Europeana DSI 2– Access to Digital Resources of European Heritage

DELIVERABLE

D6.3: SEARCH IMPROVEMENT REPORT

Revision	final
Date of submission	31.08.2017
Author(s)	Monica Paramita, Paul Clough, Timothy Hill (SHFD)
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	14/08/17	Monica Paramita, Paul Clough, Timothy Hill	SHFD	final

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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. Introduction

The document MS6.6: Search Improvement Plan[1] outlines a number of steps to improve the search and discovery in the Europeana platform. This document reports on progress made with regard to each of these steps.

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

Objective	To improve search retrieval effectiveness through automated reweighting of query fields.
Approach	Application of Learning-to-Rank framework
Success criteria	Two iterations of reweighting, resulting in an improvement of nDCG >16% compared to current scores.

In DSI-1, an updated BM25f Solr plugin was deployed to improve search ranking of the Europeana Collections site. This plugin, however, requires regular retraining using a machine-learning framework in order to be maximally effective, as the ideal field-weighting factors change over time, depending on the contents of the collection.

This retraining was scheduled to take place twice over the course of DSI-2. This requirement, however, was overtaken by an arising need for a reindex identified in May 2017 and the decision to take advantage of this opportunity to radically simplify and reduce the Solr schema (see below, Section 6.1: *Rationalisation of our datastore schema*). As this rationalisation and reindexing effort was only completed in the first week of August 2017, reweighting of the Solr plugin has been significantly delayed.

Further actions

Although rationalisation of the Solr schema has delayed reweighting, the task of retraining the Learning-to-Rank (LTR) framework is now potentially urgent. The changes to the schema are dramatic, and in fact include the removal of one of the fields the plugin previously boosted to maximise search effectiveness.

The first task, however, is to determine a new baseline for search performance. The rationalisation of the schema creates, in essence, an entirely new algorithmic environment: with less duplication of data, the performance of the default Solr search-ranking implementation should improve significantly. However, the potential effectiveness of BM25f field-weighting also increases greatly. Steps will accordingly need to be taken to determine how effective keyword search on our platform currently is, and how much the BM25f plugin can be expected to improve it in future.

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms6.6-search-improvement-plan.pdf

3. Selective Boosting by Specific Metadata Values

3.1. Metadata completeness

Objective	High-quality records should appear at the top of results listings.
Approach	Boosting of a new metadata-completeness field.
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.

Completeness of metadata is an important aspect of record quality, and the Data Quality Committee accordingly developed a new metric to measure Europeana metadata completeness over the course of DSI-2.

This measure, however, is not yet integrated into our data ingestion process. This is scheduled to occur with the development of Metis in 2018, at which point the new metric will be available for search ranking.

3.2. Thumbnail availability

Objective	Records with associated thumbnail images should appear at the top of results listings.
Approach	Boosting based on association of thumbnails with a given record.
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.

Thumbnail images convey useful information to users and are highly valued by them; accordingly, it is desirable to boost records with associated thumbnail images above those which lack them in the result list.

Such boosting was previously difficult to achieve, for two reasons: the BM25f Solr plugin could not boost on multivalued fields; and the fragility of the overall Solr server stack made the use of other boosting mechanisms (specifically, use of the EDisMax handler) difficult or impossible.

With the reduction of index size that has come with the simplification of the schema (see below, Section 6.1: *Rationalisation of our datastore schema*), new technical possibilities open up. Using the single-valued 'has_thumbnails' field may prove simpler than exploiting the multivalued 'provider_aggregation_edm_object' field; and simple search-field weighting may now be feasible.

Further Actions

Further exploration needs to be made of the best technical means of boosting records with associated thumbnail images.

4. The Entity Collection

4.1. Learning to Rank (LTR)

4.1.1 Manual (coarse) tuning

Objective	Target entities should appear at the top of the Entity Collection autosuggestion list
Approach	Rapid iteration and manual testing
Success criteria	EC relevance ranking passes basic sanity checks, as described in the user-testing document.

