ARTICLE

Adapting Machine Translation Engines to the Needs of Cultural Heritage Metadata

Konstantinos Chatzitheodorou, Eirini Kaldeli, Antoine Isaac, Paolo Scalia, Carmen Grau Lacal, and MªÁngeles García Escrivá

ABSTRACT

The Europeana digital library features cultural heritage collections from over 3,000 European institutions described in 37 languages. However, most textual metadata describe the records in a single language, the data providers' language. Improving Europeana's multilingual accessibility presents challenges due to the unique characteristics of cultural heritage metadata, often expressed in short phrases and using in-domain terminology. This work presents the EuropeanaTranslate project's approach and results, aimed at translating Europeana metadata records from 23 EU languages into English. Machine Translation engines were trained on a cleaned selection of bilingual and synthetic data from Europeana, including multilingual vocabularies and relevant cultural heritage repositories. Automatic translations were evaluated through standard metrics and human assessments by linguists and domain cultural heritage experts. The results showed significant improvements when compared to the generic engines used before the in-domain training as well as the eTranslation service for most languages. The EuropeanaTranslate engines have translated over 29 million metadata records on Europeana.eu. Additionally, the MT engines and training datasets are publicly available via the European Language Grid Catalogue and the ELRC-SHARE repository.

INTRODUCTION

Multilingual availability of metadata is an important factor that affects the browsing, retrieval, and display of digital cultural heritage (CH) collections. Multilingual metadata enables users to discover and understand more sources of information and access the knowledge and history of other cultures and less common language groups, thus making collections accessible to more diverse audiences. This is of particular interest to Europeana, the European digital library that contains more than 56 million digital items contributed by more than 3,500 different cultural heritage institutions 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, description, creator, etc. This metadata helps the users of the platform to discover and understand the objects they are interested in. However, the majority of the records contain terms only in a single language, the language of the data provider. This lack of multilingual metadata hampers Europeana's objective to offer broad access to its collections across languages for use and reuse.

About the Authors

Konstantinos Chatzitheodorou (kchatzitheodorou@ionio.gr) is Postdoctoral Researcher, Ionian University. **Eirini Kaldeli** (ekaldeli@image.ntua.gr) is Research Associate, National Technical University of Athens. **Antoine Isaac** (antoine.isaac@europeana.eu) is Research and Development Manager, Europeana Foundation. **Paolo Scalia** (paolo.scalia@europeana.eu) is Technical Business Analyst, Europeana Foundation. **Carmen Grau Lacal** (c.grau@pangeanic.com) is Computational Linguist, Pangeanic SL. **MªÁngeles García Escrivá** (ma.garcia@pangeanic.com) is AI/ML Team Lead, Pangeanic SL.

Submitted: 20 January 2024. Accepted for Publication: 15 June 2024. Published: 23 September 2024.

https://doi.org/10.5860/ital.v43i3.17247

One of the core solutions proposed in Europeana's multilingual strategy in order to improve the search and display of CH items across languages is to use English as a pivot language.[1] Machine-translating all metadata records into English, storing, and indexing them would then enable supporting multilingual display and search via the runtime translation of queries into English. Some preliminary experimentation with the eTranslation services conducted by Europeana pointed to new opportunities towards this direction but also posed certain challenges.[2] Europeana is working with collections described in not less than 37 languages and aims to make them discoverable with search terms that could come in any language. What is more, metadata is not like natural language with complete sentences and predictable grammar; it is often presented in short phrases or even single words, which means that the context required for an accurate translation is hard to find. In addition, the terms used can be very specific; they may look like a general term but they have a different meaning when used in a CH context. Finally, the data is rather rich in named entities, and these references are to persons that are not necessarily widely known.

To address these challenges, EuropeanaTranslate exploits and builds on state-of-the-art automated translation technologies with the aim to advance the multilingualism of European digital resources in CH.[3] To this end, 23 machine translation (MT) engines were trained on appropriately selected and cleaned metadata sourced from Europeana and other repositories with relevant data with the aim to achieve good-quality translations from one of the official EU languages into English. The developed tools were made openly available and interconnected with established platforms used for the management, enrichment, and display of CH data. This way, CH organizations can integrate the provided tools in their workflows to translate their metadata into English, evaluate and filter the results, and ultimately deliver them to Europeana or to their individual platforms. The MT engines have been applied to translate more than 29 million metadata records on Europeana.eu, thus improving the multilingual experience provided to its users. Moreover, EuropeanaTranslate contributes to overcoming the underrepresentation of CH-related corpora on existing repositories of language corpora, by making openly available a selection of appropriately processed multilingual data amenable for training purposes via the ELRC-SHARE repository, under a CC0 public dedication license.

