Search improvement strategy

(MS2 Search Strategy M32)

Revision	1.5
Date of submission	30 April 2021
Author(s)	Mónica Marrero, Europeana Foundation Antoine Isaac, Europeana Foundation Hugo Manguinhas, Europeana Foundation Andy Neale, Europeana Foundation
Dissemination Level	Public

Revision History

Revision No.	Date	Author	Organisation	Description
0.1	21.01.2020	Mónica Marrero	Europeana Foundation	Draft
1.0	17.12.2020	Mónica Marrero, Antoine Isaac, Hugo Manguinas	Europeana Foundation	Refinements and first complete version
1.1	17.02.2021	Mónica Marrero, Antoine Isaac, Hugo Manguinas	Europeana Foundation	Added prioritisation and impact in section Metadata Search. Added improvement M-F18, refinements
1.2	30.04.2021	Mónica Marrero, Antoine Isaac, Andy Neale, Hugo Manguinas	Europeana Foundation	Complete revision following feedback from invited experts. Added section Implementation priorities and roadmap.
1.3	14.06.2021	Mónica Marrero, Antoine Isaac, Valentine Charles, Julia Schellenberg	Europeana Foundation	Updates based on EC feedback
1.4	25.06.2021	Julia Schellenberg	Europeana Foundation	Updates based on EC feedback
1.5	23.11.2021	Mónica Marrero	Europeana Foundation	Updates related to DSI4 Implementation Plan and changes in the Multilingual Strategy. Roadmap updated and added FT-NF4.

Table of Contents

Executive Summary

The purpose of this document is to provide direction for the further development of back-end search functionality at europeana.eu .
Search is a complex area so this document drills down through search components, focus areas, outcomes, roadmap priorities, and actions to help understand options.
Familiar search components that visitors experience are the search box, predefined filters and facets, autosuggest, search for similar items, and metadata linking.
Three primary areas of focus are suggested for this strategy, covering **metadata search**, **content search (fulltext)**, and **semantic search.**
The strategy also suggests three supporting focus areas covering **documentation and procedures, multilinguality**, and **evaluation**.
For each area of focus this strategy identifies **outcomes** that are used to help group the prioritised actions in the roadmap. There are 15 outcomes like “Metadata search is improved with better ranking/sorting”.
The **Roadmap section** shows outcomes and associated actions and splits those across three stages. Stage 1 actions are envisaged to be completed within the current Europeana DSI-4 contract (up until August 2022) while Stage 2 & 3 give further direction for the future development of search functionality.
The **Actions section** lists out the details of 54 possible actions[1] that could be taken to improve search and these actions are described to understand the technical context.
The improvements that are suggested in this document will be largely supported by experimentation efforts that will be undertaken, learnt from, and factored into the features that eventually roll out to production.
Search improvement planning is a long-term effort, thus this compilation has reprised ideas that have been laid out in search improvement reports from earlier DSI projects, and brings together many years of thought and research.

[1] The action section shows a total of 70 actions. From those 54 are possible future actions while 16 actions are already completed.

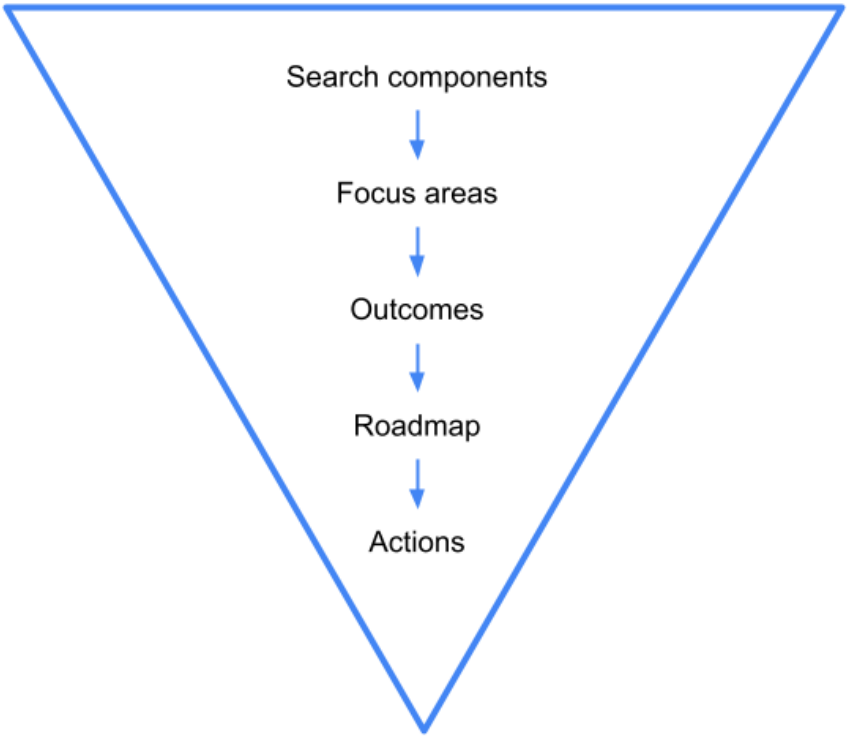
Introduction

Purpose of this document

The purpose of this document is to provide medium-term direction for the development of back-end search functionality at europeana.eu and its related services. This is ultimately to articulate the general lines of evolution, but also to provide practical guidance for the development of search services.

Document architecture

Search is a complex area, so this document has been designed to help the reader navigate from broad conceptual ideas down to deep technical actions, as per their interests. It is not expected that every reader will want to dive into the details, however the details are available for those with such a need.



Background

This strategy is the culmination of many previous research and planning efforts that are being brought together for the first time. Key points to be aware of when reviewing this information are:

- This strategy focuses on possible improvements that would have a direct impact on the search service offered by Europeana, either from the Europeana website or from the Search API.

- Note that front-end website features that provide ways for users to explore and discover the collections without entering any search term/query (such as exhibitions or browsing features on item pages) will not be addressed in this document. Neither will possible enhancements to the display of items or entities. Some are very related to search, as they use the Europeana Search API under the hood and could therefore raise search requirements that are in scope. However, in general this document is focused on the search and browsing back-end, as opposed to the front-end, which is developed and evaluated elsewhere.

- This strategy ultimately exists to inform implementation plans. A roadmap for Europeana search improvement is thus included in this strategy, similar to the one created for the multilingual strategy at Europeana[2].

- In order to draw a complete picture, included are actions that have been realised in the earlier stages of the Europeana DSI-4 project. Also, as search improvement planning is a long-term effort, our compilation has reprised ideas that have been laid out in search improvement reports from earlier DSI projects, but were not implemented.

- Community feedback on the strategy was sought and to this end the actions are provided with possible external collaborators that could participate in their accomplishment. Some improvement suggestions related to search and browse in the metadata collection have already been through a first internal review, and an estimation of the impact for end users was also included for them. The definitive priorities assigned as a result of this process can now be seen in the roadmap.

- In order to be able to trace all items, including those that are already *done*, *partially done*, *in progress*, or *scheduled*, a table is included in Appendix A. Actions status.

- Finally, to cover earlier actions not reported in previous DSI search improvement reports, the actions completed since the beginning of DSI-3 (Sep 2017) are included in Appendix B. Previous actions completed.

[2] https://docs.google.com/document/d/1pNyJom4WF9YJgq7p9A9GH8jViuFL0PdStFmaIFPEmqo/

Community feedback

This document was circulated among different experts in the field, with some of them already included as possible collaborators in some of the actions. Not all of them had enough time to conduct a full review, but we received positive feedback and interest for those who did.

We received 50 comments directly in the document, with some of those comments being threads with more than one reviewer discussing the same feature. Most of those comments were focused on:

- Multilinguality (38%)
- Metadata search (32%)

Content Search (Fulltext), *Semantic Search* and *Evaluation* also received comments, while the *Introduction* and the *Roadmap* section was the object of observations and clarifications.

The experts required clarifications (44% of the comments) for some aspects of the document that lead us to improve the readability and understanding of technical and not so technical concepts, particularly for the metadata and multilinguality related improvements. We also received observations and good pieces of advice for the implementation (40% of the comments), as well as clear explanations about the impact on users they may have (16% of the comments), which was taken into account in the roadmap.

For example, given the positive interest from reviewers about the creation of links between the collection and the editorial content (see M-F18. Editorial content in search and/or recommendation), we decided to prioritise it higher in the list and consider the editorial content not only for the regular search but also for recommendations. Also, given the discussion regarding optionality of keywords in search queries (see M-F11. Change default boolean operator from AND to OR in search), we made it clear that this is subject to validation, as the results may not be positive.

We would like to especially thank David Haskiya, Vivien Petras and Juliane Stiller for all the observations they submitted to us.

Conceptual solution

Search components

Earlier work on the Europeana DSI[3] led to the identification of various components to support users in their search and discovery activities. The search related functionality in the Europeana website is illustrated in Figures 1 and 2 below, and comprises the following components that translate directly into search features:

- Search box (the regular search of the collection)
- Predefined filters and facets (including curated filters to offer thematic collections, like for example the "World War I" or the Newspaper collections),
- Autosuggest and subsequent automatic (semantic) search with entities contained in the Europeana Entity Collection,
- Search for similar items and recommendations
- Metadata linking (which supports metadata-based browsing via item or entity pages).

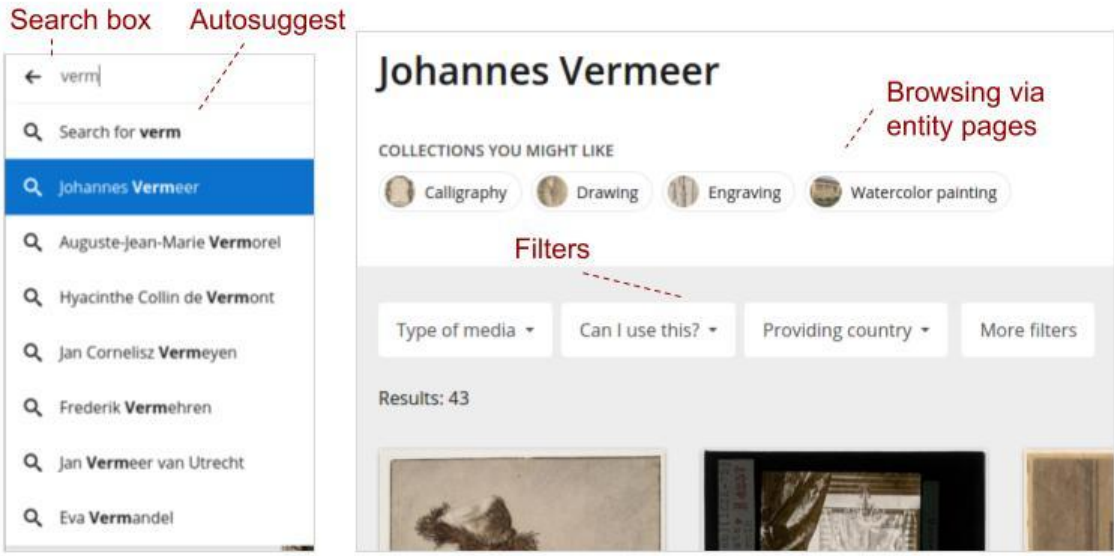


Figure 1. Search box and autosuggest components (left), browsing options in the search results page (right).

[3] *Key Performance Indicators of Search Quality for Europeana*, accessible at: https://docs.google.com/document/d/16TKUfpZVM7m3SXjgfPD1_9Z2QvScxrJ8MIpdGHbCgb4/

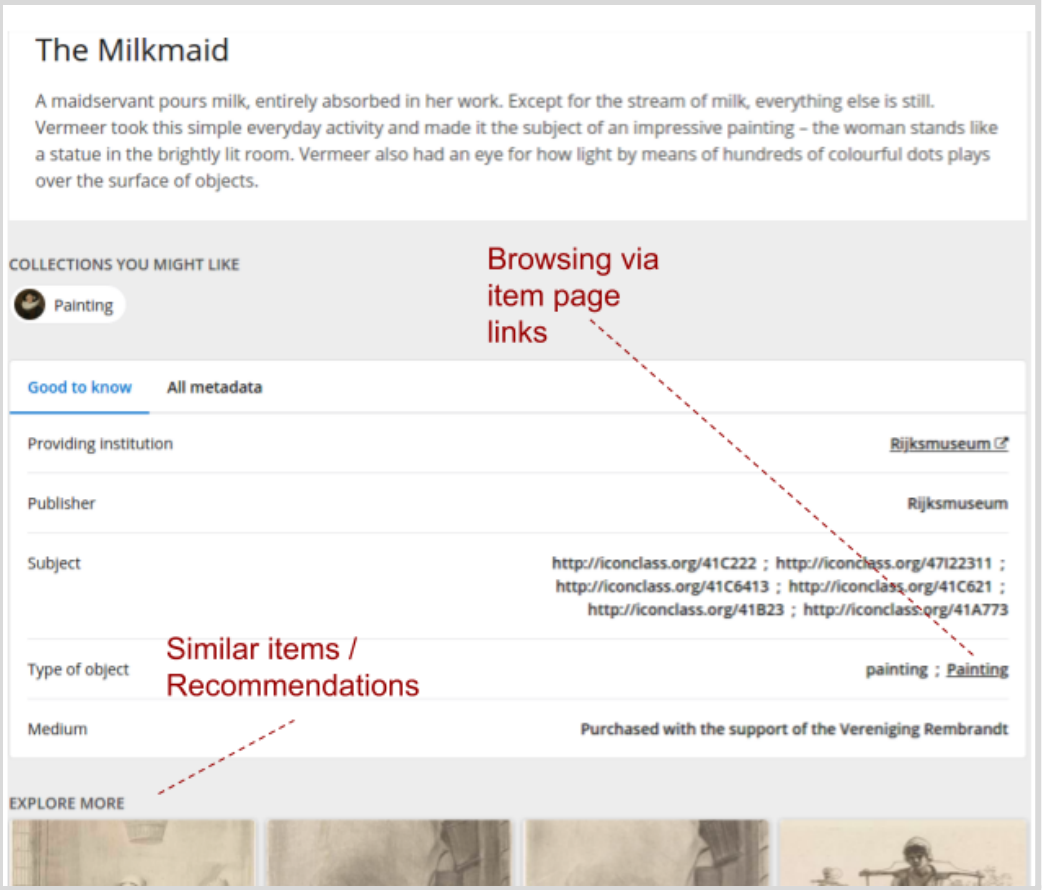


Figure 2. Item page containing links and similar / recommended items to allow users to continue exploring the collection.

Focus areas

The search components experienced by visitors are delivered by various underlying systems. The proposed actions in this strategy are therefore grouped into six main areas of focus, as seen in Figure 3.

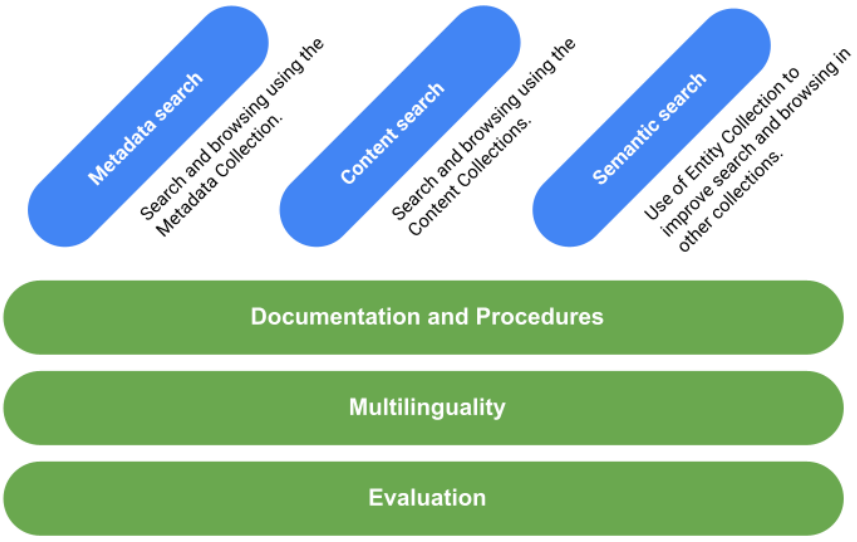


Figure 3. The actions are grouped according to these six categories.

