

# Multilingual Strategy

(MS33 Evaluate options for multilingual search and browse M18)



<b>Revision</b>	1.2
<b>Date of submission</b>	30 April 2020
<b>Author(s)</b>	Andy Neale, Europeana Foundation Antoine Isaac, Europeana Foundation Hugo Manguinhas, Europeana Foundation Dasha Moskalenko, Europeana Foundation Mónica Marrero, Europeana Foundation
<b>Dissemination Level</b>	Public

## Revision History

<b>Revision No.</b>	<b>Date</b>	<b>Author</b>	<b>Organisation</b>	<b>Description</b>
0.1	01/02/20	Andy Neale	Europeana Foundation	Draft
0.2	16/04/20	Andy Neale, Antoine Isaac, Hugo Manguinhas	Europeana Foundation	Updates
0.3	30/04/20	Andy Neale, Antoine Isaac, Hugo Manguinhas, Dasha Moskalenko	Europeana Foundation	Updates
1.0	30/04/20	Andy Neale	Europeana Foundation	Final version for DCHE Subgroup feedback
1.1	09/09/20	Andy Neale	Europeana Foundation	Updated to reflect discussion with DCHE Subgroup
1.2	23/12/21	Antoine Isaac, Douglas McCarthy, Paolo Scalia, Dasha Moskalenko	Europeana Foundation	Updates following progress in DSI year 3.

# TABLE OF CONTENTS

<b>Executive summary</b>	<b>3</b>
<b>Introduction</b>	<b>4</b>
Purpose of this document	4
Background	4
Strategic drivers	6
Use cases	7
<b>Conceptual solution</b>	<b>7</b>
Approach	7
Solution for underlying multilingual data	8
Solution for multilingual search	10
Solution for reading item text	12
Solution for reading editorial content and website copy	14
Solution for navigating the Europeana website	16
<b>Community feedback</b>	<b>17</b>
<b>Roadmap</b>	<b>18</b>
<b>Appendices</b>	<b>22</b>
Appendix A: Definitions	22
Appendix B: Summary output from Multilingualism in Digital Cultural Heritage	24
Appendix C: History of multilingual investigations	26
Appendix D: User research results	28

# Executive summary

Presented here is a medium-term strategy for the improvement of multilingual experiences on europeana.eu.

Multilingual access to extend the reach and impact of europeana.eu has been a long term goal for Europeana, stretching back to 2009.

Users also have growing needs and expectations for accessing material in alternative languages, based on their other experiences online.

The vision of a full multilingual experience, for all 24 official languages of the European Union, is now within the realms of possibility because of technological advances.

However capability, resourcing, and technical complexity are all current issues that limit the speed at which progress can be made.

Delivering on this strategy requires Europeana to break new ground for the cultural sector, and for this reason R&D and implementation will need to iterate together.

Expert solutions are proposed in this strategy, but will likely still evolve as experiments prove or disprove the proposals.

Multilingual use cases addressed in this strategy cover the ability to navigate the Europeana website, read editorial content and website copy, search, and read item text.

The core solutions for searching and reading item text across languages rely upon the:

- Use of trusted vocabularies that come with existing multilingual coverage of metadata
- Translation of all metadata and text to English, so it can act as a pivot language for the areas not covered by the trusted vocabularies

Multilingual coverage of the user interface, editorial and website copy, metadata, and full text content will be additionally supported by:

- Paid and community translators
- Use of real-time translation services from English to fill translation gaps

The roadmap identifies logical groupings and sequences of work to demonstrate that implementation of the strategy is achievable, subject to prioritisation and resources.

Definitions for terms used in this strategy are listed in Appendix A.

# Introduction

## Purpose of this document

The purpose of this document is to provide medium-term direction for the multilingual use of europeana.eu and its collected data. This is ultimately to support the mission of the Europeana Initiative, but also to provide practical guidance for the development of services.

## Background

The europeana.eu website contains material from galleries, libraries, archives and museums in all EU member countries. Visitors can currently navigate the website in all the EU's 24 official languages, and it's easy to search for items described in a visitor's own language. But things get more complicated when visitors want to see items that match a search but are described in a different language. Or when a text item is only readable in a language visitors may not be familiar with.

In total, Europeana's partners use 38 languages to describe the collections<sup>1</sup>. However, more than half of all the material (57%) uses one of just five languages - English, German, Dutch, Norwegian or French. It means that large parts of the collections have more limited language accessibility.

During October 2019, the Finnish Ministry of Education and Culture, with the Europeana Foundation, convened a two-day workshop on the topic of 'Multilingualism in Digital Cultural Heritage - needs, expectations and ways forward'. It was run for Member States, Cultural Heritage Institutions, and experts in the field, run under the umbrella of the Presidency of the Council of the European Union.

Participants identified the biggest benefit of multilingual access as being that *people can access the information of other cultures and language groups*. Followed by the point that *multilingualism promotes socially inclusive societies and mutual understanding of different cultures*.

Participants were then asked to consider the issues that prevented them from fully delivering the benefits of multilingual digital cultural heritage. The primary conclusion was that many technology solutions already exist to solve multilingual challenges, however

---

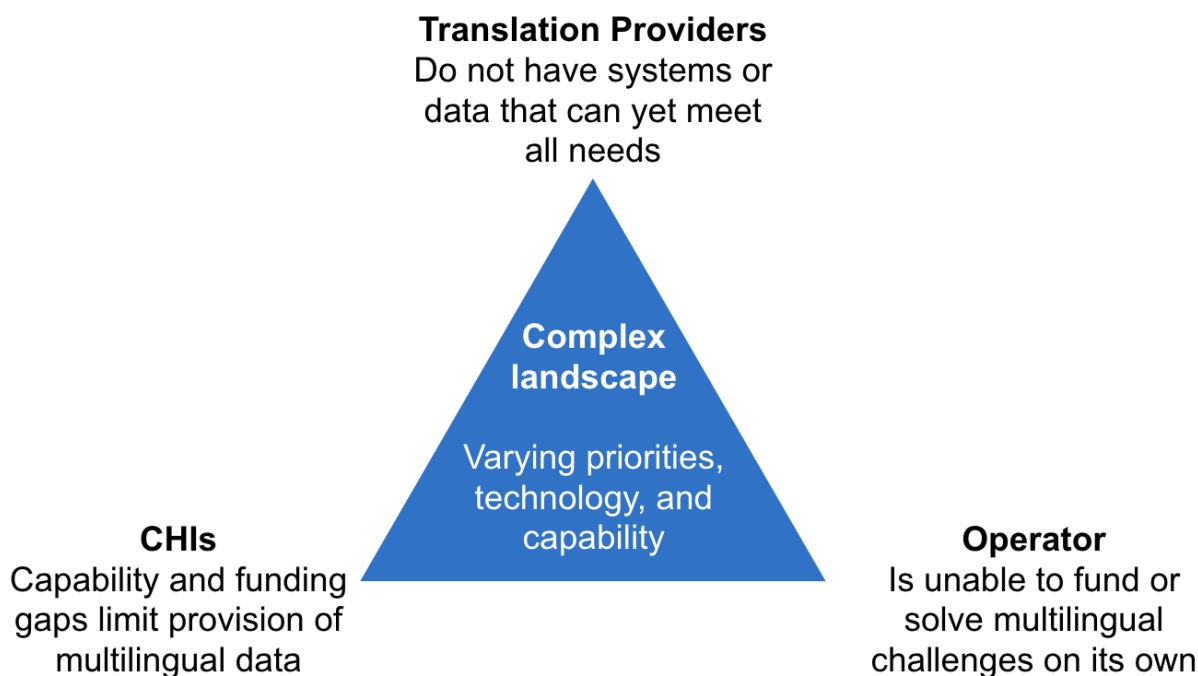
<sup>1</sup> This is the number of languages that providers declare for their datasets. On further inspection, individual metadata values (title, subject, etc.) come in many more languages.

member states felt there were often capability and resourcing gaps in institutions that slowed progress. Experts also identified that the cultural heritage sector has unique needs, and that a lack of training data meant that many existing technology solutions were not yet fit for purpose for the sector. A fuller summary of the output is detailed in Appendix B.