RELATED WORK

Access to cultural heritage collections across languages is a challenge for large digital libraries. Content and metadata are most often monolingual. Until recently, systems relied on manual translation of data, including controlled vocabularies, which constrains application to specific domains and/or selected collections, as in the case of the World Digital Library (https://www.loc.gov/collections/world-digital-library/).

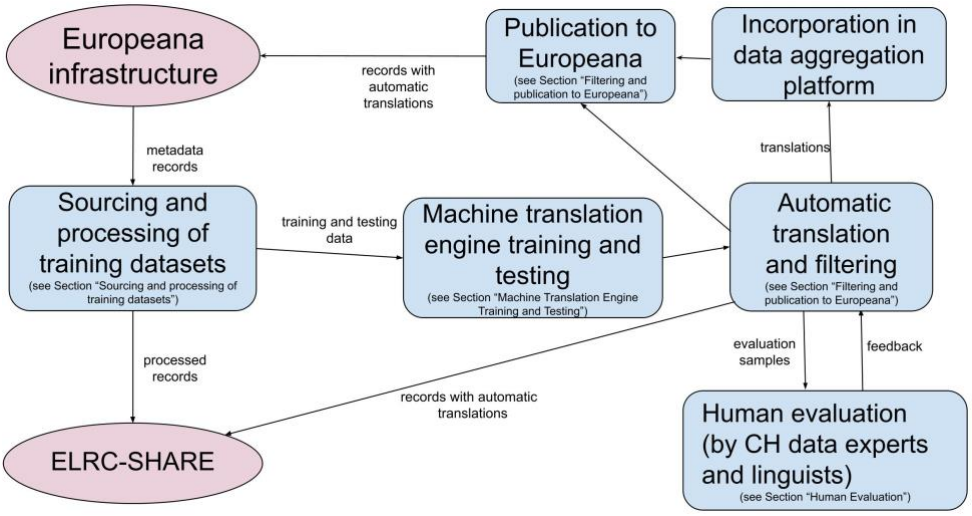
The enhanced availability of automatic translation opens perspectives for wider-scale multilingual access, especially using multilingual information retrieval.[4] Europeana has devised a multilingual strategy (https://pro.europeana.eu/post/europeana-dsi-4-multilingual-strategy) that seeks to exploit automatic translation of metadata using English as a pivot language. The approach has already been tested with some success, both in the information retrieval community and in the specific context of Europeana, which has run a pilot that applied query translation to English for the Spanish version of the Europeana website, experimenting with the Google Cloud Translation service.[5] The experiments led to promising results but also pointed to a number of open issues that have to be addressed before adopting multilingual search at a wider scale, such as the need for more good-quality metadata in English.

The domain-adapted MT engines developed under EuropeanaTranslate were built on the automated translation tools developed in the context of the NTEU Action (Neural Translation for the European Union–2018-EU-IA-0051). NTEU aimed to provide a capacity service to eTranslation by building a near-human professional quality neural engine farm which includes all 552 EU language combinations. Special attention was given to languages with fewer available resources, such as Irish or Maltese, to ensure that the language models perform well even in cases where data is limited. NTEU uses advanced technologies such as the Transformer architecture and tried a variety of techniques, such as the use of extra data via back-translation and using what is already been learned from one task to help with another, and learning without having a teacher on a single-language set of texts.[6] Each NTEU engine is constructed using approximately 12 to 15 million parallel data instances.

METHODOLOGY

We constructed 23 MT engines tailored to CH metadata within the Europeana framework. The methodology employed is outlined in the architectural overview of the Europeana Translate toolset in figure 1, which illustrates the key workflow steps involving the main target platforms and other project tools.

Figure 1. Main workflow overview



Sourcing and Processing of Training Datasets

We start by carefully acquiring and cleaning pertinent data for in-domain training. This includes both bilingual and monolingual metadata from Europeana, various multilingual CH vocabularies, and additional sources addressing languages with limited resources on Europeana.

Machine Translation Engines Training and Testing

In the subsequent phase, our attention was directed towards experimenting with different setups and finding the optimal approach to train the MT engines. To assess the performance and refine our methods, we employed automated metrics such as BLEU and TER.[7]

Human Evaluation

To validate the quality of the generated translations and confirm the automated metrics, we performed an evaluation by experts, involving both linguists and CH experts.

Automatic Translation and Filtering for Publication to Europeana and other Data Aggregation Platforms

The MT engines are exposed via an API that streamlines the production of automatic translations. This API is invoked by the Europeana platform to translate metadata records that lack English metadata. The API is also integrated into the MINT data aggregation platform, which is used by several CH aggregators to manage the metadata of their collections.[8] This enables CH organizations to integrate and manage translations in their local platforms.

The aforementioned steps are explained in more detail in the following sections.

SOURCING AND PROCESSING OF TRAINING DATASETS

Our endeavor was largely based on the premise that Europeana's metadata is a key asset for training high-quality MT engines. This section delineates the process employed to collect, select, and filter data for the in-domain training of MT engines.

