europeana

Europeana – Core Service Platform

MILESTONE

MS30 – SEARCH IMPROVEMENT PLAN

Revision	5 – Final
Date of submission	30 October 2015
Author(s)	Timothy Hill, Antoine Isaac, Valentine Charles, Europeana Foundation
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	07-10-2015	Timothy Hill	Europeana Foundation	First draft of a search improvement plan
2	15-10-2015	Timothy Hill	Europeana Foundation	Version 1 of a search improvement plan after comments from Antoine Isaac, David Haskiya from Europeana Foundation
3	29-10-2015	Timothy Hill	Europeana Foundation	Version 2 of a search improvement plan including detailed milestones
4	29-10-2015	Antoine Isaac	Europeana Foundation	First draft of MS30 folding elements of the first search improvement plan and additional R&D items
5	30-10-2015	Antoine Isaac and Valentine Charles	Europeana Foundation	Final version

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise.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

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

1. Introduction

Since the Europeana v2.0 project, Europeana has devoted a lot of effort to enhance its search services, notably by enhancing the richness of our data and the search indexes that are built on them.
Continuing this effort, this milestone presents work planned to enhance the performance of Europeana's search services as part of the remainder of the DSI project, i.e., until May 2016.
It is a result of collaboration between Europeana's R&D, Product Development and Technology teams. This effort will support the needs of end-users, re-use and data providers within the Europeana DSI platform.

NB: In line with EuropeanaTech's general openness to feedback and suggestions, we provide for the coming months a version of this milestone that is publicly open for comments at https://docs.google.com/document/d/1xRZYhl1KLIlhGitMD2fymdJ-go4XzTlbhziHbSZc8Og/

2 Core product development for improving search

Milestone 1 - Coarse Tuning of Search

The majority of the current Europeana users explore Europeana via keyword-search. This operation is well-understood by users and remains the main means to discover Europeana objects. However the retrieval of known items results is still poor. We will work on improving results retrieval by re-weighting the fields used for search. We will investigate a manual approach but also machine-learning approaches such as the one developed by 904Labs[1].

Objective	Improve result-list ordering and "More Like This" functionality within the limitations of Solr configuration changes to field-weighting and "More Like This" fields.
Approach	'Crowdsourcing' of multilingual queries; development of tool to re-weight queries
Deadline	December 2015
Success criteria	Improvement of at least 40% for nDCG measure[2], as averaged across all queries in testbed using the 904Labs Search Evaluation framework or similar.

A testbed including all languages of interest is a precondition for this milestone. The development of this testbed will be 'crowdsourced', in that a webform will be sent to selected individuals (predominantly Europeana staff) inviting them to supply multilingual queries and rank the results. Query reweighting will be tested using a specific tool to facilitate this activity, under development. "More Like This" functionality will be tested using a similar tool to apply different "More Like This" configurations, also under development. As part of this milestone the BM25f plugin originated

[1] http://904labs.com/

[2] nDCG (normalized discounted cumulative gain) is measure of ranking quality. It measures the usefulness and relevance of a document based on its position in the result list (https://en.wikipedia.org/wiki/Discounted_cumulative_gain#Normalized_DCG)

from development in the ASSETS project[3] will be integrated into Europeana's new SolrCloud setting. In addition, the Solr text-analysis chain will be reconfigured to improve stopword-handling and English-language stemming.

Milestone 2 - Fine-tuning of Search

The next step for improving the search will be to improve the query formulation and ranking of the results. Note that the ranking of the results will take into account the completeness of the metadata available in Europeana and the presence of thumbnail images within the metadata. This work will imply the definition of metadata completeness criteria which will be developed considering Europeana metadata requirements and the requirements for end-users services.

Objective	Improve result-list ordering by improving query formulation and by ensuring that ● Items with more-complete metadata are returned before those with less-complete metadata ● where applicable, items with thumbnail images are returned before those without.
Approach	Application of 904Labs query-tuning technology (or equivalent); Solr plugin development work
Deadline	February 2016
Success criteria	Improvement of at least 10% for nDCG measure, as averaged across all queries in testbed; specific criteria (see below) for ranking are met.

