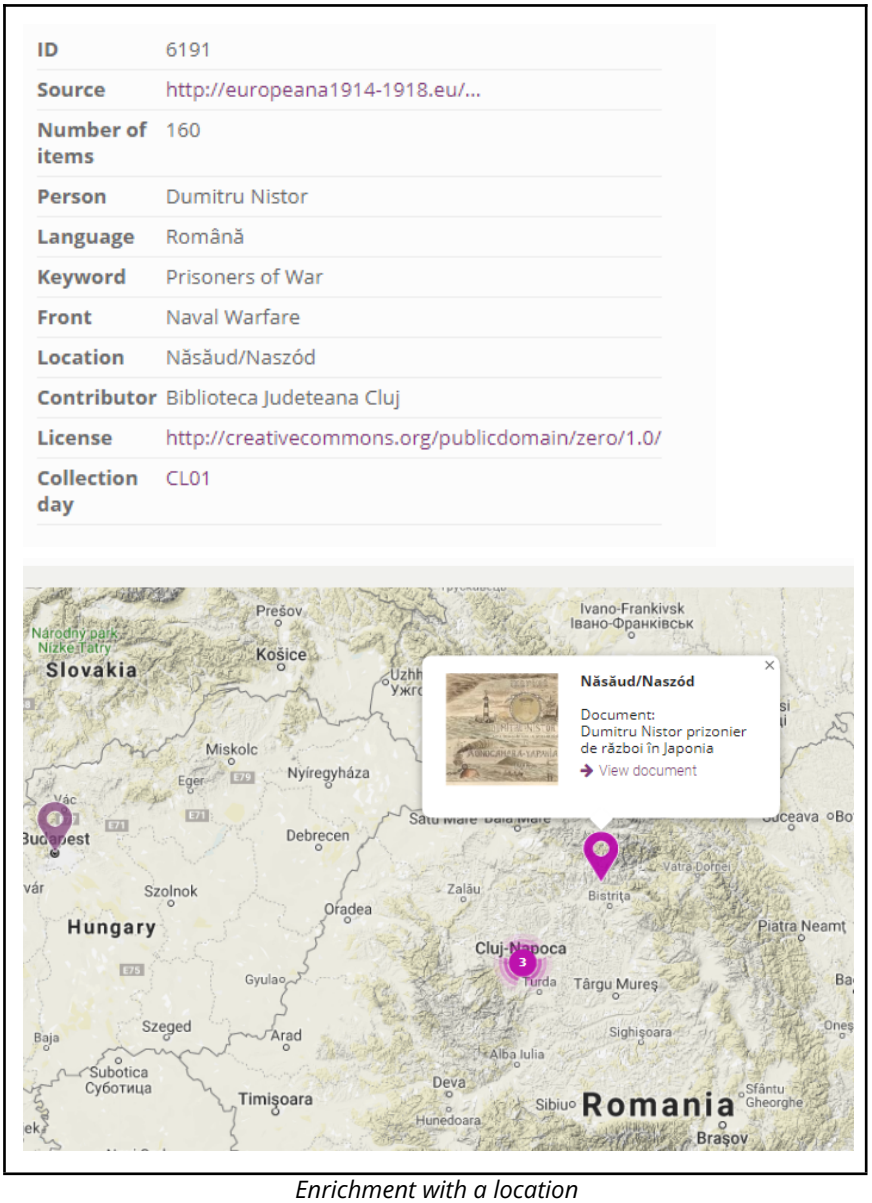
For Manual Enrichment (with Persons, Locations including geolocations, keywords) the tools on the Transcribathon platform support users to manually identify persons and locations referenced in the text. For locations, the geo-coordinates can be indicated by using a pinpoint on the map. The enrichments can be added both for the full document (using the information from the title and the description of the document) or for individual pages of the document (using the information from the transcribed text).

Automatic Semantic Enrichments with Persons and Locations are computed within the scope of the EnrichEuropeana project. This functionality will be integrated in the new version of the Transcribathon platform. The enrichments are computed by using DBpedia Spotlight and Stanford NER. DBpedia Spotlight is linking the entities with

[5] Translation of document page transcription:
... he received them from home, and if I was in his place I would have answered immediately, meaning that I would have written letters to home. Feeling myself completely healthy, I asked to go to the barracks, [...]

DBpedia resources, while named entities identified by Stanford NER are resolved against Wikidata and/or Europeana Entity Collection.

The following figure presents an example of an enrichment with a location in Transcribathon:



Enrichment with a location
(source: https://transcribathon.com/en/documents/id-6191/)

The User Roles on the Transcribathon Platform are:

- Registered User - the entry level for users of the Transcribathon Platform
- Runner - registered users that take part in a Transcribathon competition
- Champion - advanced users having the privileges to curate documents (validate transcriptions, translation and enrichment, mark document as complete)
- Administrator - powerful users that have access to the full functionality of the Transcribathon Platform

The use cases provided in this document are based on the existing functionality of the Transcribathon platform and the enhancements developed within the scope of EnrichEuropeana project (additional information is available within the project's MS2 and MS7 deliverables).

These use cases relate with the Europeana Media generic services project. The modelling of transcriptions is similar for different media types. While Transcribathon platform deals with enrichment of historical manuscripts, Europeana Media project deals with transcriptions of video recordings. In both cases, the transcriptions references a part of the media object. In Transcribathon, these are the ids of items (document pages), in Europeana Media this is expected to be based on timing information.

2.2.1. Data requirements

The system of Europeana Enrich uses four types of Annotations:

- Transcription
- Translation
- semantic tagging (including geotagging)
- linking objects

The annotations are made on Europeana data records (mainly in content from Europeana 1914-1918 collections) and use resources from:

- The Europeana Entity Collection
- Wikidata entities (persons, locations)
- DBpedia resources (persons, locations)

Below are some examples of annotations from Europeana Enrich.


```
{
       "source":"http:\/\/www.europeana1914-1918.eu\/en\/contributions\/6191",
       "item_id":"70103",
       "story_id":"6191",
       "language":"Romanian",
       "transcription":"le primea de acasă, eară eu în locul său îi și
       răspundeam, va să zică eu îișsi scrieam cărți acasă. Sămțindumă deplin
       sănătos, mam cerut la că-sarmă, ceea ce am și fost trimis Marți în 24 Dec.
       in ajunul crăciunului Ger-măn, În căsarmă era o pregătire mare de crăciun,
       toți spălau și curățau cu-rtea și căsarma. A doua zi toți neam îmbrăcat
       curat și am mers la biserică,eară la finea sfintei liturgi muzica au zis
       îmnul Austrieac ,,Gott erhalte" M.S.I.ȘI R.A.  FRANCISC-IOSIPH I ul”
}
```