Our main source of in-domain data includes bilingual metadata (from a European language to English) collected from the Europeana platform and expressed in the Europeana Data Model (EDM). A subset of all the EDM metadata fields has been selected, considering fields with textual values that are relevant for translation and that are accompanied with explicit language tags. Monolingual metadata (values without an English translation) were also collected since they can be useful in cases where there is not a sufficient number of English translations.

Subsequently, we tried to enhance the quality of the metadata. First, paragraphs were segmented into sentences that we aligned with their available translations—a crucial step given the sentence-level nature of MT engine training. Then we applied various filters to clean the data. These included de-duplicating repeated sentences and several strategies for eliminating misaligned pairs. Figure 2 provides some examples of textual values that have been discarded since they have been considered invalid translation pairs due to various reasons: value pairs that were identical in different languages; pairs of sentences whose length varied significantly between the source and target; sentences with standard phrases, such as “original language summary” and “series title,” which appear in certain languages without having an equivalent in the other language; values whose language tag was found to be wrong after the application of automatic language detection and subsequent inspection.

Figure 2. Examples of sentences that underwent the cleaning process

Identical repeated sentences
EN: Donald TUSK, President of the European Council, receives Petro POROSHENKO, President of Ukraine:- exterior, arrival and welcome, roundtable.
FR: Donald TUSK, President of the European Council, receives Petro POROSHENKO, President of Ukraine:- exterior, arrival and welcome, roundtable.

Length-discrepant sentences
EN: Christ's Head: fragment from The Last Supper, a lost mural painting on the southern wall of the Refectory of the Dominican Monastery, Ghent
FR: Tête du Christ

Sentences such as "Original Language Summary" and "SERIES TITLE" which are not translated
EN: Gymnastics festival at the Bois de Vincennes cycling arena on the 10th anniversary of the death of Léo Lagrange.
FR: Original language summary: Fête de la gymnastique au Vélodrome municipal au Bois de Vincennes pour le 10ème anniversaire de la mort de Léo Lagrange.

Sentences with the wrong label
EN (*Labelled as*): Φωτογραφία του είδους Lupinus albus (Λευκό Λούπινο) από τη φυτοθήκη του ΜΦΙΚ.
EL (*Labelled as*): Photo of the species Lupinus albus (White Lupin) from the herbarium of NHMC.

Multilingual glossaries relevant to the CH domain were also considered complementary to metadata records. Europeana integrates diverse vocabularies available as (SKOS) Linked Open Data (https://www.w3.org/2001/sw/wiki/SKOS/Datasets), which are fetched by de-referencing their URIs. We complemented our training data with the labels attached to 20,764 distinct URIs from 22 providers, including the Getty vocabularies (http://vocab.getty.edu/) and Wikidata (http://www.wikidata.org/). The final step concentrated on cleaning these terminologies (see fig. 3).

Figure 3. Examples of term cleaning

Removing multiple targets or sources
Original: grand piano <> piano à queue, de concert
Cleaned: grand piano <> piano à queue

Aligning multiple sources with multiple targets
Original: gate / fortified city <> porte / ville fortifiée
Cleaned: gate <> porte / fortified city <> ville fortifiée

Removing parentheses containing explanations
Original: Archaeology (Science) <> Archéologie (Science)
Cleaned: Archaeology <> Archéologie

Table 1 provides a breakdown of data statistics per language pair, after the application of the aforementioned cleaning process.

Table 1. Number of segments per pair of languages after cleaning.

Language pair	Bilingual segments exported from metadata	Bilingual labels from glossaries
English – Hungarian	77,292	12,200
English – Greek	70,811	11,112
English – German	67,954	10,859
English – Dutch	46,797	9,319
English – French	46,986	8,241
English – Polish	33,941	7,911
English – Spanish	36,829	7,624
English – Romanian	19,280	4,691
English – Italian	18,986	4,273
English – Czech	16,850	3,777
English – Portuguese	13,317	3,244
English – Swedish	12,088	3,231
English – Danish	11,194	3,125
English – Finnish	6,083	2,974
English – Slovenian	3,684	2,416
English – Estonian	1,402	2,398
English – Slovak	1,415	1,714
English – Lithuanian	688	1,714
English – Bulgarian	355	1,445
English – Irish	311	1,274
English – Latvian	30	1,175
English – Maltese	0	968
English – Croatian	0	257

It became evident that many languages are underrepresented when considering the availability of bilingual data on Europeana. For many languages there are less than 5,000 bilingual segments available, which is the minimum number that was considered sufficient for in-domain training in a language.[9] Therefore, we deployed strategies to complement the bilingual data. First, as English is the target language, high-quality monolingual data in English, provided by CH professionals from the UK, was processed to create synthetic data for the 23 languages. To this end, we used a neural MT model to automatically translate a selection of about 4 million cleaned English segments into

the corresponding language. We also performed experiments using synthetic data produced from source monolingual data, but this strategy did not lead to better results than the generic model. Additionally, we acquired more bilingual data from the OPUS open parallel corpus using cosine similarity between Multilingual Universal Sentence Encoder embeddings to select data close to Europeana's.[10]

This strategic application of cosine similarity not only enhances our understanding of language representation but also ensures a more thorough and nuanced approach to rectifying underrepresentation. Moreover, it stands out as a superior approach, especially considering the bilingual nature of the OPUS collection, when compared to the generation of synthetic data.

