Improving language tags in cultural heritage data: a study of the metadata in Europeana

Nuno Freire, Paolo Scalia, Antoine Isaac, Eirini Kaldeli, Arne Stabenau

SWIB 2022

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

Danse de trois faunes et trois bacchantes, Hieronymus Hopfer, Bibliothèque municipale de Lyon , Public Domain

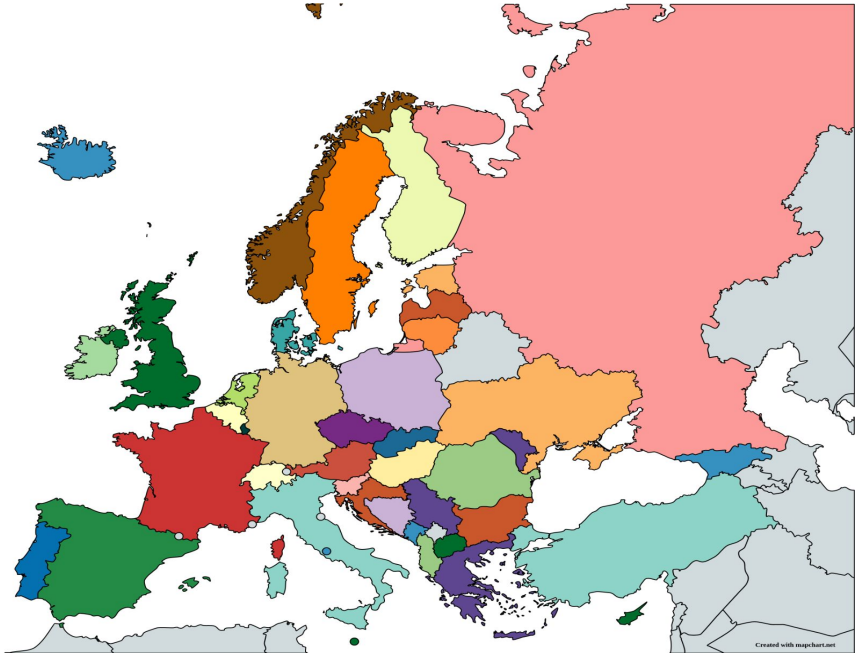
Introduction

Cats in human dress playing a variety of games, including arm wrestling and tug of war, Kunimasa IV, 1870s, Wellcome Collection, United Kingdom, CC BY

europeana

Europeana

- **Digital library,** initiative of the **European Union**, financially and politically supported by the European Commission since its launch in 2008.
- Publishes **54M+** digitized objects
- Using metadata from 4,000 libraries, archives and museums in **44** countries

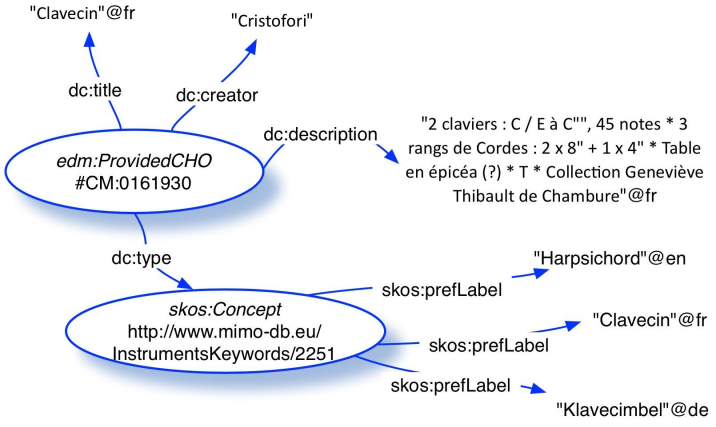


europeana

Europeana Data Model (EDM)

- Collaboratively developed
- Adhering to the Semantic Web principles
- Supports re-use of external Linked Data resources
- Supports multilingual data

