Europeana DSI 2– Access to Digital Resources of European Heritage

MILESTONE

MS6.6: Search Improvement Plan

Revision	Draft
Date of submission	2016.12.01
Author(s)	Timothy Hill (Europeana Foundation); Monica Paramita (University of Sheffield); Paul Clough (University of Sheffield); Hugo Manguinhas (Europeana Foundation); Antoine Isaac (Europeana Foundation); Valentine Charles (Europeana Foundation); David Haskiya (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	2016.30.11	Hill, Paramita, Clough, Manguinhas, Isaac, Charles, Haskiya	Europeana Foundation, University of Sheffield	First draft incorporating all comments

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This milestone contains original unpublished work except where clearly indicated otherwise.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.

Table of Contents

1. Introduction

DSI-1 and the first quarter of DSI-2 witnessed several important changes to the context in which Europeana search functionality operates. At the technical level, the shift of Europeana's Solr datastores to a cloud-based deployment is now complete, and the first iteration of the Entity Collection (EC) has been completed, with progress on the second underway. In terms of Europeana content, the work of the Data Quality Committee (henceforth, DQC) is moving towards the publication completeness measures and other outputs directly affecting search. Finally, the collaboration between Europeana, the University of Sheffield, and Humboldt University of Berlin in developing a comprehensive evaluative framework for Europeana search is underway, and is expected to yield its first outputs by the end of DSI-2.

The focus of work for the remainder of DSI-2 is accordingly on developing the technical infrastructure to complement or support these various new developments. In some cases - most notably in relation to Europeana's Solr Cloud deployment - this simply means completing work identified or initiated in the DSI-1 Search Improvement Plan;[1] where relevant, cross-references will be given to that document or to its subsequent progress report.[2] More frequently, however, this involves either the creation of new, or the adaptation of existing, technical infrastructure.

In this latter case, it should be noted that the emphasis of the Plan is largely infrastructural - on providing the technical means to support and sustain search improvements in future more than on optimising some particular metric such as nDCG. In part this reflects the iterative character of many search technologies: clickstream-based machine-learning approaches, for example, demand periodic retuning, and it is just as important that a process exist for these recalibrations as it is to achieve a one-off improvement in search performance. In part, however, it also reflects the evolving nature of Europeana, its content, and its platform: the ongoing work of the DQC and the evaluation partnership, along with the foreseeability of other initiatives with an impact on search arising in the intermediate and long term, make a reasonably generic, flexible, and adaptable search infrastructure a requirement.

Improved nDCG, then, does appear as a success criterion in relation to a number of aspects of the Plan - but not least as an indication that processes are working and producing the required outputs rather than as a maximisation end in itself.

---

[1] MS30 - Search Improvement Plan, available here:
http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms30-search-improvement-plan.pdf
[2] MS31 - Report on the improvement of search, available here: http://bit.ly/2e8TN5O

2. Europeana Solr Cloud: Completion of the Learning-To-Rank Implementation

DSI-1 saw the deployment of the updated BM25f Solr plugin to Europeana's Solr Cloud deployment. This in itself brought a 24.1% increase in nDCG values - a significant improvement, but well short of the 40% target originally set.[3]

The discrepancy in part arises from the weightings set in the BM25f handler itself. These weightings are derived from a machine-learning framework developed specifically to work with the BM25f plugin[4] - but the last time this framework was run was at some point in 2014, since which point the Europeana corpus has grown and evolved considerably.

Work accordingly needs to be focused on restoring this framework to functionality and putting procedures in place to make reweighting simple and capable of being rerun periodically.

2.1 Requirements

1. Logging is largely a solved problem, but still requires some improvement. Notably, the opening of items in new tabs needs to work in a wider selection of browsers/OSs.
2. Scripts need to be written to transform the logging format we have into that read by the framework
3. The nDCG of the current weighting needs to be calculated
4. The framework needs to be run against the formatted logs and new weightings determined and deployed
5. The nDCG of the new weighting needs to be calculated after some suitable interval (probably a month)
6. All of the above have to be runnable in a documented and straightforwardly replicable way

2.2 Success Criteria

Two iterations of reweighting, resulting in an improvement of nDCG >16% compared to current scores.