This multilingual strategy builds on many years of investigations and proposals, details of which can be found in Appendix C.

### **Problem space**

The landscape for multilingual access is complex, with stakeholders all having varying priorities, technology, and capability. The problem space can be introduced as follows.



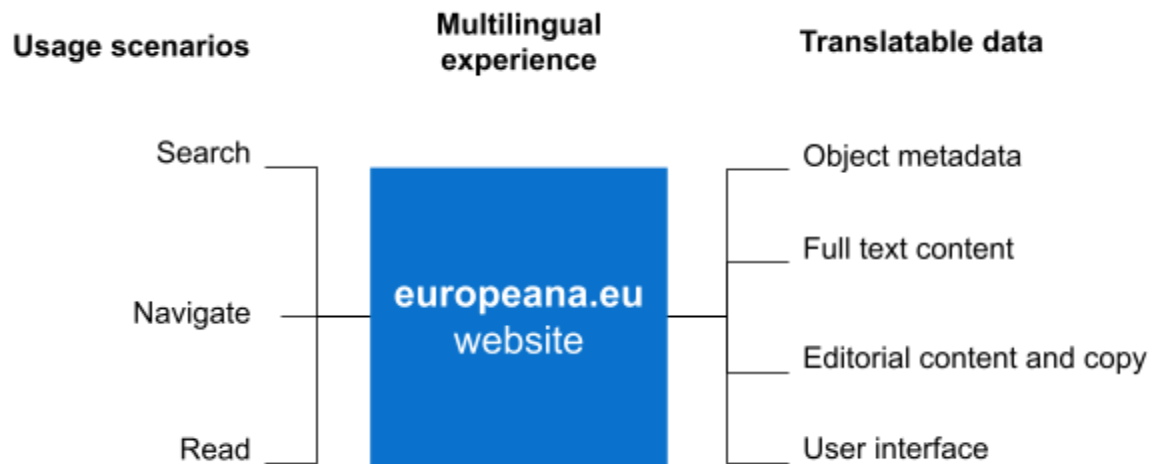
*Figure 1. Conceptual model of multilingual problem space*

### **Solution space**

Three aspects are seen as needing to be addressed in order to increase the multilingual reach of europeana.eu. These three areas build on each other, more or less covering the categories in the best practices for multilingual access<sup>2</sup>:

<sup>2</sup> <https://pro.europeana.eu/post/best-practices-for-multilingual-access>

- Translatable data (object metadata, full text content<sup>3</sup>, editorial content and copy, and user interface)
- Usage scenarios (search, navigate, read)
- Multilingual experience



*Figure 2. Conceptual model of multilingual solution space*

## Strategic drivers

Given the complex landscape, it is important to be clear about the factors that are strongly influencing the direction. These factors are driving this proposed strategy:

1. Users have growing needs and expectations for accessing material in other languages
2. Experts recognise the translation of cultural heritage material pose significant and complex challenges
3. Organisations often lack the capability to take advantage of technology solutions
4. Multilingual solutions must acknowledge constraints in data quality, technical feasibility, and funding
5. Solutions cannot be fully known until they are validated with experiments
6. No single stakeholder or solution can solve multilingual challenges on their own

<sup>3</sup> Including user-generated content

## Use cases

Four main use cases were identified after considering the practical experience of multilingual content, alongside user research results noted in Appendix D. These use cases describe more concrete needs for multilingual access by Europeana's end users, and are used in this strategy to provide more focus for solutions.

**I. Navigate the Europeana website**

Visitors choose a language they are comfortable with. They see navigation links and search filters in their language of choice.

**II. Read editorial content and website copy**

Visitors encounter engaging exhibitions, galleries, blog posts, website copy, and promotional text in their language of choice.

**III. Search Europeana**

Visitors compose a search query in a language they are comfortable with. Visitors are not confused by search results, including items that may be in different languages.

**IV. Read item text**

Visitors read the title, description, and other supporting metadata on an item or collection page regardless of the original item language. They can also read any text objects in their language of choice.

## Conceptual solution

### Approach

In order to provide an immersive multilingual experience for website visitors, this solution proposes an approach that requires less significant effort or capability from partners. Capability building by partners is welcome of course, but should not be seen as a barrier.

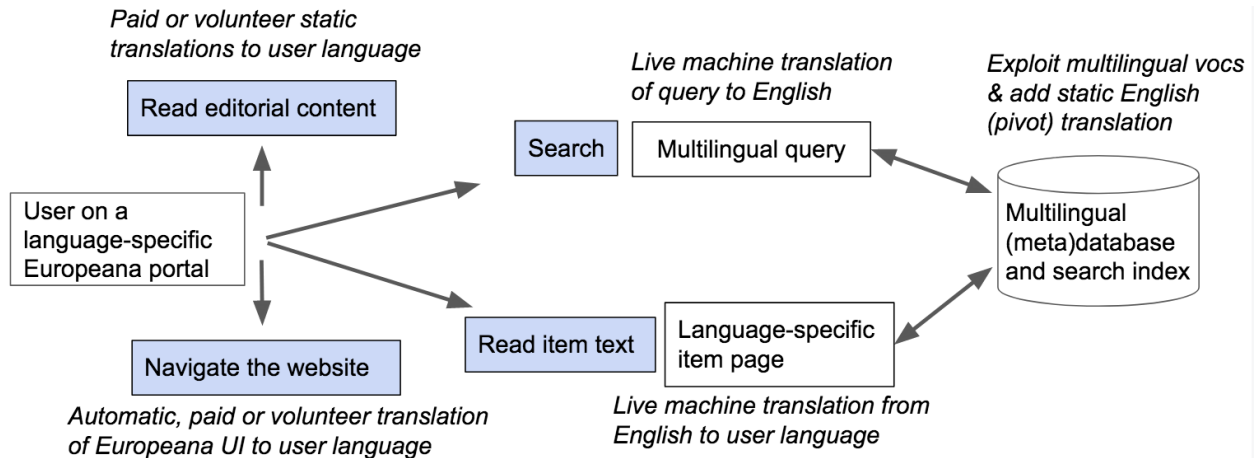


Figure 3. Main points of the multilingual approach

The core solution for searching and reading items across languages relies upon a strategy for building underlying multilingual data. This can be achieved by the:

- Use of trusted vocabularies that come with existing multilingual coverage of metadata
- Translation of all metadata and text to English, so it can act as a pivot language for the areas not covered by the trusted vocabularies

The multilingual coverage of the user interface, editorial and website copy, metadata, and full text content will be additionally supported by:

- Paid and volunteer translators (crowdsourcing community)
- Use of real-time translation services from English to fill translation gaps

Although the strategy does not rely on building network capability, stakeholders will be encouraged to partner, employ, and develop technology that better suits the specificities of cultural heritage cases. Including collaboration with DSI eTranslation and CEF partners to source training data and support digital transformation in member states.

## Solution for underlying multilingual data

The underlying multilingual data in Europeana is what needs to power the experiences for finding and using material across languages. The easiest solution would be to machine translate all metadata and full text content into all 24 official languages. But that *easy* solution would not be affordable, and not always be reliable<sup>4</sup>, so other solutions are proposed here.

<sup>4</sup> In particular, short phrases in individual metadata fields are difficult to translate, especially when their language is not certain



## Exploit trusted vocabularies

The first aspect of the underlying data solution is to exploit the existing expert translations available in trusted vocabularies used by Europeana and its partners. In the context of this multilingual strategy trusted vocabularies are used to create a knowledge graph of entities<sup>5</sup> and terms that we can source translations from. Further definition of trusted vocabulary and knowledge graphs are noted in Appendix A.

- Encourage data contributors to increase the use of terms from trusted vocabularies when providing their metadata. Vocabularies often contain multilingual translations of metadata terms
- Store and index multilingual vocabularies<sup>6</sup> to provide partial translations for metadata across Europeana to support search and display applications in the europeana.eu website
- Look for opportunities to help extend the coverage of vocabularies to support all official languages

## Use English as a pivot language

While the use of translations from trusted vocabulary terms will provide some multilingual coverage, it won't cover it all. Therefore, the second aspect of the underlying data solution is to adopt English as a pivot language to fill translation gaps. A pivot language is an intermediary language for translation between many different languages, sometimes also called a bridge language. In the context of this multilingual strategy, English is being proposed as the bridge for translating all other languages to and from.

