Europeana – Core Service Platform

MILESTONE

MS31: Report on the improvement of search

Revision	4 - Final
Date of submission	28 June 2016
Author(s)	Timothy Hill, Antoine Isaac, Valentine Charles, Nuno Freire, Hugo Manguinhas, 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	06-06-2016	Timothy Hill	Europeana Foundation	First draft
2	16-06-2016	Valentine Charles, Nuno Freire, Antoine Isaac, Hugo Manguinhas	Europeana Foundation	Minor amendments and cross-references to other documents made by Charles, Freire, Isaac, and Manguinhas
3	18-06-2016	Timothy Hill	Europeana Foundation	Incorporation of comments and information provided by Remy Gardien and David Haskiya.
4	28-06-2016	Timothy Hill	Europeana Foundation	Final draft

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise.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

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

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

1. Overview

The document MS30: Search Improvement Plan[1] outlines a number of steps to be taken through DSI1 to improve the search performance of the Europeana Collections platform. This document reports on progress made with regard to each of these steps as of 30 June 2016.

A common theme in the progress reports made below is underperformance against targets owing to current technical limitations of the Europeana platform. Where appropriate these limitations are explained for each work item and the steps being taken to address them are described.

Another repeated theme is the dependency of search upon metadata: the raw material any search algorithm operates upon is, ultimately, the content that is searched, and the higher the quality of that material, the more accurate the search operation becomes. Considerable effort has accordingly been focused on metadata quality improvement, in particular through the work of the recently-formed Data Quality Committee (DQC), the aims of which are defined in terms of discovery and information-retrieval requirements.

Note that this document deals only with relatively short-term and ongoing work related to incremental improvement of search. For longer term, larger scale directions, a more comprehensive and holistic Search Strategy[2] has been developed.

2. Core Product Development for Improving Search

Milestone 1: Coarse Tuning of Search

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,[3] as averaged across all queries in testbed using the 904Labs Search Evaluation framework or similar.

This milestone consists of three sub-tasks.

1. Creation of a multilingual query testbed for evaluation purposes

One difficulty in evaluating the performance of the Europeana Collections platform with regard to multilingual queries has been a lack of test queries in non-English languages. In order to address

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms30-search-improvement-plan.pdf

[2] http://pro.europeana.eu/publication/europeana-search-strategy

[3] 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)

this, a crowdsourcing tool called Volgus[4] was created, asking participants to supply queries in their own languages.

The results of this crowdsourcing, taken in combination with the test queries harvested from 904Labs' investigation of Europeana's logs in 2014/15,[5] means we now have sufficient query coverage for all of the core (English, French, German, Italian, Polish and Spanish) and secondary (Dutch, Hungarian, Portuguese, and Swedish) languages from the EuropeanaConnect project. In addition, we have good testbeds for Bulgarian, Catalan, Czech, Danish, Galician, Greek, Hebrew, Latvian, Maltese, Romanian, Slovene, Latin, and Welsh and some coverage of Albanian, Basque, Croatian, Danish, Finnish, and Serbian.[6]

There is currently no coverage for the following languages for which Europeana has metadata (though not all are official EU languages): Estonian, Irish, Lithuanian, Norwegian, Slovak, and Yiddish

Further actions

Outreach to partner institutions and colleagues with an interest and expertise in these languages.

2. Search-field reweighting

As planned in MS30, a simple application called RankFiddle[7] was created to enable easy reweighting of fields and visualisation of search results. Using this tool, a set of reweightings was derived.[8]

Unfortunately, implementation of these weightings required the use of Solr's EDisMax (Extended Disjunction Maximum) Query Parser,[9] which proved too heavy in its requirements for the existing Production SolrCloud deployment to support. It was accordingly decided to redeploy the weighted BM25f Solr plugin earlier developed for Europeana.[10]

Use of the plugin appears to have boosted the average normalized Discounted Cumulative Gain (nDCG) by 24.1% using the testbed created for 904Labs' search evaluation framework. This is below our target of 40%. However the fact that weightings for the BM25f plugin were derived two years ago points at immediate remedial actions (see below).

Technical limitations

As noted above, the reweightings derived from RankFiddle could not be deployed because of the inability of Europeana's SolrCloud installation to support the processing needed.