MACHINE TRANSLATION ENGINES TRAINING AND TESTING

Achieving high-quality MT engines depends on how well we handle the details of the neural framework and architecture and on the quality of the data. Navigating the challenges of data intricacies posed a significant hurdle, intensifying the overall complexity of the task.[11]

Pretrained NTEU bilingual MT engines were the basis of our training. Using pretrained MT engines is advantageous over training from scratch due to the former's ability to leverage learned representations and avoid the need for massive datasets. According to Chenyang and Luo, pretrained engines capture useful linguistic features, enabling better generalization.[12] This approach aligns with transfer learning principles, enhancing efficiency and performance.[13]

Our training (or more precisely transfer learning) employed the state-of-the-art Transformer architecture within the OpenNMT framework, featuring neural networks, a batch size of 4,096 tokens, Adam optimization, and dropout techniques.[14] Iterative refinement of the training process continued until no further enhancements were observed during evaluation with the development data. To enhance domain-specific adaptation, we conducted additional training epochs with increased learning rates, prioritizing Europeana data over the generic data that was used to train the NTEU engines.

The acquired data described in the section "Sourcing and processing of training datasets" underwent normalization and tokenization, using the Moses tokenizer, and further underwent Byte Pair Encoding to reduce unknown words and enhance vocabulary coverage.[15]

To better handle the process, we categorized the languages into four groups based on resource availability. The first group comprises highly-resourced languages (e.g., Spanish, Italian, French, Dutch) with over 5,000 parallel sentences from Europeana. The second group includes Slovenian, Estonian and Slovak, containing 1,000 to 5,000 parallel sentences. The third group involves Lithuanian, Bulgarian, Irish, and Latvian, with less than 1,000 parallel sentences. The fourth group consists of Croatian and Maltese, lacking parallel sentences from Europeana.

To assess MT system quality, we retained 1,000 sentences not included in training datasets. When data was insufficient, we supplemented it with synthetic data, glossaries or other resources, as described above. To estimate the effectiveness of the engines during the training process, we used four reference automatic metrics: chrF, TER, BLEU, and COMET.[16] chrF measures n-gram overlap, TER estimates post-editing needs for human-quality text, BLEU counts matching n-grams in candidate sentences, and COMET integrates source input and target-language reference translation information. Attempts to train engines using a large amount of English monolingual and synthetic data yielded suboptimal results. Using English-based synthetic data proved ineffective for certain low-resource languages, but led to improvements for others. Balancing with

OPUS collections notably improved outcomes in the cases of languages with the least resources available on Europeana. The use of multilingual CH glossaries exhibited limited success, with results usually being slightly worse when compared to using only bilingual or synthetic data. This can be attributed to the fact that multilingual CH vocabularies lack context, and they are usually used to guide the human translator with multiple choices, so further cleaning and context is necessary for them to be useful for training.

Table 2 displays the automatic scores for each language pair alongside the approach that achieved the highest scores.

In group 1 (light green in table 2), training the NTEU model with bilingual data yielded optimal results, with certain languages (Greek, Czech, and Romanian) demonstrating enhanced quality through the integration of glossaries. In group 2 (light yellow), using bilingual and synthetic data proved more effective; however, for Estonian, further improvements were observed by incorporating glossaries. For groups 3 (orange) and 4 (light red), using selected data from OPUS led to the most significant improvements. Specifically, Irish demonstrated superior results when using synthetic data.

The domain-adapted engines were also compared with the results achieved by the generic NTEU engines (before the in-domain adaptation) as well as with Google Translate and eTranslation. The achieved results with respect to the used metrics (https://docs.google.com/document/d/1kMwUQslRbPI2FJZ97IvvYgFQjoh_uSDNq0T-XWQZlrY/) demonstrated significantly superior results for the EuropeanaTranslate engines when compared to the generic engines as well as with eTranslation (https://ec.europa.eu/digital-building-blocks/sites/download/attachments/684630922/eTranslation%20%28dashboard%29.pdf?version=1&modificationDate=1694681090484&api=v2) for the vast majority of the languages. The competition with Google Translate was close, with EuropeanaTranslate achieving better results for most languages.

Table 2. Scores achieved by the best-performing MT model. The direction of the arrows next to the metrics indicates whether a higher or a lower value entails a better performance.

