Addressing multilingual challenges at Europeana: An update
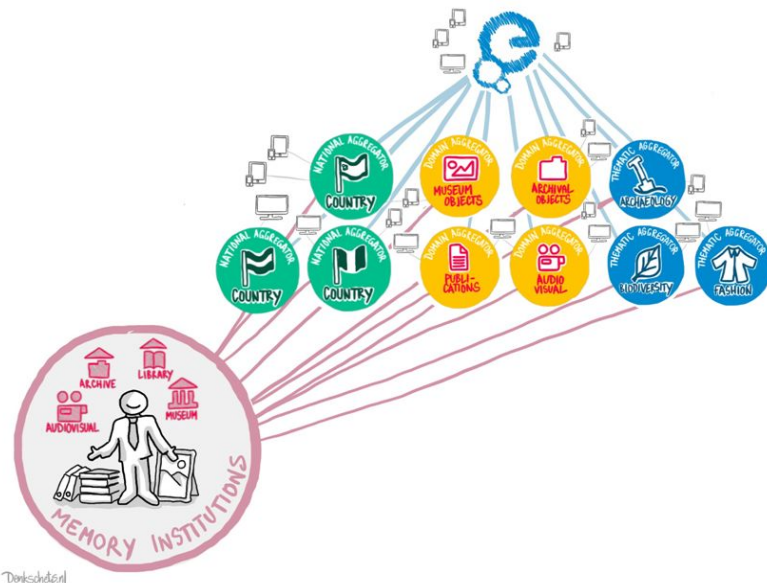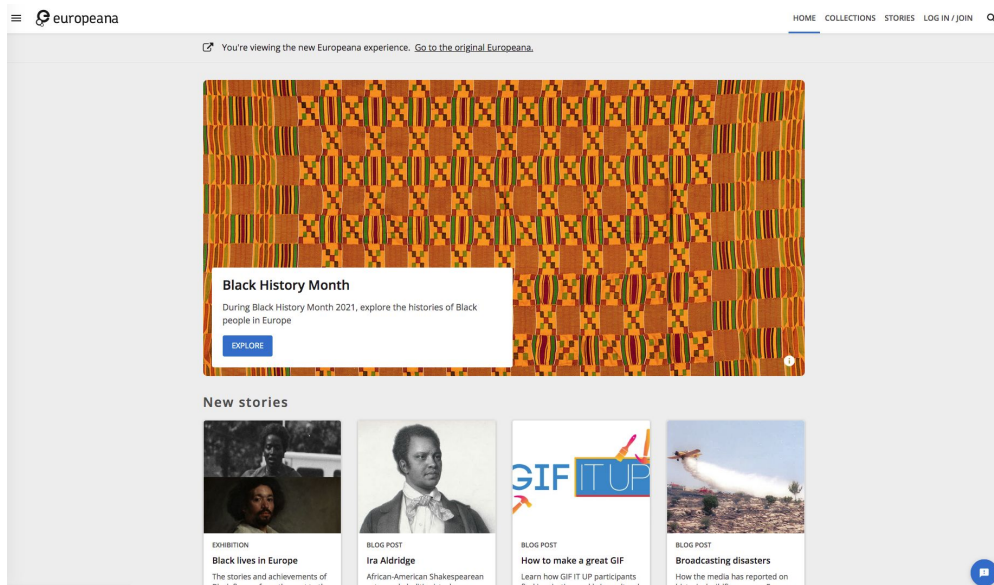
Antoine Isaac, Mónica Marrero
Europeana Foundation

DCMI Virtual 2021
13 October 2021

Fàbrica als magatzems de la Casa Teixidor, Ajuntament de Girona, Spain, Public Domain

Europeana.eu

- An initiative from the European Union
- Publishes 52M digitized objects
- Using metadata from 4,000 libraries, archives and museums in 44 countries