In addition, the application of Learning To Rank (LTR) approaches to reweighting has been hindered by the lack of an appropriate logging framework. In the first instance, this was because of the launch of Europeana Collections; subsequently, limitations to the previously-implemented logging framework were identified, and the decision was made to adopt a new framework.

---

[4] Deployed at http://144.76.218.178:8000/volgus/; code at https://github.com/europeana/search/tree/master/searchsearch/mobsource

[5] https://europeanadev.assembla.com/spaces/europeana-r-d/documents/d5e2a2ng0r5OkxacwqEsg8/download/d5e2a2ng0r5OkxacwqEsg8

[6] The list of sample evaluation queries can be found at https://www.assembla.com/spaces/europeana-r-d/documents/dic7IYniSr5R0Vdmr6CpXy/download/dic7IYniSr5R0Vdmr6CpXy.

[7] Deployed at http://144.76.218.178:8000/rankfiddle; code at https://github.com/europeana/search/tree/master/searchsearch/rankfiddle

[8] https://docs.google.com/document/d/1OnhjiIDk545bWoLEdIRY4hWpgvhISldCCyPF53v99K8/edit

[9] https://cwiki.apache.org/confluence/display/solr/The+Extended+DisMax+Query+Parser

[10] https://github.com/europeana/contrib/tree/master/bm25f-ranking

Further actions

Although the BM25f architecture allows weightings to be set arbitrarily, at the moment this cannot be done dynamically, as a tool like RankFiddle demands, without a significant expansion of our Solr server capacity. Rather than doubling (or more) our Solr commitments, it seems simplest to extend the Solr plugin to support dynamic reweighting and phrase-boosting in the same fashion that EDisMax does. We will then be in a position to use RankFiddle settings to configure our SolrCloud instance as originally intended.

Over the past three months, Europeana has moved to address its logging deficiencies, with a unified ElasticSearch - Logstash - Kibana (ELK) stack for logging. Assessment of this framework indicates that it will need to be complemented by a minimal client-side logging implementation.

3. "More Like This" Reweighting

Parallel to RankFiddle, a tool called MLTFiddle[11] has been developed and deployed. New weightings, however, have yet to be tested and agreed upon.

Further actions

New weightings need to be agreed upon and then deployed to the Europeana site.
Note that no benchmark for "More Like This" clickthroughs has been developed. The metric used might be very simple ("More Like This" item clicks), or more complex, based upon current work with searching by image similarity.[12] Determining and implementing a relevant benchmark, however, is dependent upon configuration of the new ELK logging framework.

Milestone 2: Fine-tuning of Search

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.

Fine-tuning of search was to be accomplished through increased weighting of documents which had either (a) high completeness values; (b) thumbnail images available; or (c) both.

[11] Deployed at http://14.76.218.178:8000/MLTFiddle; code at https://github.com/europeana/search/tree/master/searchsearch/mltfiddle

[12] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Creative/WP2%20-%20Infrastructure%20for%20Content%20Re-use/An%20Image%20Similarity%20Search%20for%20the%20European%20Digital%20Library%20and%20Beyond.pdf

Preliminary investigation of weighting for completeness, however, had the counterintuitive effect of substantially *lowering* nDCG. The reasons for this is at the moment unclear, though a number of causes may be hypothesised. First, the completeness metric available on Europeana records is at the moment poorly implemented. Second, the metadata displayed on the Search Engine Results Page (SERP) is minimal, and completeness may therefore be of limited relevance at this stage in the navigation process. Third, user perceptions of completeness may be a more complex phenomenon than is reflected in the metric (so that, for instance, it is not considered salient beyond some minimum threshold of acceptability).

Favoured weighting of items with thumbnails proved impossible because of limitations imposed by the BM25f-handler: specifically, the handler is at the moment unable to boost multivalued fields, such as those used to store thumbnail-image filepaths.

Also mentioned under this heading in MS30 was the implementation of spelling suggestions for search. This remains unimplemented, though the urgency of this is mitigated by the inclusion of fuzzy matching in the Entity Autocomplete functionality.

Technical limitations

As noted above, the BM25f handler is unable to boost fields that may hold more than one value, such as edm:preview, and this limitation currently acts as a blocker for thumbnail-based boosting.