Example of data of a transcription of document pages


```
{
       "source":"http:\/\/www.europeana1914-1918.eu\/en\/contributions\/6191",
       "item_id":"70103",
       "story_id":"6191",
```

```
	"language":"English",
	"field":"transcription",
	"translation":… he received them from home, and if I was in his place I
	would have answered immediately, meaning that
	I would have written letters to home. Feeling myself completely healthy, I
	asked to go to the barracks,
	therefore I also was sent on Tuesday 24 Dec. On the eve of German
	Christmas,
	there was a great Christmas preparation in the barracks, everyone was
	washing and cleaning the yard and the barracks.
	The next day we all dressed clean and went to the church,
	and at the end of sacred the liturgy the music has said the Austrian Hymn,
	"Gott erhalte" M.S.I. AND R.A. FRANCIS-JOSEPH I”
}
```


Example of data of a translation of document page transcription


```
{
	"source":"http:\/\/www.europeana1914-1918.eu\/en\/contributions\/6191",
	"story_id":"6191",
	"storyLocationName":"Năsăud\/Naszód",
	"storyCoords":"47.2921849,24.3973258",
	"storyKeyword":"Prisoners of War",
	"language":"Romanian",
	"person1FirstName":"Dumitru",
	"person1LastName":"Nistor",
	"person1DateOfBirth":1893,
	"person1PlaceOfBirth":"Năsăud",
	"person1DateOfDeath":null,
	"person1PlaceOfDeath":null
}
```


Example of data of a manual enrichment (Locations, Persons, Keywords)


```
{
	"story_id": "6191",
	"field": "description",
	"location": [
		{
		"europeanaId": null,
		"wikidataId": "http://www.wikidata.org/entity/Q48320",
		"dbpediaId": "http://dbpedia.org/resource/Kobe",
		"classificationType": "LOCATION",
		"position": 975,
		"key": "Kobe"
		}
}
```


Example of data of an automatic semantic enrichment (places/locations found in description field)


```
{
	"story_id": "6191",
	"item_id":"70103",
```

```
        "field": "transcription",
        "person": [
                {
                "europeanaId": null,
                "wikidataId": "http://www.wikidata.org/entity/Q51056",
                "dbpediaId":
        "http://dbpedia.org/resource/Franz_Joseph_I_of_Austria",
                "classificationType": "PERSON",
                "position": 460,
                "key": "FRANCISC-IOSIPH I ul"
                }
}
```


Example of data of an automatic semantic enrichment (persons found in document page transcription)

NB: for description of entities from Wikidata and DBpedia (which is not core to the annotation use case per se) the platform uses metadata elements that are similar to these used by Europeana's Entity Collection - Wikidata entities and DBpedia Resources (+ indicates mandatory elements):

- Person: id+, type+, modified, prefLabel+, altLabel, depiction, description, dateOfBirth, dateOfDeath, professionOrOccupation, country, sameAs+
- Place: id+, type+, modified, prefLabel+, altLabel, depiction, logo, description, lat+, long+, country+, sameAs+

2.2.2. Relation to EDM profile for annotations

The Enrich Europeana scenarios have been taken as leading use cases for designing the patterns for handling various kinds of annotations in EDM, as shown in the following sections of the EDM profile for Annotations:

- Section '3.3.1 Transcriptions'
- Section '3.1.3. Annotate a CHO with a semantic tag' and Section '3.1.4 Annotate a CHO with a geospatial tag'
- Section '3.1.1. Annotate a CHO with a simple tag'
- Section '3.2. Link a CHO to another'

Two use cases from EnrichEuropeana are currently pending further discussion in the EDM Profile for Annotations. User contributed descriptions and dates that reflect the content of the objects could be supported by the application scenario in Appendix 'B.2 Suggest a new statement to an existing CHO description'.

If these patterns are judged appropriate then EDM supports interoperability between Europeana and Enrich Europeana for these cases. A question to answer is that of the data expected for the semantic resources used in semantic tagging. I.e. whether the body of the annotation should cater only for the URI of a resource (be it in DBpedia, Wikidata or the Europeana Entity Collection) or whether it should include some of the metadata for this resource (labels, address, coordinates...).

2.3. CrowdHeritage - CrowdSourcing Platform for Enriching Europeana Metadata

CrowdHeritage offers the possibility to cultural heritage institutions and aggregators to design and launch ad-hoc crowdsourcing campaigns for metadata quality improvement of the content accessed via the Europeana platform. Through these campaigns, target users will be able to add annotations, depending on the type of content and missing metadata, and validate existing annotations in a user-friendly and engaging way.

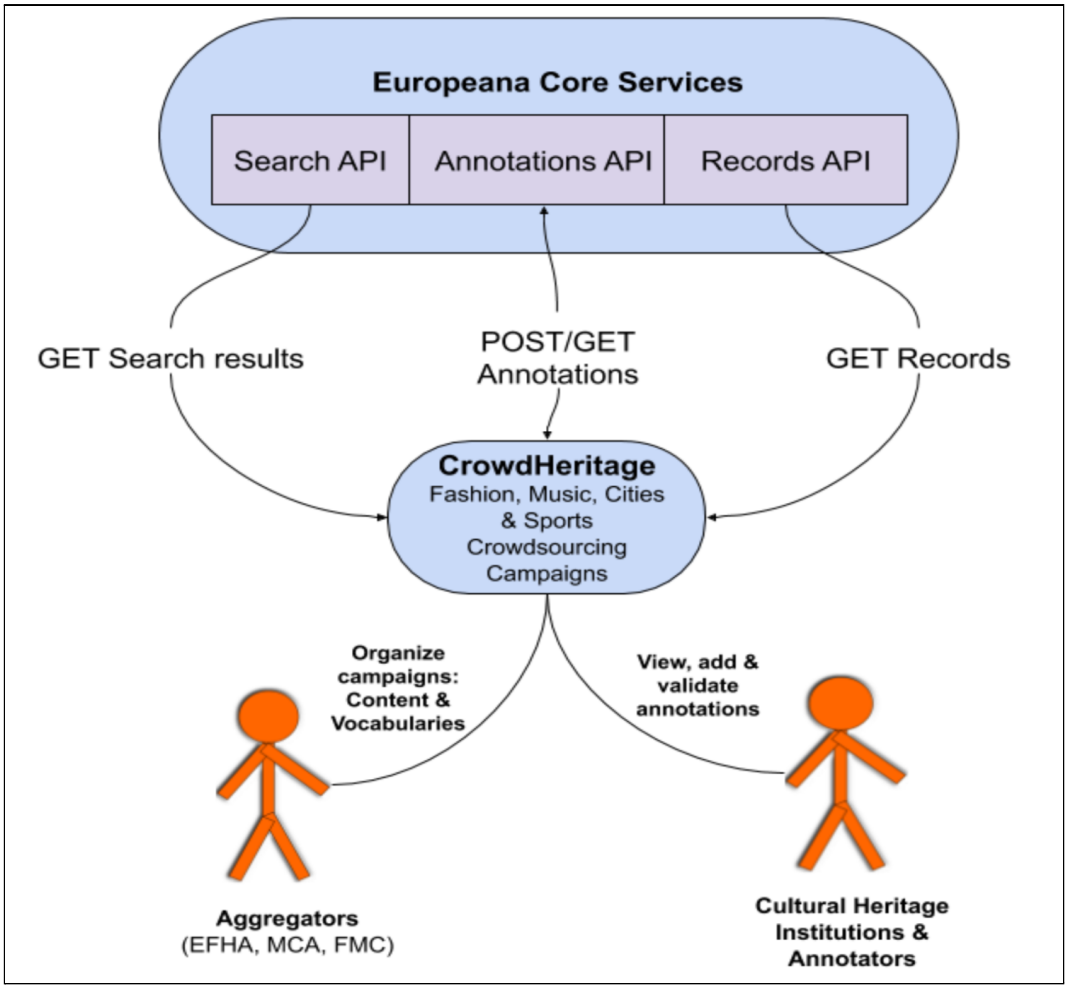
CrowdHeritage is an open standalone platform that connects to the Europeana core service platform through interaction with the Europeana API, the Annotations API, and the Entity API.