---

[3] MS31, p. 4.

[4] Code available here: https://github.com/europeana/contrib/tree/master/query-logs-analysis and here https://old.europeanalabs.eu/svn/assets/trunk/. Some exploratory work needs to be undertaken to determine the relationship between these two codebases.

3 Selective Boosting by Specific Metadata Values

Two objectives identified in the DSI-1 Search Improvement Plan involved weighting based on:

1. Metadata completeness
2. The presence or absence of thumbnail images

3.1 Metadata Completeness

Attempts to boost based on the completeness measure currently available in the Europeana corpus were frustrated by its unreliability. The work of the DQC, however, will yield a more usable measure of completeness. Once this new measure is implemented and available in our metadata, it should be used to inform relevancy ranking.

3.1.1 Requirements

1. Availability of an accurate completeness measure in Europeana metadata
2. Some mechanism (options include extension of the BM25f plugin, function queries, or other means) to boost documents based on this measure

3.1.2 Success Criteria

Inclusion of the new completeness measure as a factor in relevance ranking, as verified by automated tests. An assessment of the effect of this on nDCG should also be performed.

3.2 Thumbnail Availability

In DSI-2 it was discovered that the BM25f plugin cannot boost multivalued attributes - a restriction which precludes direct boosting of the thumbnail field, which may (and frequently does) have more than one value.

3.2.1 Requirements

This multivalued restriction of the BM25f plugin needs to be removed or circumvented - whether through extension of the plugin, use of function queries, creation of a new field, or some other mechanism.

3.2.2 Success Criteria

Inclusion of thumbnail presence as a strong factor in relevance ranking, as verified by automated tests. An assessment of the effect of this on nDCG should also be performed.

4 The Entity Collection

4.1 Overview

The Entity Collection allows us to solve several longstanding challenges for Europeana search.

1. At the most basic level, the Entity Collection (EC) allows us simply to give users what they are looking for. Analysis of our query logs from 2015 indicates that roughly 70% of searches are for Agents or Places; another 10-20% fall under the rubric of 'subject headings' and might therefore best be dealt with as Concepts. Somewhere between 80% and 90% of Europeana searches, in other words, are for entities, and the EC provides us with a very direct means of satisfying that information need.

2. The EC gives us a practical and effective means of satisfying our multilingual requirements in accordance with best practice as outlined in the White Paper on Best Practices for Multilingual Access to Digital Libraries[5]. In particular, it gives us an infrastructure enabling:

   a. Creation of multilingual vocabularies through mapping and enrichments (§2.2)
   b. Multilingual semantic enrichment (§2.3)
   c. Multilingual query autocomplete (§4.1)
   d. Multilingual browse (§4.4)
   e. Multilingual browse result filtering (§4.5)

3. The EC allows us to bootstrap relevance-ranking. Europeana's logs indicate a fairly 'flat' access pattern; that is to say, most searches are unique, and yield unique items, limiting the utility of commonly-used clickstream-driven approaches to search-weighting.[6] As further described below, however, an entity-derived approach opens up a useful range of options for search optimisation.

It is anticipated that the EC will be evaluated with regard to the above functionality as one of the components included in the new evaluation framework, as described in §6.

4.2 Learning to Rank

The implementation of the Entity Collection allows us to solve many existing problems of multilinguality and term polysemy on our platform. As the EC grows in significance, however, and

---

[5]

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

[6] Ceccarelli, D., Sergiu Gordea, Claudio Lucchese, Franco Maria Nardini, Gabriele Tolomei (2011) "Improving Europeana search experience using query logs". *Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries,* (Springer-Verlag: Berlin), p. 384-395.

through its autocomplete feature increasingly becomes the first point of entry into the platform, the importance of useful and usable EC relevance ranking becomes ever more urgent.

It is intended that the tuning of the EC's relevance ranking takes place in two stages.