MULTILINGUAL EXPERIENCE AT EUROPEANA

USE CASES FROM OUR MULTILINGUAL STRATEGY

- Navigate the Europeana website

- Read editorial content

- Read item text

- **Search Europeana**

MULTILINGUAL STRATEGY

Paid or volunteer static translations to user language

Live machine translation of query to English

Exploit multilingual vocs & add static English (pivot) translation

Read editorial content

Search

Query in user language + Query in English

User on a language-specific Europeana portal

Multilingual (meta)database and search index

Navigate the website

Read item text

Language-specific item page

Automatic, paid or volunteer translation of Europeana UI to user language

Live machine translation from English to user language

slide adapted from Paolo Scalia

europeana

WEBSITE (UI) TRANSLATION

europeana   HOME COLLECTIONS STORIES LOG IN / JOIN

europeana   ACCUEIL COLLECTIONS HISTOIRES CONNEXION/REJOIGNEZ-NOUS

europeana   INICIO COLECCIONES HISTORIAS INICIAR SESIÓN/UNIRSE

Scope and obstacles:

- Full translation coverage of 24 official EU languages (+Basque)
- We release new UI features every two weeks which need to be translated
- Professional/volunteer translation requires time and/or money and/or motivation

Approach:

- We use automated translation services to save time and money
- When in doubt we validate the translations with native speakers
- Partners submit translation updates via the feedback button and email

slide adapted from Dasha Moskalenko

Feedback

TRANSLATION OF EDITORIAL CONTENT

- Asking volunteers to translate virtual exhibitions, etc.

- Exhibitions published in more than one language increased from 50% to 64%

slide adapted from Douglas McCarthy

europeana

HOME COLLECTIONS STORIES LOG IN

Exhibitions

Explore the exhibitions

China in perspective
Discovering China through European heritage collections

Looted Library,...
How Riga Jesuit College's books found their way from Latvia to Sweden

A century of technology
Read about groundbreaking 20th-century inventions in transport, industry, media and everyday life

The Sakharov Prize, the...
Read about the Sakharov Prize for Freedom of Thought and the European Parliament's commitment to safeguarding...

The Danube: connecting...
Explore the River Danube, which has inspired musicians, artists, engineers and holders of political and military power for centuries.

The Beaver's Journey
Follow the European beaver on its journey to from Norway to Sweden in 1922.

The Charter of Fundament...
The Charter of Fundamental Rights of the European Union turns 20!

Family Matters
Explore different family concepts and discourses, looking back at the 20th century.

ITEM PAGE TRANSLATION

VIEW ITEM METADATA IN DIFFERENT LANGUAGES

Enable users of the website to translate on the fly the metadata description of an item to their language of choice

- For all 24 official EU languages

- Using real-time translation service (Google for now)

europeana

ITEM PAGE TRANSLATION

ITEM PAGE TRANSLATION

Está viendo este objeto en la nueva web de Europeana. Ver este objeto en la web original de Europeana.

Would you like to see this item in original language? Other languages ⌄

CC BY-NC-SA   1/9   ➕ ♡ ⇗ Compartir ⬇ Descargar

Clavicémbalo / fortepiano 📷

Clavicémbalo convertido en fortepiano. Rango: F1 - d3, añadido d3. Varias teclas rotas en el bajo. Cubierta de dos coros en todas partes. Una palanca de rodilla para cancelar la amortiguación. 📷

Leer más   Automated translation

SPANISH SEARCH PILOT

MULTILINGUAL SEARCH

- This pilot is meant to help test translation of search queries
- Validating implementation for one language allows us to put aside some issues raised by multilinguality on the entire process

    E.g. no translation of metadata to English

- It enables us to focus work on some components and get something visible for them much earlier

europeana

Left panel:

epidemias - Buscar

Categoría ▾ | Tipo de medio ▾ | ¿Puede reutilizarse? ▾ | País proveedor ▾ | Más filtros

Resultados: 1,559

País proveedor:
- ☐ Spain (1,518)
- ☐ Europe (11)
- ☐ Germany (9)
- ☐ Czech Republic (8)
- ☐ Austria (5)
- ☐ Netherlands (4)
- ☐ Portugal (2)
- ☐ Italy (1)
- ☐ United Kingdom (1)

REINICIAR | APLICAR

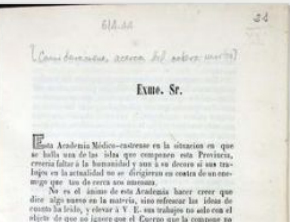Páginas luctuosas
Ruiz Bustamante, Antonio
Universidad de La Laguna

Don Esteban Mandillo Alcalde Constitucional de esta capital y presidente de su M. I. Ayunt...
Santa Cruz de Tenerife. Ayuntamiento
Universidad de La Laguna

