Annotation scenario for PAGODE: metadata creation, crowdsourcing, automated enrichment

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

Authors:
Valentina Bachi, PHOTOCONSORTIUM

TABLE OF CONTENTS

1 Introduction and rationale 3
2 Scenarios for metadata curation 4
2.1 Metadata creation for newly digitized content 4
2.2 PAGODE Annotation Pilot 5
2.3 Automated enrichment 7
3 Conclusions 8

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839

1 Introduction and rationale

This document is an extract form the MS5 Annotation Scenario that was realized in October 2020 in the framework of PAGODE – Europeana China project. Central to the project's efforts is the development of a shared and validated framework for the provision of rich information to Europeana, either in the form of descriptive source metadata for the newly digitized records or in the form of additional keywords to enrich legacy records (i.e. already existing records and collections in Europeana).

The specifications for the metadata have been developed, discussed, and finalized within *Activity 3 Content curation and digitisation*, under the coordination of sinologists from the University of Ljubljana (T3.1 Semantic background for the curation of China-related records in Europeana and T3.2 Workshop on curation of Chinese CH in Europeana). The semantic background for the curation of Chinese and China-related records in Europeana has been developed in several steps. First, the definition or the conceptual scheme of what constitutes Chinese and China-related CH in Europe was formulated, based on the concept of flows of people, objects and ideas. Secondly, the classification of various types of heritage content in that scheme, was used as the point of departure for metadata specifications in the form of keyword lists, linking to AAT and Wikidata entries. In those cases where such entries proved unavailable, new entries in Wikidata were created by the team at University of Ljubljana.

Both outcomes from Activity 3 (i.e. the conceptual framework and the lists of keywords with links to authority files) are now available for implementation by the project partners, and workflow can start in different scenarios for metadata curation.

The implementation scenarios which the project will deploy its efforts to enrich Europeana records are threefold:

- Metadata creation for newly digitized records (T.3.3, good quality metadata associated to 10,000 digital items)
- Manually curated annotations generated through two crowdsourcing campaigns focusing on Europeana records as well as new content provided by PAGODE (T2.2, target: 2,000 records)
- Automated enrichment by Artificial Intelligence (AI) and Natural Language Processing (NLP) on Europeana collections (T4.2, target: 20,000 records)

In the following chapter the three scenarios are described in the context of the work of PAGODE project.

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839

2 Scenarios for metadata curation

2.1 Metadata creation for newly digitized content

When a cultural heritage object is digitized, it is necessary to collect and attach to it the information associated with the digital object, which describe the object and its attributes. Part of this information is of an archival nature, and it is already known by the cataloguer at the time of the digitisation (e.g. description of the object and its story), part is generated through the process of digitization (e.g. technical information of the digitized file), part needs to be researched or established (e.g. the rights label to be attached to the digital record). This activity requires various levels of competency and is at the core of the activities of collections curators.

In the case of PAGODE, the metadata creation process requires specialist knowledge of Chinese and China-related cultural heritage. To this end the expert sinologists of the University of Ljubljana developed a conceptual framework and a list of keywords to support content providers in the creation of rich metadata to be associated to the newly digitized objects.

While it is complex to establish a one-fits-all scheme for the metadata curation of Chinese and China-related Cultural Heritage in the scope of Europeana, due to the variety of objects types and the granularity of information needed for various curation and reuse purposes, the conceptual framework and the long list of keywords developed in the project, with links to authority files (Getty AAT and Wikidata), will allow content providers, curators and Europeana Aggregators to use a shared, research-based and therefore authoritative set of descriptors for relevant collections.
For this reason, the semantic background developed in Activity 3 is one of the major outcomes of the project to be disseminated towards professional stakeholders and informs the whole project's implementation.

In addition, support is provided to content partners for understanding the Europeana Publishing Framework (EPF), with the recommendation to take into account the requirements of EPF metadata Tiers, soon at the stage of metadata creation, so that the dataset to be ingested in Europeana matches the highest quality standards. A summary on EPF Tier 4 for data and Tier C for metadata was shared and published on PAGODE's website.