Clavecin, Bartolomeo Cristofori
Cité de la Musique,
MIMO - Musical Instruments Museums Online
CC BY-NC-SA



europeana

About language tags

- A language tag is a standardized code or tag that is used to identify human languages in the Internet
- They allow the representations of languages and their variants for countries, regions, or scripts
- It combines codes from other standards such as ISO 639, ISO 15924, ISO 3166-1 and UN M.49
- IETF BCP 47 defines the structure, content, construction, and semantics of language tags
- The subtags are maintained by the IANA Language Subtag Registry
- Language tags are used by several standards such as XML, RDF, HTML, HTTP, and others

Motivation

Snell, A practical guide to the examination, Wellcome Collection, United Kingdom, CC BY

europeana

Europeana’s Multilingual Strategy

- Navigate the Europeana website
- Read editorial content
- Read text in item metadata
- Search Europeana
    - Use of pivot language: English

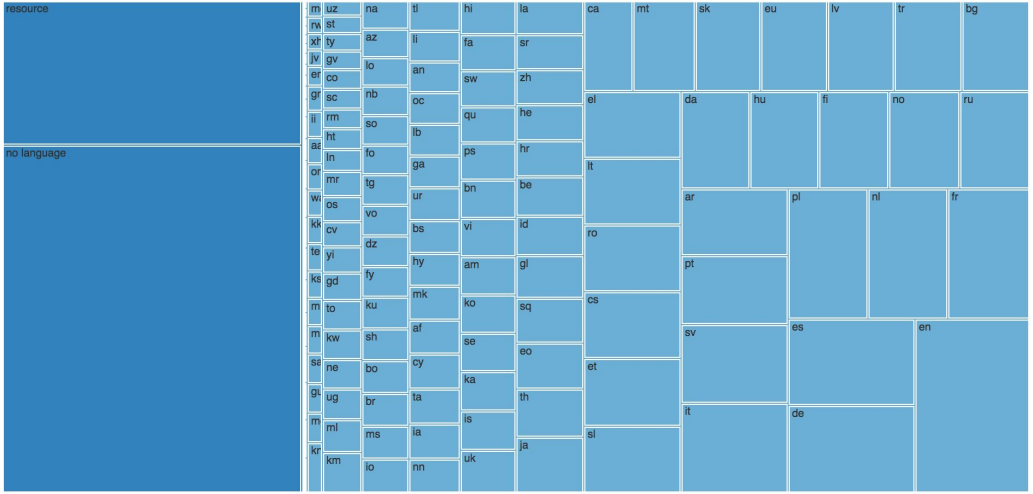
Eli from BL Royal 2 B VII, f. 48v |: 1310 - 1320 | The British Library| United Kingdom |Public Domain

europeana

Data quality problems

- Lack of language information
- Language information not normalized or incorrect
- What is the current status?
- Can we improve it?

NB: there is already a normalization process in place but we know it doesn't catch everything



All fields aggregated

1. **no field instance**: 4 272 165 045
2. **no language**: 770 143 229
3. **resource**: 320 367 666
4. **en**: 145 021 537
5. **de**: 82 024 654
6. **es**: 80 571 570
7. **fr**: 79 175 871
8. **nl**: 77 276 373
9. **pl**: 76 621 843
10. **it**: 70 659 189

388. **gez**: 12
389. **Spa**: 12
390. **efi**: 8
391. **ar-IL**: 7
392. **ar-LB**: 7
393. **la-x-medieval**: 6
394. **x-highgerm**: 6
395. **zh-latn**: 6
396. **x-aramaic-latn**: 6
397. **jut**: 5
398. **tyv**: 4
399. **RU**: 4
400. **cyr**: 4
401. **spa**: 4
402. **x-local**: 3
403. **dut**: 3
404. **ji**: 3
405. **iw**: 2
406. **x-stdgerm**: 2
407. **alb**: 2
408. **e**: 2
409. **uk-latn**: 2
410. **RS**: 2
411. **swb**: 2
412. **tr-latn**: 2
413. **gml**: 2
414. **fa-latn**: 2
415. **grc-latn**: 2
416. **tir**: 1
417. **fa-latn-x-middle**: 1
418. **dan**: 1
419. **olo**: 1
420. **phn**: 1
421. **ell**: 1
422. **src**: 1