Improvements in the Entity Collection import process mean that rapid experimental iterations for search ranking have become possible. Over the course of DSI-2, the majority of these iterations involved variations in the configuration of the Solr Suggester,[2] and/or the relative weighting of Europeana Document Count (EDC - that is to say, the number of hits with which an entity is associated in the Collections datastore) vs. Wikipedia Hit Count used for relevance ranking. Many of these experiments failed basic sanity checks; four, however, went on for more complete manual evaluation against a test set of queries niche-sourced from Europeana employees.[3] These evaluations recorded the rank position of the target entity in the result list after a given number of characters had been entered, with the aim being that the target entity should appear as close as possible to the top of the result list after a minimal number of characters had been supplied. Ranking performance is now very good for this limited test set: in all cases where the target entity exists in the Collection, it appears within the top ten rank positions after its fourth character has been entered, and very frequently at rank 1.

The single greatest improvement to ranking, after the configuration of the Solr Suggester had been decided upon, came with the adoption of Wikidata PageRank[4] rather than Wikipedia Hit Count as our external metric for relevance. The need for some exterior relevance criteria was clear, for two reasons: first, the specialised character of many of our datasets means that the relative frequency with which entities occur in our collection often diverges markedly from what the average user might expect; and second, our enrichment processes have been optimised for high precision but relatively weak recall. To compensate, our measures of Europeana Document Frequency depend not just upon a given entity's identifier, but also upon that entity's labels, and as a result our estimates for EDC will sometimes be too high; here, external relevance judgements serve as a useful corrective.

[2] https://cwiki.apache.org/confluence/display/solr/Suggester
[3] https://docs.google.com/spreadsheets/d/1y2RbOPoWRd5x_Ws4Xl3NE514HXjhK5RL1zvJJbcTJm4/
[4] As calculated by Thalhammer, A. (2017): http://people.aifb.kit.edu/ath/#Wikidata_PageRank

Wikidata PageRank, however, appears to be much more suitable for this purpose than Wikipedia Hit Count - presumably because PageRank measures expert/editor opinion of connectivity within a given knowledge domain, while Hit Count is a simple measure of general popularity.

The approach eventually adopted to ranking, then, can be expressed in the formula:

$$ranked\_weight = (\log_e(edc * (pr + 1)) * 10000) * df$$

Where *edc* is the Europeana Document Count (the number of hits the entity has in Europeana Collections), *pr* is the entity's Wikidata PageRank, and *df* is an arbitrary depreciation factor set for some entities with disproportionately high PageRank scores. The taking of a natural log and the multiplication of this log value by 10000 are simple mathematical operations to ensure the resulting figure is within a range dealt with accurately by the Solr Suggester.

4.1.2 Automated (fine) tuning

Objective	Target entities should appear at the top of the Entity Collection autosuggestion list
Approach	Application of a Learning-To-Rank framework
Success criteria	EC relevance ranking passes basic sanity checks, as described in the user-testing document.

Tuning of Entity Collection search-ranking beyond the level discussed in section 4.2.1 will require the application of an LTR framework.

Undertaking this exercise requires log data to train the LTR algorithm. Such log data cannot be gathered until the Entity API autosuggest functionality is integrated into the live site, and this work is scheduled for completion by 31 August 2017. One month's log data should be sufficient for initial LTR training, and implementation of Automated Tuning is accordingly deferred until 1 October 2017.

4.2. Coverage and Curation

Objective	Ensure the Entity Collection is capable of significantly improving user search experience.
Approach	As outlined in the Entity Collection Content Strategy and Curation plan[5]
Success criteria	A minimum of 30% of user searches should be satisfied by an entity within the EC.

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

In order to improve users' search experience, the entities held in the Entity Collection must match those queried for by users.

4.2.1. KPI 6.1

In order to assess the overlap between user queries and Entity Collection entities, the following steps were taken:

1. 1000 user-submitted queries were extracted from logs spanning 1 March 2017 - 1 Aug 2017
    a. Excluded from selection set from which these 1000 queries were drawn were: all terms used more than 10 times; all purely numeric terms; and queries targeting particular Europeana datasets. Queries matching these criteria had a relatively high likelihood of being generated by the Europeana Foundation itself; by partners from the Network (i.e. data providers testing how their collections show up in the Collections portal); by automated clients of our API; or as a result of links featured on the Europeana site or sent out as social-media promotions[6]. They were therefore considered likely to be unrepresentative of searches entered by end-users.
2. The list of 1000 queries was then reviewed and all nonsensical and/or apparently robot-generated queries were removed. Five hundred random queries were then selected from the remaining set and hand-annotated as targeting Agents, Places, or Concepts, or as a non-Entity search. 374 queries were classified as Entity searches.
3. The targeted entity was manually extracted from the query, and a query for this entity run against the Solr instantiation of the Entity Collection. If the number of hits returned was > 0, this was recorded as a success; if = 0, a failure
    a. Matching was moderately strict: all tokens in a search term had to match, though not necessarily in the order given. Query terms were left untransformed and untranslated. Matching was against the skos_prefLabel field in the case of named entities (Places and Agents). In the case of Concepts, the query was made against the general 'text' copyfield, in order to capture hits made against 'notes' fields, and broader and narrower terms.

