Europeana – Core Service Platform

MILESTONE

MS32 - Multilingual research and development plan

Revision	3 - Final
Date of submission	30 October 2015
Author(s)	Antoine Isaac and Valentine Charles, Europeana Foundation
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	28-10-2015	Antoine Isaac	Europeana Foundation	Table of content
2	29-10-2015	Antoine Isaac and Valentine Charles	Europeana Foundation	Additions and first draft
3	30-10-2015	Antoine Isaac and Valentine Charles	Europeana Foundation	Final version

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

‘Europeana DSI is co-financed by the European Union's Connecting Europe Facility’

Table of Contents

1. Introduction

Improving multilinguality is one of Europeana's top priorities. The Europeana V3.0 project has allowed Europeana to set the requirements for future work in MS12 White Paper on best practices for multilingual access to digital libraries.[1] A new version of the White Paper will be soon published, taking into account the feedback received by the community during during a public review period. In the meantime this plan and the Search Improvement (MS30) highlights the work items planned for the next months (and beyond) on multilinguality.
This document is structured following the recommendations of the White Paper. The breath of dimensions from the White Paper make it extremely difficult to find resources to address all recommendations properly. Many items will hence be considered for inclusion in the next DSI plans.

NB: in line with EuropeanaTech's general openness to feedback and suggestions, we provide for the coming months a version of this milestone that is publicly open for comments at https://docs.google.com/document/d/1s08LidGnvScjqYhLtqfb_7xO38tyjMq5UEKsS57Z4ug/

2. "Making your Data Multilingual"

2.1 Language Attributes

- We will focus on increasing the amount of language tags delivered as part of Europeana metadata in related efforts around data quality[2][3]. Language attributes can be added as part of the validation performed in EDM and weighted as part of new metadata completeness measures that we have started to investigate.
- As mentioned in MS30, if resources allow in DSI year 1, we will explore automatic language detection using tools like Apache Tika[4], especially investigate whether sufficient textual, language-dependent description is available in the source metadata for accurate language detection.

2.2 Creating Multilingual Vocabularies through Alignment and Translation

- Europeana's "semantic strategy" gives high priority to exploiting existing multilingual, linked and open vocabularies. However in order to develop its "semantic layer" and integrate in a better way the multilingual data gathered from several sources about contextual entities, Europeana will develop its own Entity Collection, as presented in

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Milestones/Ev3%20MS12%20Multilingual_Access%20White%20Paper.pdf
[2] http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies/data-multilinguality
[3] http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf
[4] https://tika.apache.org/

MS30. This Collection will gather and integrate the multilingual data about contextual entities coming from different datasets, so that it can be used to support Europeana' services such as search.

- Create alignments between multilingual vocabularies. This task can be supported by alignment tools allowing the automatic, semi-automatic or fully manual creation of semantic equivalences between resources, such as CultuurLink[5]. Europeana will promote the adoption and further development of such tools, and explore new requirements for them.
- We will create and publish a list of multilingual vocabularies available for enrichment and alignment and recommended by Europeana.
- If needed Europeana might develop new multilingual vocabularies by adding translations to selected pivot vocabularies. This approach is the one that was taken for the project Europeana 1914-1918 using the Library of Congress Subject headings (LCSH[6]) as pivot. Note that in this case Europeana may use crowdsourcing initiatives to get translations for selected terms.

2.3 Multilingual Semantic Enrichment

- Europeana will expand the language scope of the vocabularies used for enrichment, by linking object metadata elements to more multilingual vocabularies and authority files. We will also enhance enrichment process by making it better aware of the language of metadata. (mentioned in MS30)
- Increasing the amount of languages attributes in the metadata is also part of Europeana quality strategy. The selection of multilingual datasets as target for enrichment will increase the amount of labels with a language attribute in our dataset. We will especially pursue our efforts towards integration with Wikidata, by creating mappings for agents, places, concepts and works entities. Europeana will disseminate a list of criteria to ease the selection of target vocabularies[7].
- The list of multilingual vocabularies available for enrichment and alignment recommended by Europeana (see 2.2) will also be useful in the context of semantic enrichment.
- As mentioned in MS30 we will actively support aggregators and providers who are ready to provide data that includes their own semantic enrichment to multilingual datasets, and make sure Europeana can properly ingest these enrichments.