Expedientes de epidemias
Galiciana. Arquivo Dixital de Galicia

Right panel:

epidemias - Buscar

Categoría ▾ | Tipo de medio ▾ | ¿Puede reutilizarse? ▾ | País proveedor ▾ | Más filtros

Resultados: 2,017

País proveedor:
- ☐ Spain (1,725)
- ☐ United Kingdom (63)
- ☐ United States of America (59)
- ☐ Germany (26)
- ☐ Bulgaria (24)
- ☐ Europe (20)
- ☐ Finland (17)
- ☐ Netherlands (15)
- ☐ Sweden (15)
- ☐ Hungary (14)
- ☐ Czech Republic (8)
- ☐ Austria (6)
- ☐ Serbia (6)
- ☐ Denmark (4)
- ☐ Ireland (2)
- ☐ Italy (2)
- ☐ Poland (2)
- ☐ Portugal (2)

REINICIAR | APLICAR

Conquista y crisis demográfica: la población indígena del occidente de Guatemala en el sig...
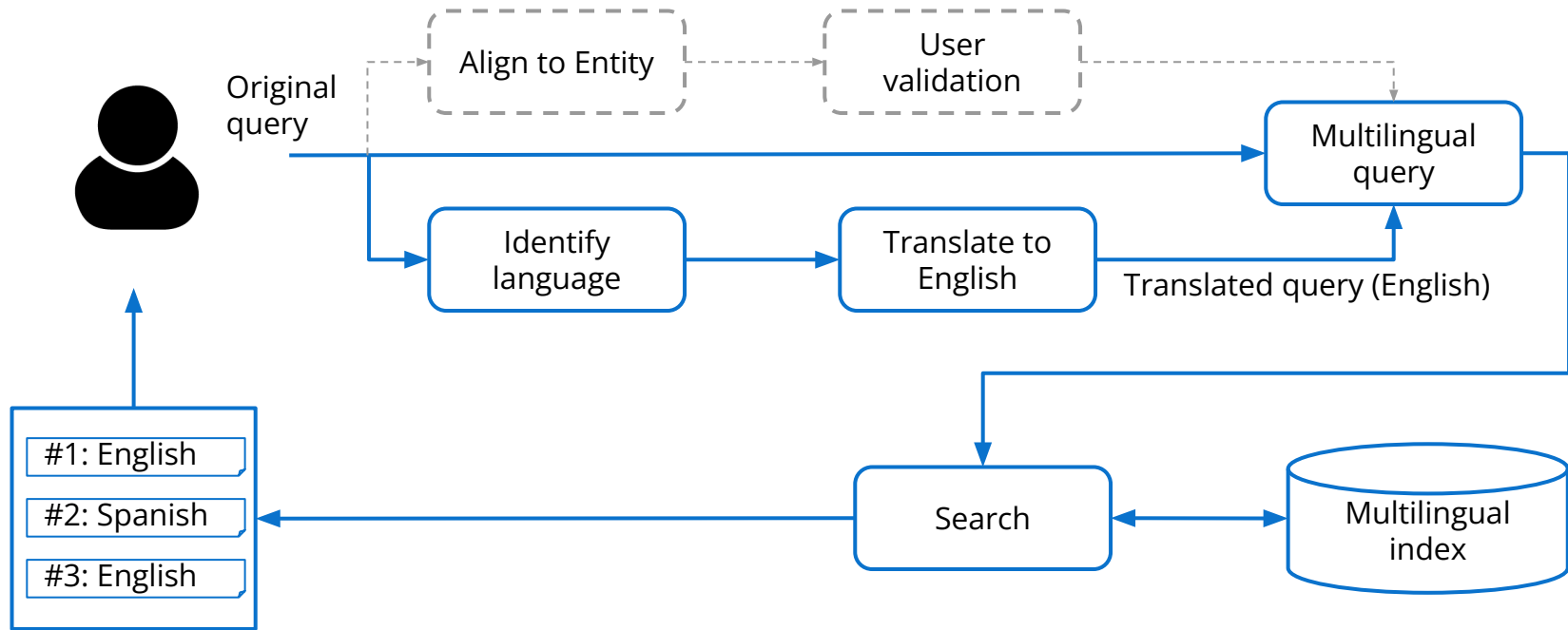idUS. Depósito de investigación de la Universidad de Sevilla