Data analysis by Péter Kiraly (Göttingen Research alliance), 2018-2019

The Europeana Translate Project

Challenges for automatic translation in metadata

- Metadata lacks context for automatic translations due to short texts
- Specific terms to cultural heritage require specialized translations

The approach of Europeana Translate

- Create specialized machine translation techniques with data from Europeana
- Translation of metadata from 24 official languages of the European Union into English
- Training data to be shared with community

europeana

Objectives of the study

- Investigated whether it was desirable to improve the normalisation of language tags performed in the ingestion system of Europeana

- For the Europeana Translate project, increase the available language tagged metadata for training automatic translation systems for cultural heritage data

europeana

Language tags in the Europeana dataset

Arrival of a Portuguese ship. Anonymous. 1660 - 1625, Rijksmuseum.. Netherlands, Public Domain

europeana

How the investigation was done

- Source data:
  - All metadata records in Europeana - over 54 million
  - Only the metadata submitted by the cultural heritage institutions
- Analysis process:
  - All the statements with literal values (approx. 2.8 billion) were checked for the presence of an xml:lang tag.
  - Contextual entities from linked data sources were processed on their first occurrence and skipped whenever they reoccurred
- Measurements:
  - Presence of language tags and their compliance to IETF BCP 47
  - Presence of language subtags (i.e., variant, region, script and extensions)

europeana

Findings - general

- Only 35.08% of the literals are language tagged
  reminder: not all literals are supposed to be textual, for example dates like ‘23-03-2004’ in dc:date, DOIs in dc:identifier, etc.

- Comparing with a 2018-2019 study* we observed:
  - A growth in distinct language tags: from 422 to 608
  - A decrease of "exotic" values with subtags such as x-arameic-latn, x-highgerm....

* Study *by Péter Kiraly (Göttingen Research alliance), 2018-2019*

europeana

Findings - validity

- Not all values of language tags are compliant:
  - Approx. 7.9 million tags (0.81%) do not comply with IETF BCP 47
  - For example:
    - Using ‘fre’ instead of ‘fr’, or ‘eng’ instead of ‘en’
    - Using an extension tag without a main language tag - ‘x-highgerm’
    - Cases like ‘Skulptur’, ‘Maleri’, ‘Jpan’
  - 7.68 million could be normalized with the current normalization procedure

Findings - subtags

- Approx. 27.7 million (2.82% of existing language tags) contain compliant subtags
  - We validated them by checking in the IANA Language Subtag Registry
    - All subtags are valid
  - Some contain extensions:
    - Around 10 thousand cases, 7 distinct tags:
      - ang-x-late
      - de-x-std
      - la-x-ancient
      - la-x-medieval
      - la-x-liturgic
      - zh-latn-pinyin-x-notone
      - zh-latn-pinyin-x-hanyu
    - Although valid, these tags may not be very useful for Europeana and data reusers

europeana

Multilingual data provision for Europeana Translate

Cats in human dress playing a variety of games, including arm wrestling and tug of war, Kunimasa IV, 1870s, Wellcome Collection, United Kingdom, CC BY

europeana

Training data requirements

- Europeana Translate trained automatic translation models for the 24 official languages of the European Union
- For training the automatic translation systems, it required:
  - Corresponding language tagged text in English and another language
  - Language tagged text in just one language (not as valuable, but still helpful)
- We used the Europeana dataset as a source of training data

europeana

Normalisation of language tags

- We investigated the non-compliant language tag values in Europeana
  - They originate from datasets that have been ingested in Europeana before language tag normalisation was implemented