- All item metadata and full text content can, over time, be translated into English
- English translations would be based on using provider's metadata, translations sourced from vocabularies, and machine translation services
- Aggregation systems and ingest tools would be updated to also support third-party provided English translations
- English metadata and full text content would then be stored and indexed to support multilingual search and translation for display
- If a visitor is searching in a language that Europeana doesn't have matching metadata and full text content for, then the query can be translated in real-time from the stored English
- Machine translation services from English to other official languages are advanced, which makes it a good choice for a pivot language. Much of the Europeana data corpus is already in English so this also makes it a good choice to build on

---

<sup>5</sup> <https://pro.europeana.eu/page/entity#entity-collection>

<sup>6</sup> This would include manually sourced translations of specialised vocabularies, such as the one Europeana uses for rights statements, which is served as linked open data at [rightsstatements.org](https://rightsstatements.org)

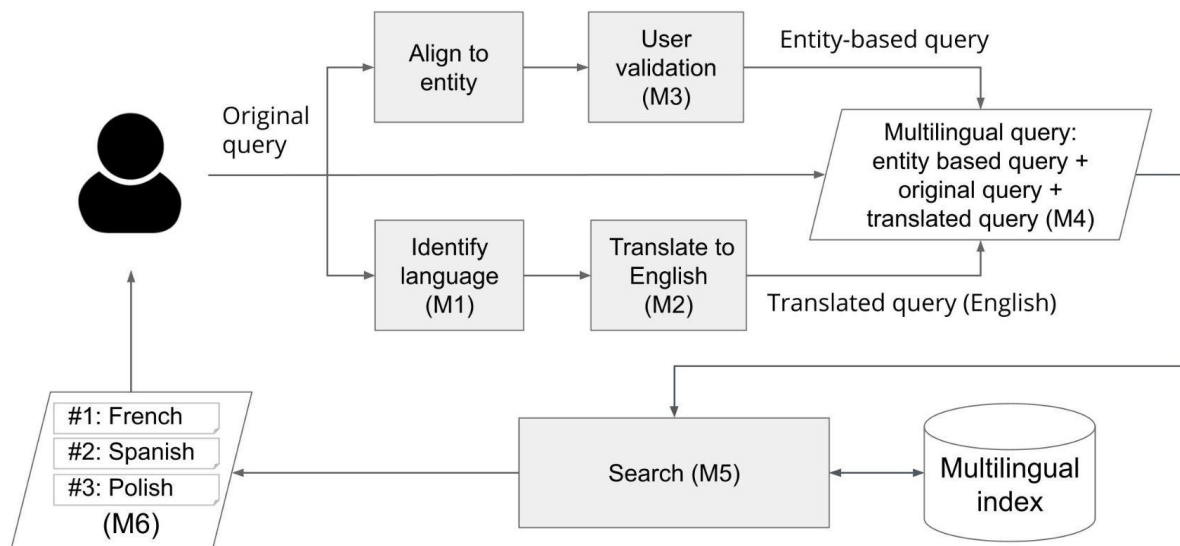
## Solution for *multilingual search*

In order to meet visitor search needs the multilingual experience must cover the entering of a search query, finding multilingual search results, and the display of results in an understandable way. The use of trusted vocabularies and the English pivot language will make this possible.

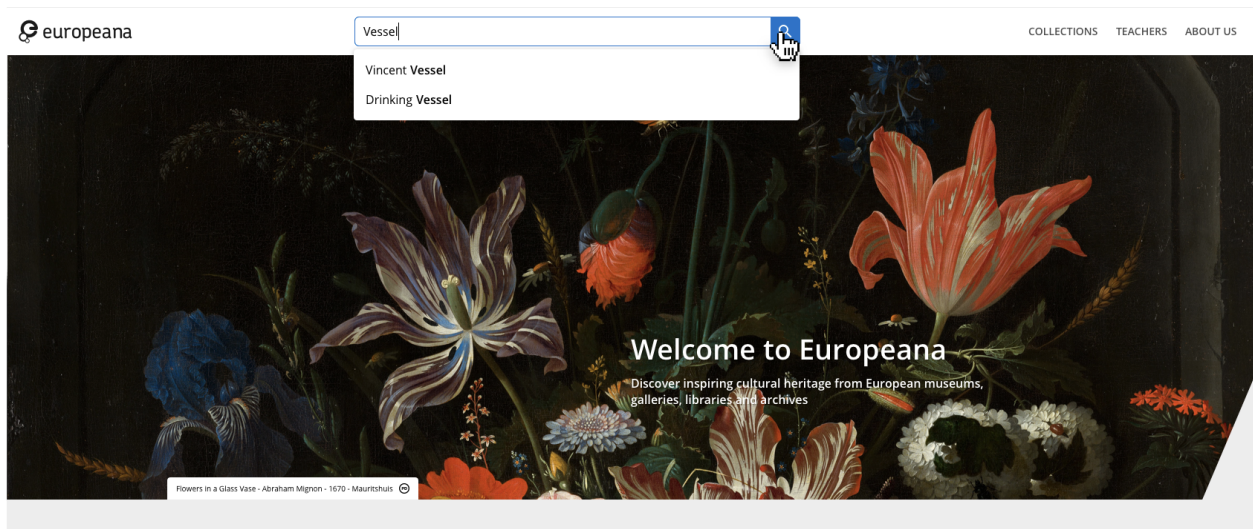
### Enable multilingual search queries

The end-to-end process for supporting search can be described in the below flow diagram.

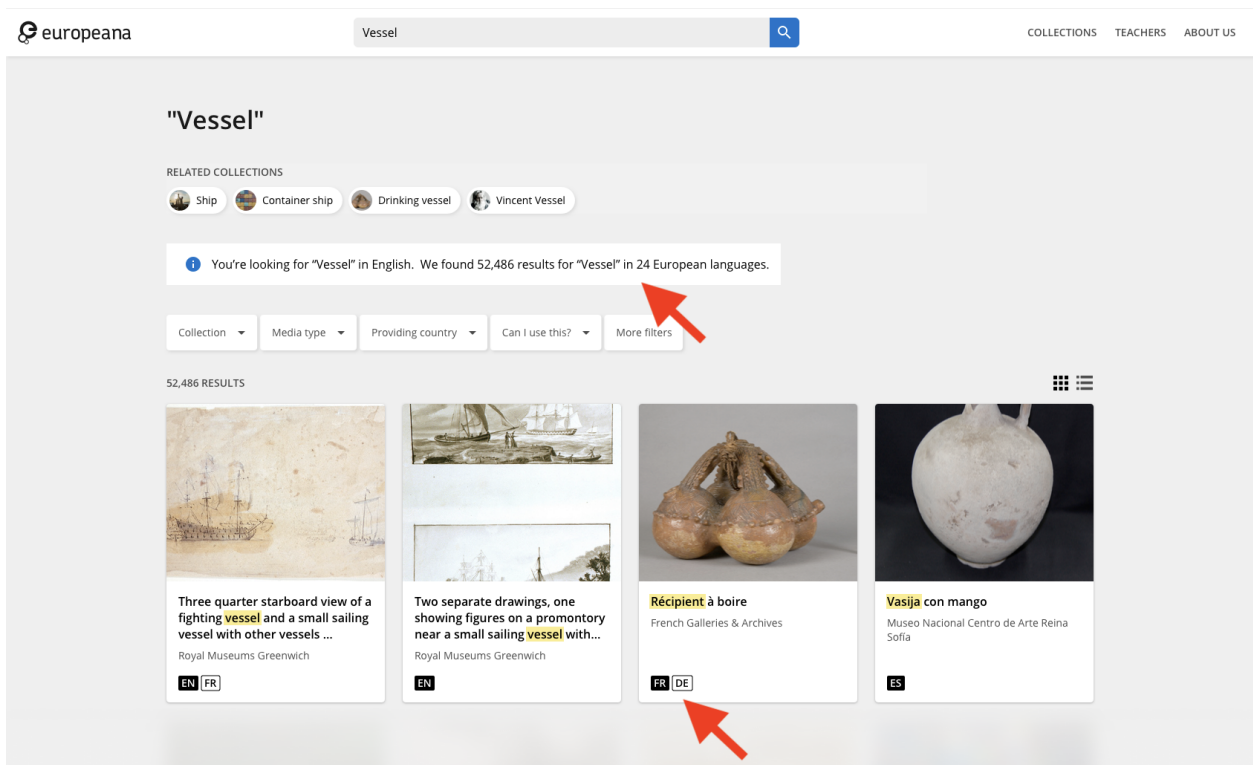
- User enters original search query in official language of choice
- Language of query is automatically identified (M1)
- Real-time translation of string to English (M2)
- Attempt to disambiguate the queries by aligning to entities in the knowledge graph (M3)
- Submit query comprising of [original search phrase] + [English translation of search phrase] + [validated entities across all languages in knowledge graph] (M4)
- Search results would have more matches because both the queries and indexes are augmented with multiple language variants (M5)
- Search result titles and descriptions present in source language (M6)