Saint Teliau, invoked durin... epidemics and fevers.
Wellcome Collection

Veterinarski

SPANISH SEARCH PILOT

SPANISH SEARCH PILOT

EVALUATION

- **Queries:** sample of 300 queries taken from the Europeana Spanish portal between 1st December 2020 and 28th February 2021

- **Collection:** snapshot of the Metadata collection (62 million documents), 17.5% documents in English, 5.5% in Spanish (according to *edm:language*, language of the provider)

- **Annotation:** manual assessment of results obtained with sample (query language, quality identification language, quality translation, and relevance of unique search results obtained in top ten positions)

europeana

SPANISH SEARCH PILOT

REAL-TIME IDENTIFICATION OF THE LANGUAGE OF THE QUERY

- **Implementation:** Google Cloud Translation API
- **Evaluation:** assessment language of the query, language identified by Google, quality translation
  - Q1. Can we assume that the language of the query is Spanish when using the Spanish portal?
  - Q2. What is the effectiveness of the language identification service offered by Google?
  - Q3. What is the impact on translation?

europeana

LANGUAGE IDENTIFICATION

Q1. LANGUAGE SPANISH?

With that assumption
we would have:

- 33% success
- 21% error
- Unclear in remaining 46%

Q2. AUTOMATIC IDENTIFICATION EFFECTIVENESS?

- 58% success
- 4% error
- Unclear remaining 38%

Q3. IMPACT ON TRANSLATION

- Less than 1 percentage point
- However, an accurate identification may be relevant for other tasks

Better results with Google language identification but no high impact on translation quality compared with assuming Spanish as source language.

europeana

SPANISH SEARCH PILOT

REAL-TIME TRANSLATION OF THE QUERY

- **Implementation:** Google Cloud Translation API
- **Evaluation:** comparison Google Cloud Translation API, and eTranslation
  - Q4. What is the most effective system?

TRANSLATION

Q4. GOOGLE VS ETRANSLATION

- Google is 20% more effective than eTranslation, with an effectiveness rate of 73%

	ok	wrong	unclear	error	% ok
Google auto	220	47	33	-	**73.3**
eTranslation auto	159	80	23	38	**53**

Using Google for the identification of the language (eTranslation does not provide that service).

Google Translation is more effective than eTranslation for translating queries.

SPANISH SEARCH PILOT

CONSTRUCTION OF MULTILINGUAL QUERY & SEARCH

- **Implementation:** (original query) OR (translated query)

carteles de la guerra civil NOT "National Library of Spain" → (carteles de la guerra civil NOT "National Library of Spain")
OR
(civil war posters NOT "National Library of Spain")

- **Evaluation:** comparing search results (original query vs multilingual query)

  ○ Q5. Difference in number of results (potential Recall)

  ○ Q6. More documents in English top ten positions?

  ○ Q7. Difference in Precision at 10 (P@10)

  ○ Q8. Effect of typos, ambiguity and non-translatable entities in queries in P@10

MULTILINGUAL SEARCH

Q5. NUMBER OF RESULTS

- More results in 35.7% of the queries, with an average of 17015 new documents found
- Queries with zero results reduced by 3%, but results were not relevant

Q6. MULTILINGUAL GAIN

- 10% more queries with at least 1 document in English (33% to 43%)
- Increase of 0.41 documents in English per query (1.43 to 1.84)

Q7. PRECISION

- 78% not affected
- 14% negative impact
- 8% positive impact
- Average: -0.026

More results in total, with more English documents in top ten positions. We lose on average one document that is relevant in the top ten for one query out of four.

MULTILINGUAL SEARCH

Q8. EFFECT OF TYPOS

Queries with typos or that are not understandable. E.g., *Julia Quin* instead of *Julia Quinn*, *Angel Pardo yPuzo* instead of *Angel Pardo y Puzo*, *HIPODAIA, ungaro*