The first three areas respond to the type of underlying data that is exploited, seen through different search features on the Europeana website:

1. **Metadata search**, based on the Metadata Collection built with the metadata aggregated from Europeana's Cultural Heritage objects. It has the code **M** for Metadata
2. **Content search (fulltext)**, based on Content Collections that gather the fulltext content that Europeana stored for some objects (newspapers, transcriptions). It has the code **FT** for Full Text
3. **Semantic search** based on European's Entity Collection, where data is gathered for a set of contextual entities that serves as reference for enriching object metadata. Is has the code **ES** for Semantic Search

The last three areas address relevant topics that are transversal to the first three:

4. **Documentation and procedures** that can help in the regular reporting and update of search. It has the code **D** for Documentation.
5. **Multilinguality** considers the actions already taken or planned to improve multilingual reach across all search features. It has the code **MUL** for Multilingual.
6. **Evaluation** of search in terms of effectiveness and efficiency, in order to make informed decisions to improve it. It has the code **EVAL** for evaluation.

The actions are further categorised into Functional **(F)**, for those dealing with the search functionality offered, and Non Functional **(NF)**, for those dealing mainly with the performance of the services implemented. For example, combined with the focus area codes, M-F stands for Metadata Functional, whilst M-NF stands for Metadata Non-Functional.

Outcomes

In order to deliver search improvements we have identified 15 outcomes (see figure 4) across the different focus areas (e.g. metadata search). The individual outcomes have associated actions which are described in this section below. The implementation priorities of the actions are available in the Roadmap section, while each action is fully explained in the Actions section.

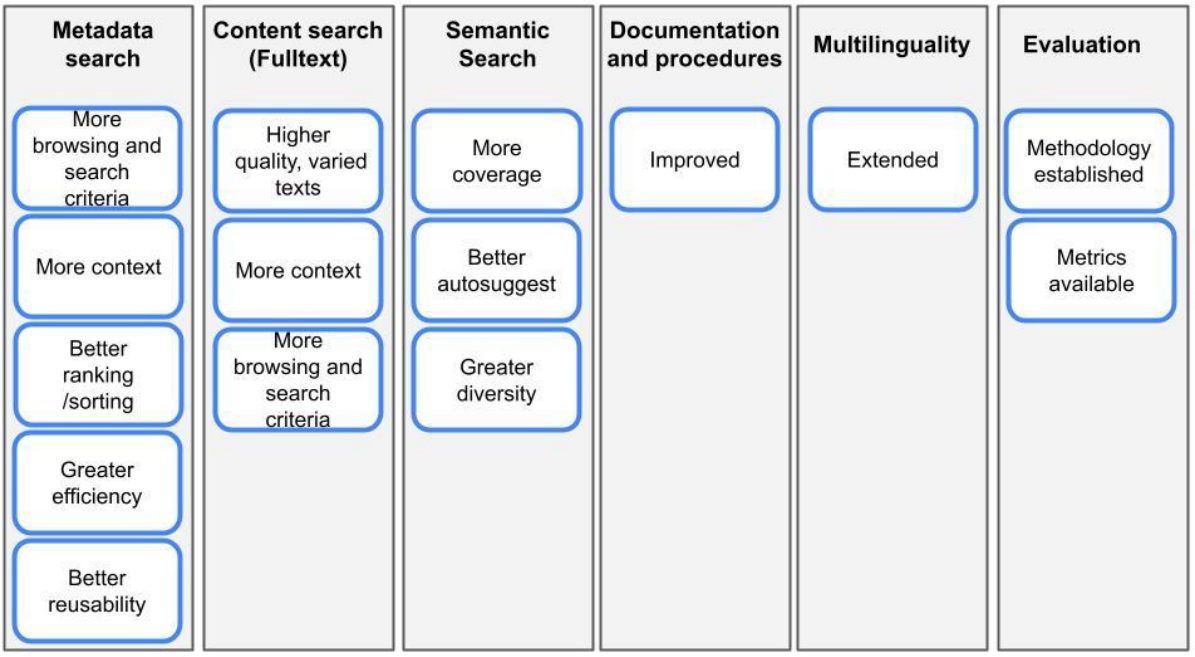


Figure 4. Expected outcomes associated with the different focus areas.

Metadata search is improved with:

- *More browsing and search criteria: new criteria are available for searching and browsing the Metadata Collection.*
  Actions: Normalized dates and georeferences of the cultural heritage objects are available for searching and browsing the collection (M-F2, M-F20). Users can browse (by faceting or filtering) along fields that are now only used for search (M-F3). Users can easily include fields and operators in the query to obtain more accurate results (M-F14).

- *More context: additional information helps contextualize the objects retrieved.*
  Actions: The search keywords are highlighted in the search results (M-F16). Not only the items in the metadata collection are displayed, but also editorial content related (M-F18). The recommendation of similar items to the one displayed to the user is improved (M-F12).

- *Better ranking/sorting of search results: the search results are sorted in a better way according to user expectations.*

Actions: Rank first items that are more usable and better understandable (M-F7), where the search keywords appear in more relevant fields (M-F8) or that are more popular according to the preferences of our users (M-F9) and/or external sources (M-F10). Test new technologies to account for contextual representations of the data in the ranking (M-F17). Offer the option to present results sorted by main fields like title or creator (M-F19). Include results which do not contain all the search keywords (M-F11). Automatically correct frequent errors in queries (M-F13). Contribute to a more effective discovery of the multiple digital representations that can be associated to each cultural heritage object (M-F1). Prevent errors from happening due to a bug related to the combined use of different functionalities associated with the ranking options (M-NF6).

- *Greater efficiency: search and browsing is faster.*
  Actions: Users see search results and browsing options faster (M-NF2). Staff can use the search engine as a source for data mining (M-NF5).

- *Better reusability: the lists of search results can be reused outside Europeana.*
  Actions: Search results can be exported in different formats (M-F15).

Content search (Fulltext) is improved with:

- *Higher quality, varied texts: more, higher quality full-text content is searchable.*
  Actions: Enable search on video and audio subtitles (FT-F5). Reduce errors in Newspapers contents derived from the use of optical character recognition (OCR) systems (FT-F3). Allow the ingestion of new and updated transcriptions live (FT-NF4).

- *More context: metadata is also considered in fulltext search (and the other way around).*
  Actions: Allow combined searches on different fulltext collections (FT-F1). Allow combined searches on metadata and (fulltext) content (FT-F7). Synchronize content and metadata in different collections (FT-NF3).

- *More browsing and search criteria: entities in the Entity Collection are also used to search and browse in the content collections.*
  Action: Named entities are identified in the textual content and links are created with the corresponding entities in the Entity Collection (FT-F4).

Semantic search is improved with:

- *More coverage: the number of entities in the Entity Collection is increased in order to better cover the data in our collections.*
  Actions: Add new entities and new types of entities to the Entity Collection (ES-F1). Expand and improve the internal enrichment process (ES-F9). Expand and exploit the external enrichment (ES-F10).

- *Better autosuggest: the functionality of the autosuggest component is improved.*
  Actions: Besides entities, autosuggestions also include selected metadata of the cultural heritage objects, such as titles (ES-F8). Autosuggest is clearer (ES-F5), more flexible (ES-F4) and the ranking of the suggestions is improved (ES-F6).

- *Greater diversity: the items retrieved when searching for an entity better represent the diversity of Europeana's collections.*
  Actions: Present more diverse results when searching by entities (ES-F3).

Documentation and procedures are:

- *Improved: changes and improvements in search are better documented, and the processes are improved.*
  Actions: Keep track of the status of the actions reported in this and subsequent documents (D-1). Collect and keep up to date the documentation related to search on a Wiki (D-2). Automatically update the curated list of items that should appear first for relevant/popular queries (D-5).

Multilinguality of search is:

- *Extended: items are easier to discover across languages.*
  Actions: Run preliminary experiments to assess the complexity of the task (MUL-F1). Normalize and translate the underlying data where necessary (MUL-F2). Assess the impact of machine learning language models in multilingual search (MUL-F6), and enable multilingual search in the Metadata and Content Collections (MUL-F3). Establish policies to guide and measure the improvements done (MUL-F5). Give support to a multilingual interface (MUL-F4), including work to make more clear the existing language filters in search (MUL-F7).

Evaluation has:

- *Methodology established: the methodology to evaluate the search performance is updated.*
  Actions: Establish a methodology for the evaluation of the search performance in different stages: implementation (EVAL-1), pre-production (EVAL-2), and production (EVAL-3). Implement the required cross-team organization and infrastructure (EVAL-4).

- *Metrics available: new metrics are provided to assess the performance in search.*
  Actions: Calculate metrics based on user behaviour to measure effectiveness (EVAL-6), usage (EVAL-7), efficiency (EVAL-8), coverage (EVAL-9), diversity (EVAL-10). Register and read the required user interactions (EVAL-5). Assess the use of other metrics not based on user behaviour (EVAL-11).

Roadmap

The roadmap information below defines the implementation priorities that will guide the Europeana Search Strategy as follows:.

- Stage 1: envisaged actions for the current Europeana DSI-4 contract (up until August 2022)
- Stage 2 & 3: envisaged for future implementation

Stage 1 actions are planned to be considered the upcoming B.1 Implementation plan M36 for Europeana DSI-4.[4] Ongoing, done, and partially done actions are highlighted in blue.

Outcomes	Stage 1	Stage 2	Stage 3
Metadata search			
Metadata search is improved with more browsing and search criteria	M-F2. Search by dates (data normalisation) M-F20. Search by location/georeference criteria (indexing implementation) M-F14. Re-activating advanced search	M-F2. Search by dates (indexing implementation) M-F20. Search by location/georeference criteria (data normalisation) M-F3. Extend faceting and filtering (subject to validation)	
Metadata search is improved with more context		M-F12. Item suggestions M-F16. Highlighting in metadata search results M-F18. Editorial content in search and/or recommendation	

[4] Note: the search strategy sets out a selection of possible envisaged actions. Final actions to be implemented will be added to the B.1 Implementation plan M36 based on recent priorities, needs and results of experimentation.

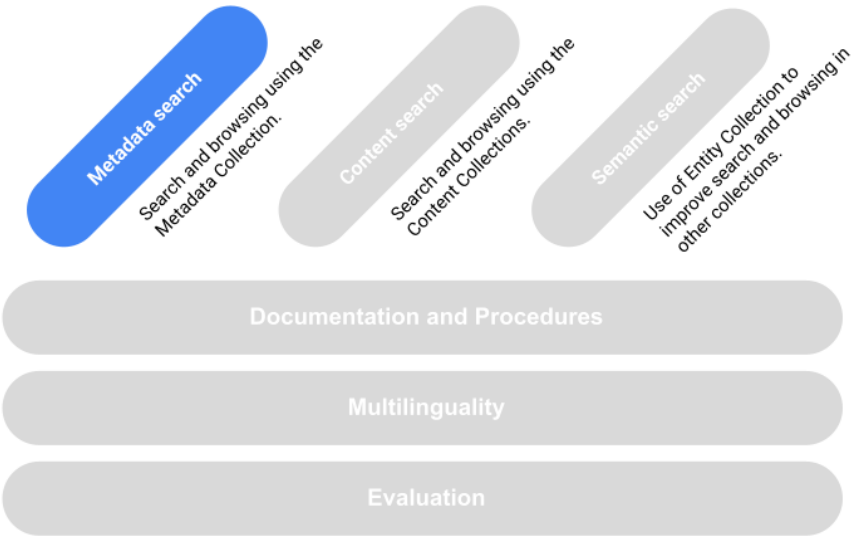
Metadata search is improved with better ranking/sorting	M-F7. Content and metadata quality is used in the ranking (subject to positive results) M-F19. Enable sorting using main fields	M-F8. Tuning the weight of different fields in ranking M-F10. Ranking by popularity (subject to positive results) M-F13. Query spelling correction M-F11. Change default boolean operator from AND to OR in search (subject to positive results)	M-F9. Learning to Rank based on user interactions (experiments) M-F17. Use of language models to improve retrieval (experiments) M-NF6. Prevent errors from elevation functionality and use of pagination (investigation) M-F1. Digital representations and hierarchies
Metadata search is improved with greater efficiency	M-NF2. Attribute docValues (also required for M-F19)		M-NF5. Streaming (use of search engine for data mining)
Metadata search is improved with better reusability			M-F15. Search results export
Content search (Fulltext)			
Content search (Fulltext) is improved with higher quality, varied texts	FT-NF4. New transcriptions are displayed and searchable live	FT-F5. Search on video and audio subtitles	FT-F3. OCR correction in Newspapers collection (experiment)
Content search (Fulltext) is improved with more context			FT-F1. Unify searches on newspapers and transcriptions FT-NF3. Content and metadata sync FT-F7. Search on metadata and (fulltext) content collections at the same time (experiments)

Content search (Fulltext) is improved with more browsing and search criteria			FT-F4. Named Entity Recognition applied to fulltext content
Semantic search			
Semantic search is improved with more coverage	ES-F1. Expanding coverage of search by entities (time and place if needed) ES-F10. Exploiting external enrichment	ES-F1. Expanding coverage of search by entities ES-F9. Improving internal enrichment	ES-F1. Expanding coverage of search by entities
Semantic search is improved with better autosuggest		ES-F5. Clearer autosuggestions	ES-F4. More flexible autosuggest ES-F6. Autosuggest ranking criteria ES-F8. Metadata and Entity Collection integrated in autosuggest
Semantic search is improved with greater diversity		ES-F3. Promote diversity when searching by entities	
Documentation and procedure			
Documentation and Procedures are improved	D-1. Document relevant changes in search	D-5. Elevation management D-2. R&D Wiki. Search section	
Multilinguality			
Multilinguality of search is extended	MUL-F2. Underlying multilingual data is established MUL-F3. Search Europeana	*These actions are scheduled in the Multilingual Strategy.*	MUL-F4. Read item text MUL-F5. Policy and plan established

	MUL-F1. Preliminary experiments with eTranslation	MUL-F6. Impact of BERT in multilinguality (experiments) MUL-F7. Clearer language filters	
Evaluation			
Evaluation methodology is established	EVAL-1. Evaluation methodology during implementation stage EVAL-2. Evaluation during pre-production stage	*These actions will begin in stage 1 and then have different aspects continue over the stages.*	EVAL-3. Evaluation after product in production EVAL-4. Cross-team effort organisation (required for EVAL-1)
Evaluation metrics are available	EVAL-5. Registering and reading user behaviour data	EVAL-6. Metrics to measure effectiveness EVAL-7. Metrics to measure usage EVAL-8. Metrics to measure efficiency EVAL-9. Metrics to measure coverage EVAL-10. Metrics to measure diversity	EVAL-11. Assess other (non-log based) modalities for use as metric sources

Actions

Metadata Search



Note: The metadata section additionally details purpose and impact for individual actions because the prioritisation of (the many) metadata search actions might be more difficult to understand otherwise.