The CrowdHeritage platform will be used and tested by the three Aggregator partners in six campaigns, targeting different audiences (e.g. researchers, teachers, students, culture lovers, etc.) on four themes: Fashion, Music, European cities and Sports. The objective is to actively engage different target communities in enhancing the metadata of selected content, evaluate the usability of the platform and assess the effectiveness of the overall approach both in terms of data quality improvement and user engagement.

The platform supports four different crowdsourcing campaign types:

- Colortagging - Users can choose the dominant colours of the picture by selecting the relative colours lors from the pallet on the right. All colours are linked with the corresponding colourURI of the Europeana Fashion Thesaurus.
- Geotagging - Users can insert the city/place by typing its name on the tagging area and then choosing the correct place from the list derived from the Geonames Resources (we also consider wikidata).
- Semantic tagging - Users can add annotations about the record by typing the relative term in the tagging area and then selecting the term from the suggestions linked to relative Thesauri and Vocabularies (e.g. MIMO vocabulary).
- Actor identification - Annotators are requested to identify if a famous person is depicted in the picture or not (we are also considering asking the annotators to indicate how well the person is depicted in the picture).

Finally, users are allowed to validate existing annotations by upvoting/downvoting them.



Interactions of the CrowdHeritage platform with the users and the Europeana services

The WithCrowd annotation tool used in this use case is also used for the "50s in Europe Kaleidoscope" use case (see section below).

2.3.1. Data requirements

In CrowdHeritage, the annotations are linked to records that already exist in the Europeana platform and are compatible with the Europeana Data Model (EDM). The project's annotation model is based on W3C's Web Annotation Model, which is a structured model and format to enable annotations to be shared and reused across different hardware and software platforms. All the annotation terms are derived from controlled vocabularies and thesauri and more specifically:

- Wikidata
- Europeana Fashion Thesaurus (EF)
- MIMO vocabulary
- Getty AAT
- Geonames Resources
- Sport and cities vocabularies with resources from Wikidata

```
{
       "@context": "http://www.w3.org/ns/anno.jsonld",
       "type" : "Annotation",
       "created" : "2018-09-21T09:12:44.105Z",
       "creator": {
              "type": "Person",
              "name": "Maria Ralli"
              },
       "generated" : "2019-07-21T09:12:44.105Z",
       "generator" : {
              "type": "Software",
              "name": "CrowdHeritage",
              "homepage" : "https://crowdheritage.eu"
              },
       "motivation" : ”tagging”,
       "body": "http://thesaurus.europeanafashion.eu/thesaurus/10416",
       "target": "http://data.europeana.eu/item/09102/_UEDIN_214"
}
```

Example of an annotation from CrowdHeritage. The annotation tags a record in Europeana with a term from the Europeana Fashion Thesaurus.

Note that the CrowdHeritage platform handles other features as mentioned in the use cases document. Especially, it includes the main label that corresponds to the body URI (semantic tag) and keeps track of up- and down-voting of annotations by users. The former is quite redundant with the body itself, as this label can be obtained by fetching the data served at the (Linked Data) URI. The latter is more relevant, but for the moment we retain a scenario for interoperability between CrowdHeritage and Europeana, where this information is not needed as part of CrowdHeritage's export. Annotations will indeed be curated (partly based on the up- and down-votes) and only those deemed reliable will be exported to Europeana. We are aware that this choice is very much application-dependent, and other scenarios may later require the inclusion of this data in the EDM annotation profile.

Stakeholders:

- National Technical University of Athens (NTUA): Project Coordinator and Technical Manager of the CrowdHeritage project. NTUA is developing the CrowdSourcing platform and is implementing the backend interaction with the Europeana Annotations API.
- Europeana Fashion Heritage Association (EFHA): Aggregator partner who will identify their digital content and the vocabularies/thesauri for the crowdsourcing campaigns.
- Stichting Europeana (EF): Facilitating the interaction of the CrowdHeritage platform with the Europeana APIs .
- MICHAEL CULTURE AISBL (MCA): Aggregator partner who will identify their digital content and the vocabularies/thesauri for the crowdsourcing campaigns.
- Ministère de la Culture (FMC): Aggregator partner who will identify their digital content and the vocabularies/thesauri for the crowdsourcing campaigns.

- Cultural Heritage Institutions (museums, associations etc): These stakeholders will use and evaluate the CrowdHeritage platform by providing annotations during the crowdsourcing campaigns.