Further actions

The Data Quality Committee has been working intensively on the problem of adequate measures for metadata completeness (see in particular the work done by Péter Király[13]). Once this work is finalised and the resulting completeness metric(s) made available on all Europeana metadata, further investigation of completeness-based boosting should be undertaken.

The BM25f-handler needs to be extended to allow for the possibility of boosting based on thumbnail availability. This should be undertaken as part of the work to allow dynamic reweighting.

Milestone 3 - Entity-Autocomplete Implementation

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

Previous evaluation of Europeana's query logs indicated that users were mostly searching for entities rather than documents *per se.* In order to support this behaviour better, Europeana has created an Entity Collection datastore.

[13] http://pkiraly.github.io/2016/01/15/second-report/

From a user perspective, the first point of entry to this datastore is an autocomplete functionality for the searchbox, supplying the names of entities in the collection as the user types. This entity-autocomplete functionality has been through three iterations,[14] and currently supports:

- Autocompletion in all official and semi-official EU languages
- Loose matching to support misspellings
- Relevance rankings returning most-popular items first
- Response times < 300ms per keystroke[15]

Technical limitations

The Entity Collection itself has not yet been integrated with the rest of the Europeana platform on a production level. Testing its effects upon search generally cannot proceed until this has been done and a strategy for semantic enrichment developed.

Further actions

The Entity Collection needs to be integrated with the Europeana platform, and further evaluation performed once logging is in place.

3. Other relevant product development items

The following items under this heading from MS30 are either complete or largely complete:

- **Diacritics in search:** improvements to the schema.xml document for Solr have been made such that the presence or absence of diacritics is now irrelevant to search in all fields where this is appropriate. This new schema.xml document is awaiting testing and will be deployed in the summer of 2016.
- **Search using the Media API:** the Media API[16] is now fully integrated into the Search API.
- **Entity Collection Database and API:** As noted above under Milestone 3, we are currently in our third iteration of work on the Entity Collection.
- **Stabilize Europeana's logging environment:** As noted above, a new logging framework has been put in place, but has not yet been assessed and adapted in relation specifically to search.
- **Document the search mechanisms employed by Europeana:** Search of the Europeana platform is complex, involving several different mechanisms. Internal and outward-facing documentation needs to be created, outlining the query algorithm used (BM25f), the procedure for creating and curating the elevate.xml file, the means by which Similar Items are generated, the role of various copyfields (e.g. ‘who’, ‘what’, ‘when’) in search, and the query syntax supported.

The following are either unattempted or in their early stages:

- **Search in Annotations**: work on this will not begin until 2017
- **Translation in Search:** The Query Translation API[17] is operational, and work on integrating its functionality into Collections Search may proceed in 2017. The existence of the Entity Collection, however, makes implementation of this less urgent than previously.

[14] Code at https://github.com/europeana/search/tree/master/autocomplete.
[15] See the evaluation at https://github.com/europeana/search/tree/master/autocomplete/evaluation
[16] Described further at http://labs.europeana.eu/api/media-search.
[17] http://labs.europeana.eu/api/query-translation

4. R&D for Improving Search

Work on the points under this heading in the Search Improvements Plan is described briefly below.

Specifications for extending the Entity Collection database: Work on this is ongoing. Implementation decisions with regard to the Entity Collection may be informed by Europeana's collaboration with other organisations focused on publishing Linked Open Data - notably Pelagios and PeriodO, both of which have representatives from Europeana on their Advisory Boards.

Basic exploitation of Entity Collection for ranking and browsing search results: Work on this requires further completion of the Entity Collection.

Identifying and exploiting semantic patterns for search: Exploration of this has begun with work on the Entity Collection. The approach so far has been ad hoc, but should become more formalised with time. In addition, work investigating triple-store implementations and their integration with existing workflows may have an impact here.

Using metadata from annotations and content/media: In addition to the pending Annotations work referred to under the 'Other relevant Product Development Items' heading, work is also scheduled to begin in DSI2 on both image- and sound-based similarity search.[18]

Search taking into account hierarchical objects: Following on from the work done by the Hierarchical Objects Task Force,[19] his has been identified as a long-term priority by the DQC,[20] and some requirements analysis has been undertaken.