M-F1. Digital representations and hierarchies

Purpose: a) all existing media is displayed in search results, b) media displayed match query submitted
Impact: Medium (mid-term) / high (long-term). Increase user satisfaction for a number of items to be measured (CHOs with more than one digital representation).
External collaborators: EuropeanaTech Data Quality Committee
(2021: work in progress)

As part of the activities in the Data Quality Committee, we have identified that part of the metadata of the different digital representations that could be associated with a CHO, is not exploited in the Europeana website. We have analysed the issues in detail and suggested several recommendations[5] to contribute to **a more efficient discovery of digital representations**. Users should be able to search by the metadata (at least type, rights and reusability) of the different digital representations instead of just the metadata associated with the CHO. For example, for a CHO with audio and image representations, a user looking for audio should be able to discover it and see it in search results, even if the image is considered the "main representation" of the CHO.

[5] *DQC: Web Resource metadata and discovery scenarios*, accessible at: https://docs.google.com/document/d/1EglM2yxK3gMQm5rZDZEskEDYomuznbMWim_RMmtjXeY/

We plan to analyse how these recommendations could be implemented from a search perspective, which could potentially lead to big changes in the schema where each digital representation is a first-class retrievable object. We are also having discussions with different teams, led by the UX team, about the modeling, display and the impact in search of specific use cases where there are other possible relations among CHOs (i.e. hierarchies), or among CHOs and entities (e.g. a Newspaper title could be modeled as an entity, connected to the corresponding Newspaper issues, modelled as CHOs).

M-F2. Search by dates

Purpose*:* Users can search/browse by specialised dates of the CHO and use ranges.
Impact: High. Increase user satisfaction (this is a frequent request).

As users suggested, review existing specialised date fields (for example the date of creation) to properly index them as *date* values. Users can then **filter and facet by *date* using ranges** (e.g. [2020-11-01 TO 2020-12-01], [1920 TO NOW])**,** in the same way they do for Newspapers with the field 'issued'. Note that in order to do this, we first need to normalise date references into a standardised format, which is not currently done, and ensure that data accuracy is appropriate. Currently only the year is indexed, and it is not possible to filter or facet using ranges. NB: this action would *not* necessarily conflate all kinds of date metadata for CHOs into one generalised date search.

M-F3. Extend faceting and filtering

Purpose: Users could facet by fields that are now only used for search (e.g. title, subject, contributor, creator, publisher, agent, place, concept), supporting browsing.
Impact: Medium. Increase browsing capabilities (although the number of values in a facet may be limited). The impact depends on the field: creator, type and subjects may have a big impact, as well as titles - if employed in a context where abstract works can be distinguished from their various "expressions"[6].

Define different fields for faceting and searching if needed. This would **avoid applying for faceting a normalisation applied for searching, which is not good for faceting**. For example, if we apply regular tokenisation, which is a normal process for text analysis, the facets will display the different tokens, instead of the whole values of the field (e.g. 'National Library' becomes 'National' and 'Library'). An alternative option to analyse is the use of the field type "SortableTextField", introduced in Solr 7.

M-F4. Improve textual normalisation (tokenisation)

Purpose: Search is more flexible (i.e. it does not require the punctuation marks to appear).
Impact: Medium. Evaluation with users (from EF) indicates it could be high.

[6] This would require enriching the metadata with work entities, a process sometimes called "FRBRisation" in reference to the FRBR model introducing the distinction between works, expressions, manifestations and items in the library domain.

(2021: work already done[7])

Improve the normalisation: in our current configuration for the Metadata Collection, only the trailing punctuation marks at the beginning and end of the default search field ('text') are removed, but not the punctuation marks in the middle of the contents (e.g. for the following title "[Funambulista] [Material gráfico]" we would need to search by "Funambulista]" instead of just "Funambulista" to find any results). These punctuation marks should be also taken into account to make a more flexible tokenisation. For example, the name "Pohl-Göns" can be splitted in two tokens instead of one, as currently happens, so it can be found with the queries "Pohl", "Göns" and any combination of those words, including "Pohl-Göns".

M-F5. Improve secondary sorting criteria

Purpose: a) same query displays same results, b) random ranking is displayed when user is just filtering the collection
Impact: Medium. a) Avoid confusing users, b) Increase diversity
(2021: work already done)

The ranking in which the records are displayed is affected by the ranking algorithm used but also by the default sorting criteria applied in the Search API, or the filtering criteria applied directly in the Europeana website. Regarding the sorting, we have recently changed the criteria adopted. Up until now, the sorting was done in the Search API considering first the presence of media, second the relevance score assigned by Solr, and finally the date of update and the metadata completeness. We have now included the identifier of the record (dc:identifier) as the last criteria to avoid slightly different order in the results when using Solr Cloud[8]. Additionally, we have also included a new field in Solr to allow random sorting of the search results, which would prevent displaying records sorted per dataset (because they all have the same date of update) when the user is just filtering the collection. The use of this new field would promote diversity, which is one of our objectives, and the Service Experience Team is currently working to see the best use of this feature.

M-F6. Content and metadata quality filters

External collaborators: Data Quality Committee
Purpose: CHOs with insufficient quality are hidden from users by default.
Impact: High. Increase satisfaction by promoting quality
(2021: work already done)

[7] The evaluation followed and the results obtained are available in the internal report *Solr reindex February 2021*- Description of changes in search, evaluation conducted and decisions adopted at: https://docs.google.com/document/d/1zL7OM6CXop0LCFIXVxxQGJ8geSKkR_27AdlT0qAovOc/
[8] See measure 6 in this table where the main changes affecting search have been described: https://docs.google.com/spreadsheets/d/1YM6mvIReSAwhOqUNwNXFl6IsM0_GY8-RIaeBogsxeb0/

Related to the filters, new filters reflecting the content and metadata quality tiers from the Europeana Publishing Framework[9] (EPF) have been automatically included by default when searching from the new Europeana website (released in March 2020). As a result, only records in content tier > 1 are displayed when searching in the thematic collections, or content tier > 0 when searching in the whole collection.

M-F7. Content and metadata quality is used in the ranking

Purpose: (1) Present first to user content that is more usable and better understandable. (2) Help fulfil the promise to our partners in the Europeana Publishing Framework ('the more you give, the more you get') in terms of visibility of higher-quality data (and consequently convince them to improve their data in)
Impact: Medium/high. It would be high in principle, but we already have a completeness measure, which helps even though incomplete. The big gap is content quality. The current ranking uses the presence of digital content as first criterion but this doesn't reflect EPF content tiers. NB: an update of the completeness measure could be useful to make higher-quality items better visible in web search engines (SEO).

There is an action, not yet planned, to replace the current completeness criteria by the use of the content and metadata tiers described in the EPF. This field is currently used as an additional sorting criteria in the search results.

M-F8. Tuning the weight of different fields in ranking

Purpose: a) CHOs ranked first in search results because the query terms appear in a more relevant field and/or in a shorter field (e.g. title instead of description), b) named entities are not transformed (e.g. *Alberts* is not transformed to *Albert,* or *Luis* to *Lui).*
Impact: Medium/high. Increase satisfaction users (to be determined by experimentation)

The previous ranking algorithm used, BM25f, was replaced in 2018 by the default Solr algorithm BM25, where the main difference is the lack of weights associated with specific fields by default. Currently the default search field is 'text' which contains a copy of the fields considered more important to search for. A big number of fields are copied there, so the same general text analyser for any language/type of data is used for all of them. An alternative would be using a virtual field pointing to the relevant fields, so the search is done field by field separately, and the score for each document results from a customisable combination of the scores obtained for individual fields. As a result:

- We take into account the length of the fields in the calculation of the score (e.g. a keyword in the title should be more relevant than the same keyword in the description).
- We can add weights to some of those fields. Those weights will be chosen following the procedures established to evaluate (see section Evaluation). **A first**

[9] https://pro.europeana.eu/post/publishing-framework

approach to the boosting has already been done and is ready to be deployed and tested in a testing environment.

This approach would imply changing the parser currently used by Solr (replacing Lucene parser by eDisMax parser) with no additional requirements. That would also solve the current problem of not taking the closeness of the search keywords in the documents into consideration in the rank.

Finally, the implementation of this action would also make it possible to apply different analysis to different fields, depending on their language or contents. This is relevant for MUL-F3. Search Europeana.

M-F9. Learning to Rank based on user interactions

Purpose: Ranking of CHOs in search results takes into account preferences shown previously by our users (e.g. manuscripts collection displays first medieval manuscripts, search by *dog* displays first media related to dogs, and not to a specific type of shell *Dog Whelk*).
Impact: High

The possibility of adding weights to the fields described in action M-F8. Tuning the weight of different fields in ranking can provide the basis to adopt a Learning To Rank schema that exploits user interactions to determine the optimal weights[10]. Those interactions are usually noisy and biased, so it is important to select the appropriate method to deal with these issues: counterfactual, where user models are created from historical data, and online, where the ranking is interactively optimised after every interaction. Depending on how noise and biased the interactions are, one method or the other is expected to have better results (see Jagerman et al: "To Model or to Intervene: A Comparison of Counterfactual and Online Learning to Rank from User Interactions"[11]).

M-F10. Ranking by popularity

Purpose: Ranking of CHOs in search results takes into account criteria based on popularity of those items among users or in external sources (e.g., most famous items in WikiArt).
Impact: Medium. Expected to be high on satisfaction of users where expert curation does not exist.
External collaborators: for a more scalable approach, possible collaboration with Prof. Paul Clough, Monica L. Paramita, Neil Ireson and Jie Jerry Gao (Univ. Sheffield).

[10] Previous Europeana (unfinished) work to re-instate a BM25F LTR framework is at https://europeana.atlassian.net/browse/RD-2. That module no longer works in the newer versions of Solr (it worked in the existing version by then, v.4.10.4), and does not use the LTR framework currently available.
[11] Available at: https://dl.acm.org/doi/10.1145/3331184.3331269

In order to improve the user experience, we have been using the elevation functionality (i.e. best bets) available in Solr for popular entity-based and well-known work title queries: given a predefined entity-based query, the results that should be displayed at the top are curated and contained in a file (elevate.xml). With the inclusion of new entities in the Entity Collection (see action ES-F1. Expanding coverage of search by entities and maybe ES-F9. Improving internal enrichment), it is possible that new queries should be included in this file. A more scalable solution should be explored, with the inclusion in the ranking algorithm of a popularity criteria that can be automatically calculated (e.g. based on number of views, the use of the items in our internal CMS, their inclusion in the user galleries, and/or their popularity based on external sources of information).
NB: this action just considers the inclusion of popularity criteria in the ranking, while M-F9. Learning to Rank based on user interactions is about the creation of a learning to rank model based on them (focused on the user interactions).

M-F11. Change default boolean operator from AND to OR in search

Purpose: Limit *no results found* message or search results with only a few items, so users have the opportunity to reformulate queries.
Impact: Medium. Increase user satisfaction , but it could well be the opposite (needs to be measured by experimentation).

In the current configuration, all the keywords that are part of the query are required in the retrieved documents. We can change the current configuration so not all of them are required and documents with consecutive terms are boosted (this approach is already in place for the Content Collections). Another option would be requiring only a percentage of keywords to appear. With these measures we would reduce the probability of not getting results after a search (i.e. higher recall). As a counterpart, we would lose absolute precision, which can be mitigated with the filtering functionality plus measures to boost relevant results at the top of result lists, like field boosting. In any case, an option to explore would be giving users the possibility to flag some of the input keywords as required if the results are not satisfactory (like Google does).

M-F12. Item suggestions

Purpose: For each item clicked, similar items are displayed based on metadata or content but also collaborative filtering (i.e. other users also viewed...).
Impact: Medium/high. Promote diversity and improve navigation.

The first iteration of item suggestions is expected to be based on image similarity of the content, as opposed to metadata. As part of the 'Europeana XX'[12] activities, we also have included a recommender system in the Europeana platform that can be considered for future iterations. This component will be used by our Editorial Team users to recommend them new items to be included in the galleries they created, as well as having the possibility of future item suggestions. Currently this system only

[12] https://pro.europeana.eu/project/europeana-xx

recommends based on the metadata similarity, instead of taking into account the interactions or profile of the user. This is already a functionality implemented in Europeana for the Metadata Collection: the similar items search. There is a pending activity to fix a bug[13] in this component, however, it could be completely replaced by the recommendation system, which is expected to be more effective given the technology behind it. For this we need to get buy-in from the design team before. Additionally, clear objectives for the recommendations should be established with them (and transparent for the users, which is one of the suggestions we received), and additional ways of recommending similar items could be also explored[14], for example searching for similar audio or images (for which we already have prototypes). Finally, it could also be interesting exploring the recommendation of items when no search results were found.

M-F13. Query spelling correction

Purpose: Frequent errors in the queries are automatically corrected, avoiding unexpected results for the user.
Impact: Medium. Increase user satisfaction (to be measured by the Service Experience team).

Previous work[15] at Europeana contemplated the use of a misspelling functionality, where the user would be suggested with the query automatically corrected (e.g. *Celso Constatini* would trigger the suggestion *Celso Costantini*). There are several technologies that might be used, notably machine learning models for natural language processing like BERT.

M-F14. Re-activating advanced search

Purpose: Users can issue more specific queries directly from the Europeana website (e.g. search by author without including description, contributors, publishers, etc.), and use advanced features to increase precision (e.g. phrase queries, where keywords must appear in sequence).
Impact: Medium. Increase user satisfaction (to be measured by the Service Experience team).

An additional action suggested by users as part of the work on a Evaluation Framework included in the Search Improvement Plan in 2017 is having a more accessible/visible advanced search. This could be achieved with a link to a form with the main fields, or just by informing the users with a help page or FAQ how they can use specific fields in the search box and how to use logical operators to combine them (we already have something similar for the Search API, but not for the Europeana website).

---

[13] See bug in https://europeana.atlassian.net/browse/RD-19 , pp.2 (Section "Variations in Metadata").
[14] Extending the Master Thesis by Karl Pineau: "The recommendation of cultural heritage objects in Europeana Collections":
https://docs.google.com/document/d/1AdmCvUsVdoGYiHBL6FUk3TtPd21lCXnJfvRAcAuHKl4/
[15] Ceccarelli, D., Sergiu Gordea, Claudio Lucchese, Franco Maria Nardini, Gabriele Tolomei (2011) "Improving Europeana search experience using query logs.

This option should be weighed against other UX priorities, especially the changes that promote browsing over search (basic as well as more advanced) as a key enabler for discovering objects.

M-F15. Search results export

Purpose: Users can export the search results (metadata) from the Europeana website. NB: this action is not directly related to the search functionality but still could have an impact on satisfaction.
Impact: Low. Increase user satisfaction (to be measured by the Service Experience team with the Europeana Research community).

The preliminary results of the Task Force on Researchers' Requirements[16], and the discussions about the Europeana case in the webinar 'Exploring Open Access Images Resources'[17] by the Frick Collection, shows that researchers are also interested in downloading the results of a search. Maybe a good option for this is making visible the query done by the Search API. This way, users could run the same queries directly from the Search API and apply custom processing to the search results. However, we have to bear in mind that social sciences and humanities researchers may not use the API, so we could find a way to allow them to export the metadata of the selected items retrieved in the appropriate formats so they can introduce that information directly in their databases and/or reference management tools.
Note that this functionality may be implemented "indirectly" via an export of search results as a user set or a gallery, where the data could be downloaded from.

M-F16. Highlighting in metadata search results

Purpose: a) users can assess faster if the document is relevant for their needs, b) users and staff know why a document was retrieved, promoting transparency
Impact: High. Increase user satisfaction (to be measured by Service Experience team)

The keywords in the query matched in the retrieved documents can be highlighted and offered to the front-end, so the user can quickly assess the relevance of a document.