2.4 Translating Multilingual Metadata and Multilingual Objects

- Europeana will further encourage providers to supply with metadata fields in several languages, including by giving these translations a bigger weight in the computation of new metadata quality measures.
- Europeana will investigate automatic translations tools to translate its object metadata. Collaboration with CEF AT will be key here.
- As an even better option, we will encourage providers to apply translation solutions prior to sharing metadata with Europeana. If translation is applied closer to the source of

[5] http://cultuurlink.beeldengeluid.nl/app/#/

[6] http://id.loc.gov/authorities/subjects.html

[7] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments

metadata in the aggregation chain, the results should be better. Europeana is currently collaborating with several partners in this direction:

  - We have signed a memorandum of understanding[8] with the Latvian Ministry of Culture and the Tilde company for ingesting Latvian records with automatic translations to English.
  - We are discussing with Galway's Insight Centre for Data Analytics and the National Library of Ireland about the perspective of translating Some of Europeana's Irish datasets.
  - We will investigate the feasibility of providers or aggregators using CEF AT, not only Europeana
  - We have started discussion with FREME on testing their localization APIs on Europeana cases.

- In the perspective of encouraging crowdsourcing of translation for metadata, we will follow and advise a project being defined between Europeana and the Wikimedia Foundation in the context of the Europeana 280 campaign. As reported in *MS10 Development Plan GLAMWiki relations*, Wikipedians will connect Europeana objects to Wikipedia, and hence to Wikidata, enabling us to "pull" more multilingual metadata (i.e., the translations of the labels attached to the multilingual concepts and authorities from Wikidata).

3. "Making your User Interface Multilingual"

3.1 Multilingual Static Pages

No R&D action is planned for this recommendation.

3.2 User Language Detection

No R&D action is planned here. The means to proceed with this recommendation are well known, and depend rather on product development decisions than on R&D investigations.

3.3 Interface Language Change

- We will explore automatic translation of the classes and properties of the Europeana Data Model (EDM) using tools such as OTTO[9] in collaboration with Galway's Insight Centre for Data Analytics. This can facilitate the use of EDM in localized user interfaces, by providing a reference for displaying names of EDM fields like "subject" or "title" in several languages, that can be used in object pages or display of facet titles.
- We will seek to establish specifications for a display of object metadata that fully exploits the translations made available as a result of semantic enrichment. E.g., to display a concept exclusively using this concept's label that matches the current user language.

[8] http://pro.europeana.eu/blogpost/latvian-ministry-of-culture-and-europeana-sign-memorandum-of-und

[9] Ontology translation system (OTTO) https://www.insight-centre.org/content/otto-%E2%80%93-ontologytranslation-system

- We will make recommendations for using controlled vocabularies as the source for displaying language and country names in the Europeana Collections portal's object pages, such as the EC controlled vocabularies[10] or Lexvo[11].

4. "Making your User Interactions Multilingual"

4.1 Query Auto-Completion & Query Suggestions

- Entity-Autocompletion will be implemented by Europeana, adhering to the recommendations of the MS12: White Paper on best practices for multilingual access to digital libraries (Europeana V3.0)[12] and as described in MS30 Search Improvement Plan. This milestone also foresees R&D actions to make that function perform better by feeding it with richer and more multilingual Entity Collection data.
- As mentioned in MS30, R&D plans on search for DSI year 1 or beyond include investigations to configure Solr's text analysis chain so that different components (or configurations thereof) are applied depending on the language of the query or the metadata being loaded in the search engine's index.

4.2 Automatic Query Translation

- As part of the product development plans exposed in MS30, Entity Auto-completion will be implemented. Linking queries to this collection amounts to translation when it matches a query with an entity that is described with metadata in different languages. The R&D plans on making the Entity Collection richer (cf. MS30) will thus contribute to better performance here.
- For these cases where there is no completion, automatic query translation should help the service to provide more results across languages. Right now MS30's product development plan foresees that the Query Translation API[13] will become part of the regular Search API. The current API is based on the Wikipedia translation API. Europeana may carry out R&D experiments with other translation services, especially the ones of the CEF AT automatic translation building block. However this will probably happen after May 2016.
- As part of MS30's product development plan, Europeana's Search Engine should have full support for different languages and dialects and should make no difference betweens search for accents and non--accents. Evaluation actions (foreseen to tackle the recommendations below) may provider further insight on the efficiency of this measure.

4.3 User-Assisted Query Translation

No R&D action is planned for this recommendation.

[10] https://open-data.europa.eu/en/data/dataset/language

[11] http://www.lexvo.org/

[12] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Milestones/Ev3%20MS12%20Multilingual_Access%20White%20Paper.pdf

[13] Király, Péter: Query Translation in Europeana. Code4Lib Journal. Issue 27. Available here: http://journal.code4lib.org/articles/10285

4.4 Browse

R&D on enriching Europeana object metadata with more multilingual contextual datasets will make it easier to implement browse features that are truly multilingual, e.g., by making available multilingual vocabularies to structure collection browsing. But we do not foresee any specific R&D actions on developing novel user interface browsing features.