## Example design of multilingual search query (nothing special to see)



## Example design of multilingual results page



## **Rationale for the solution**

Factors that support the proposed solution include:

- Use of trusted vocabularies and a pivot language would reduce the amount translation that is required, provide "more authoritative" matches across languages, and would in theory provide good multilingual search coverage
- Given the significant gaps in trusted vocabularies, the solution does mean that extensive translation to English is required, however this seems the most cost-effective solution on balance

## *Solution for reading item text*

Once a visitor has found an item of interest, they want to then go to the item page and read any details about the object, even if it is in a language they don't know. This will include the title, description, any associated metadata, and the full text content of readable items such as newspapers, books, or documents. This solution again will build on the underlying multilingual data provided for by trusted vocabularies and the English pivot language solution.


## **Extend existing index and use real-time translation**

- Default to the display of chosen navigation language on the item page, if source metadata is available, e.g., if you have selected the Spanish user interface, an item will display in Spanish if Spanish metadata is available
- Display other language options for the page where metadata is fully translated and already indexed and stored in Europeana. This can be based on translations present in the original object metadata fields or in the trusted vocabulary and knowledge graph introduced in the solution for underlying multilingual data
- Provide options to dynamically translate the page to any other language of choice, from the English pivot translation stored in Europeana. Using real-time machine translation services to make readable all displayed metadata and full text content
- Evaluate whether to add the real-time translations to the existing index and data store for the record, so that dynamic translation would not be required again for that language

## Example design of multilingual item page

europaena Search

COLLECTIONS TEACHERS ABOUT US



Public domain Download Share

This item is available in the following languages: **English** French Auto translate

Three quarter starboard view of a fighting vessel and a small sailing vessel with other vessels drawn in faintly. Copied from Greenwich 218

Three quarter starboard view of a fighting vessel and a small sailing vessel with other vessels drawn in faintly. Copied from Greenwich 218

COLLECTIONS YOU MIGHT LIKE

Ship Container ship Drinking vessel

**GOOD TO KNOW** ALL METADATA

Added by	Royal Museums Greenwich
Institution	Royal Museums Greenwich
Country	United Kingdom
Creation date	1825

## **Rationale for the solution**

Factors that support the proposed solution include:

- Some metadata records already contain translations for multiple languages, so it makes sense to give visitors the option to read the item page text in those languages
- When considering other untranslated text we are aware that dynamic translation of website content sometimes now comes for “free” in many browsers. However the coverage is not universal across all browsers/languages and not currently common for mobile experiences. Therefore it is still important to embed real-time translation features in the europeana.eu website

## *Solution for reading editorial content and website copy*

The presentation of multilingual content on the Europeana website also requires website copy ('static text') to be translated into official languages. This includes the titles and descriptions of blog posts, exhibitions, gallery and other editorial headings, the 'about us' page, and help information - but not the full text of blogs and exhibitions, due to budget constraints. .

## **Use editorial translators**

- Provide sufficient budget for regular translation of website copy ('static text') into official EU languages
- Use paid translators to ensure coverage of all gallery and editorial headings, as well as website copy linked in the footer
- Request that Generic Service projects budget for (at least some) translation of their editorial content

## Example gallery and editorial headings to translate

The screenshot shows the Europeana website interface. At the top, there is a search bar with the text 'Que cherchez-vous?' and a search icon. To the right of the search bar are links for 'COLLECTIONS', 'ENSEIGNANTS', and 'À PROPOS'. Below the search bar is a main heading 'Seasonal celebrations' with a red arrow pointing to it. Underneath is a paragraph: 'Find out how different countries and cultures celebrate the passing of the seasons and how their traditions have been shaped around nature.' with a red arrow pointing to the end of the sentence. Below this is a section titled 'Galleries' containing four gallery cards. Each card has a title and a short description. Red arrows point to the titles and descriptions of the 'Rainbows' and 'Winter Sports' galleries.

**Seasonal celebrations**

Find out how different countries and cultures celebrate the passing of the seasons and how their traditions have been shaped around nature.

**Galleries**

**Rainbows**  
Rainbows have been used as symbols for centuries, from biblical scenes to heraldry to the symbol of pride. This gallery

**Spring**  
Enjoy depictions of spring in Europeana, filled with birds, flowers and sunny skies with dappled clouds.

**Autumn in Art**  
Autumn has inspired artists for centuries. This gallery shows how artists have depicted the harvest season through colourful

**Winter Sports**  
Vintage photos and historical paintings from across Europe showing how winter sports were enjoyed in the past

## Rationale for the solution

Factors that support the proposed solution include:

- Experiments with the automatic translation of exhibitions has shown that machine translation services cannot yet meet quality criteria. Exhibition text is carefully crafted and hence manual translation is needed to maintain quality and style. As technology improves this can be re-evaluated
- Good progress on the multilingual experience can be made by prioritising which languages to manually translate editorial into. This can be based on both the expected reach of chosen languages, and the topics that may be of interest to select language groups
- A modest increase in translation funding can be supported within existing budgets to support more coverage of editorial translations
- The experience with using partners to help translate editorial has been mixed, because individual cultural heritage experts often lack the time and expertise to contribute. However, it is worth experimenting with this further to see what additional language coverage can be gained

## Solution for *navigating the Europeana website*

In order for visitors to navigate the website in their chosen language, all user interface components need to be maintained in the 24 official languages, except legal statements that remain in English. User interface components are things like the navigation bar labels at the top of the website, footer links, search filters, and other standard buttons, controls, and text used by visitors to navigate the website.

### Maintain user interface translations

- Use automatic translation services such as Google translate to maintain user interface labels as features change and validate the automatically generated translations with internal native speakers when in doubt.
- Use translation workflow tools to ensure user interface changes are easily managed<sup>7</sup>

### Example user interface translations to maintain

The screenshot shows the Europeana website search interface in French. At the top left is the Europeana logo. To its right is a search bar containing the text "Que cherchez-vous ?". Further right are navigation links: "COLLECTIONS", "ENSEIGNANTS", and "À PROPOS". Below the search bar is a message: "Vous effectuez vos recherches sur notre nouveau site web plus rapide. Affichez ces résultats de recherche dans l'Europeana original." Below this is the "Rechercher" section, which includes several filter buttons: "Catégorie", "Type de support", "Puis-je le réutiliser ?", "Pays fournisseur", and "Plus de filtres". Below the filters, it says "Résultats: 47,725,250". The main content area displays four search results, each with a thumbnail image and a title: "burnus", "jas", "herenvest", and "broek". Each result also includes the text "Heritage and Sustainability - University of Antwerp". Red arrows point to the search bar, the navigation links, the search results count, the filter buttons, and the first search result.

<sup>7</sup> Europeana currently uses an external tool from [lokalise.com](https://lokalise.com)



## Rationale for the solution

Factors that support the proposed solution include:

- Commissioning paid translations for UI components is expensive due to the number of languages we support and the number of UI components that we have, which is continuously growing.
- Automated translations are quick and easy to obtain, as opposed to paid translations that take time to commission which will disturb the frequency of our releases.
- Validation of specific translations with internal native speakers is fast and without cost and allows us to maintain a good quality of translations.