# Satisfied Entity Queries	As % of Entity Queries	As % of All Queries
171	171/374 = 45.7%	171/500 = 34.2%

It is worth noting that the proportion of entity searches (and of entity types within entity searches) diverges to some extent to that reported in earlier log analyses. The comprehensive log analysis undertaken by 904Labs of logs from 2014/15 indicated that

[6] The rationale for filtering these out at this point is that such 'canned' searches do not require any action from the user, who will simply, e.g. click on a link. The main application scenario, which corresponds to the KPI, is the 'autosuggestion' one, where a user types a query and gets suggestions from the Entity Collection. Measuring the KPI on canned searches would fail to evaluate the relevance of the Entity Collection for this scenario. That is not to say that canned queries cannot make use of the Entity Collection. On the long run we expect that canned queries will be formulated with the entities directly. But for users this will be mediated by the creator of the query (e.g. a Europeana employee or partner) in a situation much more controlled than what happens when the end users 'faces' the Entity Collection directly.

queries for entities accounted for somewhere between 80%-90% of searches, with Places accounting for 48% of searches, Agents accounting for 19%, and the remainder being supplied by the somewhat fuzzier category of Concepts. The figures for the current sample are: 11% for Places; 38% for Agents; and 26% for Concepts, yielding a total of 75% for all entity-targeting queries. The proportion of entity-focused queries, in other words, is somewhat lower than anticipated, and Agents have supplanted Places as the most popular single entity type sought. Further investigation will need to be performed to determine how much of this variance is an effect of sampling or whether these figures genuinely reflect changed usage.

A detailed breakdown of the queries used can be found in the DSI2 KPI Entity Collection Overview spreadsheet.[7]

5. An 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. Multiple forms of testing have been carried out for Europeana and activities since 2011, and are recorded in the Test Master Catalogue[8]. This includes evaluation of (i) the *user experience*; (ii) the *usability* of the user interface; (iii) the *quality of content and metadata* quality and (iv) *search quality*. In this deliverable we focus on identifying key performance indicators for *search quality*.

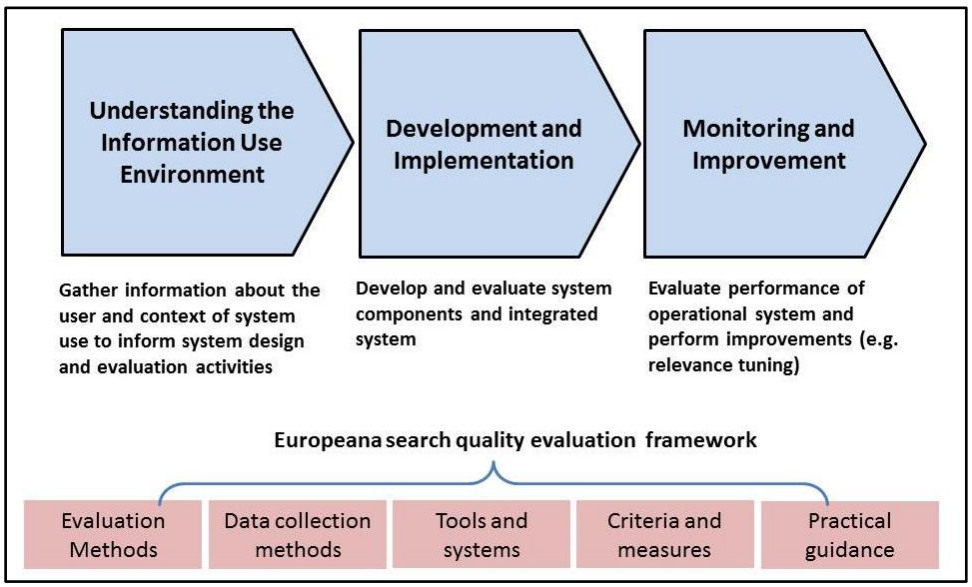

Figure 1. Evaluation framework for improving search quality

[7] https://docs.google.com/spreadsheets/d/15xSicsev5v1JIK3an8JOXGvFWkM75G56IkBtVLfNFkQ/
[8] Test Master Catalogue: https://docs.google.com/spreadsheets/d/1G2tIJLTO4mY-slGIQK6kbleKuWMgFgepqrY4_KJ8jKo/