904Labs has developed a brute-force engine that tests identical terms with multiple query configurations to determine the configuration that maximises nDCG scores. This or similar FLOSS technology will be used to improve query formulation. Where specific ranking criteria are given (more-complete metadata preferred over less, thumbnail-image presence preferred over absence), automated tests must be created to ensure they are being met.
This milestone will require analysis of query logs for the new portal, with particular attention paid to geographical queries. We will also explore questions of core configuration and implement a spelling correction (‘Did You Mean …?’) functionality.

Milestone 3 - Entity-Autocomplete Implementation

A previous study run by 904Labs showed that Europeana’s users are mostly looking for entities (Agents, Concepts, Places, Time periods, etc.) when searching in the Europeana Collections portal. In terms of search this means that while improving traditional means of search (keyword-based), Europeana will also develop a more semantic search. The first step will be to implement an Entity based search autocompletion. These efforts will be enabled by Europeana’s underlying data model which allows for the description of contextual entities[4].

[3] D2.2.1 SPECIFICATION OF POST QUERYING PROCESSING FUNCTIONALITIES http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ASSETS/Deliverables/D2.2.1%20Specification%20of%20post%20querying%20processing%20functionalities.pdf

[4] http://pro.europeana.eu/edm-documentation

Objective	Refactoring, extension, and integration of entity-autocomplete functionality
Approach	Reworking of code at https://github.com/europeana/entity-autocompletion
Deadline	April 2016
Success criteria	Entity-autocomplete functioning, as measured by unit and integration tests, and with most-relevant entities appearing topmost

Entity-autocomplete should function as shown in the demo realized in the past months together with CNR-ISTI[5], but with:

- improved multilingual support as outlined in the White paper on Best Practices for Multilingual Access to Digital Libraries[6] (and with particular reference to the section on autocompletion)
- inclusion of Best Bets (reference objects that are presented as first result for specific queries)
- loose matching to support correction of misspellings
- a maximum response time of 300ms per keystroke

As part of this milestone, we will also progress on implementing an Advanced Search page incorporating Entity Autocomplete.

Evaluation
Implementation of the above improvements should be followed by an evaluation phase spanning the final month of the plan period (31 May 2016).

Evaluation is complicated by the fact that the work on autocompletion involves changing not just the technical implementation of search, but how users formulate searches in the first place. As a result, like-to-like comparison of query responses before and after implementation (or with another platform) is insufficient, and more coarse-grained metrics must be used.
The most important of these are:

- before-and-after average nDCG of new-portal result lists
- proportion of searches made using autocomplete entries, and nDCG of such searches
- proportion of searches made using the Advanced Search form, and nDCG of these
- 'bounce rate' (the proportion of users who leave a page without interacting with it) from the search results page, as indicated by Google Analytics
- qualitative analysis of our query logs, in particular of query reformulation

3. Other relevant product development items

While the core of the development work directly targeted at improving search performance will take place in the above described milestones, we will also devote efforts to other development work that will base and/or accompany the core effort. These items will impact the way search is

[5] Temporary available at http://node5.novello.isti.cnr.it:8888/

[6] This paper has been previously published as a Europeana V3 milestone (MS12, available at http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Milestones/Ev3%20MS12%20Multilingual_Access%20White%20Paper.pdf). It has been re-worked and will be published in the coming weeks. A temporary draft can be seen at https://docs.google.com/document/d/1pxp3JS72odvUc2PYn3-0gwXY-8WrmJhzbcKm9wnNueM/.

made (e.g., diacritics) or make available new components to enhance existing search processes (such as Entity API or Media API)

Search-related development defined in D6.1 Requirements for the platform backend[7]:

- Diacritics in search: the Search Engine should have full support for different languages and dialects and should make no difference betweens search for accents and non-accents. (May 2016)
- Translation in search: the Query Translation API should become part of the regular Search API (March 2016)
- Search using Media API: The API should include the ability to search, filter on and retrieve technical metadata for images, texts, sounds and video. (Fall 2015)
- Search in annotations: the Annotation API must support indexing of annotations to allow for search of annotations (May 2016)
- Entity Collection Database and API: the Entity API should use the search index which contains the Europeana Entity Collection (May 2016)

“Meta” development on search:

- Stabilize Europeana's logging environment (February 2016)
- Document the search mechanisms applied by Europeana, e./, by extending the current documentation effort on identifying duplicates for best bets. (May 2016)

4. R&D for improving search

In this section we list more explorative items, which we will investigate as R&D actions to contribute to the above milestones or to future product search improvement development.