## Community feedback

A technical paper<sup>8</sup> outlining the strategy presented here was circulated in advance of the Finnish presidency event on Multilingualism in Digital Cultural Heritage. Presentations by Europeana staff at this event explained the main proposals in our strategy. Participants were positive about the material, and expressed interest for these options to be further explored.

In parallel, Europeana had issued a call for feedback on the technical paper especially targeting our technical community (EuropeanaTech). We did not receive many reactions, and no negative ones<sup>9</sup>. The community can be seen to endorse our general direction, which does not come as a great surprise considering that it has been previously aware of and often involved in elaborating key components of the strategy, as noted in the history of multilingual investigations in Appendix C.

### Feedback of note

One respondent highlighted that our approach of relying on multilingual vocabularies for the most sensitive specialised translation needs, and on automatic translation for a bigger mass of less-sensitive needs, was quite similar to the one employed for the aggregator platform for egyptology Cleo<sup>10</sup>.

Others warned about the difficulties and risks of some specific aspects of it, such as using IP detection for identifying a user's language (as opposed to using their browser's

---

<sup>8</sup> <https://pro.europeana.eu/post/help-build-multilingual-systems-for-digital-cultural-heritage>

<sup>9</sup> We would like to thank Heleen Wilbrink, David Haskiya, Andreas Maier, Johanna Monti and Ulrich Kampffmeyer for the written observations they submitted to us.

<sup>10</sup> <https://www.cleo.aincient.org/pages/en/>

preferences or dedicated Natural Language Processing technology). Or that some "language-neutral" fields (like dates and measures) could include some country- or culture-specific aspects (calendars or units of measures).

The feedback received also reiterated points from the Finnish presidency event, that building sector capability to exploit machine translation technologies needs appropriate attention.

Sourcing enough data to properly train machine translation tools for cultural heritage cases is crucial, as is the creation of robust consortiums of cultural heritage institutions, technology providers and researchers, which can rise to the challenges of solving multilingual issues using these tools. Related to network capability, one respondent suggested that national aggregators could be more involved in the sourcing of manual translations for UI components.

A last piece of feedback identified the possibility of exploiting language resources gathered by communities, such as the community for Linguistic Linked Open Data<sup>11</sup>. These resources include terminologies and vocabularies that can be used to extend the language coverage of translations that can be sourced from trusted vocabularies and knowledge graphs, even though they are less cultural heritage specific. These language resources can also boost the efficiency of Natural Language Processing technology for cultural heritage data, which include the running of automatic translation services like the eTranslation DSI<sup>12</sup>. Initiatives to gather more and better-fit resources in this area, like the Nexus Linguarum COST action<sup>13</sup>, may fit well with the capability building aspects of the roadmap.

## Roadmap

The roadmap starts to identify, organise, and sequence the types of tasks that are required to deliver on this multilingual strategy. A significant number of experiments are identified because this approach is breaking new ground for the cultural sector, and there are many details that cannot be known in advance of the work. The ultimate implementation plan will need to factor in that solutions and tasks may change significantly based on the outcomes of the experiments. Scheduling of activity will happen in implementation planning, subject to prioritisation and resources.

---

<sup>11</sup> <https://linguistic-lod.org/lod-cloud>

<sup>12</sup> <https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>

<sup>13</sup> <https://www.cost.eu/actions/CA18209/>

Outcomes	Stage 1	Stage 2	Stage 3
<p>Policy and plan established</p>	<p>Prioritise languages to support if resourcing does not allow coverage of full 24 official languages</p> <p>Update Europeana policy to account for support of other non-official EU and European regional languages</p> <p>Confirm a set of metrics and KPIs to define both quality targets and desired performance improvements to the multilingual experience</p> <p>Assess language coverage of entities used in search and source data</p>	<p>First implementation and evaluation of the metrics and KPIs, focusing on entities</p>	
<p>Underlying multilingual data is established</p>	<p>Language detection of metadata is validated (experiment)</p> <p>Prioritise normalisation of not-yet normalised tags (in original and dereferenced data)</p>	<p>Agree on evaluation methodology and quality thresholds for translations</p> <p>Candidate machine translation services for metadata are evaluated (experiment)</p> <p>Work with communities and data partners to extend language coverage of entities (vocabularies) where necessary</p> <p>Decide whether translation of metadata fields should focus on a selection of fields, such as discovery-enabling fields or metadata record Tier 2+ objects</p> <p>Language detection of full text is validated (experiment)</p> <p>Candidate machine</p>	<p>Machine translation pipeline translates all metadata to English, allows for quality control, then stores and indexes data</p> <p>Enhance coverage of multilingual knowledge graph over Europeana collection objects by improving semantic enrichment</p> <p>Machine translation pipeline translates all full text to English, allows for quality control, then stores and indexes data</p> <p>Evaluate options for handling full text that is embedded within IIIF (especially for language detection)</p>

		translation services for (static) full text content is evaluated (experiment)	
Capability is grown across the network of stakeholders and partners	<p>Meet with eTranslation-related actors such as eTranslation DSI, Crosslang, and Pangeanic to agree on collaborations</p> <p>Consider how multilingual strategy supports CHI digital transformation</p>	<p>Encourage the cultural heritage sector to contribute training datasets</p> <p>Encourage data providers to share their vocabularies and link them to spine/broader vocabularies such as VIAF, Wikidata, AAT, etc.</p> <p>Disseminate information to Europeana Network to aid the knowledge building e.g. via EuropeanaTech Insight publication</p>	<p>Organise and participate in project consortiums that can contribute to improving quality of multilingual services, including multilingual vocabularies, use of machine translation services, and BERT-based models</p> <p>Evaluate EPF updates for measuring and rewarding multilingual data</p>
Navigate the Europeana website	Maintain user interface translations in supported languages	Maintain user interface translations in supported languages	Maintain user interface translations in supported languages
Read editorial content and website copy	Use editorial translators to translate editorial features and website copy in supported languages	Use editorial translators to translate editorial features and website copy in supported languages	<p>Assess costs of fully translating website legal text</p> <p>Use editorial translators to translate editorial features and website copy in supported languages</p>
Search Europeana	<p>Real-time detection of search query language is validated (experiment)</p> <p>Construction of multilingual search string is validated (experiment)</p> <p>Multilingual search designs prove to be usable and understood by users (user research)</p> <p>Stop applying English text</p>	<p>Real-time translation of search query is validated (experiment)</p> <p>Route queries to specific language fields (metadata or full text separately) instead of issuing them against all data (experiment)</p> <p>Design of ranking for multilingual search results is validated (experiment)</p>	<p>User can enter search query in chosen language and get multilingual results (implementation)</p> <p>Users get better multilingual search results based on the inclusion of full text translated to English in search indexes</p> <p>Improve detection of</p>

	analysis to all languages in Solr	<p>Ranking for multilingual results (implementation)</p> <p>Review handling of languages in Entity API suggester method to meet expectations of new multilingual search UX</p>	<p>entities in phrase queries</p> <p>Stop applying language analysis (e.g. stemming) to entities in metadata and full text (experiment)</p>
Read item text	Multilingual item page designs prove to be usable and understood by users (user research)	<p>Real-time translation of item page metadata from English is validated (experiment)</p> <p>Real-time translation of full text objects from English is validated (experiment)</p>	<p>Users can view item pages in language of choice (implementation)</p> <p>Evaluate whether to add the real-time translations to the existing index and stores for the record so that dynamic translation would not be required again for that language</p> <p>Users can view full text content in language of choice (implementation)</p>

# Appendices

## Appendix A: Definitions

Browse	Topical pages of collection items organised by subjects, people, and places.
Chosen language	The language that the visitor chooses, or prefers, to experience.
Discovery-enabling metadata	Metadata fields that play an essential role in helping users find the objects they are interested in, such as title, subject and description.
Display	The display of a collection record on an item page.
Editorial content	Europeana content such as blogs, galleries, and exhibitions.
Field	A single element of a metadata record describing an object. For example title, creator, date.
Full text content	Digital content (e.g. from an article, document, or book) available as plain text, as opposed to in an image or audio-visual form.
Index	A system component that holds all a copy of metadata that can be queried via the search experience.
Knowledge graph	In the Europeana context, a knowledge graph is a network of related places, persons, concepts extracted from trusted vocabularies, coming with terms in various languages. It is part of a network of data sources available in the wider Linked Open Data cloud.
Object	Digital content such as an image, document, video, audio, or 3D item.
Object metadata	The textual information and hyperlinks that serve to identify, discover, interpret and/or manage a content object. An object's metadata record is structured into fields like title, creator, subject, etc.
Official EU languages	Official languages as per Europa.eu website <sup>14</sup> . Note that variants to these official languages are not within scope of this strategy e.g. the service will support European Spanish, but not necessarily Spanish of the Americas.
Pivot language	Is a language used as an intermediary language for translation between many different languages, sometimes also called a bridge language. In the context of this multilingual strategy, English is being proposed as the proposed bridge for translating all other languages to and from.

