# Europeana Translate: Providing multilingual access to digital cultural heritage

**Eirini Kaldeli[1], Mercedes García-Martínez[2], Antoine Isaac[3], Paolo Sebastiano Scalia[3], Arne Stabenau[1], Iván Lena Almor[2], Carmen Grau Lacal[2], Martín Barroso Ordóñez[2], Amando Estela[2] and Manuel Herranz[2]**

[1] National Technical University of Athens, Greece, [2] Pangeanic SL, Spain,
[3] Europeana Foundation, The Netherlands

## Abstract

Europeana Translate is a project funded under the Connecting European Facility with the objective to take advantage of state-of-the-art machine translation in order to increase the multilinguality of resources in the cultural heritage domain.

## 1 Europeana Translate Mission

The Europeana platform[1] provides access to European digital cultural heritage (CH). It currently contains more than 58 million digital items contributed by more than 3,500 different museums, libraries, archives and galleries from all EU member countries. Each item is described via a set of metadata fields that convey essential information about it, such as its title, free text description, creator, etc., and help users to discover and understand the objects they are interested in. Currently, the majority of records contain terms only in a single language, the data providers' language. This lack of multilingual metadata hampers Europeana's goal of offering broad access to its collection across languages.

In order to address this challenge, the Europeana Translate project (May 2021 until Apr 2023) seeks to exploit and build on state-of-the-art machine translation (MT) services to advance the multilinguality of European digital CH. The project proposes a sustainable workflow and accompanying toolset which can be used to enrich CH datasets with multilingual metadata. The consortium includes: the National Technical University of Athens, the Europeana Foundation, Pangeanic SL, the European Fashion Heritage Association, the Netherlands Institute for Sound and Vision, and the Michael Culture Association.

A selection of CH metadata resources in various languages will be used to train and improve the accuracy of translation algorithms in this specific sector. The proposed solution will be applied to produce automatic translations from the 23 official EU languages to English for at least 25 million metadata records on the Europeana platform. Moreover, Europeana Translate will make openly available a number of multilingual resources from the CH sector, a domain of public interest which is currently under-represented in existing repositories of language corpora. To this end, the project will publish to the ELRC-SHARE[2] repository CH metadata in parallel languages and monolingual records under a free reuse license (CC0).

## 2 Architectural Overview

Figure 1 provides an overview of the overall Europeana Translate architecture.

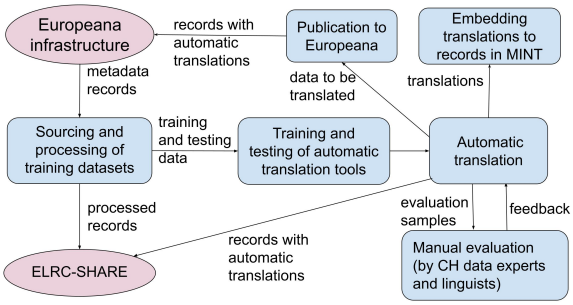


**Figure 1:** Europeana Translate Workflow

*Sourcing and processing of in-domain training datasets*: In this step (detailed in Section 3) we select and process all the data that will be used for the



[1] https://www.europeana.eu/

[2] https://www.elrc-share.eu/

in-domain training of the translation tools and apply all the necessary processing and cleaning, so as to bring them to the formats expected by the translation tools.

*Training and testing of automatic translation tools*: The training phase will use 12 million translation segments from existing generic linguistic corpora, enhanced with the CH-specific data resulting from the previous step.

*Automatic translation*: The in-domain trained MT engines will be deployed and their capabilities exposed via an API interconnected with the Europeana platform as well as the MINT[3] (Metadata INTeroperability services) aggregation platform, which is used by several CH organisations for uploading and managing metadata records.

*Manual evaluation*: Two complementary evaluation methods will be performed to assess the produced automatic translations: evaluation by linguist experts using the Machine Translation Evaluation Tool MTET;[4] and evaluation by CH domain experts using CrowdHeritage,[5] a platform for organising online crowdsourcing campaigns in the CH domain.

*Publication to Europeana*: The translations retrieved by invoking the in-domain MT engines will be ingested, indexed, and presented on the Europeana platform. To save on indexing space and technical complexity, the idea is to use English translations as a pivot that acts as the bridge for translating all other languages (and search queries) to and from.

*Embedding translations to MINT records*: The automatic translations can also be inserted as enrichments to datasets uploaded in MINT. The augmented records can then be published to Europeana or be further exploited by CH organisations' own platforms.

## 3 Selection and filtering of training data

The main source of training data is metadata records with parallel languages retrieved from the Europeana platform. In the cases where the amount of bilingual data is not adequate, monolingual data will also be used to specialise the models via the generation of synthetic data. Complementary to the training data corresponding to metadata records, multilingual vocabularies relevant to the CH domain and used by Europeana for semantic enrichment are also exploited for the domain adaptation of the translation engines.

Figure 2 provides an indication of the amount of monolingual and bilingual (English–EU language) metadata fields in Europeana across different languages. For many languages there are more than 100,000 bilingual metadata fields, an amount which is considered a sufficient for in-domain specialisation. At the same time, some languages, such as Hungarian and Slovakian, are significantly underrepresented.

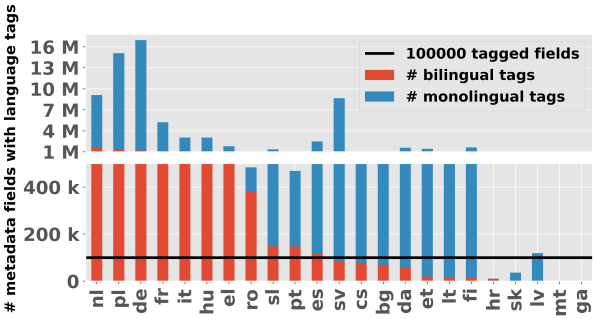


**Figure 2:** Raw number of mono- and bi-lingual (English–EU language) metadata fields on Europeana. An horizontal line marks the threshold of 100,000 fields.

Note that the plot only considers a subset of all metadata fields and only the values that are provided with explicit language tags. It may be possible to obtain more data by applying advanced data analysis, especially language detection. The numbers indicated here refer to unfiltered data that need to undergo further processing, since only a fragment of the raw data is actually suitable for training. To retain relevant data, multiple processing steps are applied, including the de-duplication of metadata field values repeated across many records, segmentation, and various types of cleaning, such as identification of incorrect language tags and pruning of incompatible value pairs.

In conclusion, Europeana Translate has the potential to significantly improve the multilinguality of CH items. The project builds on a well-defined architecture and has conducted an investigation of available raw data that can be leveraged for in-domain training. Preliminary experiments for translating metadata from French to English demonstrate an improvement of results compared to generic models. Several challenges still remain, such as acquiring additional training data for underrepresented languages and adopting appropriate methods for evaluation.

[3] http://mint.image.ece.ntua.gr/
[4] http://mtet.pangeamt.com/
[5] https://crowdheritage.eu/