Specifications for extending the Entity Collection database

We will explore how the Europeana Entity Collection database mentioned in the previous section can be populated with new, richer datasets of contextual entities. This work will likely include defining new mappings and maybe extensions to the Europeana Data Model (EDM). It will also require to automatically enrich the Europeana dataset with links to the newly introduced resources, which in turn depend on making appropriate updates to the Europeana semantic enrichment framework. The main candidate target datasets for inclusion are:

- Wikidata[8]
- PeriodO dataset for time periods[9]
- Pleiades for historical places[10]

Next to these candidates, we will have to select a reference dataset for cultural objects.
We will prioritize our choices based on mining what users are really interested in (using logs) and according to the criteria identified in the newly published report of the EuropeanaTech Task Force on Evaluation and Enrichments[11].

[7] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d6.1-requirementsfortheplatformbackend.docx.pdf

[8] https://www.wikidata.org/wiki/Wikidata:Main_Page

[9] http://perio.do/

[10] http://pleiades.stoa.org/

[11] http://pro.europeana.eu/taskforce/evaluation-and-enrichments

Basic exploitation of Entity Collection for ranking and browsing search results

Two options have been retained for exploring a disruptive approach of exploiting the entity collection in ranking and browsing search results:

- Aggressive matching for search results: any time there is a 50% match between a query and a (Wikidata) Collection entity, put the corresponding entity on top of our results.
- Browse entry points based on Wikidata: clustering the entity collection based on domains and entity types (e.g., ‘Paintings’, ‘Letters’, ‘Pottery’, ‘Photographs’ for CHO, ‘Composers’, ‘Libraries’, ‘Gods’ for Agents, etc.), to provide with representative entry points on selected domains (e.g. for thematic collections like Art History and Music)

Identifying and exploiting semantic patterns for search

The information contained as semantic networks in the Entity Collection needs to be massaged into Solr-like search indexes. There are several alternatives, as exemplified for example in the options (and issues) for the index 'augmentation' that follows semantic enrichment identified in the report of the EuropeanaTech Task force on Multilingual and Semantic Enrichment Strategy.[12] We will seek inspiration here on the functional side in the approaches of, e.g., the old Semantic Search Thought Lab[13], the Crotos[14] demonstrator, or the more recent ResearchSpace[15], Food&Drink semantic demo[16] and Linked Open Images prototypes[17]. The German Digital Library (DDB) also links from their metadata structures to ‘front end’ ontologies.

Technology-wise, the work by Pedro Szekely at ISWC 2015 on using ElasticSearch with RDF[18] could be relevant, as well as the combination of the Apache tools Stanbol[19] and Marmotta[20]. From the perspective of data patterns, the work could start with enriching indexes using alignment between entities (owl:sameAs, skos:exactMatch, etc). This work could be connected to the requirements of the ‘similar items’ function and modeling of roles in EDM.

Using metadata from annotations and content/media

When a significant amount of human ('crowdsourced') annotations and technical data attached to web resource (a.k.a 'content' or 'media') become available to Europeana, components used for improving search such as completeness measure could be extended by elements that are derived from this data.

Search taking into account hierarchical objects

Similar to semantic networks in the Entity Collection database, there are also networks in the Europeana 'basic' object metadata, which could also be exploited for getting more complete search results. This especially includes hierarchical links between objects, where some metadata fields could be copied in the search index from parents to children or the other way around.