Figure 1 provides an overview of evaluation activities that begins with gathering further data about user's context for search including search tasks and information uses (see Section 5.1). Europeana is an operational system and therefore will involve online testing for monitoring and improvement. Evaluation activities will also be centred on developing new components and features that will be tested independently and integrated into the operational system. An evaluation framework will need to provide the capability to encompass different evaluation and data collection methods, tools and systems, criteria and measures and provide various forms of guidance, such as formal recommendations.

A helpful way of identifying links between evaluation activities in Europeana and placing evaluation of search quality in context is to consider the model of information systems success from DeLone and McLean (2003), see Figure 2. Many activities in Europeana are focused on aspects of this model, such as the work on improving data quality, the work on user experience and user satisfaction and the impact of Europeana on cultural heritage practices more widely. In the context of evaluating search this is a useful model to relate activities to as it shows that search quality (system quality) is one of the factors that will impact on users' satisfaction and their use of the system (i.e. higher search quality will lead to greater user satisfaction and therefore increase system use). The model also helpfully shows the relationships between search quality, user satisfaction and the impact of using the system (i.e., its utility), such as improved task performance or an increase in user's knowledge.

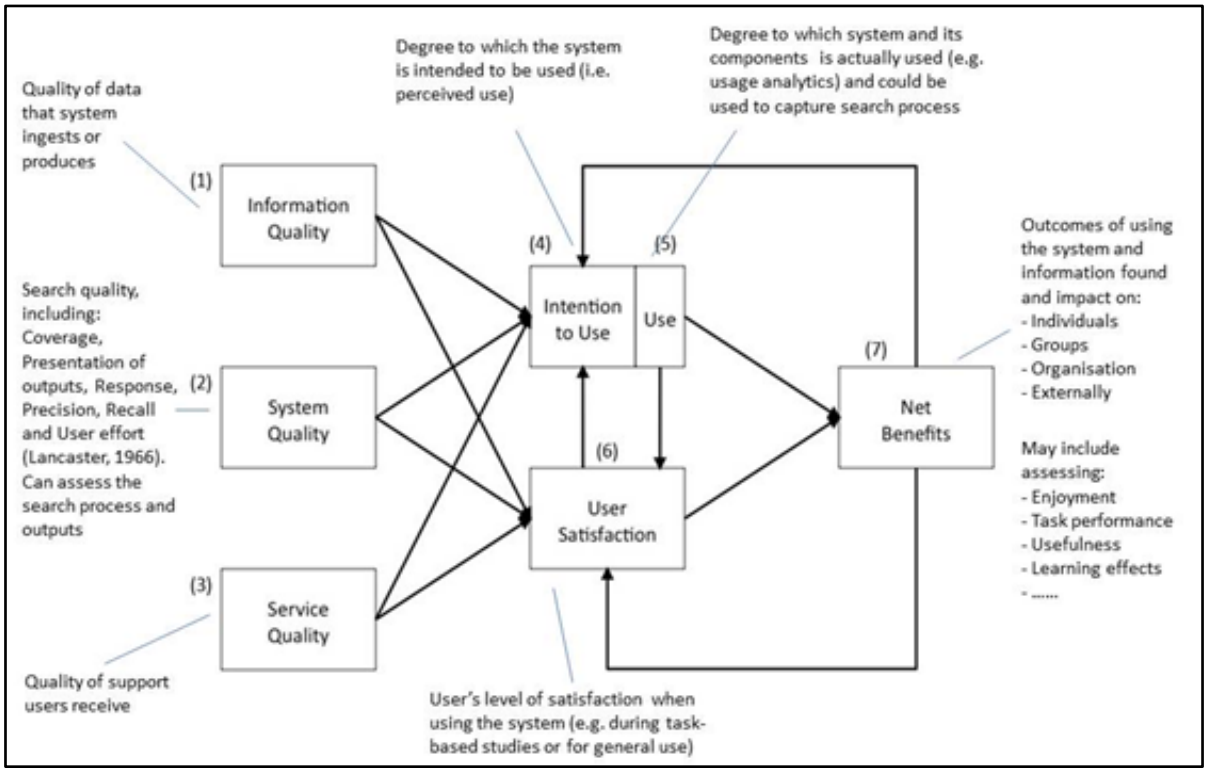

Figure 2. A holistic view of success in Europeana based on an information systems perspective (DeLone & McLean, 2003)