Group	Language	Approach	CHRF↑	TER↓	BLEU↑	COMET↑
1	Czech	bilingual + glossary	64.3	50.4	38.9	0.61
	Danish	bilingual	64.8	46.1	42.5	0.67
	German	bilingual	64.7	53	43.2	0.81
	Greek	bilingual + glossary	67	46.6	44.4	0.59
	Spanish	bilingual	67.1	47	40	0.64
	Finnish	bilingual	60.2	52.1	34.2	0.58
	French	bilingual	79.3	30.3	63.2	0.96
	Hungarian	bilingual	57.4	67.8	35.7	0.17
	Italian	bilingual	67.1	44.4	42	0.76
	Dutch	bilingual	69.3	41.2	52.6	0.86
	Polish	bilingual	68.9	44.4	50.2	0.90
	Portuguese	bilingual	76.8	35	57.9	0.85
	Romanian	bilingual + glossary	71.8	40.9	46.1	0.83
	Swedish	bilingual	61.4	48.3	31.9	0.57
2	Swedish	bilingual + synthetic (monolingual) + glossary	54.2	67	27.4	0.14
	Slovak	bilingual + synthetic (monolingual)	59.4	58.9	33.2	0.5
	Slovenian	bilingual + synthetic (monolingual)	59.1	54	38.6	0.59
3	Bulgarian	data selection (OPUS)	61.3	53	37.7	0.81
	Irish	synthetic (monolingual)	18.8	92.7	12.8	-0.89
	Lithuanian	data selection	54.6	80.6	14.8	-0.13
	Latvian	data selection	52.7	75.1	28.9	0.35
4	Croatian	data selection	48.3	66	26.8	0.22
	Maltese	data selection	60.2	58.4	35.6	0.35

HUMAN EVALUATION

We adopted two complementary human evaluation methods to assess the produced automatic translations: evaluation by linguist experts and by CH domain experts. As pointed out by previous initiatives, insights by CH experts complement the assessments by professional translators: the latter may not be well aware of specific art–historical or "local" terminology, while the former may be less versed in terms of fluency, grammar, and syntax.[17]

In both human evaluation methods, participants were asked to rate the automatic translations into English on a scale from 0 to 100, considering aspects such as accuracy (general meaning), adequacy (proper use of terminology), and fluency (grammatical correctness).[18] In addition to entering a rating, participants were also highly encouraged to provide supplementary information about the type of errors they spotted as well as post-edit translations so that they are improved.

In the case of the evaluation by linguist experts, participants were invited to evaluate the automatic translations from 22 European languages into English (Maltese was not considered, since there are no records at all in that language on Europeana). In the case of the evaluation by CH domain experts, participants were invited to evaluate the translations of records sourced from the Europeana platform coming from three representative CH domains/Europeana aggregators (fashion, audiovisual, and museum heritage) in three of the most common source languages on Europeana (French, Dutch, and Italian).

The target for each evaluation method was to have at least 500 translated metadata field values evaluated per source language. The segments represent the textual values of metadata fields that were selected based on their relevance in terms of multilingual searchability and presentation for the Europeana platform (https://pro.europeana.eu/post/publishing-framework). Furthermore, the representation of these metadata fields in the evaluation sample reflects their frequency of occurrence and significance in the Europeana metadata. For example, translations of the field "dc:description" represent roughly 30 percent of the sample while translations of "dc:format" represent about 2 percent. The different metadata fields also include a rich diversity of textual values (single words, sentences, presence of named entities of various types, time periods, complex formatting, etc.).

Evaluation by linguist experts

In the evaluation campaign involving linguist experts, each language combination was evaluated by two different participants. All recruited evaluators were professional translators with experience in MT post-editing. Besides providing a rating, linguist experts were also asked to provide free-text comments about issues they detected during the review as well as a correction of the commented segment concerning the identified issue. The feedback was provided by using the Machine Translation Evaluation Tool evaluation platform (https://wiki.pangeanic.com/index.php/MTET_User_Manual).

In total, 44 freelance linguists remotely participated in this campaign. Each linguist evaluated a dataset consisting of 500 metadata field values automatically translated into English. All of these values, each corresponding to a translation unit (TU) in the MTET platform, were scored from 0 to 100 and were commented (if not rated with 100 points) with a specification of the detected error types, examples of the respective errors, and possible corrections. With regards to inter-annotator agreement, we observed that cases of significant disagreement involved mainly TUs with low scores. Table 3 shows an overview of the results collected for all languages.

Table 3. Evaluation by linguist experts for the 22 languages. The Average evaluator X column shows the average score over all segments given by evaluator X. The Overall average column shows the overall average of the scores given by the two evaluators, per language.