[12] http://pro.europeana.eu/taskforce/multilingual-and-semantic-enrichment-strategy

[13] http://labs.europeana.eu/apps/searchengineeuropeana

[14] http://zone47.com/crotos/

[15] http://www.researchspace.org/home/project-information/design

[16] http://efd.ontotext.com/app/search?query=&limit=24&offset=0&dataProvider=Alinari&place=Roman_Empire&category=Beverages

[17] Other relevant prototypes include data.bnf.fr, ECLAP linked open graph, Fundacion Larramendi Polymath virtual library, OCLC-Europeana clustering experiments (paper, slides), Work done by PATHS (in different facets of their work might be interesting in this context), FRBRization prototypes like Linked Jazz or Scherzo), Prototype Linking Lives from the Archives Hub , SNAC project, CultureSampo from SeCo group, JocondeLab,

[18] http://usc-isi-i2.github.io/dig/

[19] http://stanbol.apache.org/

[20] http://marmotta.apache.org/

Ranking representing diversity of collections

Search results in Europeana often present large chunks of very similar objects or objects coming from a same collection, hiding the complexity of the entire Europeana object base. We plan to explore options to make the current search results more representative. This could include data provider information, after they have been mapped to a reference list, either one maintained by Europeana as part of its CRM system, or (most likely) a third party one derived from e.g. the ISNI[21] Linked Dataset. Plans for such alignments are likely to be decided in the context of WP1 later in this project, which would make this item more suitable for investigation in a follow-up to this project.

Language detection for queries and metadata

Recognizing the language of strings used as queries or values in object metadata is a key success factor for search in multilingual context such as Europeana's (see MS32 Multilingual research and development plan). Too often search and metadata enrichment have to be applied in a setting where language is unknown, e.g. if there is no language tags in textual metadata fields for object. If time allows it in the remaining months of this project, we will explore using Apache Tika or others (e.g. LOCloud) services for language detection[22], and then exploit results for better ranking of searches (e.g., privileging objects for which there was a match of language).

Configure Solr text analysis chain in a language-specific way

Solr NLP functions are currently deployed in the same setting for every language[23]. Ideally, we could use different analysis chains for different languages. As this is most likely not possible for resource reasons, we could at least provide multilingual stopwords, and perhaps exploit Solr's synonyms.txt field for multilingual support. The desirability of implementing the ICU Normaliser 2 for Solr should be assessed. This will however probably require too much work to be investigated over the remaining months of this project and will have to span over follow-up initiatives.

Meta-ranking

Instead of fine-tuning one single ranking mechanism, which require a lot of research and may lead to incomplete coverage of enhancements, improving search rankings can be done by combining different rankers, possibly by machine learning, as experimented by 904Labs. This task is promising, even though in the remainder of this project our activities will most likely be focused on finding a suitable partner to investigate these matters in a follow-up project.

Evaluation improvement

Several options to fine-tune the evaluation methodology should be researched:

- Tune evaluation relevance measures. Currently the nDCG measure mentioned in previous sections may punish more explorative queries too much, while these have been identified to meet relevant user needs.
- Explore different alternatives to measure user satisfaction on the portal, notably by identifying successful sessions, an issue that has hindered previous evaluation efforts in Europeana[24].
- Keep maintaining a list of relevant Europeana-related evaluation work.

Maintain a list of (anecdotal) search issues

Maintaining a list of specific search issues might help Europeana to identify deeper concerns from stakeholders. For instance the European Commission would like searches for canonical

[21] http://www.isni.org/

[22] https://europeanadev.assembla.com/spaces/europeana-ingestion/tickets/1589-investigate-feasibility-of-using-apache-tika-for-language-detection/details#

[23] https://www.assembla.com/spaces/europeana-npc/tickets/643-reconfigure-solr-text-analysis-chain/

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

creative works to have recall across all European languages, also with partial matches. E.g., a search for 'ragazza perla' should return the famous Vermeer painting even if the full title is 'Ragazza con l'orecchino di perla'.

Explore ways to mine and exploit query reformulation strategies from users
The analysis of the Europeana Logs done by 904Labs reported lots of cases of query reformulation made by users to refine their search and filter results.

```
jugendstil vaas (7 reformulations, numfound: 176)
        jugendstil vase   (1719 numfound)
        jugendstil vase   (1719 numfound)
        jugendstil klei vazen   (0 numfound)
        bakeliet jugendstil vazen   (0 numfound)
        jugendstil vazen   (16 numfound)
        jugendstil vazden   (0 numfound)
        jugendstil vazen   (16 numfound)
```

These cases are interesting as they can be used as a motivation for further work on contextualisation, using semantic links and multilingual data to augment indexes or queries. In the same way this work could be used to fine tune facets or clusters of results to help a user finding what she wants in a faster manner.