An excellent overview of a decade of evaluation activities conducted within Europeana is provided in the TPDL'17 paper by Juliane Stiller and Vivien Petras (Humboldt University)[9] and a summary of past evaluation activities for search and discovery can be found on the Europeana Assembla site[10].

NB: We also provide a document that complements this deliverable by describing IR evaluation more generally and makes specific suggestions for implementation an evaluation framework for Europeana: https://docs.google.com/document/d/1d7FxjabJvLYbBQBLBwBA16OIJRLNKk6VmEwpg5K7vno/ *(work in progress)*.

5.1. Understanding User Needs and Requirements for Search

Objective	To understand users' needs and search requirements for Europeana
Approach	A literature review, a query-log analysis, and user surveys
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.

To inform evaluation activities we first sought to better understand the searching behaviours of Europeana users, specifically their search tasks, goals and uses of information found. Three approaches were used: (i) a review of existing Europeana user studies (Section 5.1.1), (ii) an analysis of search logs (Section 5.1.2), and (iii) a user survey (Section 5.1.3).

5.1.1. Literature review

Many previous studies have been carried out in Europeana to gather user requirements and inform the design of the system[11]. For example, in 2016 a Europeana user survey identified the majority of users as coming from the educational (e.g. teacher, student, academic, researcher) and cultural sectors. This and other studies lead to the specification of two distinct types of Europeana users (see DSI1 D3.1[12]): (i) 'culture vultures' and (ii) 'culture snackers.' The former group are dedicated enthusiasts and professionals: they have domain expertise and likely lifelong enthusiasts of cultural heritage (likely to be returning users and mainly wanting to use Europeana to find resources to use in their own work, gain knowledge, expertise or inspiration). The latter group are more representative of the novice or general user who come with lower levels of technical/domain expertise and typically engage out of general interest. A number of personas have also been created to further capture these different types of user, specifying examples of occupations (e.g., 'microbiologist with interest

[9] Link forthcoming
[10] Search evaluation activities for Europeana available in Assembla: https://europeanadev.assembla.com/spaces/europeana-r-d/wiki/Search_evaluations_at_Europeana
[11] List of previous user studies are listed in: https://docs.google.com/spreadsheets/d/1G2tIJLTO4mY-slGIQK6kbleKuWMgFgepqrY4_KJ8jKo/
[12] Europeana DSI D3.1 Creative Industry Reach Report. http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d3.1-creative-industries-reach-report-2016.pdf

in history'), goals (e.g., 'to find interesting World War 1 history related items') and their behaviours (e.g. 'explorative behaviour, high search literacy'). Other studies focused on identifying the purposes of their users when carrying out a search using Europeana. In a 2014 survey, users specified that their last visit to Europeana was to explore within a topic (32%), to find out more about Europeana (30%), and to find a specific item (17%). Others (15%) visited Europeana because it was linked from another site. When a similar survey was carried out two years later (Europeana, 2016), the purpose of Europeana users was more frequently to search for a specific item (73%). Users also use Europeana to search for resources for research/academic studies (42%), reuse items found in Europeana (37%), search for WW1 related content (26%), collect inspiration (22%) and search for music related items (18%).

However, we could find no prior study of the work and search tasks of Europeana users that would provide information that could help inform the design of evaluation instruments, such as simulated work tasks for controlled task-based user testing. Section 5.1.3 describes work we carried out to gather information on user's' search tasks.

5.1.2. Query log analysis

Query logs provide behavioural signals (e.g. queries and clicks) regarding user-system interactions and can be used to analyse searching behaviour and inform evaluation, e.g. testing with 'popular' queries to assess search output (Dumais et al., 2014). Although log data is available via Google Analytics, access is limited and not focused on analysis of searching behaviours. Therefore in DSI-2 further work has been undertaken to implement a custom-built interaction logging framework (the Europeana Search Log or ESL) to track user's searching (and browsing) activities. The ESL records the type of user-system interaction (e.g. searching using a query, viewing search results, clicking on an item page, etc.) and statistics such as number of results for a query (see Section 5.2.1.2). This information can be used for analytical testing of Europeana search, such as the usage of system features and user's overall engagement.