Language	Average evaluator 1	Average evaluator 2	Overall average
Slovak	95.09	94.75	94.92
Croatian	92.47	94.55	93.51
Polish	95.64	90.06	92.85
Romanian	89.67	95.99	92.83
Italian	88.59	96.06	92.32
Swedish	89.44	94.41	91.93
Bulgarian	91.25	91.39	91.32
French	94.15	86.46	90.30
Spanish	88.75	89.13	88.94
Czech	88.87	87.52	88.20
German	85.90	88.60	87.25
Latvian	87.74	78.88	83.31
Greek	90.21	74.94	82.58
Finnish	73.54	91.35	82.44
Dutch	74.92	85.86	80.39
Hungarian	80.31	78.67	79.49
Danish	85.11	67.78	76.45
Slovenian	70.54	81.28	75.91
Estonian	85.55	62.52	74.03
Polish	80.42	61.84	71.13
Lithuanian	72.83	45.48	59.11
Irish	44.67	42.06	43.36

As can be seen in table 3, MT into English from Bulgarian, French, Croatian, Italian, Polish, Romanian, Slovak, and Swedish obtained an average rating of above 90 percent. The lowest average results are found in the evaluations having Irish and Lithuanian as source languages. This performance was to be expected, considering the lack of data in those languages for training the engines, as well as the often low quality of the original data. Regarding the rest of the languages, the results were satisfactory, especially when considering that the training data were rather few and often of poor quality.

Evaluation by CH domain experts

The evaluation under this method took place via three "niche-sourcing" campaigns, one for each of the three considered source languages. The campaigns were conducted via CrowdHeritage (https://crowdheritage.eu), a platform that hosts the crowdsourcing campaigns for the

enrichment and validation of CH metadata. Each campaign involved CH items coming from all three considered Europeana aggregators/domains (fashion, audiovisual, and museum heritage), each one being represented with about 100 CH records that include segments covering all relevant metadata fields (more than 750 metadata values per language have been evaluated, amounting to 2,815 TUs in total). These 100 CH records constitute a subset of the records that were evaluated by expert linguists.

The recruited participants had both adequate domain expertise and mastery of the target and source languages. They were either domain experts via their profession or otherwise familiar with the domain and its terminology (e.g., as students). The requirement we set for the participation of members of the CH community was that they were proficient English users, i.e., C1 level according to the Common European Framework of Reference for Languages standard (https://www.coe.int/en/web/common-european-framework-reference-languages/table-2-cefr-3.3-common-reference-levels-self-assessment-grid) and had a minimum of C2 proficiency level in the source language.

Overall, 29 members of the CH community participated voluntarily in the evaluation campaigns. In total, users completed 3,327 ratings. No significant difference in the human-perceived quality of translations was observed across the considered CH domains.

It is interesting to compare the ratings assigned by linguist experts with those of CH professionals for the three languages evaluated by both groups. However, a direct comparison between the average ratings resulting from the two groups is not completely accurate, given that the segments evaluated by the linguist experts for those three languages is a subset of the segments evaluated by the CH experts. Table 4 gives an impression about how strict each group was with their evaluation.

Table 4. Average ratings assigned by linguists and CH experts for the three languages evaluated by both groups.

Language	Overall Average assigned by CH experts	Overall Average assigned by linguist experts (from table 3)
Dutch	90.08%	80.39%
French	95.65%	90.30%
Italian	86.57%	92.32%

FILTERING AND PUBLICATION TO EUROPEANA

The derived evaluation data were used to fit a quality estimation (QE) model so that the MT engines could predict confidence scores for each produced translation. Automatic QE of MT is a complex research issue and finding an optimal solution is challenging.[19] In the context of EuropeanaTranslate, we developed a method to obtain a QE score based on the automatic evaluation metrics word error rate (WER) and character error rate (CER), which indicate the percentage of words and characters, respectively, that were incorrectly predicted, and F1-score, which combines recall and precision metrics.[20] Due to the differences across languages concerning the number of sentences and unique words included in the samples evaluated by the expert groups, we used a model based on the aforementioned metrics and second order polynomial

regression coefficients to adjust the weights on the basis of language (https://github.com/Pangeamt/europeana-translate-mt/blob/main/get_quality_score.ipynb). After obtaining the automatic QE score per translation unit, we calculated the average score for each language. Finally, to analyze how human rating and automatic QE scores correlate to each other, we also calculated Pearson's correlation. We found that human and automatic scores are strongly correlated in most languages, with only Croatian and French having correlation below 0.5.

The automatic scores function as estimators for the quality of the translations and were used to determine appropriate thresholds per language as to which translations to filter out and which ones to retain for publication to the Europeana platform. With the aim of estimating such thresholds, we performed a linear regression analysis between the human evaluation average scores and the automatic QE scores, per language.

Given that there is a correlation between human and automatic scores, and based on the linear regression plots per language, we predicted the automatic QE score that corresponds to the score assigned by humans to a translation that is believed to be of satisfactory quality for publication. We decided to use a 75 percent human score target for determining this publication threshold.