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839

2.2 PAGODE Annotation Pilot

The PAGODE Annotation Pilot is dedicated to performing curation and enrichment on selected content already available in Europeana as well as on new content from partner collections, to be achieved through the crowdsourcing campaigns. The crowdsourcing activities, through which the campaigns are deployed, make use of the WITHCrowd environment.

PHOTOCONSORTIUM is in charge of the coordination of the Annotation Pilot.

The Pilot, unfolding in two campaigns (Autumn 2020 and Spring 2021) enables user engagement with Europeana content: different groups of participants are invited to look at Europeana records and annotate them either by adding terms selected from the PAGODE keyword list, either by up-voting or down-voting existing tags.

The crowdsourced annotations will be then checked and validated by the PAGODE Content Team, and finally sent to Europeana for publication via the Annotation API.

Participants include:

- Sinologists and expert curators
- Colleagues from Europeana Aggregators
- Metadata experts
- University students
- General public and culture lovers

With such a varied target audience, the campaigns are expected to deliver both very specific as well as more general annotations.

In terms of selection of Europeana content to be used in the crowdsourcing action, two different themes were identified:

- 'Scenes and People from China'.
  The focus of the campaign is on heritage photography.
  This theme is used for the campaign of Autumn 2020.
- Chinese Artefacts
  The focus of the campaign is on collections of museal content.
  This theme is used for the campaign of Spring 2021.

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839

The setup of the WITHCrowd environment was discussed with the CrowdHeritage expert during the summer time and initiated actually in September 2020.
A demo of the WITHCrowd tool customised for the need of PAGODE was produced and discussed with partners involved in the Annotation Pilot.
A review of the graphics of the webpage and the functionalities of the PAGODE Annotation was conducted in the occasion of the PAGODE Project Management Board on 6th October 2020, in sight of the 'annotation sprint' on 26th October 2020 at the PAGODE Technical Workshop. The time in between was used to work further on content selection and curation as well as on the controlled vocabularies, to provide additional partner demonstrations and to discuss a workflow for possible future campaign refinements.

In addition to these activities - which correspond with what was set out in the Grant Agreement - on 21st October 2020 a first meeting took place with KU Leuven students of the Master in Digital Humanities, who will be promoting the crowdsourcing activities in the frame of PAGODE, as well as organizing their own (physical or online) crowdsourcing event in 2021. These students have actively opted in for this project, which is a part of their 'Cultural Policy' course, out of interest in either Chinese/China-related heritage, digital cultural heritage, digital curation or user engagement strategies. With the business plan scheduled to be presented in January 2021 and the promotional activities and event to be produced in Spring, the students will contribute to the PAGODE project by generating more visibility for Europeana and diversifying the community involved in the annotation campaign.

The PAGODE crowdsourcing campaign is available at the address https://crowdheritage.eu/en/china.

The two campaigns of the Annotation Pilot are expected to deliver the target of min 2,000 records enriched by crowdsourced annotations.

The new tags generated via the Annotation Pilot are sent to Europeana via the Annotation API. Representatives of Europeana Foundation confirmed that the annotation provided via the Annotation API will be displayed in Europeana as 'Keywords from the community' in association to the source metadata originally provided by the content holder.