4.5 Search and Browse Result Filtering

- The above recommendations for using controlled vocabularies (see section 2.2) as the source for displaying language and country names will also help enhance multilingual filtering mechanisms (facets).

4.6 Language-Independent Access Options

- Europeana will seek to participate in an "integration pilot" with the Pelagios team in the Peripleo[14] prototype. The integration of the Europeana data with the Pleiades dataset will allow the development of new multilingual and time-and-place-based features.

4.7 Site Structure and Search Engine Landing Pages

No R&D action is planned for this web site design recommendation.

4.8 Multilingual User-Generated Content

- As part of EuropeanaSounds[15], tagging applications of Europeana are planned. Europeana R&D facilitates this effort by providing best practices for representing such annotations, making sure they are multilingual. I.e. ensuring that annotation client allow users to enter appropriate language tags for simple text tags or can rely on suitable multilingual vocabularies for semantic tagging.

5. "Overcoming Challenges in Achieving Multilinguality"

5.1 Avoiding the Language Mix

No R&D action is planned for this (UX) recommendation.

5.2 Distinguishing between the Object and the Metadata Language

No R&D action is planned for this (UX) recommendation.

5.3 Managing Expectations for Automatic Processing

[14] http://pelagios-project.blogspot.co.at/2015/07/peripleo-sneak-preview.html

[15] Europeana Sounds D2.2 Functional design of semantic enrichment http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D2.2-Functional-design-of-semantic-enrichment-v1.0.pdf

- The evaluation actions reported in the next section should provide with useful data for managing expectations. This concerns both the *performance and trust levels for the automatic processes* applied to make the data and UI more multilingual, and a straight measurement of the *'multilingual quality' of the resulting data*. For example, if Europeana or data re-users want to perform search services that use the Entity Collection planned in MS30, at the multilingual scale Europeana must handle, this Entity Collection will have to consistently provide labels in 20 to 30 languages for the contextual resources it includes. Managing users' expectations should then be tackled by providing the necessary UI features that warn them about, say, a display that is based on less precise automatic translation. This falls in scope of regular product (UI) development activities.
- In the Europeana context, i.e., that of a platform, the recommendation to manage expectations shall not be only about end-users. We also need to target data re-users who need reliable information when they develop new services on top of Europeana's. This will require R&D actions on expressing and sharing quality metadata about Europeana's dataset. Such investigations have already started as part of our involvement in the W3C Data on the Web Best Practices Working Group[16].

5.4 Providing Sustainable Multilinguality

- Europeana will have to continuously maintain its focus on multilinguality, for example by continuing its evaluation efforts for multilingual UI features or automatic translation processes. We will make provision for language-related activities in coming project proposals as well as the future plans for the coming DSI years.

- Our commitment to sustainability will also be reflected in our continuous focus on using open data and open software, avoiding vendor or data locking that could make it more difficult to plan for longer-term sustainability and live up to expectations.

6. "Evaluating your Multilingual Components"

6.1 Evaluating your Data

- Performing systematic evaluation of the automatic processes that contribute to multilingual features of the Europeana services (enrichment, translation, etc) is crucial. We will research the definition of metrics for quantifying the performance of these processes, and appropriate quality publication channels so that these assessments, when available, become as transparent as possible to end users and data re-users. In this context we will reuse the evaluation methodology defined and tested by the Task Force on Evaluation and Enrichment[17].
- We will develop criteria to evaluate the “level of multilinguality” in the Europeana dataset. These criteria could include for instance the number of objects with metadata with a specific language tag, the ratio between translated data and data in its original language,

[16] http://www.w3.org/2013/dwbp/wiki/Main_Page

[17] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments

the number of translation call for object pages etc… More specific metrics on the linkage or alignment could be also relevant[18].

6.2 Evaluating your User Interface

This recommendation is planned to be handled through UX product development and evaluation.

6.3 Evaluating your User Interactions

This recommendation is partly planned to be handled through UX product development and evaluation.

- Evaluating search performance with a finer detail will require some more fundamental research, to help develop suitable metrics . We will start by consolidating Europeana's state-of-the-art on search evaluation methodology[19], notably by reflecting on the estimation of what counts as a relevant results and possibly fine-tuning the metrics used in recent evaluations. Based on this, we will provide suitable guidance for the design of evaluation tasks planned in MS30.

[18] http://w3c.github.io/dwbp/vocab-dqg.html#express-the-quality-of-a-linkset

[19] https://europeanadev.assembla.com/spaces/europeana-r-d/wiki/Search_evaluations_at_Europeana