- To provide language tagged metadata to Europeana Translate, we have applied language tag normalisation on a data dump of the Europeana dataset

europeana

Normalisation of language tags

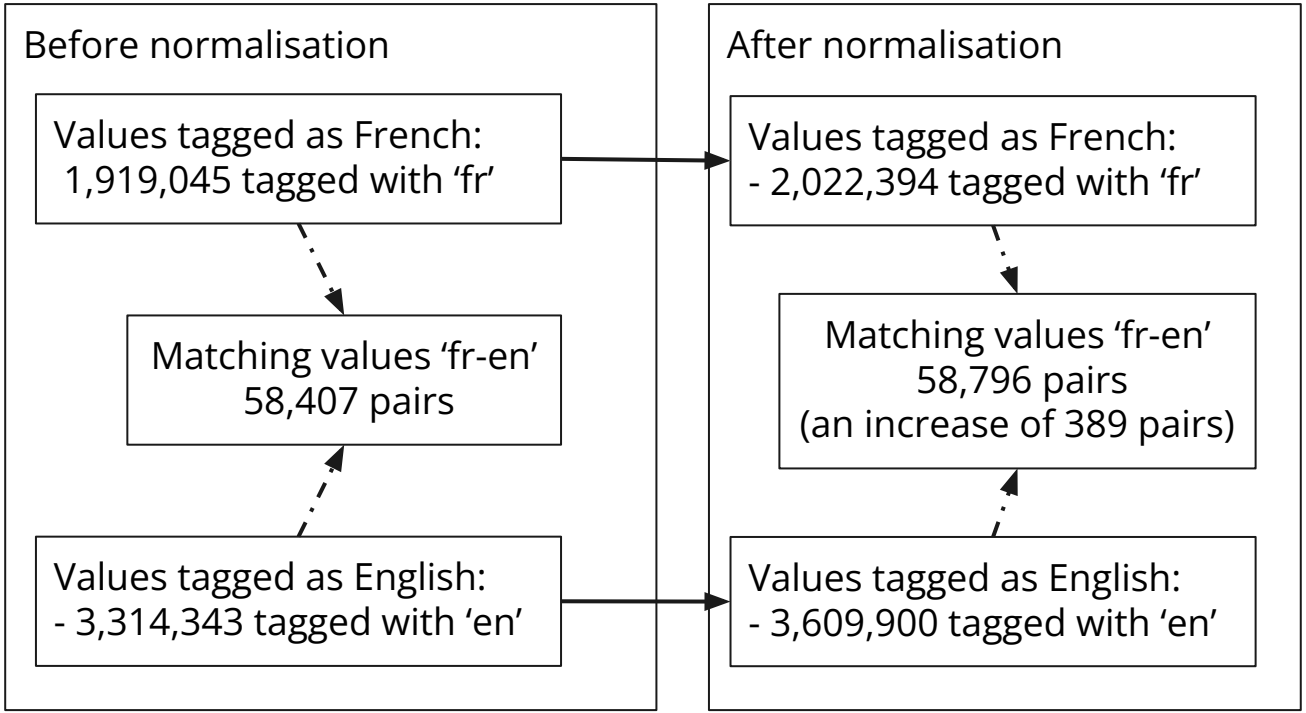
- The normalisation is based on the Language Named Authority List (NAL) from the Publications Office of the European Union
- NAL provides for each language:
  - ISO 639-1 codes (alpha-2) - i.e. ‘en’, ‘fr’
  - ISO 639-2/B and ISO 639-2/T codes (alpha-3) - i.e. ‘eng’, ‘fra’
  - ISO 639-3 codes - covers all the languages and macro-languages of the world (alpha-3) - i.e. ‘fsl’ for French Sign Language
  - The names of the languages in multiple languages
- The correspondence between the various codes and the languages names allows Europeana to normalise many language tag values into the code registered in the IANA Registry (typically the ISO 639-1 code)

Data provision process

- It starts with obtaining a data dump of the Europeana.eu dataset
- Language tags are normalised on the data dump
- Language tagged text is extracted:
  - Only from a set of properties that is deemed important to be translated.
  - Only considering the 24 official languages of the European Union
- Two types of groups of text are selected:
  - Metadata values in only one language
  - Pairs of values in two languages (one of them being English)