<sup>14</sup> [https://europa.eu/european-union/about-eu/eu-languages\\_en](https://europa.eu/european-union/about-eu/eu-languages_en)

Real-time translation	Query that returns an instant translation using an automated machine translation service, often from third-party providers.
Search	Search experience covering the search query and results page.
Source language	The original language of material provided by the provider.
Static translation	Batch translation process that relies on machine translation services with validation and/or human translation services.
Trusted vocabularies	A selection of agreed words and phrases used in metadata to describe an item, often within a specific domain. Sometimes referred to as a <i>controlled vocabulary</i> , knowledge organisation system, or (name) authority list. They allow different providers to use metadata terms consistently. In the context of this multilingual strategy they are used to create a knowledge graph of entities and terms that we can source translations from.
User interface	Website interface text such as navigation, buttons, graphics, and search filters.

## **Appendix B:** Summary output from *Multilingualism in Digital Cultural Heritage*

### **What was talked about?**

The event was a mixture of speeches, case study presentations, workshop sessions and a panel discussion. The event covered a broad range of topics, including multilingual policy, user experience design, learnings from automatic translation projects, multilingual metadata, linked vocabularies, automatic subject indexing services, and measuring success.

Working in groups, participants were then invited to share their experiences of opportunities and challenges related to multilingualism. They identified benefits of multilingualism and discussed what solutions or changes are needed to address the challenges.

### **Benefits of multilingualism**

During the speeches and workshop exercises, the following benefits of multilingualism emerged:

- Access to more sources of information, and to the knowledge and history of other cultures and less common language groups
- Promotion of socially inclusive societies and mutual understanding of diverse cultures
- Increased usability of digital cultural heritage in education and research
- Outreach to more diverse audiences, attraction of more visitors and increased exposure of collections
- Contribution to a stronger European identity.

### **Challenges facing the advancement of multilingualism**

To realise the identified opportunities, a number of challenges need to be addressed. The issues the sector faces in relation to multilingualism were identified as follows:

- Lack of understanding of the benefits of multilingual digital cultural heritage and the opportunities it brings to the sector and to society, causing a lack of unified multilingual/translation policy
- Lack of awareness and failure to share, disseminate and promote competences and knowledge in the sector, leading to a shortage of expert resources and training



- Lack of tools, technologies and digital resources that are readily adapted to digital cultural heritage and able to tackle the intricate nature and constant evolution of linguistic concepts related to our domain
- Lack of critical mass for applying machine learning to less common languages
- Wider issues in cultural heritage that also have an impact on handling multilingual issues: lack of quality (translations of) metadata/content, lack of interoperability/standardisation, institutions not aware of or not making use of existing tools, lack of R&D in future technologies e.g. AI.

### **Potential solutions to advance multilingualism**

It became clear that addressing these issues is a shared responsibility of the Member States' ministries of culture and cultural heritage institutions, aggregators and data providers, Europeana and the European Commission. Digital innovation hubs, the Europeana Network Association and the Europeana Aggregators' Forum, the EuropeanaTech Community, domain representatives, associations, ontology providers, developers and the DCHE, can all positively contribute to such advancement.

Solutions and actions identified include:

- Co-operation at European, national and local level among all parties involved, including content providers and collection managers
- Making more funding available to institutions for investment in the improvement of multilingualism
- Providing standards/frameworks for multilingual data cataloguing practices, crowdsourcing, curated translation, and mass translation for the cultural domain
- Supporting the development of more expertise in the sector
- Raising awareness and facilitating the transfer of existing tools, standards and frameworks, and R&D by Europeana and/or the language technology industry
- Raising awareness about the benefits of good quality content and metadata
- Improved ingestion of metadata from aggregators to Europeana, supported by relevant services such as automated data cleansing tools

### **Full output from the event**

Presentations<sup>15</sup>, images<sup>16</sup>, video<sup>17</sup> and the final report<sup>18</sup> from the event are available on Europeana Pro.

---

<sup>15</sup> <https://www.slideshare.net/Europeana/tag/finnish-presidency>

<sup>16</sup> <https://www.flickr.com/photos/europeanaimages2/albums/72157711667364238>

<sup>17</sup> <https://vimeo.com/372582901/dd1e668bc0>

<sup>18</sup> <https://pro.europeana.eu/post/benefits-challenges-and-solutions-for-multilingual-digital-cultural-heritage>

## Appendix C: History of multilingual investigations

Work to address multilinguality issues in the context of Europeana started as early as the EuropeanaConnect project<sup>19</sup> (2009-2011). The project showed the potential for Natural Language Processing (NLP) technology to be applied across languages, but the technological landscape was too fragmented, with too much effort to adapt and run software (as well as enabling language resources) for all languages. At the same time, the project developed the alternative strategic option of relying on a "semantic layer" of multilingual vocabularies<sup>20</sup> to act as mediator between users and original metadata, bringing context and multilinguality. This coincided with involvement with the part of the Semantic Web community<sup>21</sup> that was trying to facilitate the publication of multilingual resources as part of the then nascent Linked Data cloud.

This work on identifying issues and possible solutions to multilingual issues continued in the projects Europeana V1.0, V2.0, V3.0, which developed a stream of reports on Best Practices for Multilingual Access, culminating in the publication of a whitepaper<sup>22</sup> during the Europeana DS11 project (2016). As a parallel effort, Europeana partners, especially the Humboldt University in Berlin, ran evaluations for (query) translation in coordination with other EU projects, such as Galateas<sup>23</sup>, and in academic initiatives, like the CHiC (Cultural Heritage in CLEF) evaluation lab<sup>24</sup>.

In the meantime, Europeana implemented its first solutions for tackling multilingual issues with semantic resources. Especially, in 2011 Europeana began applying the automatic semantic enrichment process that is currently in service<sup>25</sup>. We also called for providers to contribute links to multilingual vocabularies such as the Getty AAT<sup>26</sup> (2014). The EuropeanaTech community organized discussions to evaluate and refine this approach in two task forces, one on Multilingual and semantic enrichment strategy<sup>27</sup> (2013-2014), and one on Evaluation of enrichments<sup>28</sup> (2015). The latter showed that Europeana's solution performed rather well considering the difficulties of enriching the kind of metadata it gathers.