- 5% queries
- No clear effect found on translation quality
- No significant impact in $\Delta P@10$ (P-value = 0.5821)
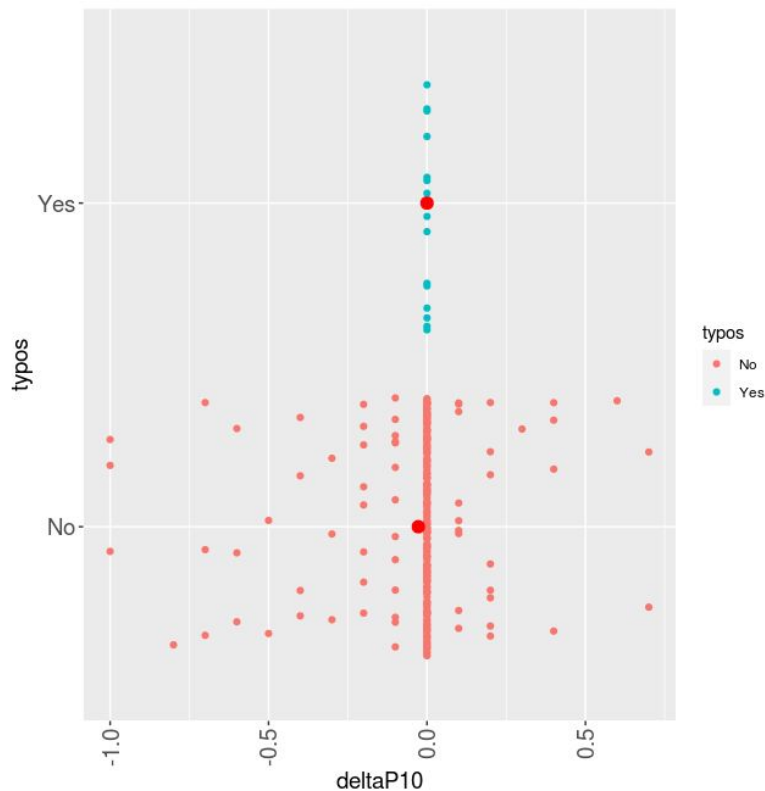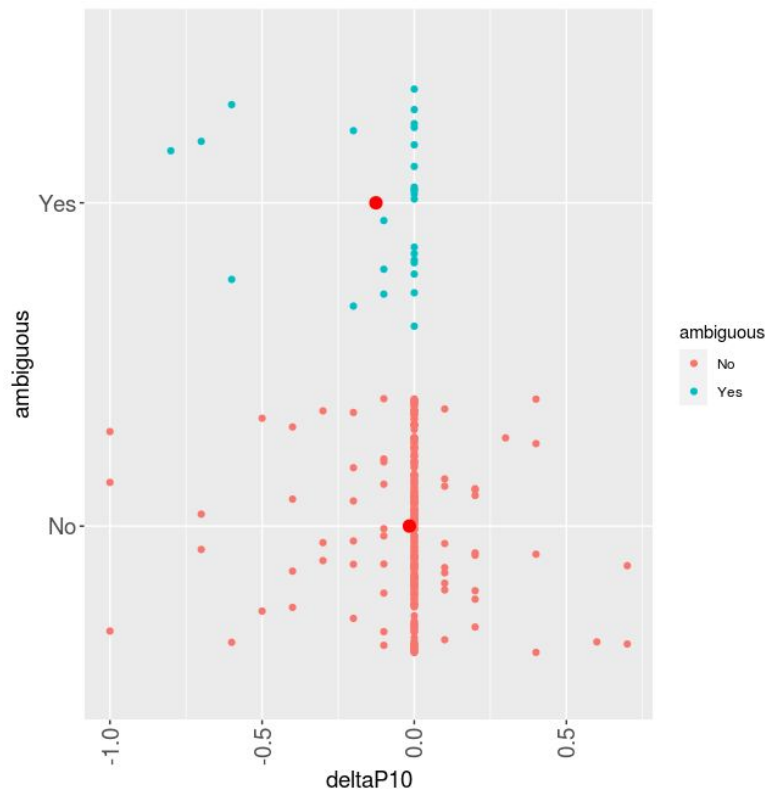

These queries maintain the relevance of the top ten results, but the effect is not significant.

MULTILINGUAL SEARCH

Q8. EFFECT OF AMBIGUITY

Queries that can have different translations depending on the intention of the user. E.g., *granada* could be a Spanish city but also a Latin American city, a country and island, or even a fruit; *CUEVA PINTADA* can be an archaeological park with that name or a general painted cave; *historia* can refer to history or a story.