4.2.1 Manual (Coarse) Tuning

At the moment relevance in the EC is calculated based upon click-data measures of popularity from Wikipedia, the frequency of an entity's labels in the Europeana corpus, and the number of enrichments in the document corpus for that entity. This is, however, a relatively crude measure. Some coarse preliminary adjustments to how relevance is calculated accordingly need to be explored. For example:

1. The relative weighting of the contribution of Wikipedia popularity and Europeana term frequency might need to be adjusted.
2. Entity-type boosting might be required (i.e., different boosts might be needed for Places, Agents, and Concepts)
3. Other intrinsic features, such as birth date and profession (in the case of Agents) or population / place hierarchy (in the case of Places) might be used to inform the relevance measure.

Fortunately it is possible to iterate rapidly for these kinds of changes.

4.2.1.1. Success criteria

EC relevance ranking passes basic sanity checks, as described in the user-testing document.

4.2.2 Automated (fine) tuning

As in the case of the Europeana corpus and the BM25f plugin, the number of parameters and their possible permutations available for search weighting in the Entity Collection defies attempts at fine-grained manual tuning. Indeed, the Entity Collection offers us the potential of more and rather different kinds of parameters for optimising search than the relatively narrow range of single-valued fields the BM25f framework offer us, for example:

1. Field-weighting
2. User-derived information, such as clickstream data
3. Intrinsic characteristics, e.g. birthdate (Agents), population (Places), hypernymy (Concepts), etc.

Correctly determining the contribution each of these features ought to make to relevance ranking requires a machine-learning approach.

4.2.2.1 Requirements

1. The capacity to log the rank of clicked entities in the EC autocomplete dropdown needs to be implemented.

2. An assessment needs to be made of available open-source Machine Learning frameworks and their suitability for the EC LTR task.
3. Scripts must be written amalgamating all potentially-relevant features (fields, user data, and intrinsic data characteristics) into a format usable by this framework
4. The nDCG of the current EC search configuration needs to be calculated.
5. The Machine Learning application needs to be run against the feature set and a new configuration determined and deployed.
6. After some suitable interval (most likely a month), the nDCG of the new configuration needs to be determined and compared with the previous nDCG score.
7. The above process must be well-documented and readily iterable.

4.2.2.2 Success criteria

Two iterations of EC search reconfiguration, resulting in a positive effect on nDCG scores.

4.3 Coverage and Curation

Preliminary evaluation of the EC has indicated that its contents require further curation, in particular:

1. The support for Time Periods needs to be added. At the moment there is no data about them on the EC.
2. Longer- range exploratory work needs to be carried out regarding the implementation of Events and of Works (in the FRBR sense of the term).
3. The number of Concepts covered needs to be increased substantially from its current number of 265.
4. Agents and Places need in some cases to be whittled down (the EC holds far too many minor pop stars) and in other areas to be expanded (politicians, American cities).
5. Some policy needs to be developed for rank-differentiation between entities with identical names but different levels in Place hierarchies (e.g., the City of London and London; Wexford and Wexford County; Amsterdam and Amsterdam Gemeente).

These questions are to be addressed in the Entity Curation Plan[7] document.

4.3.1 Requirements

Our existing harvesting approach evidently needs to be modified, and to some extent streamlined and standardised. It is anticipated that our partners at AIT will undertake the technical implementation (as opposed to the curatorial and design) aspect of this work.

4.3.2 Success criteria

A minimum of 30% of user searches should be satisfied by an entity within the EC.

---

[7] https://docs.google.com/document/d/1A5Rb3Oe9edin5gdRpqFILlR0YPUodVOel3SdcBP00dA/edit

5. A New Evaluation Framework for Europeana Search and Discovery

Central to the work of DSI-2 is the development of an evaluation framework for Europeana search and discovery which moves beyond tightly-focused measures of rank-relevance in the Search Engine Results Page (SERP) such as nDCG, and captures the full range of search, discovery, and exploratory behaviours in which users engage and the features/functionalities that can support them. Research will accordingly be undertaken and an evaluation framework produced that addresses (a) the full spectrum of user needs and requirements in relation to search and discovery; (b) the role of each search-related UI and IR component in supporting these requirements; (c) the interaction of these components to form a unified search experience; and (d) user-based evaluation of this search experience in relation to relevant information tasks.