M-F17. Use of language models to improve retrieval

External collaborators: 'Europeana XX' partners.
Purpose: Feeding search engine with more contextual information may improve search results
Impact: Medium. Increase user satisfaction (to be measured by experimentation)

In the context of the project 'Europeana XX', our partners will use a state-of-the-art language model (BERT[18]) for the recommendation system that will facilitate the creation

[16] https://pro.europeana.eu/project/research-requirements
[17] https://www.frick.org/interact/exploring_open_access_image_resources
[18] Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, pages 4171–4186.

of galleries with related content. This model, created using machine learning techniques, can be used to improve the search results by using contextual representations of the data in the CHO in the form of vectors. BERT was already used in the project 'Culture Chatbot'[19], and our partners already created a prototype for Information Retrieval using our *World War 1* collection. The discussions to integrate this model into our platform (which may also reduce language barriers, see MUL-F6. Impact of BERT in multilinguality) are paused for the moment, as well as the plans to run an A/B test to evaluate its impact on our users.

M-F18. Editorial content in search and/or recommendation

Purpose: Not only the metadata that matches a query from a user is displayed, but also editorial content related (e.g. blogs and galleries).
Impact: Medium

As part of a workshop organised by the Service Experience team, one of our colleagues indicated that there should be more links between the data in our website, and suggested the possibility of displaying editorial content as part of the search and/or recommended items. This has to be assessed first by the Service Experience team, but from a technical perspective, an option could be to set up a search engine for the editorial content (and a parallel indexing process), and simultaneously get the editorial content that matches the query launched by the user in the Europeana website, or display that content as part of the related content.

M-F19. Enable sorting using main fields

Purpose: Allow sorting of search results using fields like title or creator.
Impact: High

Most fields are now configured as multivalued fields, and that prevents using them for sorting the search results. That is the case for example for the field *title* or *creator.* The implementation of the action M-NF2. Attribute docValues is expected to solve this issue.

M-F20. Search by location/georeference criteria

Purpose: enable spatial search
Impact: High

Presently, the latitude and longitude coordinates for georeferences are indexed as separate fields which makes it difficult to accurately search using coordinates or bounding boxes, especially if more than one location is indicated in the metadata. In the scope of the new 'Jewish History Tours' project[20], we will be reviewing the geo information present in the metadata and its indexing on the specialised fields using Solr GIS support so that spatial search can be fully offered.

[19] https://pro.europeana.eu/post/the-culture-chatbot-engaging-visitors-with-your-collections
[20] https://pro.europeana.eu/project/jewish-history-tours

At the end of this action (and based on the lessons learnt from its implementation) we will seek to assess whether place entities could play a bigger role in next versions of spatial search, possibly in combination with time information[21]. This could lead to a follow-up action in the Semantic search focus area.

M-NF1. Solr upgrade for Metadata Collection

Purpose: Users see results faster, new features are available for staff to be exploited
Impact: Low (for external users). Increase user satisfaction, staff can use new features to make improvements, Solr is better supported
(2021: work already done)

Additionally, we have also made an assessment[22] to update Solr from the current Solr v6.6.2 to Solr v7.7, which could lead to improvements in the schema used (where the data is defined). As a result, we could replace Trie* fields, already deprecated in v6.6.2, with *PointFields, which are **more performant**.

M-NF2. Attribute docValues

Purpose: Users see results faster, especially when faceting, and can sort by any field (e.g. currently we can not sort by *title, subject* or *publisher*), c) staff can use Solr for data mining
Impact: Medium. Increase user satisfaction, staff can run faster calculations over data

Activate attribute 'docValues' for most of the fields. This action will increase the size of the index, but it will also **improve efficiency when faceting and sorting**, and moreover, it will allow **sorting by multivalued fields** (see M-F19. Enable sorting using main fields**).** This attribute is also required to export data from Solr in streams (see action M-NF5. Streaming (use of search engine for data mining)), which would be **useful for data mining** operations (it would already be **useful for the broken link detection** that the Aggregation Systems team is currently implementing).

M-NF3. Improving performance filtering by Content Tier

Purpose: Users see results faster when filtering by content tier (e.g. thematic collections)
Impact: Medium. Increase user satisfaction
(2021: work already done)

[21] In 2017 we already worked in a pilot to display the objects in our collection in maps and timelines, allowing users easily accessing them based on geoposition and time criteria. Cf Europeana DSI-2 D6.4 Pilot for Time and Place Discovery at https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.4-pilot-for-time-and-place-discovery.pdf
[22] See assessment done: https://docs.google.com/document/d/1LHz9NbkL1VEF9aVvWao317-_eUczmW6GPHR82Wvpk_0/

Measures have been taken[23] in the Solr configuration to improve the response time for the filtering by content tiers (see action M-F6. Content and metadata quality filters).

M-NF4. Monitoring Metadata Collection

Purpose: Staff can detect possible issues (e.g. attacks, lack of resources) and take solutions before the problem becomes bigger
Impact: Low (for external users)
(2021: work already done)

Together with the Platform Engineer, we have already put in place the monitoring of the Metadata Collection to detect possible issues. The parameters monitored are:

- Memory consumed (Java heap).
- Number of queries
- Time per query
- Cache performance

Additionally, an alarm is triggered when the memory consumption is above a predefined threshold.

M-NF5. Streaming (use of search engine for data mining)

Purpose: Staff can use Solr as a source for data mining
Impact: Low (no impact for external users). Faster calculations over data are available to other services (e.g. detection of broken links by METIS).

It is possible in Solr to launch queries in stream mode, so a big amount of records are retrieved at once without incurring memory issues. This would allow a quick implementation of efficient data mining processes. This is technically possible by configuring a proper export handler[24], and using Streaming Expressions[25], where the queries have to meet strong conditions:

- Sorting field (at least one), and list of fields to be exported ("fl") are mandatory. Only basic field types are accepted (int, long, float, double, string, boolean and date). For the sorting, only fields with single values are accepted.
- Only fields with attribute "docValues" activated can be exported (see action M-NF2. Attribute docValues).

[23] See measures 3 and 4 in 'Relevant Changes in Search':
https://docs.google.com/spreadsheets/d/1YM6mvIReSAwhOqUNwNXFl6IsM0_GY8-RIaeBogsxeb0/
[24] https://lucene.apache.org/solr/guide/6_6/exporting-result-sets.html#exporting-result-sets
[25] https://lucene.apache.org/solr/guide/6_6/streaming-expressions.html#streaming-expressions

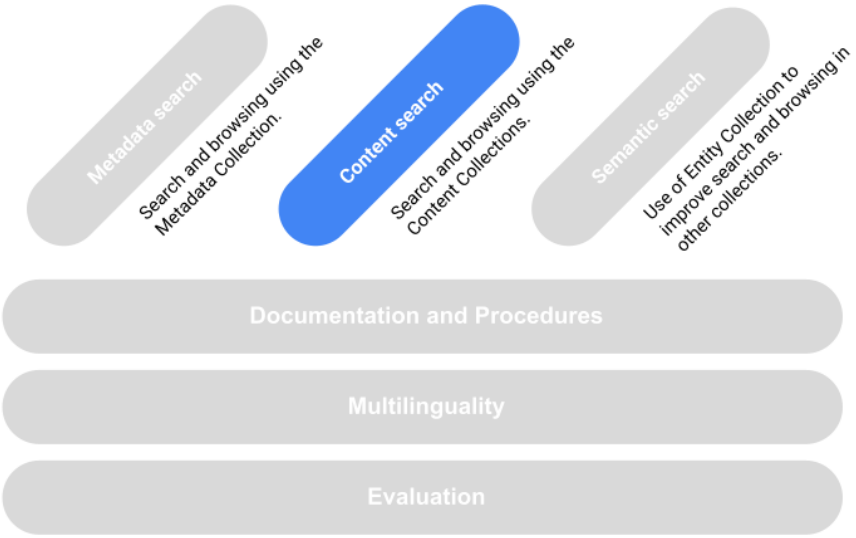
M-NF6. Prevent errors from elevation functionality and use of pagination

Purpose: User do not see an error when launching a query that includes that specific functionality
Impact: Medium
(2021: work in progress)

Elevation does not work when using advance pagination[26], therefore the elevation functionality had to be disabled in that case to prevent errors. We have reported this issue internally (https://europeana.atlassian.net/browse/EA-1364), as well as publicly through Apache Solr JIRA (https://issues.apache.org/jira/browse/SOLR-11921). It has been already fixed in Solr and will be included in Solr v9.

26 https://lucene.apache.org/solr/guide/6_6/pagination-of-results.html#fetching-a-large-number-of-sorted-results-cursors

Content Search (Fulltext)



The Content Collections (fulltext) contain only the Newspapers collection for the moment, where the fulltext content can be searched by users (as opposed to only the metadata). It was created at the end of 2018 with the newspapers dataset, containing not only the metadata but also the fulltext obtained by OCR of more than 800K newspaper issues. We have recently included a new collection, Transcriptions, which it is obtained by manually transcribing the text contained in images that are part of our collection (for the moment obtained in the context of the project 'Enrich Europeana'[27]). This new collection is currently being tested.

FT-F1. Unify searches on newspapers and transcriptions

(2021: work partially completed)

Both collections, Newspapers and Transcriptions, will share in Solr the same schema and will be hosted in the same Solr instance (different from the one that hosts the Metadata Collection), so the same query can be issued in both collections at the same time, and the results can be merged. In order to make this happen, we have adapted the schema used for the Newspapers index, so it can be also used for Transcriptions. This schema is published on Github[28].

FT-F2. Highlighting in content search results

(2021: work already done)

[27] https://pro.europeana.eu/project/enrich-europeana
[28] https://github.com/europeana/search/tree/master/solr_confs/fulltext/conf

We are also currently working on making fully functional the highlighting for the newspapers' contents in the Europeana website. This is already implemented and functional in Solr, but there is still a gap between the highlighting in the text and the highlighting in the image containing that text that is displayed to the user. In order to solve this gap, we are working on a solution that combines Solr and a Mongo database containing the coordinates. An alternative option to explore is storing additional information in Solr (payload) that links directly or indirectly each lexical token indexed in Solr to its corresponding coordinates in the image, or at least, to the specific page.

FT-F3. OCR correction in Newspapers collection

External collaborators: Clemens Neudecker (Berlin State Library).

Multiple improvements[29] can be explored on the Newspaper data to enhance the search functionality, one of them being dealing with the errors introduced by the OCR (e.g. adopting fuzzy search, more error-tolerant language analysers) in order to reduce noise.
Some progress has already been done on OCR correction in our community: Clemens Neudecker (Berlin State Library), partner in the Europeana Newspapers Project, is applying automatic correction of OCR output.

FT-F4. Named Entity Recognition applied to fulltext content

External collaborators: Clemens Neudecker (Berlin State Library).

Applying named entity recognition techniques, and specific analysers to the named entities contained in the fulltext to reduce errors and therefore improve the precision (i.e. applying stemmers to named entities brings noise). Some progress has already been done on Named Entity Recognition in our community: Clemens Neudecker (Berlin State Library), partner in the Europeana Newspapers Project, is applying Named Entity Recognition[30] [31], and Entity Linking to Wikidata resources for historical documents, including newspapers. A future collaboration to apply those processes in our Newspapers Collection would be highly desirable.

FT-F5. Search on video and audio subtitles

Europeana Foundation has received subtitles from partners in the Europeana Media[32] project and made them available to be displayed in the new media player. We are working in the 'Europeana XX' project to continue that work on enriching our collection with fulltext information associated with the media, now through efforts in crowdsourcing. In the future, we would like these to also be searchable together with other fulltext content (see FT-F1. Unify searches on newspapers and transcriptions).

[29] See the ideas at https://docs.google.com/document/d/1qXcBhv43JGw5g4Ky9bVLyrP3zkkUY7R41zsKFs2CNSY/
[30] https://lab.kb.nl/dataset/europeana-newspapers-ner
[31] https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_4.pdf
[32] https://pro.europeana.eu/project/europeana-media

FT-F6. Search on content within a single (IIIF) Item

(2021: work already done)

An additional task that could be worth exploring is the implementation of search on IIIF (International Image Interoperability Framework) resources. We are currently using this framework to display the information contained in the Newspapers Collection. The next step will be the implementation of the IIIF Search API[33].

FT-F7. Search on metadata and (fulltext) content collections at the same time

Currently at Europeana it is possible to search in the metadata collection or in the (fulltext) content collections separately but we do not offer the option to automatically search in both collections at the same time. This feature would allow the user to search in any collection in Europeana (possibly also in the editorial content, see M-F18. Editorial content in search and/or recommendation), and display the results combined or grouped by collection. An analysis of a possible integration of the metadata and the fulltext content of the Newspapers collection was done in the past from a technical point of view[34]. This functionality is subject to validation by the Service Experience team.

FT-NF1. Solr upgrade for Content Collections

(2021: work already done)

We have already upgraded the Solr instance containing the Content Collections from v6.6.2 to v7.7, in order to be in sync with the Solr Metadata Collection, useful from a software development and maintenance perspective, in preparation for its update as part of action M-NF1. Solr upgrade for Metadata Collection.

FT-NF2. Monitoring Content Collections

(2021: work already done)

Similarly to the monitoring of the Metadata Collection (see action M-NF4. Monitoring Metadata Collection), in order to be informed and assess possible performance issues in the Content Collections, we need to set up the monitoring of this collection too.

[33] https://iiif.io/api/search/1.0/
[34] *Newspapers storage: overview and notes*, accessible at: https://docs.google.com/document/d/1C3cOD80sMhY5ZbPjs1frQukfBvB_pwAYZ5osSDES7Zo/

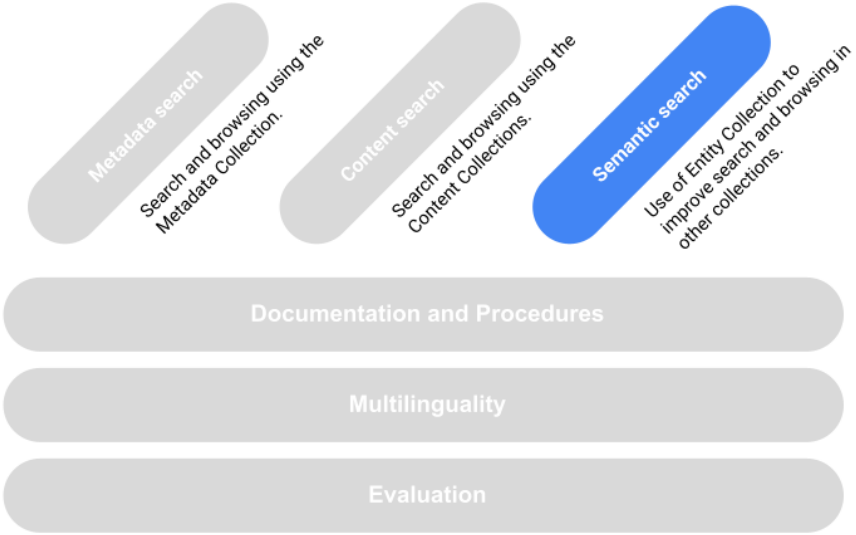
FT-NF3. Content and metadata sync

We have already developed new services[35] in order to update and synchronise our Content Collections with the data in the Metadata Collection, and we have recently synchronised the newspapers data, which has been currently tested successfully. Synchronisation processes should be run periodically in order to keep up to date the metadata contained in the Content Collections with the Metadata Collection. In order to support this more effectively, we need first to harmonise the modules for the indexing of metadata, used by our Ingestion Team, and the ones used for the indexing of the full text content. Ultimately, the solution adopted will depend on the implementation of FT-F7. Search on metadata and (fulltext) content collections at the same time.