Create a EuropeanaTech task force on search and the evaluation thereof
So far very few discussions about search took place within the EuropeanaTech community. Creating a new Task Force would allow Europeana to get more insight on the issues encountered (and solutions found) by other data providers within their digital libraries.

5. Other relevant R&D work

Some R&D activities that are carried out independently of search product development will nonetheless eventually benefit to search performance. Especially these who result in better metadata and measures to be used in Europeana's search functions.

Involvement in helping aggregators send better quality data
The involvement of data providers in improving the quality of their data will be key to increase the repercussions of the search improvements. As described in *D1.3 Description of planned work for the aggregators on improving the data*, we will work on the following:

- publication and use in EDM of (multilingual) vocabularies as (SKOS) linked data for DSI aggregating partners such as EFG, MUSEU, EUscreen, OpenUp.
- methodological help as well as Europeana's ingestion of automatic or manual metadata enrichment made by DSI aggregating partners (Europeana Sounds, Europeana Food&Drink, CARARE)
- facilitate evaluation of enrichment by DSI aggregating partners
- play a key role in a Data Quality Committee soon to be launched by WP1

Support DPS team with improvement of metadata on the Europeana side
Work on enhancing the current metadata enrichment process and on developing data normalisation rules will further improve the quality of the data in Europeana and therefore enhance the search.
We will particularly work on defining criteria and metrics to measure the quality of data including:

- measures for the quality of enrichments,
- a new data completeness measure,
- a list of vocabularies used in Europeana data

- an evaluation methodology and a corpus of data to be used for evaluation purpose

Push production of alignments between contextual entities.
In order to increase the amount of links between metadata and vocabularies or between vocabularies themselves, Europeana will work on the creation of more alignments between contextual entities. For instance, experiments with the CultuurLink[25] tool with the Europeana Sound project will demonstrate whether we are in a position to align multilingual datasets onto a pivot vocabulary (the MIMO vocabulary for musical instruments in this case).
We will continue population the Entity collection in order to enable services like the entity based search. On a longer term the semantic networks created with Europeana “semantic layer” should support a more network/graph-based search.

Metadata translation experiments
Europeana will investigate translations tools to translate its metadata and will also encourage providers to apply automatic translation solutions prior to sharing metadata with Europeana - the closer translation is applied to the source of metadata, the better the results may be. Europeana is currently collaborating with several partners in this direction:

- We have signed a memorandum of understanding[26] with the Latvian Ministry of Culture and the Tilde company for ingesting Latvian records with automatic translations to English.
- We are discussing with Galway's Insight Centre for Data Analytics[27] and the National Library of Ireland about translating some of Europeana's Irish datasets.
- Feasibility of providers or aggregators using CEF AT, not only Europeana
- We have started discussing with the FREME project on testing their localization APIs on Europeana cases.

Europeana data in Peripleo[28]
Europeana will seek to participate in an "integration pilot" with the Pelagios team. This pilot will be an opportunity for Europeana to try linking its dataset to the Pleiades dataset. If succeeding this new alignment will enrich the Europeana dataset with geographic information (ancient place names, coordinates...).

6. Conclusion

While presenting clear and achievable milestones, this plan reflects a more ambitious strategy, which on a longer term will help Europeana fully move from a traditional search to a more semantic search. Improvements on the search mechanisms are needed but to be successful they will need to be supported by R&D efforts on the quality of the data, the refinement of the data model and more importantly on the data enrichment. This plan shows that many efforts have already started and will continue as part of fruitful collaboration with experts in the domain.

[25] http://cultuurlink.beeldengeluid.nl/app/#/
[26] http://pro.europeana.eu/blogpost/latvian-ministry-of-culture-and-europeana-sign-memorandum-of-und
[27] https://www.insight-centre.org/
[28] http://pelagios-project.blogspot.co.at/2015/07/peripleo-sneak-preview.html