5.1 Understanding User Needs and Requirements for Search

To create a suitable search evaluation framework for Europeana, users' needs and requirements of the system first need to be well-understood. To this end, a three-pronged research approach will be undertaken.

1. A literature review, including:
   a. Internal Europeana documentation (e.g. user scenarios and personas)
   b. Other relevant literature regarding search behaviour in relation to Europeana and to cultural-heritage systems more generally
2. Query-log analysis, including:
   a. Identification of popular queries
   b. Identification of popular clicked items
   c. Analysis of user sessions (e.g. session length, number of interactions, etc.)
3. User surveys, administered through:
   a. On-site pop-ups
   b. the EuropeanaTech community

5.1.1 Requirements

1. Query logs capturing a wide range of user interactions, including at a minimum:
   a. Session information (identifier)
   b. Query terms
   c. Query reformulations
   d. Items returned
   e. Items clicked
   f. Dwell time
2. The capacity to administer user surveys through the Europeana site

5.1.2 Success Criteria

Identification of relevant query samples and use cases (e.g. in the form of "simulated work tasks") that can be implemented in the evaluation framework.

5.2 Evaluation Framework for Individual Search Components

5.2.1 Components: System-Testing

A site such as Europeana's typically provides a number of components to support users' search, exploration, and discovery activities; examples include recommendations and links to similar items, clusters, facets, query suggestions and autocomplete, visualisations and collection overviews, etc.[8] A process of formative evaluation will be necessary to design and optimise each of these components individually.

The subtasks for creation of such a formative evaluation framework include:

- Literature review for component-level testing of search systems
- Comprehensive listing of Europeana components (UI and backend) eligible for testing
- For each component, provide guidance on potential methods, criteria and measures that can be used to assess performance
- Investigation of query logs to determine the extent to which behavioural data can be used to assist in component-testing (e.g. click-throughs used to compute click entropy for assessing diversity)
- Comparison of Europeana components with those offered by other digital cultural heritage systems (i.e., comparative analysis or informal feature analysis)

5.2.1.1 Requirements

1. Other approaches for component-based evaluation may also be investigated:
   a. A web application capable of showing users the Europeana site as it currently stands, and the site with modified search behaviour implemented. There are various options here:
      i. Extension of the existing A/B testing apparatus. At the moment this currently supports only UI-level changes.
      ii. An extension of the existing various "Fiddle" applications. This is technically simple, but means for obvious reasons that we couldn't A/B test through our regular site - meaning some other means of recruiting test users would be necessary (described in point 2).
   b. To evaluate the components in isolation, a specific task related to the component to be evaluated needs to be defined, and an interface to be developed accordingly. A number of options can be explored:

---

[8] Note that diversification of Similar Items and the use of clustering have specifically been flagged as complex areas demanding further investigation in MS31, Milesone 1.3 and as DQC User Scenario 13 (http://bit.ly/2dnt5U6), respectively.

     i.   An interface to enable users to perform a pair-wise comparison of two different search results/interfaces.

    ii.   A results page showing interleaving results from two different searching methods that will the collection of relevance judgments and assessment of preferred method. This will require data about their queries and their clicked results.

    iii.   Other interfaces may need to be developed depending on the component. For example, to assess the "auto-complete" accuracy, an interface can be developed only showing the query box and the list of auto-complete results, for users to evaluate.

5.2.1.2 Success Criteria

This task should result in the following outputs:

1. A table describing all components of Europeana and guidelines on approaches to testing (e.g. related work, methods, criteria and measures).
2. A proposed framework/method for evaluating components (e.g. crowdsourcing, use of side-by-side online evaluation and preference judgments).
3. Where feasible, an initial evaluation carried out according to this framework. This evaluation is primarily intended as a means of 'road-testing' and refining the framework itself, but may also serve as a preliminary benchmark for future component-testing.