FT-NF4. New transcriptions are displayed and searchable live

Although there is already a collection with transcriptions that were delivered by the Europeana Enrich project which are available for search, the software required to update and add new transcriptions needs to be refactored to streamline the workflow from data submission up to storage and indexing. The new pipeline will now include the fulltext MongoDB storage which will make it possible for transcriptions to be accessible via IIIF manifests and this way be displayed in the IIIF viewer on the Europeana website. Besides the functional impact, the pipeline will better isolate the steps and will be flexible to accommodate new use cases such as indexing of subtitles (and newspapers) and ultimately make the process faster.

Semantic Search



[35] https://github.com/europeana/search-tools

The semantic search in Europeana allows us to search by entities, which are collected from different linked open data repositories. The index containing this collection is in a different Solr instance, in this case version 7.7.2. The Entity API implements the autosuggest functionality, consisting in suggesting an appropriate entity after several keystrokes while the user is typing the query to search in the Metadata Collection.

Last year, the Europeana Foundation invested big efforts in developing a new website. One of the main changes is the higher focus in the search and browsing based on entities. As a consequence, several tasks have been done or are underway.

ES-F1. Expanding coverage of search by entities

In previous evaluations with sampling[36], approximately between 30% and 50% of the entity-based queries were about an entity already existing in the Entity Collection, so we will work to increase the number of entities. On the other hand, approximately half of the records in the Metadata Collection are enriched with entities, meaning that only those can be searched and browsed using entities. We will work to make sure that at least the entities provided by our partners are also covered in the Entity collection, especially using co-reference links that enable us to link more objects to entities from the Entity Collection, when these objects have been enriched by providers (which will also contribute to expanding the multilingual coverage, see section Multilinguality). Additionally, we will assess the inclusion of new entities referred to in properties into the collection (e.g. profession/ occupation of agents) which are expected to create more relations among CHOs, thus improving the browsing experience of the user. Features like integration of metadata in auto-suggestion (ES-F8) and multilingual search (MUL-F3) could also benefit from the availability of historical (named) events and work entities in the Entity Collection. Corresponding requirements should be directed towards (and evaluated) as part of the Entity Collection curation plan[37].

ES-F2. Search by entity

(2021: work already done)

When a suggested entity has been selected by the user, an automatic query is launched on the Metadata Collection. Previously, that query contained the preferred label of the entity. After a recent update the search is more precise and requires that the records also contain the URI of that entity. The URIs should not be included in the query itself, but in the filters, so the elevation functionality can be triggered (the queries expected to trigger the elevation functionality are only the labels of the entities, not their URIs).

[36] See experiment and results in DSI-2 D6.3 Search Improvement Report, pp. 8-9, accessible at: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.3-search-improvement-report.pdf

[37] The live version of this plan is available at https://docs.google.com/spreadsheets/d/1e3_UNpHLqD6fMbYkyckwIqnn_VQuA_oaDhk1pcKYYvQ/ and is based on this document https://docs.google.com/document/d/1A5Rb3Oe9edin5gdRpqFILlR0YPUodVOel3SdcBP00dA/

ES-F3. Promote diversity when searching by entities

We can promote diversity not only in the regular search results, but also in the items retrieved related to a specific entity selected by the user. Besides random sorting, (see M-F5. Improve secondary sorting criteria) another option could be applying (Solr) grouping (e.g. by Country) so only a representative item per category is shown. We are also currently assessing an approach where we offer power users the ability to manually curate the first page of results for entities which will make intensive use of the elevation functionality (i.e. best bets) for this type of queries.

ES-F4. More flexible autosuggest

We need to add more flexibility to the autosuggest functionality by supporting misspellings and different order of the words.

ES-F5. Clearer autosuggestions

As already included in a previous report[38], we can make it easier for the user to choose among different suggestions by indicating the type of the entities suggested.

ES-F6. Autosuggest ranking criteria

Possible external collaborator: researchers involved in Wikidata PageRank

The ranking criteria used for the autosuggest is based on how often the entity is linked to objects in the Metadata Collection (intrinsic measurement) and the Wikidata PageRank (extrinsic measurement). Additional intrinsic popularity measures based on the preferences of our users could also be explored. A disambiguation algorithm to decide what type of entity is more appropriate (e.g. based on the context of the query or preferences of the user) would also contribute to improve this ranking.

ACTION: We will reach out to the designers of Wikidata PageRank for this line of work (also to inform them on our use of their research).

ES-F7. Basic autosuggest language handling

(2021: work already done)

Different approaches have been considered to select the language of the labels to be suggested[39]. Recently, we have chosen to implement one of the simplest approaches, where only the labels in the language of the website, assumed to be the language of the user, are considered. This approach, while easy to design and deploy, could be less

[38] C.2 Usage pattern report 2, 30 April 2018: "Entity types should be made available to users.. This information will help users in choosing the correct entities in the suggestions"
[39] https://docs.google.com/document/d/1iDuxcSb-7gzRsW5CqWn0SCUuaYYfsY3i-JX0ZHa0rh8/

satisfying and perhaps confusing for users querying in different languages, so the use of other alternatives is open and will be investigated as part of the Multilingual Strategy (see section Multilinguality).

ES-F8. Metadata and Entity Collection integrated in autosuggest

The inclusion of additional information (e.g. title) contained in the Metadata Collection as suggestions, together with the entities from the Entity Collection, could be useful for users issuing non entity-based queries, and is also considered an interesting option to explore, even if it does not necessarily count as extending the semantic search functionality, from a strict technical perspective.

ES-F9. Improving internal enrichment

(2021: work in progress)

We are working to improve our enrichment framework[40]. The improvements, as well as the preliminary planning, are collected in the document 'Improving Semantic Enrichment'[41]. The work items related to evaluation, e.g. the evaluation of item recommendations (see M-F12. Item suggestions and EVAL-5. Registering and reading user behaviour data) could be relevant to the work here, as they can help build gold standards to evaluate and/or train future enrichment systems.

ES-F10. Exploiting external enrichment

External collaborators: partner projects where enrichment is done.
(2021: work in progress)

Several ongoing projects can contribute to the semantic enrichment of the Metadata Collection. That is the case of the project 'Saint George on a Bike'[42], where we expect to automatically enrich images in our collection with descriptions adapted to the cultural heritage domain, or 'Europeana XX', which will contribute with a new thematic collection focused on the 20th and its enrichment with different types of entities. These enrichments may be ingested in Europeana following two paths: via the Annotations API, and recently, via the new changes introduced in the EDM model to support this[43]. An additional consideration for this task, with a big impact in search, is how to assess the quality of those enrichments and what should be the threshold to include them or not in the index. In order to help with this task, we have defined a template to fill in by the project partners[44].

---

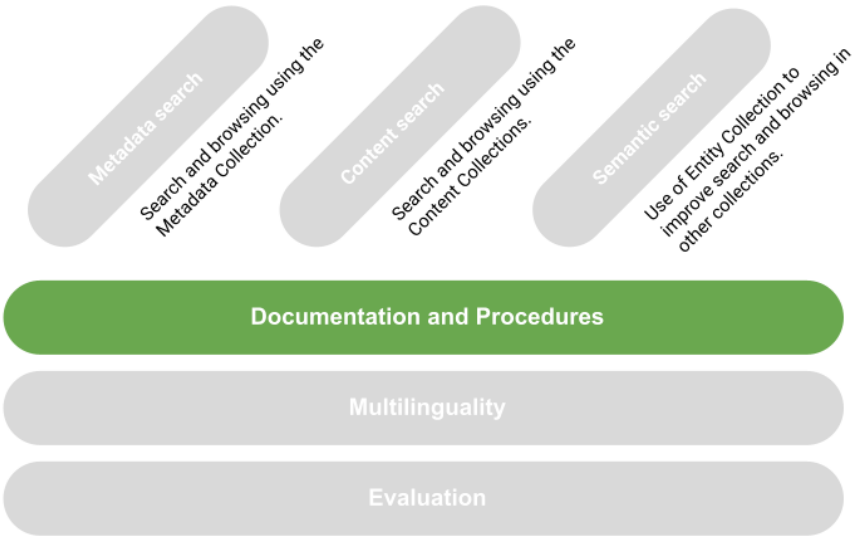
[40] *Assessment and recommendations for enrichment tools and rules*, accessible at: https://docs.google.com/document/d/10p89ap5pfgCxs2YCsB3Fg2W3U-KPdmclDxdgu58707A/
[41] *Improving semantic enrichments at Europeana*, accessible at: https://docs.google.com/document/d/1BCriz4T6XeuITrmZtePUAqVF0wowIPBuVJKddACpjro/
[42] https://pro.europeana.eu/project/saint-george-on-a-bike
[43] *Provision of metadata enrichments and translations*, accessible at: https://docs.google.com/document/d/1AFxfr7BESYGd_1sFOLWBWs7C6uGXcAchom8TE2ClvTM/
[44] *Europeana Template : Enrichment evaluations carried out by the project*, accessible at: https://docs.google.com/document/d/11y7L78_QopELNMZXbqcn-WyFaba0ip9HFckEdAEB3uU/

ES-NF1. Migration of the Entity Collection to Solr Cloud

(2021: work already done)

The Solr instance containing the Entity Collection has been migrated recently to a distributed mode.

Documentation and Procedures



D-1. Document relevant changes in search

(2021: work partially done)

In order to keep track of the improvements done in different teams affecting the search, we created a new document, "Relevant Changes in Search"[45]. That excel sheet was expected to list the main changes in the Europeana website, APIs or Solr, that have an impact in the search results, UX, efficiency or monitoring. However, it became outdated very quickly as the people involved were not using it. Now the actions listed there are also included in this document, but we still need to find a proper way to keep track of the status of all the actions reported.

[45] https://docs.google.com/spreadsheets/d/1YM6mvIReSAwhOqUNwNXFl6IsM0_GY8-RIaeBogsxeb0/

D-2. R&D Wiki. Search section

(2021: work partially done)

We have reorganised the Search section in the R&D Wiki[46], while its update is still pending.

D-3. Github repository

(2021: work already done)

The Github version-control repository for search has also been organised, splitting the repository in two, where one of them hosts the different Solr configurations and plugins we have in production[47], and the other hosts search tools[48].

D-4. Thematic collection updates

(2021: work already done)

Regarding the procedures, we have established a clear process to periodically update the thematic collection filters (contained as alias in the Solr collections). Currently the changes to the alias are received via pull-requests in the Github repository where the configuration for the Metadata index is tracked, and we have implemented a script to automatically update that configuration in production (after testing). If the changes are successful, the pull-request is accepted and the repository updated. While the implementation of thematic collections may be revisited, this process can be reproduced for any change in the configuration of the search engine that does not require indexing.

D-5. Elevation management

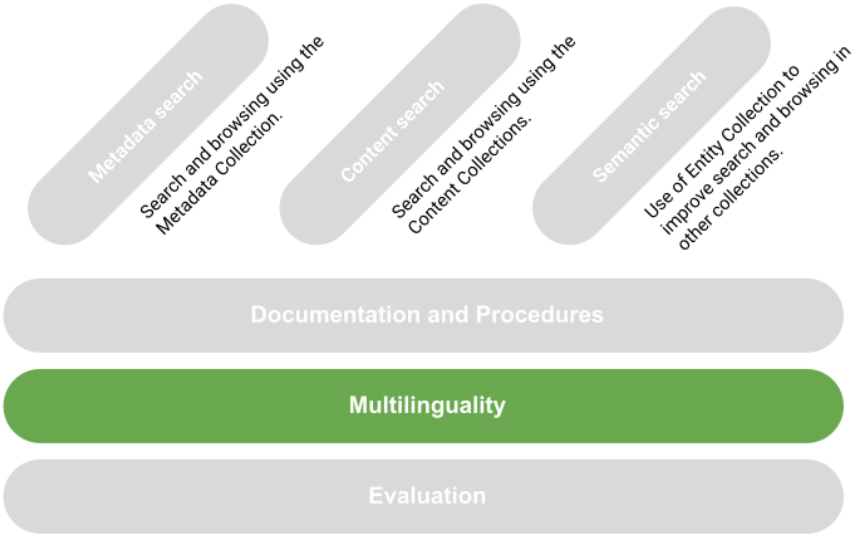
(2021: work in progress)

A process should be in place in order to easily and effectively update the elevation file described in M-F10. Ranking by popularity. Ideally the changes introduced here by the editorial team should be directly reflected in the elevation file[49], and then updated in the Solr configuration following the same process described in D-4. Thematic collection updates.
It would be also wise to trace the impact on search when there is a significant change in the elevation (e.g. number of entries edited/introduced above a threshold).

[46] https://europeana.atlassian.net/wiki/spaces/RD/pages/20971521/Search
[47] https://github.com/europeana/search
[48] https://github.com/europeana/search-tools
[49] Taking into account also possible changes in the record identifiers as registered in the Redirection API.

The new recommendation system implemented in Europeana could also be used to help in the process of deciding the best bets, showing similar items to relevant known items.

Multilinguality



Currently, the Europeana collection contains more than 57 million records in 38 languages. In most cases, only one language is used for the content and metadata of those records, although it could happen that the language of the content is different from the language of the metadata, or that more than one language is used for each of them. Additionally, the enrichment post-processing applied by our ingestion team results in records containing named entities (person, location and concept) described in multiple languages. We also normalise language information (tags) given by providers, which sometimes do not align with existing standards (BCP47)[50]. All that information is stored and indexed in our search engine in order to provide users with a search functionality over the collection. To that end, fields with specific language tags are used (e.g. fulltext.de, fulltext.en) unless the language is not provided by our providers, in which case generic fields are used (e.g. fulltext).

On the other hand, users from all around the world access our website and issue queries in their native language, expecting to find any type of records, from textual documents, like archives or newspapers, to multimedia contents like audio, videos and paintings, whose contents and/or metadata are not necessarily described in that specific language. This scenario poses several challenges to solve, from the discovery to the display of the information contained in the records.

[50] As part of supporting actions for multilinguality, we will also continue this work, addressing the pending issue of sub-tags (language variants and scripts), cf https://europeana.atlassian.net/browse/RD-65

Although the question about introducing specific textual analysis per language has been discussed previously[51], currently the queries are issued against the whole data (metadata or fulltext separately) instead of routing them to specific languages. That approach results in missing relevant metadata and content in languages different from the one used in the query, but also in noisy results when one word exists in more than one language with different meanings, or when different words in different languages are reduced to one common string after normalisation.

In order to increase the discoverability across languages, we first run preliminary experiments using the eTranslation[52] service (see action MUL-F1. Preliminary experiments with eTranslation), and we recently outlined a strategy[53]. The strategy is organised along several outcomes, where one is central to search (see action MUL-F3. Search Europeana) and the others are key enablers (see actions MUL-F2. Underlying multilingual data is established, MUL-F4. Read item text, MUL-F5. Policy and Plan establised). As part of the enable work, we have been working intensively to increase the enrichment, especially by mapping vocabularies used by Europeana providers with our Entity Collection, so we can directly assign those entities (and possible additional properties) to those records (see action ES-F1. Expanding coverage of search by entities).

Preliminary experiments instructing Multilingual Strategy

MUL-F1. Preliminary experiments with eTranslation

(2021: work partially done)

We have run two experiments with the eTranslation service: first we tested the languages supported[54], and then we run an experiment to compare the results obtained in a monolingual system, with those obtained in a cross-lingual system when using eTranslation to translate queries and fulltext content to English[55]. The results obtained indicate that we can discover up to 67% of the translations that are more likely to be relevant, but as expected, at the cost of introducing noise in the results (precision could go down to 49%). We observed that a good portion of the queries were badly translated, in some cases because of the lack of context, in others because the original language did not match the language of the website (which was assumed to be the language of the query), and in a good portion of them because of the wrong translation of the entities contained (61% of queries were entity-based).