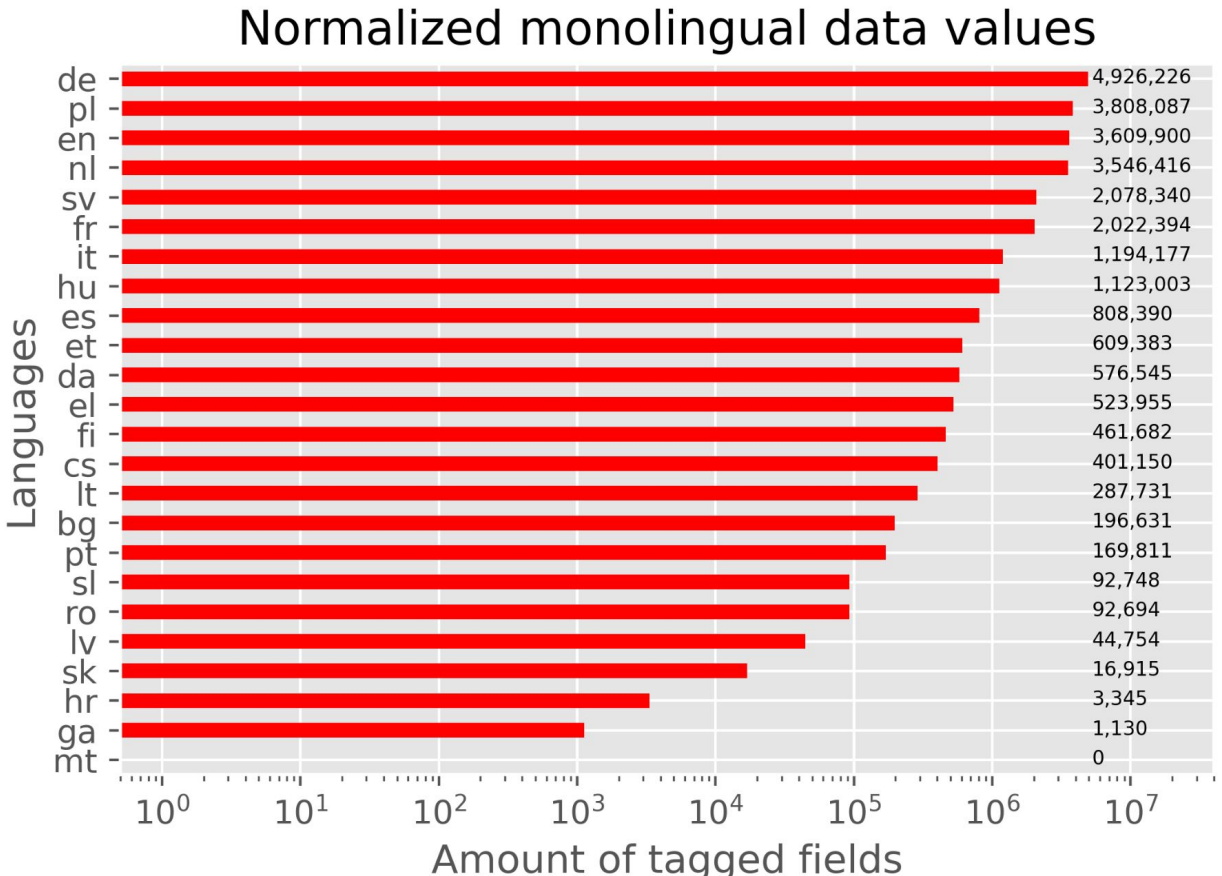
Data provision process (example)

An example using the case of French-English value pairs:



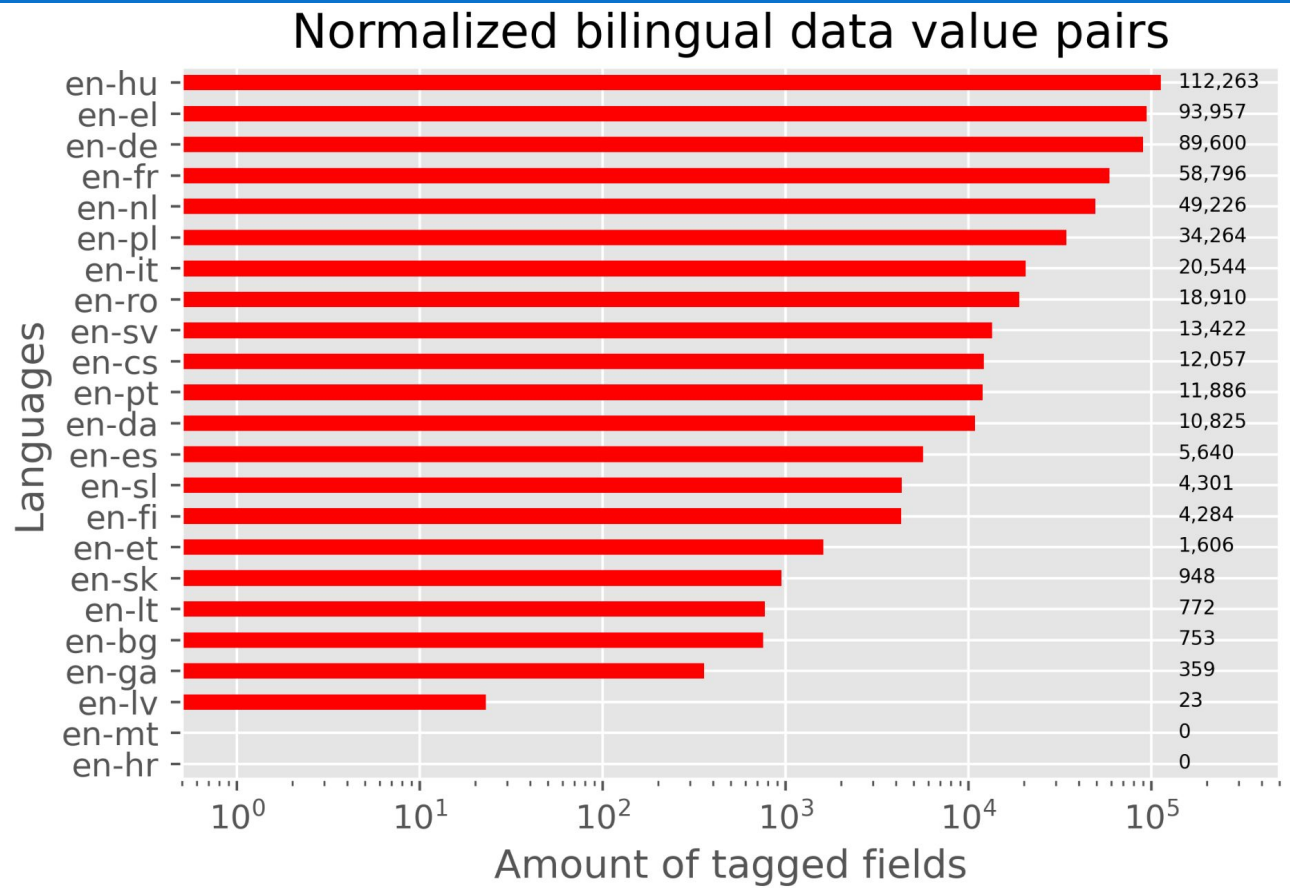
europeana

Results



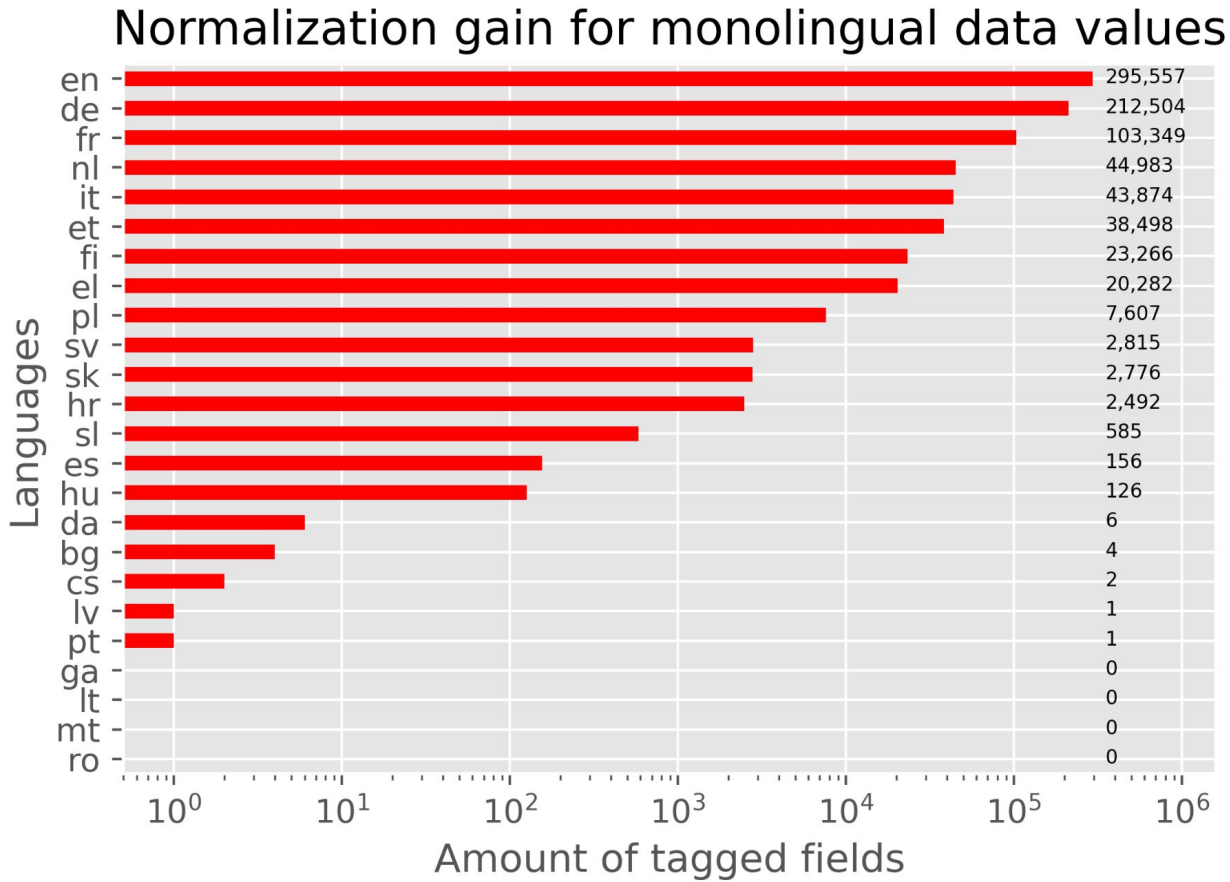
europeana

Results



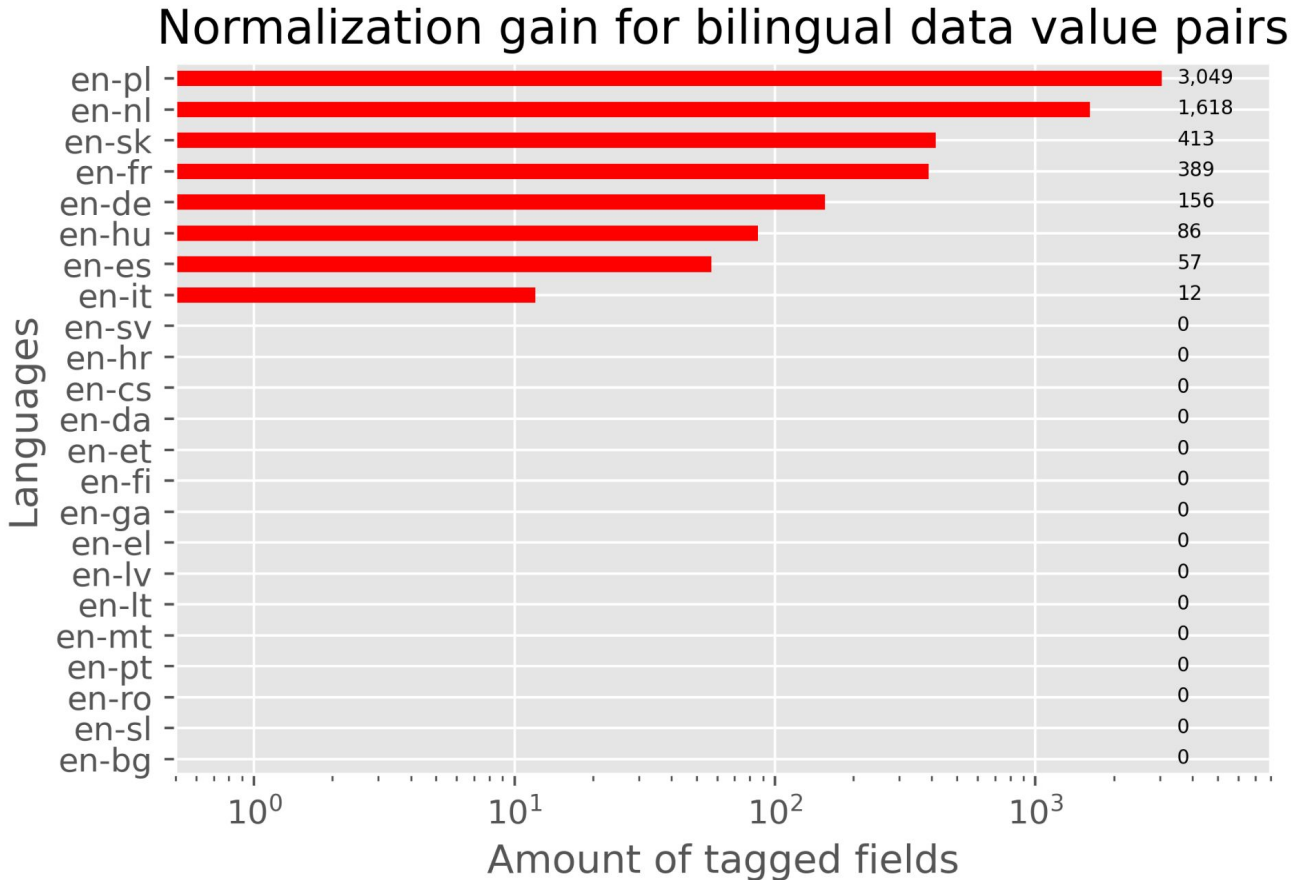
europeana

Results - Improvements obtained by normalisation



europeana

Results - Improvements obtained by normalisation



Conclusions

- Cultural heritage institutions often provide language tagged values in the metadata
- The values don't always comply with the IETF BCP 47 standard, however. For example::
  - Language name used instead of the ISO 639 code
  - ISO alpha-3 code used when an ISO alpha-2 code exists
- Europeana could improve language tags in its metadata by extending its normalisation procedure
  - 7.68 million tags could be made compliant cross all languages
  - 800 thousand tags can be made compliant across the 24 official languages of the European Union, with some significant consequences for training of translation engines.

europeana

Thank you

nuno.freire@europeana.eu

Europeana's language normalisation implementation in Java is available at

https://github.com/europeana/rd-metis-language-normalization

Please contact us if you want to try it on your data!

europeana

europeana.eu

@EuropeanaEU

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

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