Ranking representing diversity of collections: The need to represent more fully both the semantic range of search terms, and the different kinds of CHOs that satisfy them (in terms of, for example, medium or object type) in our search interface has been highlighted as a priority by the DQC,[21] and some requirements and implementation analysis have been undertaken.

Language detection for queries and metadata: Preliminary trials[22] indicate that Solr's built-in LangDetect capabilities prove to be roughly 89% reliable on Europeana metadata with regard to the EuropeanaConnect core languages (English, French, German, Italian, Polish, and Spanish).[23] This provides a useful basis for testing with additional languages and integration of LangDetect into Europeana's ingestion chain.

---

[18] Planned work on image-based similarity search is described in Subtask 6.3.4 in the DSI2 Description of Work; audio-similarity is part of the DSI2 work for Europeana Sounds. Both tasks are to be undertaken by partners at the Austrian Institute of Technology.

[19] http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Hierarchical_objects//TF%20report%20V1.0%20PDF.pdf

[20] See specifically the 'Metadata Analysis' for Discovery Scenario 13: https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/edit#heading=h.qht5ygtqvera.

[21] https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/edit#heading=h.qht5ygtqvera

[22] Evaluation code and results can be found at https://github.com/europeana/search/tree/master/language_detection.

[23] For the concepts of 'Core' and 'Secondary' supported languages, see http://www.europeanaconnect.eu/documents/D2.7.1_eConnect-Facilitation%20and%20exchange%20of%20multilingual%20access%20strategies%20to%20digital%20libraries_v1.0.pdf

Configure Solr text analysis chain in a language-specific way: Preliminary analysis has been completed. A test environment for this will be in place by the end of Autumn 2016, with implementation occurring in 2017.

Evaluation improvement: As outlined in the Search Strategy,[24] this has been highlighted as an area of fundamental importance to the platform, and we will begin collaborating with University of Sheffield, Humboldt University (both in DSI2) and others in the definition of a new search evaluation framework for the cultural heritage sector.

Maintain a list of (anecdotal) search issues: As noted in the Introduction, the root cause of problems with search often prove to be problems with metadata quality. This item has accordingly evolved into an ongoing list of problematic data patterns.[25]

Explore ways to mine and exploit query reformulation strategies from users: Work here is dependent upon further progress with the logging framework.

Create a EuropeanaTech task force on search and the evaluation thereof: As noted in the Introduction, this has evolved into a task force on metadata quality, the Data Quality Committee (DQC).

5. Other relevant R&D work

Involvement in helping aggregators send better quality data: Recommendations here are expected to emerge from the work of the DQC, as described above.

Support the Metis Scrum team with improvement of metadata on the Europeana side: As described under subtask 1.6.4 of the DSI2 Description of Work, Europeana is committed to undertaking:

- Normalisation and deduplication of organisations providing data to Europeana (edm:provider, edm:dataProvider), incl. ensuring all organisations have a unique identifier
- Normalisation of the values in dc:language
- Normalisation of date information for specific datasets or data partners, (e.g. eFashion)

Push production of alignments between contextual entities: Consideration of this question forms part of the ongoing work on the Entity Collection curation plan.

Metadata translation experiments: A multi-pronged approach has been taken here, and work has progressed along a number of fronts. A number of collaborations - notably with the Latvian Ministry of Culture, the National Library of Ireland, and others - are enabling Europeana to explore possibilities for strong support of particular languages, while the utility of digital tools such as BabelNet has also been investigated. We also started discussions with the CEF AT DSI. Also the Entity Collection will play a key role with respect to metadata translation: it gives us a powerful mechanism for maximising the value of translated metadata, and could be used to support translation services.

Europeana data in Peripleo: Europeana's partnership with Pelagios has strengthened, with the addition of Antoine Isaac to the Pelagios Commons Advisory Board and Tim Hill acting as Coordinator of the Pelagios Commons Linked Pasts Special Interest Group (SIG). The focus of this Group is specifically upon data-modelling for interoperability, and the SIG is expected to provide the framework for further work integrating Europeana data with Pelagios.

[24] http://pro.europeana.eu/publication/europeana-search-strategy
[25] https://docs.google.com/spreadsheets/d/1zoU-1uPk2O5t5zRC1-MD3LakBQGJ2hsWlSnp3XS2iAk/edit#gid=0