2.3.2. Relation to EDM profile for annotations

The example above can be almost entirely handled with the current EDM annotation patterns. The motivation and body of the annotation follow the pattern used in the application scenario '3.1.3. Annotate a CHO with a semantic tag' of the EDM profile for Annotations. The metadata produced by CrowdHeritage (creator, generator and dates) are also in line with what is supported in the basic annotation pattern, which is not a surprise as the EDM profile is largely based on the Web Annotation model. This includes the generated attribute, which is present in the EDM profile as the underlying property `dcterms:issued`. NB: In the current setting for publishing CrowdHeritage annotations to Europeana, the generator attribute refers to CrowdHeritage, the created one to the original date of annotation on the CrowdHeritage platform and the generated one to the date at which the annotation is sent to Europeana's API.

2.4. SOCH - Semantic Annotations

SOCH has since 2012 held an independent Semantic annotation platform that as of 2021 contains over 2.5 million annotations (made with properties that are a superset of this list) created by both users and bots. An annotation consists of a triple where the subject URL always is an SOCH URI. The objects of those triples can be other SOCH URIs as well as other external URIs. Common platforms to link to, include Wikidata and Europeana. Each triple has a set of metadata associated with it such as its creator, created data and comments.

There is no defined scope for relations types. They can be relations between objects, between objects and actors, between actors, between objects and concepts, etc.. Today this data is accessible through an API as non-RDF JSON and XML. Because SOCH is a LOD by default platform it would be preferable to migrate its Semantic annotation platform part to RDF as well and if possible follow the same practice as for example Europeana.

2.4.1. Data requirements

An annotation in SOCH contains the following elements (elements marked with a '+', are mandatory):

- id+
- uri+ (subject)
- userName+ (creator)
- createdDate+
- updatedDate
- relationType+ (predicate)
- relatedUri+ (object)
- commentValue

- applicationName

SOCH does not limit the types of relations for the annotations. Most applications using the annotations are based on a set of relations types, but additional relations can be added by using arbitrary predicates as fully qualified IRIs. The list of relation types (given here as plain strings) includes but is not limited to:

- sameAs
- isDescribedBy
- visualizes
- hasPart
- isPartOf
- isVisualizedBy
- isContainedIn
- author
- authorOf
- hasBeenUsedIn
- isRelatedTo
- architectOf
- architect
- user
- userOf
- child
- mother
- father
- photographerOf
- photographer
- isMentionedBy
- mentions

The SOCH API is based on XML data, but these relations exist as RDF properties with their URIs in the SOCH data model. See also a - slightly outdated - documentation.

```
<relation>
  <id>28259238</id>
  <uri>http://kulturarvsdata.se/jlm/item/324610</uri>
  <userName>carwash</userName>
  <applicationName>Kringla</applicationName>
  <createDate>2019-04-16</createDate>
  <updateDate/>
  <relationType>visualizes</relationType>
  <relatedUri>http://kulturarvsdata.se/raa/fmi/10252801060001</relatedUri>
  <commentValue/>
</relation>
```

Example of an annotation from SOCH.

`commentValue` is usually used when creating relations from third party sources and can state things like "Imported from Wikidata Q472975".

The SOCH UGC platform is not capable of handling changes of the URIs of the subjects and objects of statements, and has no knowledge or understanding of predicates such as `owl:sameAs` or `dcterms:replaces` (in contrast to the SOCH platform proper, which understands and can account for such changes). When the SOCH team becomes aware of large-scale changes to URIs in SOCH, a manual intervention must be made to update statements using those URIs in the UGC component.

2.4.2. Relation to EDM profile for annotations

Regarding the body of the annotation, the SOCH requirements seem to be manageable using the pattern described in Section '3.2.2. Link a CHO to another using a qualified link' of the EDM profile for Annotations. This pattern makes it possible to annotate a cultural object with any statement linking that object to another, using a relation deemed relevant by the annotator or the designer of the annotation application.

The pattern could be used here since it enables creation of RDF statements with predicates having a property URI like the one used in SOCH. It is a bit more verbose than the original SOCH pattern (especially, the URI of the annotated CHO appears twice in EDM, once as the target and once as the subject of the body's RDF statement).

The main hurdle, however, is that the current pattern restricts the RDF properties that can be used - using a small list of properties currently in EDM. This reflects a fundamental difference, as the Europeana case is one of linking across CHOs, while SOCH has a case of general statements about an object, potentially linking it to persons, especially.

Full compatibility between Europeana and SOCH, more work is needed. Either the Europeana profile for annotations would have to be updated to be more flexible, or existing SOCH properties would have to be mapped to the EDM ones if the SOCH data would be ingested in Europeana.

Regarding the metadata of the annotation, the base for the current EDM annotation profile allows to express the names of user and software agents (see Section '2.2. User and Software Agents'), as well as the creation and modification dates (see Section '2.1. Annotation'). The one issue here could be the expression of SOCH's commentValue, which could be represented as a general comment on the annotation.

2.5. Burckhardt Source correspondence text analysis

The Pundit Annotator[6] is a semantic annotation tool, in the case study of Burckhardt's correspondence. In this context the occurrences of persons and places names as well as references to works of art in the letters are being semantically annotated and linked to the Web of Data.

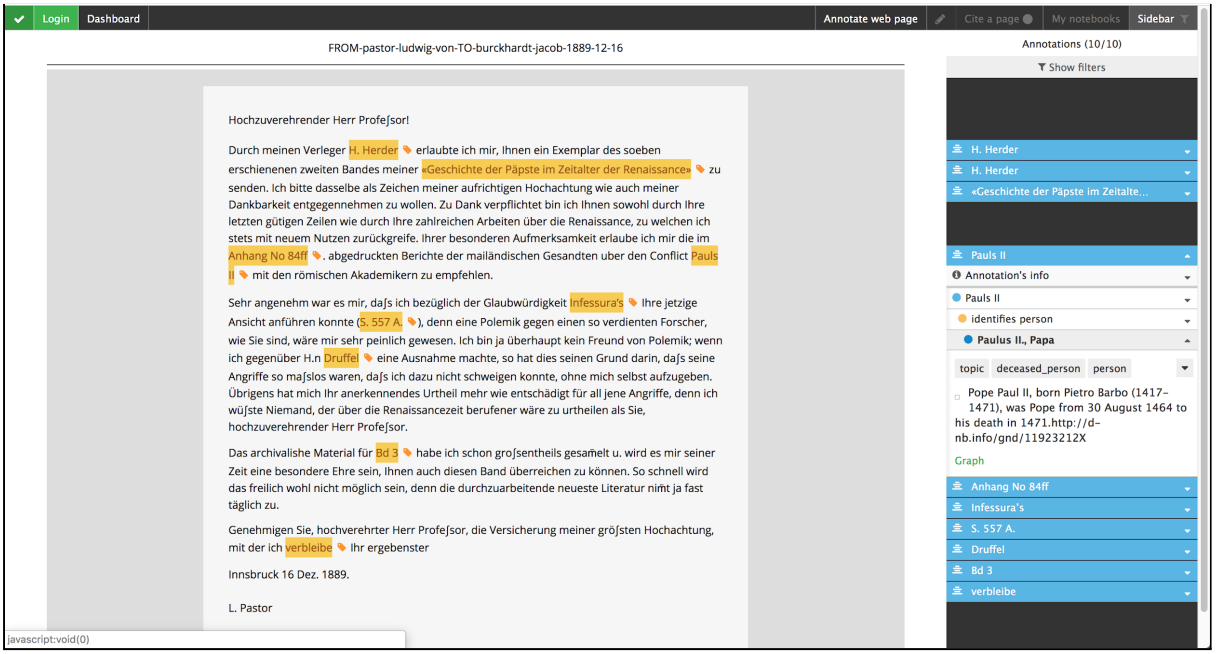
Pundit has two stakeholders:

