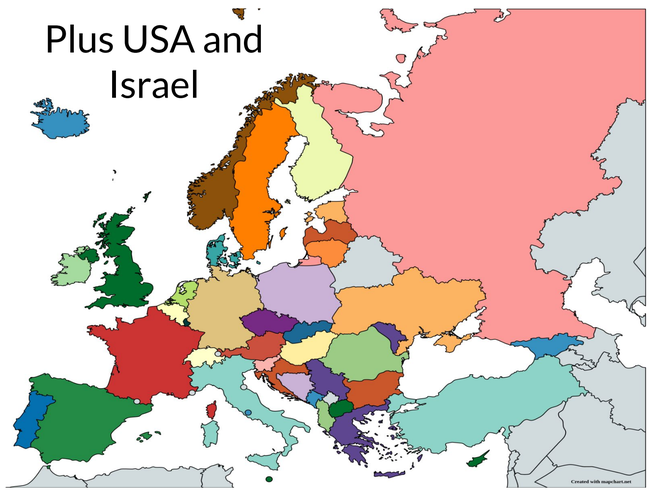
Automatic Translation and Multilingual Cultural Heritage Retrieval

A CASE STUDY WITH TRANSCRIPTIONS IN EUROPEANA

Mónica Marrero, Antoine Isaac and Nuno Freire

EUROPEANA

More than **60 million digital objects** in more than **40 languages** from over 3500 libraries, archives and museums, mainly across Europe.



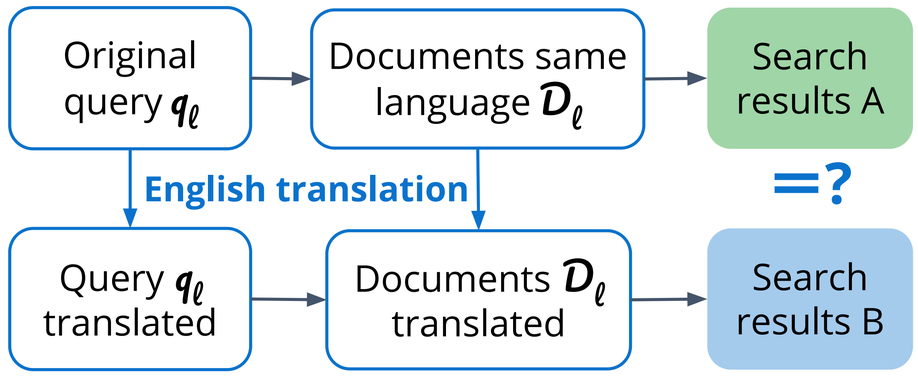
CURRENTLY	>	FUTURE
Monolingual Information Retrieval		**Multilingual** Information Retrieval

METADATA AND CONTENT

EXPERIMENT: Similar results to original query when using translations?

STRATEGY

- Translate query ← Noisy, specially for not so common languages
- Translate content ← Not scalable for so many languages
- **Translate query and content to a pivot language (English)**



DATA & RESOURCES

Content: 18,257 manual transcriptions of handwritten documents from the Europeana 1914–1918 collection. The source language is known.

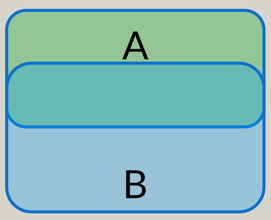
Queries: a sample of 68 non-English user queries issued on that collection. The source language is assumed to be the one of the Europeana portal where the query was issued.

Translation service: CEF eTranslation, which is intended as a free, secure service provided by the EC.

FINDINGS

- ⅓ of the documents are not retrieved, and almost half are new!
- The language of the portal IS NOT the language of the query in 26% of the cases
- Entities to be left unchanged accounted for 86% of the cases, and we obtained wrong translations in almost half of them. Example: ‘Antonio Sordi’ IS NOT ‘Antonio Deaf’!
- 9% of the queries contain typos, which makes even more difficult the translation

⅓ documents are not retrieved when using translations



Almost half documents are new when using translations:
even if part of them could be explained by (quasy) synonymy, the risk of applying this system is high.

NEXT STEPS

- Better language identification: test tools, use other signals (e.g. language portal + IP)
- Identify queries with entities and limit or avoid its translation
- Introduce spelling-correction systems to mitigate issues caused by typos
- New experiments with relevance assessment for results and balanced data per language

DATA & MORE INFO AVAILABLE AT ZENODO: https://doi.org/10.5281/zenodo.5045066

europeana