At this time, Europeana re-started exploring NLP and automatic translation, using the Wikipedia Translation API to perform query translation<sup>29</sup> and Microsoft Bing to dynamically

---

<sup>19</sup> <https://www.europeanaconnect.eu/>

<sup>20</sup> <https://pro.europeana.eu/post/knowledgeinformation-in-context>

<sup>21</sup> <https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=12362>

<sup>22</sup> <https://pro.europeana.eu/post/best-practices-for-multilingual-access>

<sup>23</sup> <https://cordis.europa.eu/project/id/250430>

<sup>24</sup> <http://www.promise-noe.eu/chic-2013/home>

<sup>25</sup> <https://pro.europeana.eu/page/europeana-semantic-enrichment#automatic-semantic-enrichment>

<sup>26</sup> <https://pro.europeana.eu/page/europeana-aat>

<sup>27</sup> <https://pro.europeana.eu/project/multilingual-and-semantic-enrichment-strategy>

<sup>28</sup> <https://pro.europeana.eu/project/evaluation-and-enrichments>

<sup>29</sup> <https://journal.code4lib.org/articles/10285>

translate website pages - both were discontinued after a while as technical and cost barriers to using them were raised too high in the light of their perceived value. Discussions with the Language Technology community also took off, for instance in the context of the LT-innovate events. Europeana was suggested as a prime application case for language technology. However, in the light of the specific difficulties (resourcing, data sparseness) in cultural heritage, this community hinted that working with semantic enrichment and trusted multilingual vocabularies was the best Europeana could do at the time. We continued however to explore partnerships around language technology, for example in the context of the Riga summit on the Multilingual digital single market, where a Memorandum of Understanding on applying automatic translation was signed between Europeana and the Latvian ministry of culture<sup>30</sup>

Discussions on using the automatic translation service CEF.AT (later eTranslation DSI) began during the Europeana DSI-2 project. Perspectives were exchanged with experts in automatic translation and language technology at the Europeana Commission and beyond (for example Crosslang). In recent years, the technology has become much more mature, and eTranslation provides one API that services many languages at once. Applicability remains to be fully tested though, as cultural heritage metadata and requirements remain rather hard to handle. In Europeana DSI4, first experiments happened with automatic translation of virtual exhibitions, queries, and document transcriptions. Together with the 2019 Finnish presidency event on Multilingualism in digital cultural heritage, these experiments helped assess where automatic translation could benefit Europeana most, and contributed to the elaboration of the strategy presented here.

During all these years, Europeana has sought to make its work transparent, and tried to encourage all its community to tackle multilingual issues. Europeana Foundation and its partners have presented in many academic and professional external events about our multilingual challenges. Multilinguality has also been featured in many talks at Europeana and EuropeanaTech events. The very first issue of the EuropeanaTech Insight publication, in 2015, was precisely about multilinguality in our sector<sup>31</sup>.

---

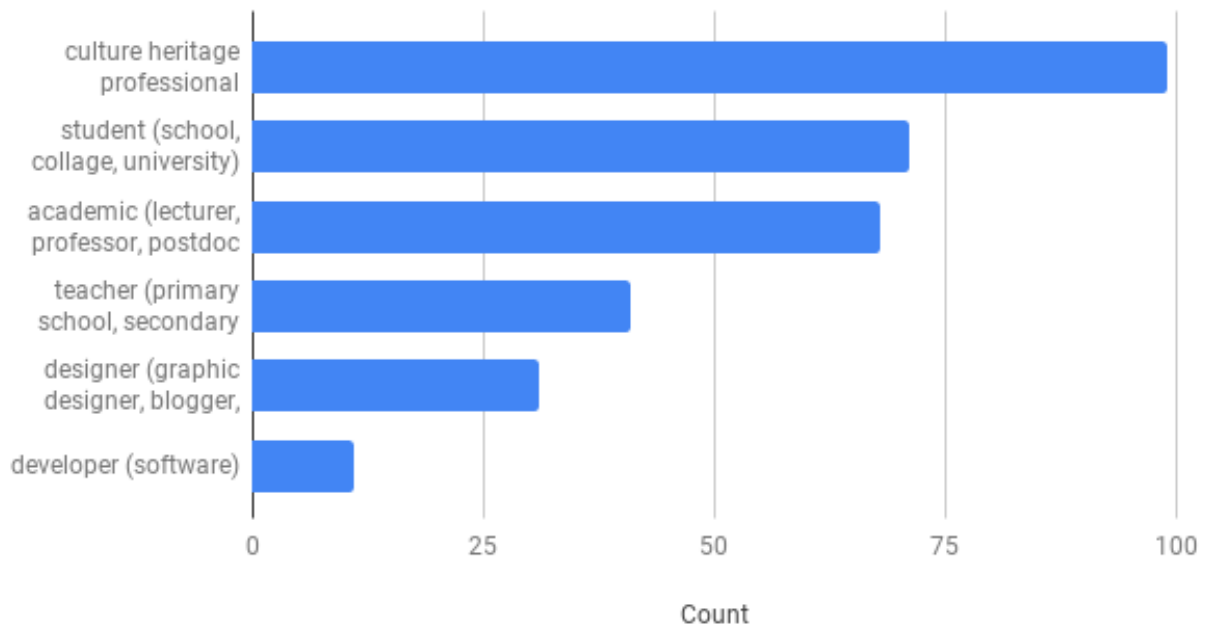
<sup>30</sup> <https://pro.europeana.eu/post/latvian-ministry-of-culture-and-europeana-sign-memorandum-of-und>

<sup>31</sup> <https://pro.europeana.eu/page/insight-issue1-multilinguality>

## Appendix D: User research results

User research results are from study of 309 visitors, undertaken on europeana.eu, in October 2019.

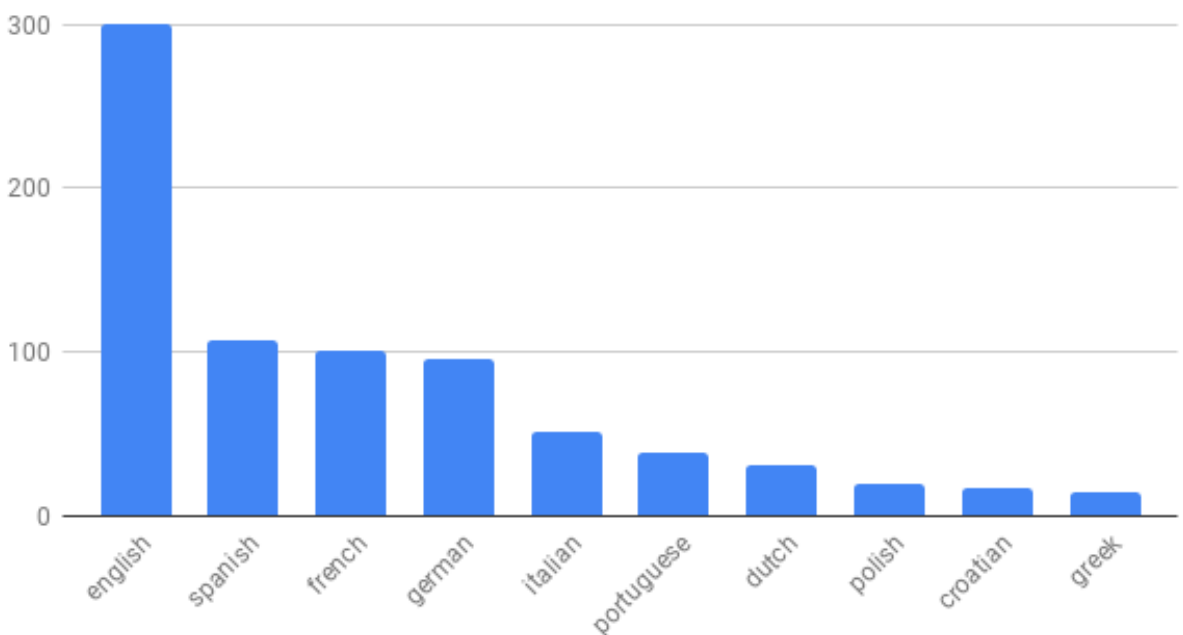
### How do you identify yourself?