- Data producer - The research team at the Scuola Normale Superiore di Pisa which consist of 5 researchers involved to analyse the correspondence to Jacob Burckhardt.
- Data consumer - Students and scholars interested in studying Jacob Burckhardt.

[6] Semantic annotation with Pundit: A case study and a practical demonstration https://www.researchgate.net/publication/262358453_Semantic_annotation_with_Pundit_A_case_study_and_a_practical_demonstration

The research team (the data producer) has done a deep text analysis of the correspondence to Jacob Burckhardt. The analysis consisted in spotting any reference to Persons, Places, Artworks and Bibliographic items in the text of the letters. The Pundit Annotator allowed them to select portions of text and create semantic annotations that relate the text to an entity. Semantic annotations are built with a predicate that describes the type of relation and relate to Freebase entities (the project started when Freebase was active). The text analysis activity was the core activity of the project.

Data consumers can use the annotations made on the text in 3 main ways: (1) annotated entities are highlighted in the transcription of each letter so that the user can easily see the reference and get some more info about the entity by clicking on it and navigate to the original URL; (2) entities present in the letters can be used to filter the letters in the Browse interface with a faceted system which allows to filter by Persons, Places, Artworks and Bibliographic items; (3) semantic enrichment is used to display the relations in a graph view (powered by the tool lodlive).



Pundit web annotator used to annotate the letters in the backend

Frontend view of a letter with semantic enrichment made with Pundit Annotator Pro

2.5.1. Data requirements

NB: this section takes into account the data model and architecture for a more modern version of Pundit than the one originally used for the Burckhardt Source use case. Especially, the data model follows more closely the W3C Web Annotation model, and the middleware application Korbo has been replaced by an Annotation Server[7] developed for the TRIPLE research project.

In Pundit annotations are grouped in notebooks, which act as containers of annotations.

Semantic annotations were built following this pattern:

- subject: the subject of the annotation is always a text fragment pointed using an Xpath selector on the XHTML document of the letter.
- predicate: we use a selection of predicates from a controlled vocabulary to describe the types of allowed relation. .
- object: objects in Pundit can be :
    - a literal (e.g. a string);
    - an URI identifying an external resource or an entity defined in a Semantic Web ontology;
    - another fragment of text pointed using an Xpath selector on the XHTML of, possibly a different web document;

[7] TRIPLE Deliverable: D5.5 Report on the Open Annotation Tool. Available at https://doi.org/10.5281/zenodo.5653219 .

- custom entities, created on the Pundit back-end (a functionality not available yet for the general public).

For the Burckhardt project, custom entities were defined through a previous "semantic proxy" application named Korbo, which allowed users to define and link them to the corresponding entities in Freebase".

Subject, predicate and object are all mandatory to build a semantic annotation in Pundit.

Some examples of predicates used in Pundit are the following:

```
label: "my comment"
url: "http://www.w3.org/2000/01/rdf-schema#comment"
description: "add a comment to a part of text in a web page"

label: "was written by"
url: "http://purl.org/dc/terms/creator"
description: "say that a part of text in a web page was written by a specified
person"

label: "is this date"
url: "http://purl.org/dc/terms/date"
description: "say that a part of text in a web page refers to a exact year or
month or day or hour"

label: "has type"
url: "http://purl.org/dc/terms/type"
description: "say that a part of text in a web page represents an entity of a
specified type"

label: "cites"
url: "http://purl.org/spar/cito/cites"
description: "The selected text fragment cites another text fragment, or a Work
or a Person"

label: "replies to"
url: "http://purl.org/spar/cito/repliesTo"
description: "say that a part of text in a web page is a reply to another part of
text"

label: "describes"
url: "http://purl.org/spar/cito/describes"

label: "is related to"
url: "http://purl.org/pundit/ont/oa#isRelatedTo"
description: "The selected text fragment is someway related to another text,
Entity, Person or any other kind of concept"

label: "is defined by"
url: "http://www.w3.org/2000/01/rdf-schema#isDefinedBy"
```

```
label: "was said by"
url: "http://purl.org/spar/cito/includesQuotationFrom"
```

In Pudit, notebooks collect annotations in Web Annotation Pages:

```
{
  "id": "https://purl.thepund.it/notebook/id/4895cc41",
  "type": "AnnotationCollection",
  "total": 11,
  "label": "Burckhardt annotations",
  "first": {
       "id": "https://purl.thepund.it/notebook/id/4895cc41?page=1",
       "type": "AnnotationPage",
       "next": "https://purl.thepund.it/notebook/id/4895cc41?page=1",
       "items": [
       {
       "id": "https://purl.thepund.it/annotation/id/5b43148c",
       "type": "Annotation",
       "creator": {
[...]
```

The following example contains an Annotation in Pundit:

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "https://purl.thepund.it/annotation/id/5b43148c",
  "type": "Annotation",
  "creator": {
    "id": "https://purl.thepund.it/user/id/xxxxxx",
    "name": "xxx xxx",
    "givenName": "xxx",
    "familyName": "xxx"
  },
  "created": "2015-09-09T13:01:09.160Z",
  "modified": "2015-09-09T13:01:09.160Z",
  "generator": "pundit-client",
  "target": {
    "source": "https://burckhardtsource.org/letter/682?semantic",
    "selector": [
            {
              "type": "RangeSelector",
              "startSelector": {
                "type": "XPathSelector",
                "value":
"/div[1]/div[1]/div[2]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/di
v[1]/div[1]/div[1]/div[1]/p[2]/span[2]"
              },
              "endSelector": {
                "type": "XPathSelector",
```