5.2.2 Components: User-Testing

In addition to evaluating the effectiveness of each component from a systems perspective, an assessment of the contribution each component makes to the search experience and process overall ('whole-page relevance') needs to be made. Technical and procedural recommendations accordingly need to be made regarding how such an evaluation is to be performed.

It is anticipated that the School Assignment Satisfaction Index (SASI)[9] will form an early focus of research here.

5.2.2.1 Requirements

1. A web interface showing the whole page of results from the Europeana search system and enabling a component (or interface feature)-level evaluation, based on SASI or a similar protocol, and which allows evaluators to indicate the relevance of individual components in their search activities. Implicit in this requirement are further requirements for,
   a. Evaluators, whether crowd-sourced from Amazon Mechanical Turk or drawn from e.g. the EuropeanaTech mailing list

---

[9] For a full description of the SASI framework, see http://research-srv.microsoft.com/en-us/um/people/ryenw/papers/BaileySIGIR2010.pdf

    b. A list of search components to be evaluated.

    c. A list of queries and their corresponding SERP as evaluation data.

5.2.2.2 Success criteria

This task should result in the following outputs:

1. If possible, a SASI-based rating of component relevance for the Europeana search system, based on varying information needs and search tasks.
2. A preliminary evaluation of the Europeana search system in accordance with these recommendations. As with item 3 in §5.2.1.2 above, the primary purpose of this is to assist in refining the recommendation. Secondarily, however, it may serve as a benchmark for future evaluations.

5.3 Task-Based Evaluation of Europeana Search and Discovery

While the evaluation of individual components and their interaction allows optimisation of particular aspects of the search system, the ultimate aim of such work is to support users in satisfying their information needs. Compared to the testing of components, this work provides a formative evaluation of Europeana for supporting users with searching and discovery activities. Accordingly the proposed evaluation framework takes user-testing as its evaluative endpoint.

It is envisaged that the user-evaluation framework will be based in large part on that previously developed for Europeana within the context of the PATHS project[10], and informed in large part by the outputs described in §5.1 and §5.2.

Because of the fine-grained and comprehensive character desired for the evaluation framework, it is expected that user-testing will be performed using (a) controlled lab-based observation and evaluation, and (b) remote testing.

5.3.1 Requirements

1. An initial benchmarking and exploratory evaluation based on a simulated work task derived from existing personas, user scenarios, and previous work in e.g. PATHS.
2. A number of "simulated work tasks" created from the existing Europeana personas and use cases and evidence from the query logs (described at §5.1.1).
3. Controlled lab-based observation and evaluation, in turn requiring:
    a. Usability lab (lab-time procurable from the University of Sheffield).
    b. Evaluators: University of Sheffield students or staff with interest in cultural heritage systems (specifically Europeana)

---

[10] http://www.paths-project.eu/, specifically D5.2: Evaluation of the second PATHS prototype:
http://www.paths-project.eu/eng/content/download/5237/39882/version/1/file/D5.2+Evaluation+of+second+prototype.pdf

5.3.2 Success Criteria

1. A framework usable in future optimisations of the Europeana search infrastructure, including new formal Key Performance Indicators (KPIs) for search evaluation.
2. An initial evaluation produced using this framework. For the role and application of this evaluation, see item 3 in §5.2.1.2, above.
3. Technical and other recommendations regarding the application of this framework in the absence of a usability lab - that is to say, as a remote-testing framework.

6 Other Work Ongoing

6.1 Infrastructure/architectural work

Over the start of DSI2 the limits of our current datastore infrastructure have started to become clear: the responsiveness and reliability of our Apache Solr instances are suffering because of the sheer scale of the data we store.

A number of approaches are being taken to address this issue.

Rationalisation of our datastore schema. Solr currently stores considerable redundant or superfluous information. Its schema.xml document needs to be reviewed to make the data model as lean and efficient as possible. This effort will be led by Timothy Hill.

Investigation of ElasticSearch as an alternative to Solr. While being built around the same retrieval library (Lucene) as Solr, ElasticSearch is for many use-cases considered a more scalable and responsive platform than its older sibling, while offering most of the same functionality. We will accordingly be evaluating possible ElasticSearch architectures and how viable each of these is for the Europeana use-case.