- 9% queries
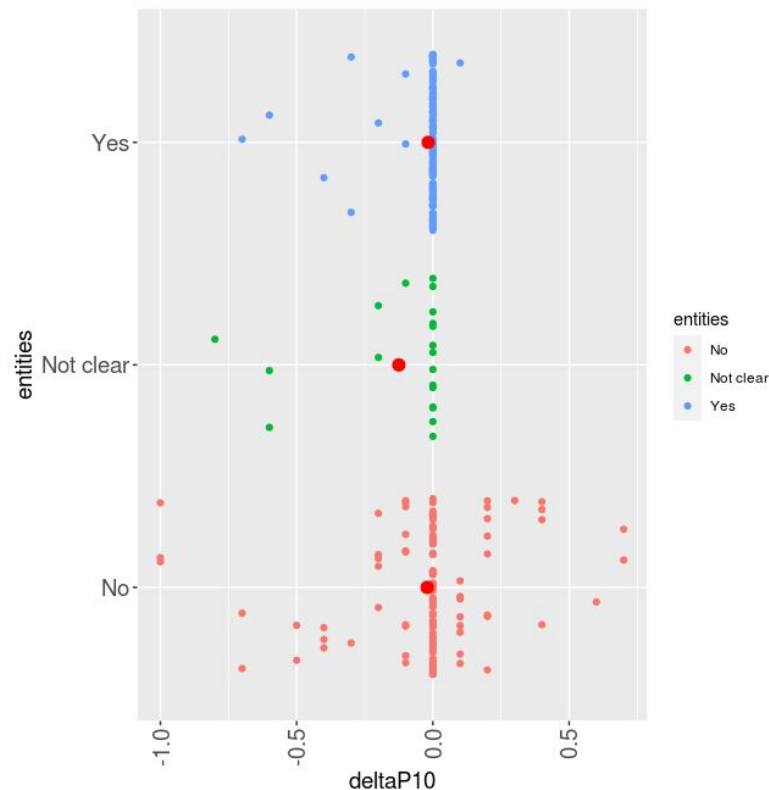- Significant impact in ΔP@10
  (P-value = 0.003378)

These queries decrease or maintain the relevance of the top ten results, and the effect is significant.

MULTILINGUAL SEARCH

Q8. EFFECT OF ENTITIES

Queries with entities that should not be translated. E.g., *Juan Gris* can not be translated as *John Gray*, or *Antonio Sordi* as *Antonio Deaf*. Also book titles when there is no English translation available or they are in English.

- At least **51%** queries
- Most of Google's wrong translations come from queries containing entities
- Significant impact in ΔP@10 (P-value = 0.04785)

These queries mainly decrease or maintain the relevance of the top ten results, and the effect is significant.

SPANISH SEARCH PILOT

SUMMARY

- Promising results
- Google more effective than assuming language of the portal for the identification of the language of the queries
- Google more effective than eTranslation for the translation of the queries
- Ambiguity: transparency, more options to the user to limit the results to specific fields or topics
- Entities: use of controlled vocabularies
- Increase precision: routing multilingual query to fields in specific languages, better enrichment

CONCLUSIONS: PROGRESS AGAINST USE CASES

Navigate the Europeana website

Users can benefit from translation of UI in their own language (24 EU official languages + Basque)

Always require adaptation to new features and to feedback

Read editorial content

Users of the Europeana website have access to more and more editorials in their own languages.

Percentage of exhibitions published in more than one language increased from 50% to 64%.

Read item text

Users of the website can translate on the fly the metadata description of an item in their language of choice.

COMING OCTOBER 2021

Search Europeana

With the Spanish search pilot, a user searching the Spanish version of Europeana will obtain more results described in English next to the Spanish ones.

COMING QUARTER 4 2021

slide adapted from Valentine Charles

europeana

CONCLUSIONS: COMING WORK

- Continuing all the "ongoing" tasks...

- Tests and evaluations

  - Further test item page and Spanish search with users

  - Later, test in other portals

- Work on underlying data

  - Implement indicators for actions on making underlying data more multilingual

  - Normalization of language info - and more language detection

  - Translation of metadata to English: Europeana Translate project (ongoing)

    Aligned corpora from the CH domain are very welcome!

europeana

Thank you!

europeana

europeana.eu
@EuropeanaEU

The Chinese Market, 1767 - 1769, Rijksmuseum, Netherlands, Public domain

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