```
                "value":
"/div[1]/div[1]/div[2]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/di
v[1]/div[1]/div[1]/div[1]/p[2]/span[2]"
              }
            },
            {
              "type": "TextPositionSelector",
              "start": 1730,
              "end": 1738
            },
            {
              "type": "TextQuoteSelector",
              "exact": "Jenseits",
              "suffix": "» in Ihre verehrten Hände gelegt",
              "prefix": "und mein letzthin erschienenes «"
            }
          ]
  }
  "motivation": "linking"
  "body": [
        {
            "http://purl.org/spar/cito/cites": {
                "@id": "http://purl.org/net7/korbo2/item/103450"
            }
        },
        {
            "type": "TextualBody",
            "value": "Nietzsche",
            "purpose": "tagging"
        }
  ]
}
```


2.5.2. Relation to EDM profile for annotations

We have here annotations with RDF-like statements using predicates (URI) that are specific to a project. Handling it using the current EDM patterns, especially the one in Section '3.2.2. Link a CHO to another using a qualified link' of the EDM profile for Annotations, would face the same issues as mentioned for the SOCH case above. Moreover, the statements may not necessarily reflect a linking "motivation", thus other motivations should be allowed for graph-bodied annotations (describing, commenting).

More fundamentally, Pundit supports multiple bodies in one annotation, which is not supported in Europeana yet, but is under discussion for the extension scenarios. In the case of multiple (single-)statement bodies, this could however be handled by grouping the single statement of each body into one graph - using the pattern for graphs discussed in Appendix B of the EDM profile document.

In addition, linking to parts of external documents is not supported.

Pundit's annotations target parts of textual documents in XHTML, using XPath expressions expressed as a selector in Web Annotation. These can be compared to the implicit fragment selectors pencilled in the advanced EDM annotation scenario Q5 in Appendix 'A. Possible extensions to existing application scenarios', but implicit fragment selections cannot express an XPath. Therefore support for explicit fragment selectors would be needed. Targeting parts of XHTML documents is also not covered by the EDM profile, which only addresses targeting of parts of images and video resources. This is justified by the fact that XHTML documents are not currently displayed within the Europeana portal.

Pundit also introduces additional complexity in the textual bodies. The bodies composed of simple text are supported in the EDM profile, for example in Section '3.1.2. Annotate a CHO with a language-specific tag'. But Pundit also can indicate the purpose of each textual body as specified by Web Annotation (i.e., with an `oa:purpose` such as `oa:tagging`). The example annotation in the previous section includes such a case. While not supported by the EDM profile, it is close to one of the proposals under discussion for case Q3 of Appendix 'A. Possible extensions to existing application scenarios'.

Pundit Notebooks are represented in Web Annotation using Annotation Collections. These collections are currently not supported in EDM's annotation profile. However, one could argue that keeping track of them is slightly less important than exchanging the 'core' annotation data, outside of the original environment where users create and manage their annotations.

Finally, the main elements of the annotation metadata (creator, creation date) can be handled in the current EDM pattern.

2.6. 50s in Europe Kaleidoscope - Dealing with multiple layers of annotations

In the course of this CEF 2018 project (2018-2020) we were dealing with multiple layers of annotations for datasets and individual items. The objective of the project was to enrich Europeana with photographic materials about the 1950s, and to provide tools and methodologies for user engagement.

The project has the following stakeholders:

- Data producers
- Content holders
- Professional curators
- Non-professional curators

As part of the outcomes, the project developed a visual similarity matching algorithm, that - going across million training cycles and refinements - will eventually be able to generate numerical "fingerprints" that can be used to suggest similar items to annotators based on visual similarity. In the end, these fingerprints are not designed

for export to Europeana, but the tool was a good exercise in an area (the use of AI for improving user experience when browsing vast amounts of digital content) that is highly relevant.

In addition to this, the project's content providers curated their specific datasets relating to the visual imagery of the 1950s by adding relevant metadata information that aimed at granting a meaningful searchability in Europeana, given the project did not foresee the creation of a thematic/featured page or collection on Europeana portal. Such metadata enrichment happened either on providers source datasets (then imported via MINT and mapped/published), or via the WithCrowd platform[8]. This latter tool was noted to be possibly improved with a semi-automated batch annotation system, so that the content provider can upload a batch of images where fields with a common text (e.g. content provider or rights etc) are batch-annotated, while specific metadata fields (e.g. title, description etc) would need to be populated afterwards on a one-by-one basis.

Finally, as the thematic focus is about the 1950s, and in order to implement user engagement, the project organised engagement actions with the general public and users, for curation of a number of individual photographs on the WithCrowd platform. In fact, the 1950s is close to the memory of everyone and the input from general users can be very interesting for enriching the records and for collecting stories from various backgrounds. This action is to similar other crowdsourcing actions that collect user-generated tags (in terms of additional metadata)

In summary, the project delivered:

- computer-generated metadata
- professional curation and annotations
- annotations by general users, that are not necessarily super-relevant for curation but may be interesting for capturing users' interactions with cultural heritage

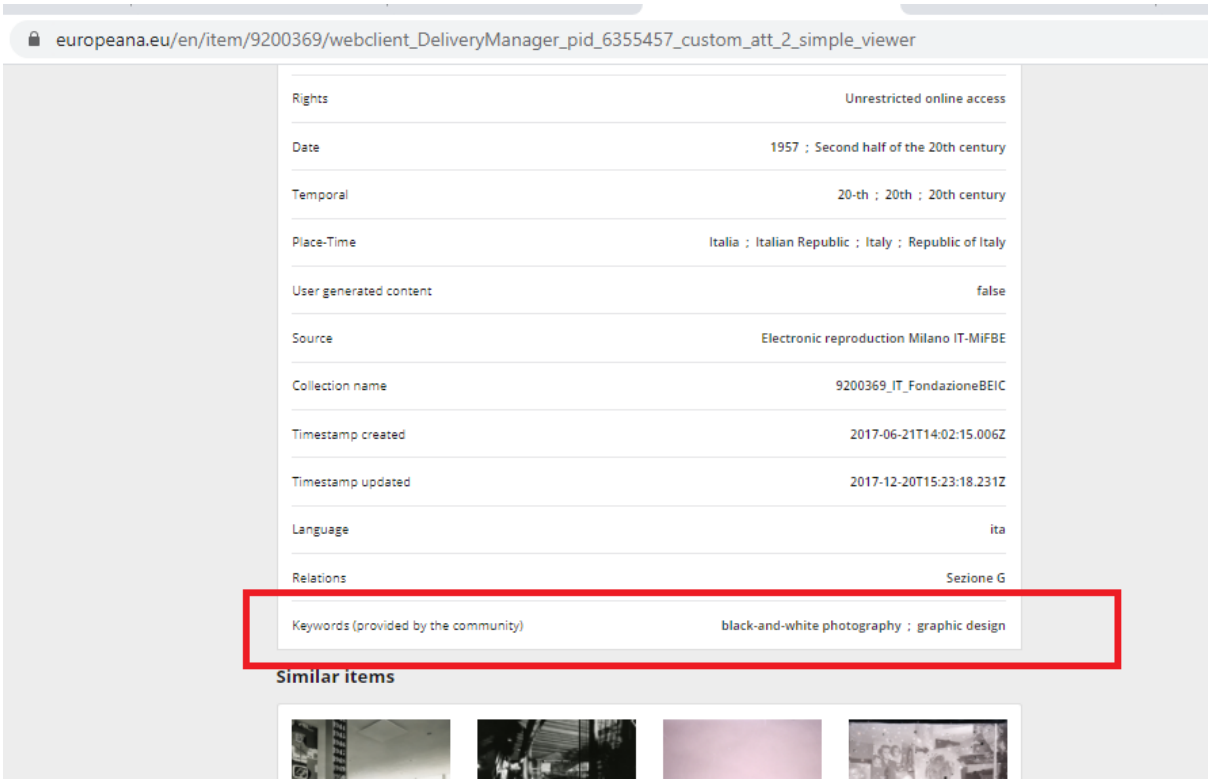
We understood that there are issues to be solved, possibly in the context of future projects:

1. How can we validate the relevance of the metadata labels generated by the algorithm? (mechanisms of validation)
2. How do we verify that the user-generated annotations are relevant or interesting? (mechanisms of ranking)
3. How can we push all these annotations in the Europeana records? (interoperability)
4. How can the content providers get their datasets back, enriched with all these different annotations? (interoperability)

The annotations created by the project were successfully delivered to Europeana via the Europeana Annotation API (ca. 8.000 annotations) and displayed as "Keywords provided

[8] https://withcrowd.eu. A custom version of the WithCrowd annotation tool is also used for the CrowdHeritage use case – see previous section.

by the community" at the bottom of each enriched record. This enables users to clearly identify that these tags have been added by other sources than the content providers, in an effort of granting transparency.



Nevertheless, the project identified that the potential of these annotations is not yet fully exploited by the Europeana portal. The annotations are not searchable via the search engine in Europeana and thus, while they offer additional information to the user who sees the record, they do not improve the record's searchability. Improved mechanisms and technical developments in order to enable the enrichments to be searchable in Europeana are currently ongoing.

2.6.1. Data requirements

The records to be linked with annotations are stored already in the Europeana platform (legacy records) or were published by the end of the project. The project's annotation model is based on W3C's Web Annotation Model, which is a structured model and format to enable annotations to be shared and reused across different hardware and software platforms. All the annotation terms are derived from controlled vocabularies and thesauri and more specifically:

- The Photo Vocabulary that was specifically developed in the Europeana Photography project (2012-2015) and hosted by KU Leuven/LIBIS.
- Wikidata.
- Getty Art & Architecture Thesaurus.

The following example contains an Annotation from the project:

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "type" : "Annotation",
  "created" : "2018-09-21T09:12:44.105Z",
  "creator": {
    "type": "Person",
    "name": "Maria Ralli"
  },
  "generated" : "2019-07-21T09:12:44.105Z",
  "generator" : {
    "type": "Software",
    "name": "WithCrowd",
    "homepage" : "https://withcrowd.eu"
  },
  "motivation" : "tagging",
  "body" : "http://bib.arts.kuleuven.be/photoVocabulary/13001",
  "target": "http://data.europeana.eu/item/09102/_UEDIN_214"
}
```