6.2 Image and Audio Similarity search

DSI-2 inaugurates a partnership between the Austrian Institute of Technology (AIT) and Europeana to develop search-by-image-similarity and search-by-audio-similarity for the Europeana platform. This development will be carried out in parallel with the workplan described in this document, though the two obviously overlap in their concerns.

More detail concerning the image-similarity service can be found in An Image Similarity Search for the Europeana Digital Library and Beyond[11]. Work on audio similarity is of a lower priority, but

---

[11]

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

a pilot has already been developed in the context of the Europeana Sounds project; see further the Music Information Retrieval Pilot Delivery Report[12].

6.3 Horizon scanning

There are a number of recent developments in the areas of open culture and cultural informatics that merit tracking, even if their implications are unlikely to fall amongst the immediate concerns of DSI2.

Mediachain Mediachain's efforts to create a 'decentralized, universal media library', and their particular focus on cultural heritage, aligns their interests closely with those of Europeana. Antoine Isaac and Timothy Hill are currently following their blog and Slack projects, and the former is in discussions with them.

Persistent Reproducible Identifiers Maarten Zeinstra's work on Persistent Reproducible Identifiers to some extent parallels the efforts of Mediachain, and should be tracked alongside it.

Commons Machinery The work of Jonas Öberg and the Commons Machinery team aims to contribute directly to the Commons and Open Access ecosystem Europeana also aims to foster and contribute to.

Pastec The potential for this open-source image-recognition platform for Europeana's vast holdings of visual content needs to be monitored and assessed.

6.4 Work carried over from MS30 and MS31

A number of activities discussed previously in MS30 and MS31 are continuing through DSI2. Those not already dealt with above are summarised briefly here.

Document the search mechanisms employed by Europeana. Given the ever-changing character of Europeana's technical and search infrastructure, this is an ongoing challenge. Documentation of now-stable elements such as the elevate.xml file and the BM25f configuration, however, should now be undertaken.

Search in Annotations. Work on this cannot proceed until the Annotations framework is mature. In the first instance this will consist of monitoring the annotations produced by users and performing a basic assessment of their utility for enhancing search. It should be noted that a similar approach should be taken to the User Sets feature, once this is completed.

---

[12] Music Information Retrieval Pilot Delivery Report.

Translation in Search. The Translation API remains available, but unintegrated with the Search API. The existence of the Entity Collection reduces the desirability of such integration considerably, and whether there should be any further work here will need to be assessed.

Language Detection. Preliminary experimentation with Solr's built in LangDetect functionality indicated acceptable (c. 89%) accuracy even on very short text snippets such as is found in Europeana metadata. The exploitation of LangDetect to support automatic language-tagging of metadata fields during the Metis ingestion process accordingly requires assessment and exploration.

Search taking into account hierarchical objects. The status here is as described in MS31.

Configure Solr text analysis chain in a language-specific way. It has become clear that, while language-specific text-analysis is feasible at the Solr level, implementation would involve non-trivial changes to code at other layers of the stack. A preliminary analysis of these changes and their feasibility needs to be undertaken before work can proceed further here.

Explore ways to mine and exploit query reformulation strategies from users. The logging framework that would support exploration of this question is largely complete. For the moment, however, such exploration should be only preliminary and *ad hoc.*

Metadata translation experiments. A variety of services exist which might be used to provide query translation and/or metadata and EC enrichment. Of particular interest are:

1. Connecting Europe Facility - Automatic Translation (CEF.AT), which is to be tracked closely
2. BabelNet, which should be of particular interest for EC curation and multilingual expansion if problems of licensing can be resolved
3. Crowdsourcing, leveraging the Annotations API.

Europeana data in Peripleo. Simon Rainer of Pelagios and AIT will be working to integrate Europeana data with the former's Peripleo platform. Antoine Isaac and Timothy Hill are in regular contact with Pelagios, and will be monitoring progress here and providing advice and assistance where appropriate.

Remaining items have in most instances been subsumed into the work of the DQC, as noted in Milestone 31.