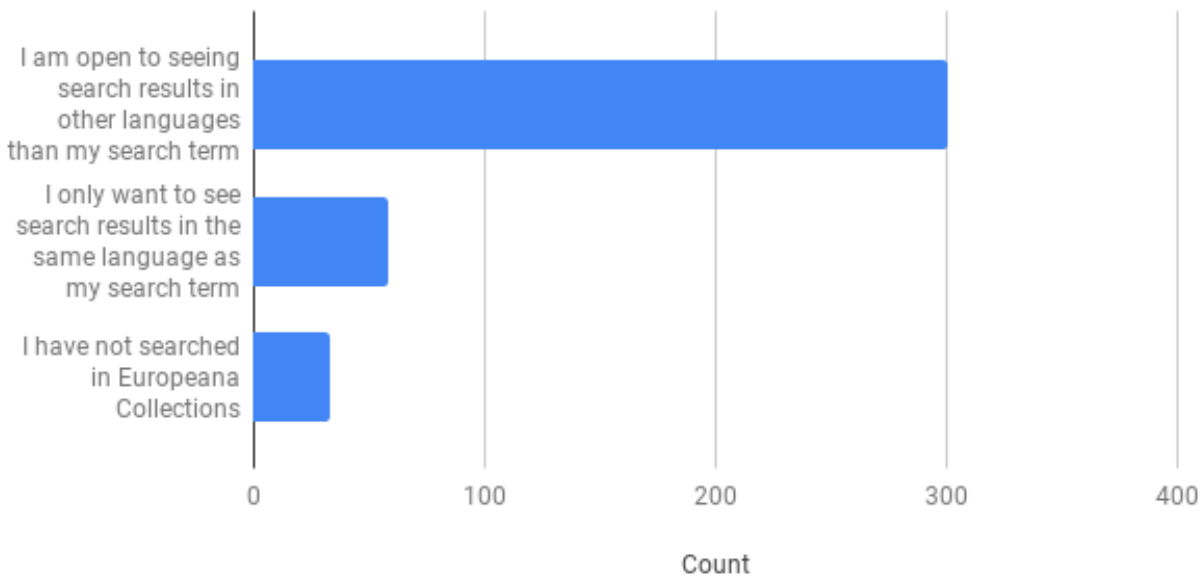
### Average number of languages used to search (per person)

4

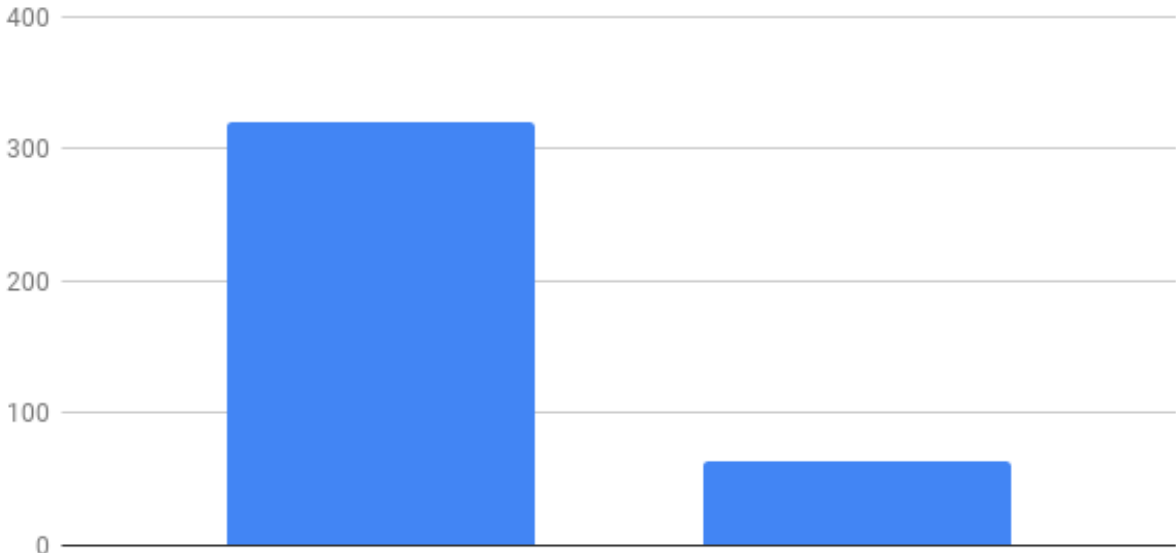
### Which language(s) do you primarily use to search?



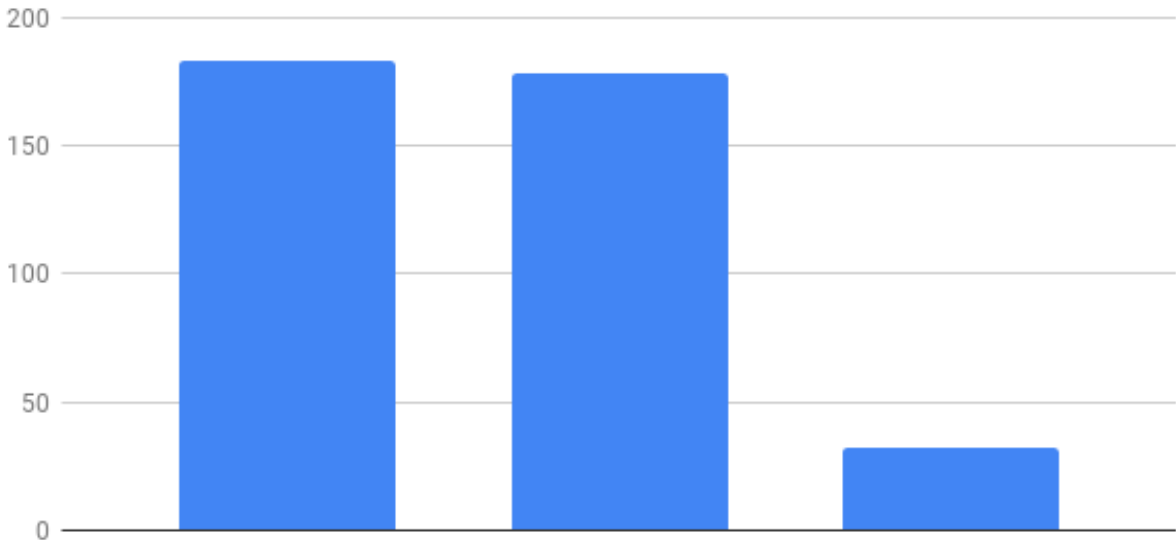
### How particular are you about the language of your search result?



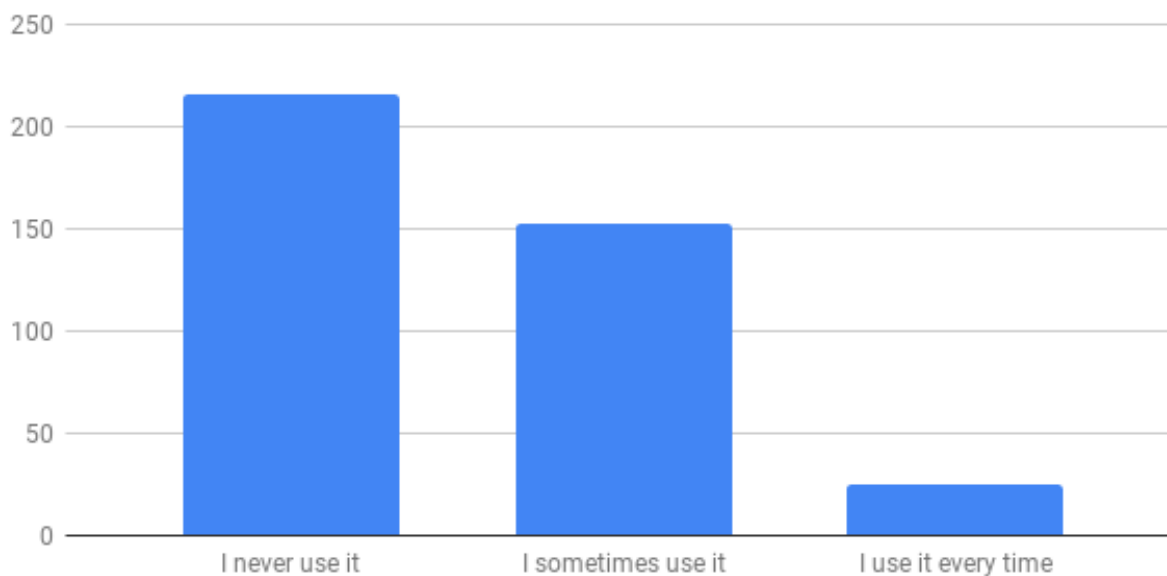
### How would you prefer to search?



### How frequently do you use the language selector in the top right corner?



## How frequently do you use the language filter on the left-hand side of the search results page?



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

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

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.