Note that the project uses the CrowdHeritage platform mentioned above, with a very similar functionality, i.e. crowdsourced annotations added by selecting terms from dropdown lists presenting pertinent entries included in the photo Vocabulary, Getty Art & Architecture Thesaurus and Wikidata.

2.6.2. Relation to EDM profile for annotations

The data requirements expressed in this section currently do not differ from the ones presented for the CrowdHeritage case (see Section 2.3). The case matches the scenario in Section '3.1.3. Annotate a CHO with a semantic tag' of the EDM profile for Annotations, with tags taken from the photo Vocabulary, Wikidata and Getty Art & Architecture Thesaurus.

2.7. The Europeana Media Generic Services project - Transcriptions, subtitles and captions for A/V content

The Europeana Media Generic Services project aims to increase the appeal, visibility, reuse, research and interaction with Europe's audiovisual heritage in Europeana Collections and third-party platforms that use Europeana content. The main outcome of this project will be a media player to be integrated in Europeana Collections that offers a single/consistent user experience for play-out of videos (possibly audio as well) resources.

One of the key features will be the display of captions/subtitles and transcriptions that can be either provided by the heritage institutions or contributed by users. This results in three sub cases:

- Audio transcriptions - These are text documents most often without any styling.

- Captions - Captions link text to a specific time in a document. Some captions additionally include information on styling and layout. Caption files sometimes contain metadata about them.
- Subtitles: Even though there is a conceptual difference between captions and subtitles, in practice both use cases are not so different when it comes to representation which is reflected on the fact that the same data formats are used to encode either of them. The only difference is that captions may have non-speech information, typically denoted within parenthesis e.g. "(sighs)" while subtitles typically do not.

This project was influenced by the recommendations Audiovisual Media Task Force[9]. Detailed information about the use case can be found in the document Supporting audio transcriptions, closed captions and subtitles at Europeana.

2.7.1. Data requirements

Audio transcriptions are text documents that most often come without any styling. The only information needed to be conveyed is therefore the media resource from which the transcription was made and the actual transcribed text with the respective language(s).

The Europeana Media project analysed the available formats for captions and subtitles and considered the following aspects to distinguish the existing formats and also identify the relevant information features that need to be conveyed in the model:

- Unit of time and precision: elapsed time is the most common unit of time, but some formats rely on the frames instead of a fixed time range, therefore varying depending on playout. Precision slightly varies between formats with 1 millisecond being the most granular and common for elapsed time, and a frame as the unit for frame based formats.
- Text styling: most HTML based formats (or equivalent) use CSS for styling (or equivalent, e.g., TTL), such as: font-size, font-family, colour, padding, etc.
- Text layout: some formats specify where and how the text should be presented, such as its position, orientation (horizontal/vertical), reading direction (left-to-right, right-to-left, top-to-bottom), and alignment.
- Text language: some formats specify language either generally or directly at each subtitle. Most HTML based formats use xml:lang attributes.