Given these results, we should be cautious when applying automatic translation, as already recommended in the *White Paper on Best Practices for Multilingual Access to Digital Libraries*[56]. If applied, they recommend that users should be able to turn on / off

[51] MS30. Search Improvement Plan MS30, 2015 (p. 8): https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms30-search-improvement-plan.pdf
[52] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[53] https://pro.europeana.eu/post/europeana-dsi-4-multilingual-strategy
[54] https://docs.google.com/document/d/1tzr3U3NfDMuICsMXWspKOIbWCjDOepLO7sJzysFwuC4/
[55] https://docs.google.com/document/d/1K7NaqpFVNl7nIFbOLlEftsLPlDUsTVom5Efwy5htBm4/
[56] https://pro.europeana.eu/files/Europeana_Professional/Publications/BestPracticesForMultilingualAccess_whitepaper.pdf

this functionality, and Named Entity Recognition (NER) should be incorporated in the process. Named entities should be neither translated nor normalised (e.g. stemmed). That would mean not only applying NER to the original queries, but also to the text before it is indexed in the search engine. That way, given a query, we would search for entities and the rest of the text, original and translated, separately. For this approach to work, we need metadata and contents tagged with the proper language. As it is not always provided, we would need to apply language recognition. According to the literature, this could be especially challenging for the metadata given the lack of context in some fields[57], however a small experiment[58] conducted in 2016 with Solr's built in LangDetect functionality shows an acceptable (c. 89%) accuracy even on very short text snippets. As an alternative, we can explore first if just by searching the entities and the original and translated query keywords in any field, independently of the language, works well enough. As a final step, we need to decide if the results obtained in different languages are still merged together (which requires that relevance scores would be comparable across languages) or if, as recommended in the White Paper mentioned before, there should be a clear separation of languages in the results list.

Implementation of Multilingual Strategy

All the actions in this category are highly desirable and have been already done, or planned for the short and medium term as part of the multilingual strategy roadmap. Detailed reporting on their progress is expected to be available in the coming DSI-4 MS3 ("Outcomes of multilingual experiments, and their contribution to the multilingual strategy").
NB: this listing currently includes items that are less directly related to search (e.g. on user design and translation for visualisation, as their implementation may turn to be related with that of core search items).

Possible external collaborators for future iterations: Vivien Petras (Humboldt University) and Juliane Stiller.

MUL-F2. Underlying multilingual data is established

(2021: work in progress)

This part of the Multilingual Strategy includes work items focused on metadata:

- Prioritise normalisation of not-yet normalised tags (in original & dereferenced data)
- Language detection of metadata is validated (experiment)
- Agree on evaluation methodology and quality thresholds for translations
- Candidate machine translation services for metadata are evaluated (experiment)
- Work with communities and data partners to extend language coverage of entities (vocabularies) where necessary

[57] See for example this paper: https://edoc.hu-berlin.de/bitstream/handle/18452/21160/Stilleretal_CLEF2010.pdf?sequence=3
[58] https://github.com/europeana/search/tree/051952fad7af9aeba3dab8b97c2dacd5627575b0/language_detection

- Decide whether translation of metadata fields should focus on a selection of fields, such as discovery-enabling fields or metadata record Tier 2+ objects
- Language detection of full text is validated (experiment)
- Candidate machine translation services for (static) full text content is evaluated (experiment)
- Machine translation pipeline translates all metadata to English, allows for quality control, then stores and indexes data
- Enhance coverage of multilingual knowledge graph over Europeana collection objects by improving semantic enrichment
- Machine translation pipeline translates all full text to English, allows for quality control, then stores and indexes data
- Evaluate options for handling full text that is embedded within IIIF (especially for language detection)

MUL-F3. Search Europeana

(2021: work in progress)

- Stop applying English text analysis to all languages in Solr
- Real-time detection of search query language is validated (experiment)
- Real-time translation of search query is validated (experiment)
- Construction of multilingual search string is validated (experiment)
- Multilingual search designs prove to be usable and understood by users (user research)
- Route queries to specific language fields (metadata or fulltext separately) instead of issuing them against all data (experiment)
- Review handling of languages in Entity API suggester method to meet expectations of new multilingual search UX
- Design of ranking for multilingual search results is validated (experiment)
- Ranking for multilingual results (implementation)
- User can enter search query in chosen language and get multilingual results (implementation)
- Improve detection of entities in phrase queries
- Stop applying language analysis (e.g. stemming) to entities in metadata and full text (experiment)
- Users get better multilingual search results based on the inclusion of full text translated to English in search indexes

MUL-F4. Read item text

- Real-time translation of item page metadata from English is validated (experiment)
- Multilingual item page designs prove to be usable and understood by users (user research)
- Users can view item pages in language of choice (implementation)
- 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
- Users can view full text content in language of choice (implementation)

MUL-F5. Policy and Plan established

- Confirm a set of metrics and KPIs to define both quality targets and desired performance improvements to the multilingual experience
- Assess language coverage of entities used in search and source data
- Prioritise languages to support if resourcing does not allow coverage of full 24 official languages
- Update Europeana policy to account for support of other non-official EU and European regional languages
- First implementation and evaluation of the metrics and KPIs, focusing on entities

Backlog Items

MUL-F6. Impact of BERT in multilinguality

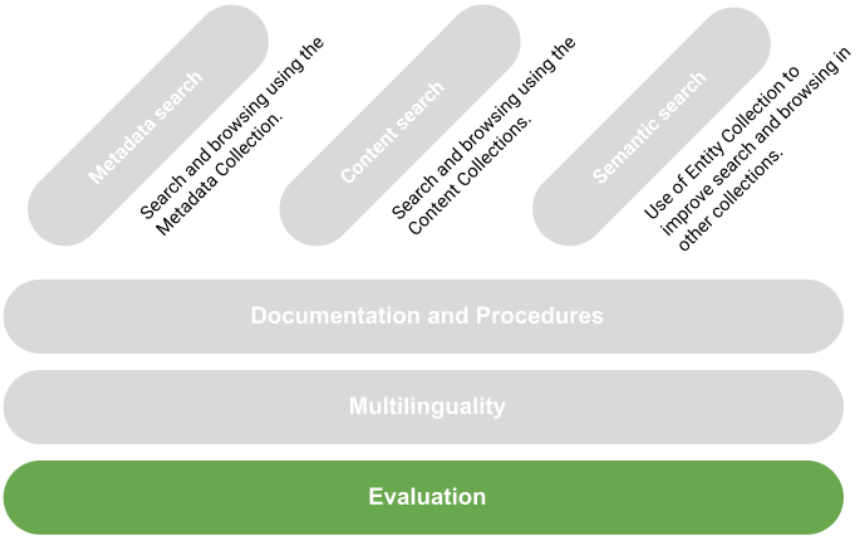
As part of the experiments in order to improve multilinguality, we discussed the possibility of evaluate an information retrieval system prototype built with BERT[59], and developed in the context of The Culture Chatbot project (see action M-F17. Use of language models to improve retrieval). BERT is a language model that is state-of-the-art in multiple Natural Language Processing tasks, and has proved to help in information retrieval tasks. The multilingual version used in those experiments (M-BERT) has the ability to overcome language barriers, and its results in terms of cross-lingual search are promising.

MUL-F7. Clearer language filters

The *White Paper on Best Practices for Multilingual Access to Digital Libraries* indicates that users tend to refine results by language, but it also states that they do not properly understand the language filter we provide. Ongoing design work is trying to address this. We may need extra work to fulfil (future iterations of) this design, or to enhance the way EDM language-related fields (edm:language and dc:language) can better contribute to more effective interactions, either based solely on the metadata language (or the language of the provider's institution when no language is provided) or the language of the content in case it applies (e.g. text, audio).

[59] https://arxiv.org/abs/1810.04805

Evaluation



An evaluation framework focused on search quality was proposed in a DSI Search Improvement Report from 2017. This proposal guided the subsequent design of periodic evaluations over the search components in production, in which we have measured usage and effectiveness based on the behaviour of the users. In Appendix C we review the methodology, metrics and KPIs proposed in these previous efforts, and the main findings obtained from their implementation.

These proposals and findings influence the proposals for following evaluations, which we lay out in this section.

Evaluation methodology and support

To begin with, the evaluation conducted should be different depending on the specific stage we have defined: implementation, and post-implementation (divided in pre-production and production).
Although we do not include here a pre-implementation stage, note that for all the stages, the topics used for evaluation should represent those issued by Europeana users. That means that for the evaluation we should make a proper sampling in case of queries taken from logs, or we should conduct proper studies to drive the selection of topics in user evaluations, such as the one described in Appendix C.1.3 (see Figures 5 and 6). Other analysis can be conducted during the pre-implementation stage in order to identify user patterns that can drive our decisions. See for example the work by Tessel Bogaard et al. ”Searching for Old News” and “Metadata categorisation for identifying search patterns in a digital library”.

EVAL-1. Evaluation methodology during implementation stage

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).
(2021: work in progress)

During the improvement of the existing functionalities or the implementation of new ones, an expert evaluation should be in place, so the search specialist (and eventually a search cross-team defined in EVAL-4. Cross-team effort organisation) can assess the impact of the changes made, at least from the perspective of the system's effectiveness. In order to do this, we need to know the relevance of documents for specific queries. This information can be obtained from:

A. Existing gold standards.
B. An estimation of the relevance of documents based on the logged behaviour of users.
C. User evaluations where users assess the relevance of the documents returned for predefined queries or search tasks.

The last approach should only be considered at the end of the stage, when we want to compare the previous system with the new one, and no more fitting is required. Actually, this type of evaluation could be combined with the environment set up in Europeana to run integration and performance tests when software or infrastructure changes are deployed. This infrastructure has already been deployed, and the first user-oriented evaluations have been done[60], for the moment with volunteers from different teams at Europeana. We expect we can improve the infrastructure to include two different test websites instead of just one (and compare with the one in production), as well as the organisation of the people participating, following the creation of a more stable multidisciplinary team with this objective (see action EVAL-4. Cross-team effort organisation).

The first two approaches are more flexible and therefore suitable for testing and fitting the system iteratively. Regarding option A, currently at Europeana we do not have gold standards that can be used for the improvement of the search engine in general. The one created for CHiC'2013 (see C.2.1) is largely outdated: the document identifiers are now broken because of Europeana data updates, and it does not contain all the current metadata fields, including some of the ones most relevant for search, like contextual fields (e.g. agents). We could seek to retrieve the dataset and subject them to our semantic enrichment to populate these fields anew. As an alternative, we can use as the gold standard the collection of curated top representative items for some of the main cultural creators, art works, concepts or places. We internally call this collection 'best bets', and it is currently used in the search engine to rank those items higher. The evaluation using this collection is however limited, as it would only be valid for queries containing those (necessarily few) famous entities. As another alternative we could

[60] Evaluation and results accessible here: https://docs.google.com/document/d/1zL7OM6CXop0LCFIXVxxQGJ8geSKkR_27AdlT0qAovOc/edit#heading=h.vrww9f29n4bc

create a gold standard via crowdsourcing, although as explained in the Europeana Evaluation Framework report, creating gold standards is expensive and, in our case, not so useful for our changing collection.

Another option is to make use of the logs (option B) to dynamically create gold standards for evaluation, which would be much more flexible and useful in the long term. In order to do that, we need an estimation of the relevance of the documents based on the behaviour of the user. We have published a Master Thesis proposal in the department of Intelligent Systems in the Delft University of Technology to do this. With these estimations, we could skip, in some cases, conducting user evaluations in this stage, and probably we could also skip the pre-production stage described next. Related to this approach, we should improve Europeana's tracking of user sessions, and find some means of measuring user dwell time (that is to say, the amount of time spent examining a given page) precisely.

EVAL-2. Evaluation during pre-production stage

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Once the changes have been internally accepted, it would be desirable to run tests with users, ideally before it goes fully on production. In order to do so we can run A/B tests, where the new search system is exposed randomly to some users (they may be only a few or half of our users) during a period of time, so we can analyse which one is better based on their behaviour. In order to do this we could use low-level signals of success in user behaviour (see Appendix C.1.2). This type of test is being considered for the evaluation of a pilot for search across Spanish and English collection scheduled in 2021, where we are putting into practice some of the actions described in the Multilingual Strategy. The pilot is expected to be ready by the end of the current DSI.

EVAL-3. Evaluation after product in production

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).
(2021: work in progress)

Once the system is in production, feedback channels should be established to get input from users. This is already in place, and we have recently created an internal JIRA board in order to track the relevant feedback obtained related to search (see EVAL-4. Cross-team effort organisation). A periodical evaluation with user tests should also be conducted to measure user satisfaction (see Appendix C.1.3 Figure 11 for a good example of a user satisfaction evaluation).

EVAL-4. Cross-team effort organisation

(2021: work in progress)

As part of the evaluation process, we need to be aware of the issues/opinions/requirements our users have. In order to fulfill this objective we have

informally articulated a team with the task of collecting all the feedback related to search from our users (see EVAL-3. Evaluation after product in production). The team in charge of the feedback is integrated with the people responsible for processing the feedback from external users, plus other people in the organisation who either are direct users of the search functionality (the Editorial Team), or give support to specific groups of users (Collection Engagement team, which gives support to educators, and Community and Partner Engagement team, which gives support to researchers). They will send this internal and/or external feedback to the search specialist so she can organise the issues and requirements in the JIRA board mentioned in EVAL-3, and present it to be assessed by the management team. This board shall also be used for issues being discussed (and raised) across teams.

In the same vein, we should also create another multidisciplinary team, this time with the task of evaluating the search in the first stages of the implementation (see action EVAL-1. Evaluation methodology during implementation stage). At the end, the quality and quantity of the data provided and enriched, the way the information is displayed, the informational needs of our users, and the way the Europeana website and APIs make use of the search engine are all crucial aspects in the search, and they shouldn't be tackled by a person or even a team in isolation. This team, as opposed to the one for the collection of feedback, is more flexible in terms of the components, and we should aim to include as many participants as possible. We have already put this into practice with a group of volunteers across teams as mentioned in EVAL-1.

EVAL-5. Registering and reading user behaviour data

Some of the data required to run the current evaluation (described in Appendix C.2) is no longer available with the launch of the new Europeana website. Previously the clicked documents and its rank was registered for specific queries and sessions. **If this information is not available, we can no longer run the current evaluation on the search box, the autosuggest and the similar items**. Additionally, we need to analyse the data required for the metrics described in actions EVAL-6 TO EVAL-10 that will be finally used, and take the necessary steps to register that data and make it available, be it in Google Analytics or in the Europeana logging infrastructure.

Part of the new logging data that must be registered is the one related to the use of the recommender system (see M-F12. Item suggestions). The logging of the user behaviour and preferences (queries, viewed items, recommended items marked as favorites) can be used for the retraining of the system, as well as for the training and/or evaluation of the enrichment system.
Additionally, the infrastructure used to access the logs registered in our logging system has been changed, while the old one will be decommissioned by April 2021. Changes in the software interacting with the old infrastructure may be needed. This is the case for the software used to get the data necessary for the current evaluation described in Appendix C.2.