A connection via API has been established between the Europeana infrastructure and EuropeanaTranslate MT services for retrieving automatic English translations of EDM metadata fields, following the aforementioned filtering methodology. A language detection mechanism was also used as part of the workflow, in case the original metadata did not include language tags. The process was applied to metadata records that lacked English translations on Europeana, leading to the translation of more than 29 million records by July 2023. The acquired translations were indexed and displayed on the respective item views on the Europeana website, along with a dedicated tag that indicates that the translations result from an automatic tool.

DISCUSSION/CONCLUSION

Automatic translation of CH metadata records opens new perspectives for making digitized collections accessible to a wider audience. However, the idiosyncrasies of these records make them more difficult to process than "traditional" text for existing translation engines. Also, it is often not possible—or desirable—for CH organizations to use best-of-breed commercial solutions. In EuropeanaTranslate we have devised, implemented, and evaluated a methodology and toolset that enables the enrichment of CH metadata by Europeana or cultural institutions with automatic translations to English. All 23 domain-adapted MT engines set via the adopted methodology have been made available as Docker containers on the ELG repository (https://live.european-language-grid.eu), so that CH organizations and other interested stakeholders can deploy and use them for their own objectives. The evaluation conducted shows satisfactory results that have been published on the Europeana platform, after establishing appropriate quality filters.

With the completion of the EuropeanaTranslate workflow, users of the Europeana platform who issue search queries in English are now able to access more relevant CH items, which used to be out of their reach due to their metadata being represented in another language. The availability of high-quality static metadata translations in English also paves the way for the realization of fully-fledged cross-lingual search, as envisaged by the Europeana multilingual strategy.

The next step to enhance the multilingual experience of Europeana users is indeed to expand the pilot done for Spanish by applying real-time translation into English to search queries that are issued in more European languages.[21]

ACKNOWLEDGEMENTS

The work is co-funded by the European Union, under the projects "EuropeanaTranslate: Providing multilingual access to digital cultural heritage" (action number 2020-EU-IA-0084) and "AI4Culture: An AI platform for the cultural heritage data space" (101100683). We would also like to thank Mercedes García-Martínez, Martín Barroso Ordóñez, Iván Lena Almor, Hugo Manguinhas, Arne Stabenau, Panagiotis Tzortzis, and Spyros Bekiaris for their indispensable contributions to completing this work.

ENDNOTES

[1] Andy Neale, Antoine Isaac, H. Manguinas, and D. Moskalenko, *Multilingual Strategy*, Europeana, https://pro.europeana.eu/post/europeana-dsi-4-multilingual-strategy.

[2] Mónica Marrero, Antoine Isaac, and Nuno Freire, "Automatic Translation and Multilingual Cultural Heritage Retrieval: A Case Study with Transcriptions in Europeana," in *Linking Theory and Practice of Digital Libraries: Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries,* Lecture Notes in Computer Science (LNCS) vol. 12866 (Springer, Cham, 2021): 133–38.

[3] Eirini Kaldeli, Mercedes García-Martínez, Antoine Isaac, Paolo Scalia, Arne Stabenau, Ivan Lena Almor, Carmen Grau Lacal, Martin Barroso Ordóñez, Amando Estela, and Manuel. Herranz, "Europeana Translate: Providing Multilingual Access to Digital Cultural Heritage," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (European Association for Machine Translation, 2022): 297–98.

[4] Maristella Agosti, Erika Fabris, and Gianmaria Silvello, "On Synergies between Information Retrieval and Digital Libraries," in *Proceedings of the Italian Research Conference on Digital Libraries* (Pisa, Italia: 2019): 3–17.

[5] Jacques Savoy and Martin Braschler, "Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF," in *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*" (Springer, Cham, 2019): 177–200.; Mónica Marrero and Antoine Isaac, "Implementation and Evaluation of a Multilingual Search Pilot in the Europeana Digital Library," in *Linking Theory and Practice of Digital Libraries: Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries,* Lecture Notes in Computer Science (LNCS) vol. 13541 (Springer, Cham, 2022): 93–106.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc., 2017), https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html; Mercedes García-Martínez, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis, "Neural Translation for European Union (NTEU)," in *Proceedings of Neural Machine Translation XVIII: Users and Providers Track* (Association for Machine Translation in the Americas: 2021): 316–34, https://aclanthology.org/2021.mtsummit-up.23; Rico Sennrich, Barry Haddow, and Alexandra Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* vol. 1: Long Papers, (Association for Computational Linguistics, 2016): 86–96,

https://doi.org/10.18653/v1/p16-1009; Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight, "Transfer Learning for Low-Resource Neural Machine Translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2016): 1568–75, https://doi.org/10.18653/v1/d16-1163; Mikel Artetxe, Gorka Labaka, and Eneko Agirre, "Unsupervised Neural Machine Translation, a New Paradigm Solely Based on Monolingual Text," *Natural Language Processing* 63 (September 2019): 151–54, RUA (Institutional Repository of the University of Alicante), https://rua.ua.es/dspace/handle/10045/96620.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, 2001): 311–18, https://doi.org/10.3115/1073083.1073135; Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas* (Association for Machine Translation in the Americas, 2006): 223–31.