For the display of captions, subtitles and transcriptions, the Europeana Media project has defined its data model based on the EDM Fulltext profile for (full-text content of) newspapers.

For the first use case of modelling audio transcriptions of entire A/V resources, the media resource is represented as an edm:WebResource, while the textual transcription of the audio is represented as an edm:FullTextResource. The link between these two information resources is represented using an oa:Annotation.

[9] https://pro.europeana.eu/project/audiovisual-media-task-force

```
<!-- audio resource -->
<edm:WebResource rdf:about="AUDIO_URL">
...
</edm:WebResource>

<!-- the complete transcribed text -->
<edm:FullTextResource rdf:about="FULLTEXT_URL">
  <rdf:value xml:lang="en">… complete audio transcription in English …
  </rdf:value>
  <dc:language>en</dc:language>
<edm:FullTextResource>

<!-- annotation -->
<oa:Annotation rdf:about="ANNOTATION_ID">
<oa:motivatedBy
rdf:resource="http://www.europeana.eu/schemas/edm/transcribing"/>
  <oa:hasBody rdf:resource="FULLTEXT_URL">
  <oa:hasTarget rdf:resource="AUDIO_URL"/>
</oa:Annotation>
```

An example of an annotation for the transcription use case

For modelling captions and subtitles, the model considers the fact that both captions and subtitles divide the text transcript into time frames. The model represents the whole transcript as an edm:FullTextResource (the same way as for an audio transcript) and the frames are represented as annotations that refer to the time frame that the portion of text reflects.

The model also specifies the unit of time and precision. The reference to a period of time in a media resource or a single frame is done with a reference (i.e., URL) used within the target of an annotation. The period of time is indicated by appending the start and end times to the URL of the media resource the start and end times, in the form of "`#t=START,END`", following the W3C Media Fragment recommendation for temporal fragments[10]. The start and end times are encoded using the format "`HH:MM:SS.FRACTION`". The model does not consider the use of frame based formats.

```
<oa:hasTarget rdf:about="VIDEO_URL#t=00:00:05.780,00:00:07.000"/>
```

An example of a reference to a period of time in a media resource

The full-text model covers the representation of styling information associated to the text. This is achieved by referring to a CSS style class (via the oa:styleClass property and using a class that is defined within a CSS document and defined generally using the oa:styledBy property) associated to the oa:hasBody.

```
<oa:hasBody>
  <oa:SpecificResource>
    <oa:hasSource rdf:resource="http://..."/>
    <oa:styleClass>.font-12-bold</oa:styleClass>
  </oa:SpecificResource>
</oa:hasBody>
```

An example of representation of the styling information associated with the text

[10] https://www.w3.org/TR/media-frags/

For specifying the text layout, the model considers the positioning of the subtitle on the still frame (or viewport) of a video in combination (in combination with the temporal dimension). Regarding the direction of the text, the oa:textDirection is used on the `edm:FullTextResource` to indicate the text direction of the complete text or an `oa:SpecificResource` is used to indicate that only a specific portion of text reflects a different text direction than the rest.

The following example shows a full-text resource indicating that text needs to be displayed left-to-right.

```
<edm:FullTextResource rdf:about="FULLTEXT_URL">
  <rdf:value xml:lang="en">… complete audio transcription in English …
</rdf:value>
  <oa:textDirection rdf:resource="http://www.w3.org/ns/oa#ltrDirection"/>
  <dc:language>en</dc:language>
<edm:FullTextResource>
...
<oa:hasTarget rdf:about="VIDEO_URL#t=00:00:05.780,00:00:07.000"/>
```

An example of a full-text resource indicating that text needs to be displayed left-to-right.

The full-text model covers the representation of language information associated with the text. There are two levels where language can be indicated, either within the edm:FullTextResource corresponding to the most predominant (or only) language of the text, or within the body of the annotation to indicate that a specific portion of text is written in a different language than the remainder.

```
<oa:Annotation rdf:about="ANNOTATION_ID">
  <oa:motivatedBy rdf:resource="oa:transcribing"/>
  <oa:hasBody>
    <oa:SpecificResource rdf:about="FULLTEXT_URL#char=100,200">
      <oa:hasSource rdf:resource="FULLTEXT_URL">
      <dc:language>en</dc:language>
    </oa:SpecificResource>
  </oa:hasBody>
  <oa:hasTarget rdf:resource="VIDEO_URL#t=00:00:05.780,00:00:07.000"/>
<oa:Annotation>
```

An example of a caption of a video resource written in English.

2.7.2. Relation to EDM profile for annotations

The most relevant requirements from Europeana Media are supported by the EDM annotation profile. The profile defines three annotation motivations that address requirements of Europeana Media: `edm:transcribing`, `edm:captioning` and `edm:subtitling` (see Section '2.3 Motivation'). The language of the annotation's text is also addressed in Section '2.5.1 Textual Body'. The specification of temporal dimensions as required for captions and subtitles is currently addressed in Appendix 'A. Possible extensions to existing application scenarios'. The requirements for text styling and layout are not currently supported by the EDM Annotations profile, but they are fully addressed in the EDM Full-text profile.

3. Conclusion

The Task Force on Interoperability of annotations and user sets has gathered use cases on the use of annotations in projects related to Europeana. From these use cases, we have extracted data configurations that should be supported by Europeana so that interoperability can be achieved between these projects and the core Europeana service. This work has guided the specification of a profile of the Europeana Data Model for annotations, and its implementation in the Europeana Annotation API. The EDM profile for annotations is available at https://pro.europeana.eu/page/edm-profiles#annotations-profile.

The specification and service developed by Europeana do not support all the possibilities mentioned in the use cases inventorized here. This most often results from a combination of lack of business priority (e.g., annotations that are complex for the users and/or seldom produced) and technical obstacles (e.g. sophisticated data patterns). However, when an annotation case is not supported at the data level by Europeana, it is still mentioned in the EDM profile for annotations, together with one or several proposals to support it.

In the future, Europeana will monitor projects in the area to see if new cases emerge or if existing ones gain more traction. Especially, while no use cases in this Task Force were submitted about sets of objects created by users, these cases have appeared after the creation of the Task Force. In fact Europeana supports the creation of user sets on its own platforms. Also, there are projects interested in contributing such sets to Europeana. The EuropeanaTech community shall be ready to support these efforts in terms of data modelling if necessary.

europeana
network association

pro.europeana.eu

@EuropeanaEU

Co-financed by the Connecting Europe
Facility of the European Union