Evaluation metrics

In parallel to the methodology and support work above, we can use metrics to estimate the effectiveness, usage, efficiency, diversity, and coverage of the search components (see actions EVAL-6 to EVAL-10 below). The metrics described are limited to those applied to the log data, and most of them are based on previous reports, especially including those that are easy to understand, and for which an increase or decrease clearly indicates an improvement. Other approaches and metrics are possible, like the use of gold standards, with metrics like precision and recall, and the employment of user surveys, with metrics based on satisfaction. We envision assessing and implementing some of them. In particular, user surveys, while being very relevant for evaluation, have already been addressed in the D6.3 Search Improvement Report[61], and will be planned and implemented together with the Service Experience team.

EVAL-6. Metrics to measure effectiveness

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Table 5. Effectiveness metrics

Search component	**Effectiveness metrics**	**Increase (↑) / Decrease (↓)**
Search box	% queries with clicked results	↑
	% query sessions with clicked results	↑
	% query sessions ends with item view (as opposed to search)	↑
	% query reformulation (including adding/removing filters)	↓
	Avg. dwell time per items clicked	↑
	Avg. rank clicked items per query	↓
Filters	% queries with filters no clicks	↓
	Avg. dwell time per items clicked in query with filters	↑
	Avg. rank clicked items per query with filters	↓
Autosuggest	% autosuggest queries with no clicks	↓
	Avg. dwell time per items clicked in autosuggest query	↑
	Avg. rank clicked items per autosuggest query	↓
	Avg. rank of entity selected from autosuggest	↓
Similar items	Avg. dwell time per items clicked in similar items	↑

[61] D6.3 Search Improvement Report, pp. 19-29, accessible at https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.3-search-improvement-report.pdf

	Avg. rank clicked items in similar items	↓

Additionally we can apply standard metrics like nDCG and Mean Reciprocal Rank[62]. However, note that we have deliberately excluded the number of clicks (Click-through rate) as an estimation of effectiveness of the search. It could be considered that the more clicks in the search results, the better, because users show interest in those documents. However a previous study shows that there is a negative correlation between the number of interactions (including clicks) and the satisfaction of the user (see Appendix C.1.3), because it could also be that they can not find what they are looking for. Additionally, this number can be also determined by the type of query: very specific queries for which we expect to find one or two documents at most (and therefore a reduced number of clicks), or informational queries where we want to explore a broad concept and for which the number of relevant documents is expected to be higher (and so the clicks).

EVAL-7. Metrics to measure usage

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Table 6. Usage metrics

Search component	**Usage metrics**	**Increase (↑) / Decrease (↓)**
Search box	Ratio search sessions per sessions starting in home page	↑
	% subsequent user search sessions	↑
Filters	% queries with filters	↑
	% search sessions with use of filters	↑
	Distribution of queries per type of filters	N/A
	Distribution of queries per values of filter	N/A
Autosuggest	% autosuggest queries	↑
Similar items	% sessions with clicks in similar items	↑

EVAL-8. Metrics to measure efficiency

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Table 7. Efficiency metrics

Search component	**Efficiency metrics**	**Increase (↑) /**

[62] Other metrics could be more suitable when we work with such estimations though (incomplete relevance assessment).

		Decrease (↓)
Search box and filters	Avg. query time response	↓
Autosuggest	Avg. time autosuggest service (per number of characters introduced)	↓
Similar items	Avg. query time response similar items search	↓

EVAL-9. Metrics to measure coverage

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Table 8. Metrics to measure coverage

Search component	Coverage metrics	Increase (↑) / Decrease (↓)
Search box	% of queries with entities included in Entity Collection	↑
	% queries with no results	↓

EVAL-10. Metrics to measure diversity

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

Table9. Metrics to measure diversity

Search component	Diversity metrics	Increase (↑) / Decrease (↓)
Search box	% items with no clicks (excluding tier 0)	↓
	Distribution of clicks in the collection	**N/A**
	% items never appear in top 100 search results	↓

EVAL-11. Assess other (non-log based) modalities for use as metric sources

Possible external collaborators: Paul Clough and Monica L. Paramita (Univ. Sheffield).

The metrics presented in this document are based on the user interactions registered in the logging system. However, other metrics are possible using as a source user studies, which can more directly measure the user satisfaction on the different dimensions considered.

Conclusions

In this document we have described the changes that can be made in order to improve the search functionality offered in Europeana, specially the one offered from the Europeana website, as opposed to the one offered to our users through the use of the APIs.

Some of the actions described are straightforward, whilst some others require experimentation, where in some cases it would be wise to count on external collaborators. This is specially relevant in the areas of multilinguality and evaluation.

Key aspects identified that need to be improved are the registering of user data, the exploitation of that data in order to get useful insights through clear metrics, the internal collaboration required to agree on relevant changes and directions, and the existence of documentation and established procedures for testing and tracing possible issues.

APPENDICES

APPENDIX A. Actions status

The table below contains the status of all the actions reported in the document (B: in progress, C: partially done, D: done)

Focus Area	Action	Purpose	Status
Metadata search	M-F1. Digital representations and hierarchies	a) All existing media is displayed in search results, b) media displayed match query submitted	B
	M-F2. Search by dates	Users can search/browse by date of the CHO and use ranges.	B
	M-F3. Extend faceting and filtering	Users could facet by fields that are now only used for search (e.g. title, subject, contributor, creator, publisher, agent, place, concept), supporting browsing.	C
	M-F4. Improve textual normalisation (tokenisation)	Search is more flexible (i.e. it does not require the punctuation marks to appear).	D
	M-F5. Improve secondary sorting criteria	a) Same query displays same results, b) random ranking is displayed when user is just filtering the collection	D
	M-F6. Content and metadata quality filters	CHOs with not enough quality are hidden from users by default.	D
	M-F7. Content and metadata quality is used in the ranking	(1) Present first to user content that is more usable and better understandable. (2) Help fulfil the promise to our partners in the Europeana Publishing Framework ('the more you give, the more you get') in terms of visibility of higher-quality data (and consequently convince them to improve their data in)	
	M-F8. Tuning the weight of different fields in ranking	a) CHOs ranked first in search results because the query terms appear in a more relevant field and/or in a shorter field (e.g. title instead of description), b) named entities are not transformed (e.g. *Alberts* is not transformed to *Albert,* or *Luis* to *Lui).*	
	M-F9. Learning to Rank based on user interactions	Ranking of CHOs in search results takes into account preferences shown previously by our users (e.g. manuscripts collection displays first medieval manuscripts, search by *dog* displays first media related to dogs, and not to a specific type of shell *Dog Whelk*).	

	M-F10. Ranking by popularity	Ranking of CHOs in search results takes into account criteria based on popularity of those items among users or in external sources (e.g. most famous items in WikiArt).	
	M-F11. Change default boolean operator from AND to OR in search	Limit *no results found* message or search results with only a few items, so users have the opportunity to reformulate queries.	
	M-F12. Item suggestions	For each item clicked, similar items are displayed based on metadata or content but also on collaborative filtering (i.e. other users also viewed...).	B
	M-F13. Query spelling correction	Frequent errors in the queries are automatically corrected, avoiding unexpected results for the user.	
	M-F14. Re-activating advanced search	Users can issue more specific queries directly from the Europeana website (e.g. search by author without including description, contributors, publishers, etc.), and use advanced features to increase precision (e.g. phrase queries, where keywords must appear in sequence)	
	M-F15. Search results export	Users can export the search results (metadata) from the Europeana website.	
	M-F16. Highlighting in metadata search results	a) Users can assess faster if the document is relevant for their needs, b) users and staff know why a document was retrieved, promoting transparency	
	M-F17. Use of language models to improve retrieval	Feeding search engine with more contextual information may improve search results	B
	M-F18. Editorial content in search and/or recommendation	Not only the metadata that matches a query from a user is displayed, but also editorial content related (e.g. blogs and galleries).	
	M-F19. Enable sorting using main fields	Allow sorting in fields like title or creator	C
	M-F20. Search by location/georeference criteria	Enable spatial search	

	M-NF1. Solr upgrade for Metadata Collection	Users see results faster, new features are available for staff to be exploited	D
	M-NF2. Attribute docValues	Users see results faster, especially when faceting, and can sort by any field (e.g. currently we can not sort by *title, subject* or *publisher*), c) staff can use Solr for data mining	D
	M-NF3. Improving performance filtering by Content Tier	Users see results faster when filtering by content tier (e.g. thematic collections)	D
	M-NF4. Monitoring Metadata Collection	Staff can detect possible issues (e.g. attacks, lack of resources) and take solutions before the problem becomes bigger	D
	M-NF5. Streaming (use of search engine for data mining)	Staff can use Solr as a source for data mining	D
	M-NF6. Prevent errors from elevation functionality and use of pagination	User do not see an error when launching a query that includes that specific functionality	B
(Fulltext) Content search	FT-F1. Unify searches on newspapers and transcriptions		C
	FT-F2. Highlighting in content search results		D
	FT-F3. OCR correction in Newspapers collection		
	FT-F4. Named Entity Recognition applied to fulltext content		
	FT-F5. Search on video and audio subtitles		
	FT-F6. Search on content within a single (IIIF) Item		D

	FT-F7. Search on metadata and (fulltext) content collections at the same time		
	FT-NF1. Solr upgrade for Content Collections		D
	FT-NF2. Monitoring Content Collections		D
	FT-NF3. Content and metadata sync		
	FT-NF4. New transcriptions are displayed and searchable live		
Semantic Search	ES-F1. Expanding coverage of search by entities		
	ES-F2. Search by entity		D
	ES-F3. Promote diversity when searching by entities		B
	ES-F4. More flexible autosuggest		
	ES-F5. Clearer autosuggestions		
	ES-F6. Autosuggest ranking criteria		
	ES-F7. Basic autosuggest language handling		D
	ES-F8. Metadata and Entity Collection integrated in autosuggest		
	ES-F9. Improving internal enrichment		B
	ES-F10. Exploiting external enrichment		B

	ES-NF1. Migration of the Entity Collection to Solr Cloud		D
Documentation and procedures	D-1. Document relevant changes in search		C
	D-2. R&D Wiki. Search section		C
	D-3. Github repository		D
	D-4. Thematic collection updates		D
	D-5. Elevation management		B
Multilinguality	MUL-F1. Preliminary experiments with eTranslation		C
	MUL-F2. Underlying multilingual data is established		B
	MUL-F3. Search Europeana		B
	MUL-F4. Read item text		B
	MUL-F5. Policy and Plan established		B
	MUL-F6. Impact of BERT in multilinguality		
	MUL-F7. Clearer language filters		
Evaluation	EVAL-1. Evaluation methodology during implementation stage		B
	EVAL-2. Evaluation during pre-production stage		

	EVAL-3. Evaluation after product in production		B
	EVAL-4. Cross-team effort organisation		B
	EVAL-5. Registering and reading user behaviour data		
	EVAL-6. Metrics to measure effectiveness		
	EVAL-7. Metrics to measure usage		
	EVAL-8. Metrics to measure efficiency		
	EVAL-9. Metrics to measure coverage		
	EVAL-10. Metrics to measure diversity		
	EVAL-11. Assess other (non-log based) modalities for use as metric sources		

APPENDIX B. Previous actions completed

The table below contains the actions done between the beginning of DSI-3 (September 2017) and the beginning of DSI-4 (September 2018), which have not been reported in previous DSI search improvement reports. The actions done after that date (thus performed in DSI-4) have been already included in this report.

Table 11. Action done during DSI-3

Focus Area	**Action**	**Source**
Entity Search	Autosuggest	C.3 Data access pattern report 1 (M4), 22 Jan 2018
Metadata Search	Update search server infrastructure	C.2 Usage pattern report 2, 30 April 2018
	New solution for User Generated Content platform	
	Embed functionality for images displayed in item page	C.2 Usage pattern report M12, 31 August 2018
	New ranking which prioritises objects with available media over objects without media.	
	Replace custom ranking algorithm BM25f by default Lucene algorithm BM25.	C.2 Users and usage report M5, 31 January 2019
	Update Solr to 6.6.2	
Evaluation	Autosuggest evaluation	C.2 Usage pattern report 2, 30 April 2018
	In-depth evaluation of the Entity collection	
	Established metrics for search performance	
	Study on the correlation between users' search behaviour and their satisfaction with Europeana website	C.2 Usage pattern report M12, 31 August 2018
	Included Precision and Reciprocal Rank to measure search performance	C.2 Users and usage report M5, 31 January 2019
	Consider only queries with	

	keywords for measures nDCG, Precision, and Reciprocal Rank	
Content Search	Newspapers browse pages and IIIF viewer	C.2 Usage pattern report M12, 31 August 2018
	Newspapers collection in production	C.2 Users and usage report M5, 31 January 2019

APPENDIX C. Previous evaluation frameworks

C.1 2017 Evaluation Framework proposal (2017)

An evaluation framework focused on search quality was proposed in 2017's Search Improvement Report[63]. We review here the methodology proposed, the metrics and KPIs suggested, and the main findings obtained from its implementation.

C.1.1 Methodology

The methodology includes the approaches needed to evaluate Europeana search and its components at three stages in the development lifecycle:

1. Pre-implementation: understanding the user and their context for search (the information use environment).
2. Implementation: search component development and integration into Europeana.
3. Post-implementation: monitoring and improving search (e.g. through tuning the relevance ranking).

At that moment, the evaluations carried out at Europeana comprises (see Figure 5):

- Expert evaluation, where experts assess possible issues during the development and optimisation of the search components.
- User-oriented evaluation following controlled tasks (remote).
- User-oriented evaluation using clickthrough analysis from the logs.

63 https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.3-search-improvement-report.pdf

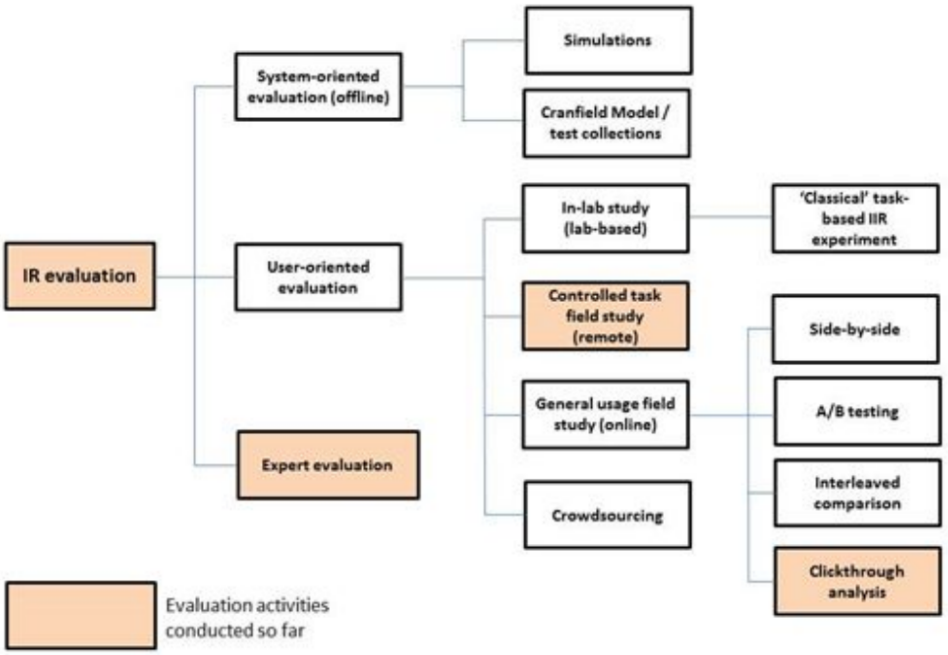


Figure 5. Search Evaluation approaches used so far in Europeana (as of 2017)

They propose and run user-oriented evaluation specially for the pre- and post-implementation (see C.1.3 to see the main findings). They indicate that test collections can be used in early stages of component development, ideally with our own collection and topics. Those resources could be created by assessing the relevance of the items retrieved for predefined queries (e.g. via crowdsourcing). Once the search components are in production, we should evaluate via online user tests and using the logs. The latter approach is the one taken to date in Europeana on a regular basis.