[8] Valentine Charles, Antoine Isaac, Vassilis Tzouvaras, and Steffen Hennicke, "Mapping Cross-Domain Metadata to the Europeana Data Model (EDM)," in *Research and Advanced Technology for Digital Libraries* (Springer, January 2013): 484–85, https://doi.org/10.1007/978-3-642-40501-3_68.

[9] Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico, "Multi-Domain Neural Machine Translation through Unsupervised Adaptation," in *Proceedings of the Conference on Machine Technology (WMT)* (Association for Computational Linguistics, 2017): 127–37, https://aclanthology.org/W17-4713.pdf.

[10] Jörg Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resource and Evaluation* (European Language Resources Association (ELRA), 2012): 2214–18, https://aclanthology.org/L12-1246; J. D. Cortés, "What Is the Mission of Innovation?—Lexical Structure, Sentiment Analysis, and Cosine Similarity of Mission Statements of Research-Knowledge Intensive Institutions" *PLoS ONE* 17 no. 8 (2022): e0267454, https://doi.org/10.1371/journal.pone.0267454.

[11] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, and Ondřej Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, et al., "Findings of the 2021 Conference on Machine Translation (WMT21)," in *Proceedings of the Sixth Conference on Machine Translation* (Association for Computational Linguistics, 2021): 1–88, https://aclanthology.org/2021.wmt-1.1.

[12] Chenyang Li and Gongxu Luo, "Improving Zero-Shot Multilingual Neural Machine Translation for Low-Resource Languages," arXiv, 2110.00712 (October 1, 2021), https://doi.org/10.48550/arXiv.2110.00712.

[13] Biao Zhang, Ivan Titov, and Rico Sennrich, "Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP–IJCNLP) (Association for Computational Linguistics, 2019): 898–909, https://doi.org/10.18653/v1/d19-1083.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need" in *Advances in neural information processing systems*, pages 5998-6008, 2017; Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart, "The OpenNMT Neural Machine Translation Toolkit: 2020 Edition," in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas* vol. 1, Research Track (Association for Machine Translation in the Americas, 2020): 102–109, https://aclanthology.org/2020.amta-research.9; Sébastien Martin and Martin Weiß, "A Proof of Local Convergence for the Adam Optimizer," in *2019 International Joint Conference on Neural Networks (IJCNN)*, (2019): 1–8, https://ieeexplore.ieee.org/document/8852239.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* vol. 1, Long Papers (Association for Computational Linguistics, 2016): 86–96, https://doi.org/10.18653/v1/p16-1009.

[16] Maja Popović, "ChrF: Character N-Gram F-Score for Automatic MT Evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation* (Association for Computational Linguistics, 2015): 392–95, https://doi.org/10.18653/v1/w15-3049; Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate" in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.; Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation" in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.; Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie, "COMET: A Neural Framework for MT Evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2022): 2685–2702, https://doi.org/10.18653/v1/2020.emnlp-main.213.

[17] Alexander Soetaert, Luc Truyens, and Henk Vanstappen, "Brugge Grenzeloos Digitaal: Ondersteuning Meertaligheid: Eindrapport," *Datable BV*, Antwerp, 2021, https://www.projectcest.be/w/images/2021_Grenzeloos_eindrapport.pdf.

[18] John White and Theresa O'Connell, "Evaluation in the ARPA Machine Translation Program, in *Proceedings of Human Language Technology* (Association for Computational Linguistics, 1994): 135–40, https://doi.org/10.3115/1075812.1075840.

[19] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn, "QuEst – A Translation Quality Estimation Framework," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Association for Computational Linguistics, 2013): 79–84, https://aclanthology.org/P13-4014.

[20] Dietrich Klakow and Jochen Peters, "Testing the Correlation of Word Error Rate and Perplexity," *Speech Communication* 38, no. 1–2 (2002): 19–28, https://doi.org/10.1016/s0167-6393(01)00041-3; Andrew Cameron Morris, Viktoria Maier, Phil Green, "From WER and RIL to

MER and WIL: Improved Evaluation Measures for Connected Speech Recognition," in *Proceedings of Interspeech* (2004): 2765–68, https://doi.org/10.21437/interspeech.2004-668.

[21] Mónica Marrero and Antoine Isaac, "Implementation and Evaluation." In: Silvello, G. et al. *Linking Theory and Practice of Digital Libraries*. TPDL (2022) Lecture Notes in Computer Science, vol 13541. Springer, Cham.