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839

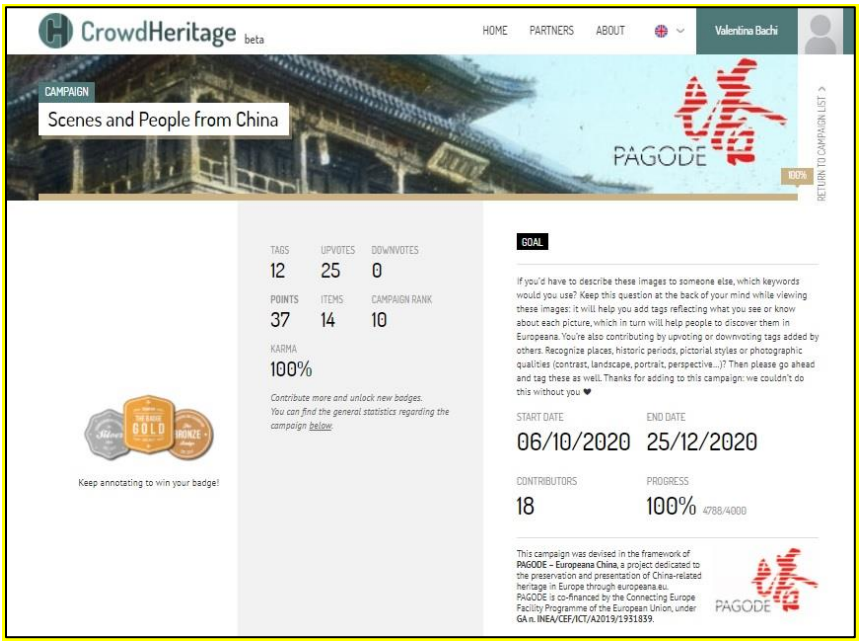

Fig. 2: The current annotation campaign on the theme 'Scenes and People from China'

2.3 Automated enrichment

Another possibility for metadata curation and enrichment leverages on NLP techniques and AI.

This is explored with various research and Generic Service projects, and in particular with the sister project Europeana XX – Century of Change, in the frame of the meeting with representatives of Europeana Foundation, mentioned in the following paragraph, where PAGODE participated and the enrichment processes was further discussed.

A combination of AI and NLP allows to recognize automatically information that are embedded in the content and in the metadata, and to add relevant terms and links to authority files in the existing metadata. This has the advantage of enabling bulk enrichment of large datasets, which would require a huge effort and extended timescale if done manually.

In the scope of PAGODE, automated enrichment will be performed by use of algorithms trained with the list of keywords developed in Activity 3, running on metadata from various Europeana collections. Automatic enrichment is part of Activity 4. The algorithms are expected to recognize the terms and enrich the metadata with the link to the respective entries in Getty AAT or Wikidata, and the enrichments will be made available to Europeana, via the Annovation API, for display in the respective records.

A call was organized by Europeana technical team on the 22nd of October 2020, to present recent developments in the Europeana engine that will allow for proper display of machine-created metadata. The call was planned in the scope of the parallel GS project Europeana XX: Century of Change, and Photoconsortium was invited also representing PAGODE, where automated enrichment is planned as well. Collaboration between the two GS projects is in facts aimed at generating synergies where possible, and at stimulating cross-exchange of information and knowledge.

Currently, work is ongoing at Europeana in order to provide new features and modifications in the Europeana Data Model in order to accommodate the algorithm-generated information and to offer the user an assessment on the level of confidence of such information. For the time being however this feature is not yet available, and thus it was confirmed that the enrichments performed in PAGODE is pushed to Europeana via the Annotation API.

3 Conclusions

This document outlines different scenarios for metadata curation and enrichment as identified in the context of PAGODE project.

Specifications for the creation of rich metadata for the description of Chinese and China-related CH in Europeana are available in the form of the PAGODE conceptual scheme and of the keyword list developed in Activity 3. These specifications will also be shared widely for dissemination and uptake from other stakeholders in – and beyond - the Europeana family.

The effort performed in the development of the PAGODE conceptual scheme and keyword list is at the basis of the creation of high quality and rich metadata for newly digitized records offered to Europeana, and for enriching existing records that are already published in Europeana. This is the objective indicated in the PAGODE Grant Agreement. However, the work done is not limited to this objective: it aims to have a wider exploitation and to go beyond the PAGODE project. The guidelines and reports made available online as well as the public workshops, namely the Semantic Workshop in July 2020 and the Technical Workshop in October 2020, represent an occasion of knowledge transfer and capacity building, offering resources that are made available to many other stakeholders.

Co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/1931839