C.1.2 Proposed metrics and KPIs

The proposed framework suggests that the evaluation of search should be focused on four components: search box, autosuggest, filters and similar items search. For each of them, the authors propose a list of evaluation criteria and related metrics (see Table 1), and the KPIs associated (see Table 2).

Table 1. Metrics for search component evaluation in the Evaluation Framework proposed.

Components	Evaluation criteria	Metrics
Search box	Retrieval effectiveness	-Relevance-based metrics (nDCG,MRR, precision, recall and F-measure) -% searches retrieving 0 hits -% queries with clicked results -Nº items viewed in SERP -Depth of SERP pagination -Number of query reformulations -Subsequent session duration -Task-completion value (actual and perceived) -Average rank of clicked items
	Diversity	-MMR -Mean Marginal Relevance-

		-alpha nDCG
	User satisfaction	-User satisfaction rate -Subsequent bounce rate
	Efficiency	-Average response time for search
	Effort	-Average query reformulation before clicks in a search session -Average removable query terms without any clicks in a search session
	Coverage of collection	-% user searches matched to an entity from the Entity Collection
Autosuggest	Effectiveness	-Relevance-based metrics at n characters -% completions with no clicks
	User satisfaction	-User satisfaction rate
	Usefulness	-Informational value -Task-completion value (actual and perceived)
	Efficiency	-Response time
filters	Effectiveness	-Relevance-based metrics
	User satisfaction	-User satisfaction rate -Time to subsequent filter removal -% removed filters after zero click -Subsequent bounce rate
	Usage	-Frequency of use -Clickthrough to objects
Similar items	Effectiveness	-Relevance-based metrics
	Usage	-Frequency of use -Subsequent session duration

Table 2. KPIs for search component evaluation in the Evaluation Framework proposed.

Components	KPIs	Measures/Metrics
Search	Increase retrieval effectiveness by 5%	• nDCG • MRR
	Improve diversity of search results	• Maximal Marginal Relevance or Alpha-nDCG
	Improve user satisfaction of the search results	• User satisfaction rate of search quality
	Improve search efficiency	• Responsiveness (average response time for search)
	Reduce search effort	• Average # query reformulation before clicks in a search session
	Increase coverage of Entity Collection to 30% of queries or more[12]	• % of all user searches matched to an entity from the Entity Collection
Auto-complete	Improve the effectiveness of the auto-complete component	• MRR at n characters
Filters	Improve the effectiveness of filters	• Subsequent nDCG
Similar items	Improve the effectiveness of similar items	• Precision

They also include a list of metrics that can be used as low-level signals of success (i.e. can be noisy):

- Abandonment rate (% queries with no click) - decrease
- Reformulation rate (% queries followed by reformulation , including filters?) - decrease
- Queries per session (30 min. sessions) - decrease
- Clicks per query (number of clicks) - increase
- Cicks @ 1 (clicks on top results) - increase
- pSkip (probability of skipping): decrease
- Max Reciprocal rank (1/rank highest click) - increase
- Mean Reciprocal rank (mean 1/rank for all clicks) - increase
- Time to first click (seconds before first click) - decrease
- Time to last click (seconds before final click) - increase

C.1.3 Main Findings

The 2017 evaluation framework led to two user-evaluations run: a web-based survey that was filled in by a sample of 240 search users of the Europeana website, and a controlled task evaluation with 51 participants. Using surveys, they analysed the type of search task (Figure 6), the motivation and its relation with the task type (Figure 7), the perceived importance of the features offered to help in searching activities (Figures 8 and 9), and the perceived quality of the search results (Figure 10). The user-evaluation with controlled tasks aimed at assessing the success perceived when completing search

tasks (Figure 11). For more detail, see the 2017's Search Improvement Report[64] and the paper Clough, Paul, et al. "Europeana: What users search for and why." (International Conference on Theory and Practice of Digital Libraries. Springer, Cham, 2017).

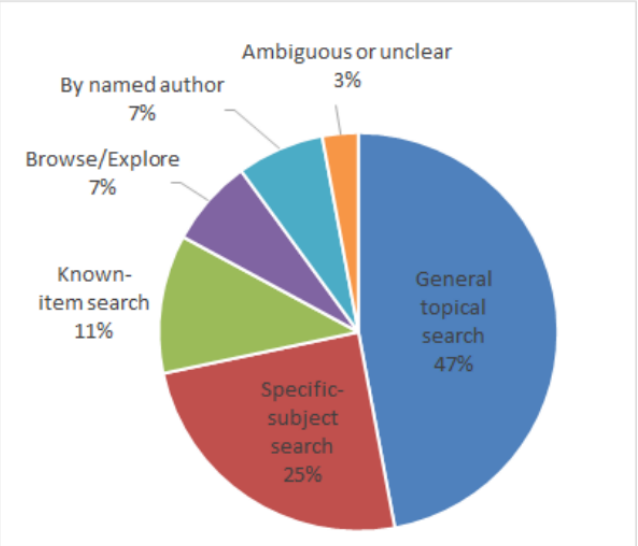


Figure 6. Percentage of users per search tasks types.

	Browse/ explore	By named author	General topical search	Specific-item search	Subject-specific search	Total
Create new work	17.6%	23.5%	37.2%	48.1%	42.4%	37.1%
Personal interest	35.3%	29.4%	22.1%	11.1%	44.1%	27.5%
Professional activity	11.8%	41.2%	26.5%	22.2%	5.1%	20.8%
Teaching	17.6%	5.9%	5.3%	14.8%	8.5%	7.9%
Other			0.9%			0.4%
Ambiguous / unclear	17.6%		8%	3.7%		6.3%
Total	100%	100%	100%	100%	100%	100%

Figure 7. User's motivation vs search task type.

64 https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.3-search-improvement-report.pdf

ID	Feature	Mean	SD	N
F1	Features to allow you to refine your search (e.g. filters)	7.92	1.97	234
F2	Availability of high-quality images	8.06	2.11	233
F3	Detailed information about an object (e.g. provenance, provider institution, etc.)	8.21	2.00	240
F4	The availability of information about an object in your own language	5.73	2.93	221
F5	Links to download an object	8.05	2.00	235
F6	Access to content you can freely re-use	8.21	2.26	235
F7	Links to an object provider's site (e.g. library, archive, etc.)	7.91	1.97	234
F8	Links to similar items	7.62	1.9	236
F9	Links to categorised collections (e.g. Art Nouveau posters, Irish folk music, Fashion illustrations)	6.76	2.31	226

Figure 8. Perceived importance of features to help in search (1:not important;10:extremely important).

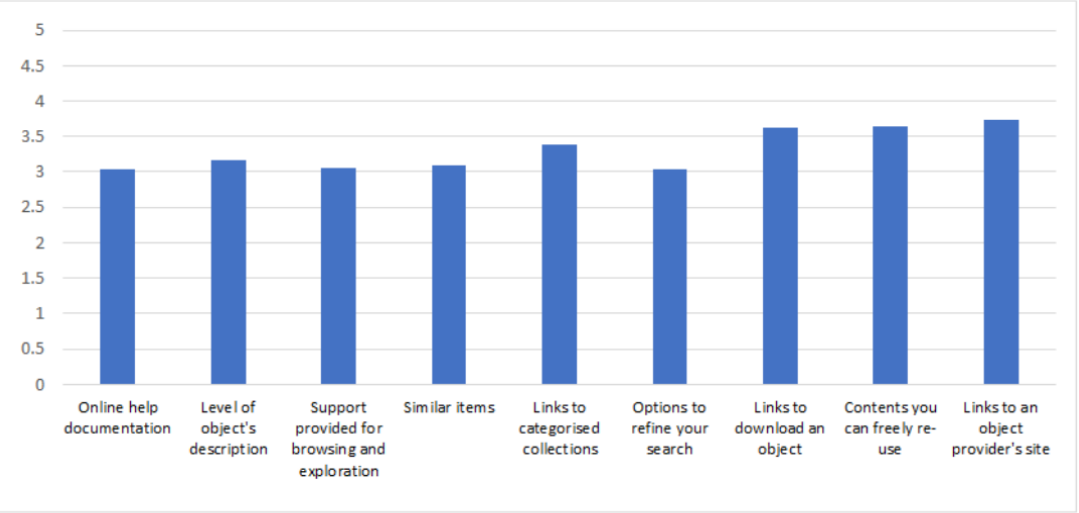


Figure 9. User assessment on the degree of support for search.

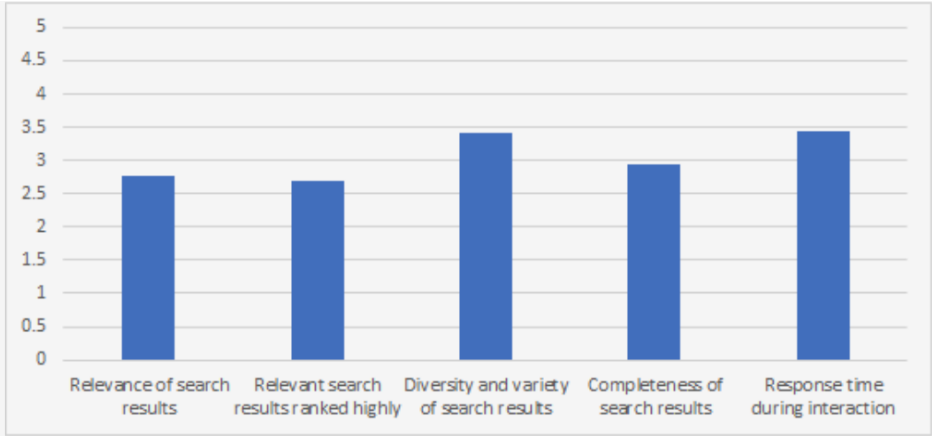


Figure 10. User assessment on the quality of search.

Question	Task 1	Task 2	Task 3
How familiar are you with the given topic?	1.92	3.39	3.39
How easy was the task to complete?	3.55	3.25	3.37
How useful was the information you found?	3.51	3.08	3.18
How would you rate the success of Europeana in helping you complete the task?	3.39	3.06	3.02

Figure 11. Average score (mean) of post-task questionnaire to analyse the search success. Predefined topics per task: 1) specific item search, 2) specific subject search, 3) exploring.

As part of the controlled-task evaluation carried out, the correlation between users' search behaviours and some of the questions concerning satisfaction was also analysed [65] (How easy was the task to complete? How useful was the information you found? How would you rate the success of Europeana in helping you complete the task?). Several specific logged behaviours were found to be correlated (statistically significantly[66]) with the last question related to the success in carrying out the task in Europeana. The results can be seen in Table 3. This information is highly relevant in order to evaluate the satisfaction of a search component from the logs.

Table 3. Statistically significant correlation between logged behaviours and success (the values are between -1 and 1, -1 and 1 meaning strong correlation, and 0 no correlation).

All tasks	Time spent carrying out the tasks	-0.31
	Number of interactions	-0.27
	Number of searches	-0.25
	Number of result pages viewed	-0.22
	Number unique queries	-0.32
	Removal collection filters	-0.26
	Number of multiple searches carried out consecutively	-0.22
	Number multiple results pages viewed consecutively	-0.25
	Number of activities viewing results pages followed by search	-0.25
Task 1	Time spent	-0.42
	Nº interactions	-0.37
	Nº result pages viewed	-0.36
	Nº multiple result pages viewed consecutively	-0.38

[65] Accessible at: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-3/Europeana%20DSI-3_C.2%20Usage%20pattern%20report_M12.pdf

[66] At the time evaluation, the strength of these statistically significant correlations was assessed as 'weak' or 'moderate' at best. In the document here we refrain from such interpretation, as thresholds may vary across evaluation frameworks.

Task 2	Removal collection filters	-0.46
Task 3	Nº unique queries	-0.45

C.2 Current evaluation

C.2.1 Methodology and datasets

Currently we run periodic evaluations over the search components in production, and we measure usage and effectiveness based on the behaviour of the users (see section C.2.2).

In the past, we also created gold standards to measure effectiveness. One was created by the 904Labs in 2016[67]: based on Europeana's logs, they estimated the relevance of the documents given a query by taking into account the click-through rate (ratio of users clicking on the same item for the same query) and the position of the results (the higher in the rank, the higher the probability to be clicked). Unfortunately this corpus is no longer useful as we don't have a snapshot of the collection when it was created: many objects are no longer in Europeana (or the identifiers have changed), and many others have been added.

Another gold standard was created for the Cultural Heritage track in CLEF Workshop (CHiC 2013[68]), which, among other tasks, was used to measure the effectiveness in cross-lingual search. It comprises a snapshot of our Metadata Collection (but with only a few selected fields) grouped by language, a set of topics (description and queries), also in different languages, and the relevance of records per topic. This gold standard may still be used for new experiments on cross-lingual search at Europeana.

Finally, last year we developed code and created dashboards in Google Analytics to get information about the queries our users issue, for the whole collection or by thematic collection. In general this information is crucial for the pre-implementation stage of any component, and actually the results have been already used for the experiments on ranking using BERT (section 2.5), and for the experiments on multilinguality (section 3).

C.2.2 Current metrics and KPIs

The current evaluation follows the recommendations given in the 2017 Evaluation Framework, and reports on several metrics, focused on the effectiveness and usage, for the main search components (see Table 4). The data used to calculate those metrics mainly come from the logs, and the main assumption to calculate performance metrics is that a clicked document is always relevant. As using low-level signals, especially just clicks, for tasks such as tuning relevance ranking is problematic, Europeana

[67] https://drive.google.com/drive/folders/1QNHcJjN_onLCaI_3GRMZPBv66WER8MAr
[68] https://link.springer.com/chapter/10.1007/978-3-642-40802-1_23

complements this evaluation with user-oriented evaluations (conducted and reported by the Service Experience team[69]).

Table 4. Current metrics for search component evaluation, taken from the 2017 Framework (table 1).

Components	Evaluation criteria	Metrics
Search box	Retrieval effectiveness	- Relevance-based metrics (nDCG,MRR, precision@10) - % queries with clicked results
Autosuggest	Effectiveness	- Relevance-based metrics at n characters (nDCG)
Filters	Effectiveness	- % queries with filters with clicked results
	Usage	- Frequency of use
Similar items	Effectiveness	- Relevance-based metrics (nDCG)
	Usage	- Frequency of use
Entity Collection	Coverage	- % user searches matched to an entity from the Entity Collection - records in Metadata Collection with at least one entity

C.2.3 Main Findings

The results of the evaluation carried out on a periodic basis during the last two years show similar results. The effectiveness is similar for the general search (nDCG slightly above 0.5), and slightly worse for the similar items search (nDCG dropped from 0.66 in April 2018 to 0.51 in April 2020). The usage of the similar items search is also worse (from 0.24% to 0.054% of the users). The usage of the filters remains similar, as well as the coverage of the Entity Collection for our queries (in around 53% of the queries with entities, those entities are included in the Entity Collection).

In view of these results, we should review the design of the similar items feature (already included in action M-F12. Item suggestions), as well as analyse why the number of queries with clicks is traditionally very low (10.2% in April 2020). Besides that, we suspect that we are also reaching the limits of the simple evaluation method we have used so far, so additional options are proposed in this report to overcome this issue.

[69] See the test catalogue at https://docs.google.com/spreadsheets/d/1G2tIJLTO4mY-slGIQK6kbleKuWMgFgepqrY4_KJ8jKo/