At the time of analysis (M6), the query logs contained information about the queries, clicked results and timestamps only. Due to a technical difficulty, the query log did not capture any user and session identifier, which prevented any significant analysis to be carried out at the time. This issue has now been resolved; information recorded in the current query logs is described in Section 5.2.1.2.

5.1.3. User search task survey

To complement data collected from the ESL and help better understand 'why' people search using Europeana (i.e., their underlying search goals and information needs) we designed and implemented a novel pop-up survey instrument (a form of intercept survey) to further understand user needs; specifically their search tasks and information use.

In this study, we aimed to gather responses from Europeana users as they carried out their searching activities. The survey contained 10 questions in English (see Q1-Q10 in Appendix 1) that could be shown to users at any point during their interaction with Europeana. Using the Hotjar service employed by Europeana for conducting past surveys, it was shown to 30% of users (later increased to 66% to increase response rates) who visited Europeana using

desktop or tablet devices. The survey was triggered when users scrolled halfway down either a search results page, or a Europeana item page. Users who completed the survey were given the opportunity to enter a prize draw to win a €50 Amazon voucher. In addition to the questions posed, Hotjar also captured the date and time of submission and the respondent's country of origin. The study was approved by the University of Sheffield's Ethics Committee.

The pop-up survey provides a template for eliciting users' search tasks. The open questions Q4 ("What are you looking for in Europeana?") and Q5 ("Why are you looking for this information?") provided key insights into users' information needs. We used qualitative content analysis to analyse responses for Q4 and Q5 and the Shatford-Panofsky mode/facet analysis technique. We categorised search tasks against existing types and developed a novel scheme for information use. The methodology (incl. categorisation schemes) and findings are likely be useful to the wider cultural heritage community, as witnessed by the acceptance of a paper describing it for the TPDL'17 conference (Clough et al., 2017)[13].

[13] http://paramita.staff.shef.ac.uk/papers/cloughetal-TPDL2017.pdf

Results

The pop-up survey ran for 2 weeks (21 March - 4 April 2017) and elicited responses from 240 Europeana users[14]. The survey respondents came from 48 different countries (Spain 12.9%, US 8.8%, Italy 8.8%, France 7.1%, Germany 6.7%, UK 6.3%, Netherlands 4.2%, Sweden 3.3%, Hungary 3.3%, Brazil 2.9%). 27% of respondents were first-time visitors to Europeana, 23% visited less than once a month, 26% visited at least once a month, 20% visited at least once a week, and 4% visited Europeana every day. A third of the respondents identified themselves to be academics (e.g. lecturer, professor), 25% cultural heritage enthusiasts (e.g., hobbyist, genealogist, amateur historian) and 18% cultural heritage professional (e.g. curator, historian, archivist). The remaining ones were students (13%), school teachers (5%), and others (9%). Finally, almost half of the respondents knew about Europeana already and came directly to the site and a third visited Europeana through a link from a search engine. Around 11% came to Europeana through a link from social media and teaching resources, and 6% from other external resources (e.g. newsletter link, recommended by a friend).

In the pop-up survey, Europeana users were asked to specify their *current search tasks*, i.e. what they were looking for (Q4), why they were looking for this information (Q5), what they would do after finding this information (Q6) and their level of knowledge (a scale of 1-10) in their task (Q7). Some examples of the answers are shown below:

Table 1. Survey answers (Q4-Q6)

Q4	"I am trying to explore images of objects and monuments from ancient Italy and the Roman Empire."
Q5	"I am a librarian teaching a session for students in an Art & Archaeology of Ancient Italy class at a university. They need to find an object or monument that has not been covered in class to write a research paper."
Q6	"Look for more information using other resources"

We categorised the search tasks (Q4) using the following categories (see our TPDL paper mentioned above):

- *known-item search* (e.g., "I am looking for the 1919 film `Les fetes de la victoire.'"),
- *by named author* (e.g., "to look for paintings by Henriette Ronner"),
- *specific-subject search* (e.g., "I am looking for pictures of Stuttgart"),
- *general topical search* (e.g., "Italian medieval illuminations"),
- *browse/explore* ("I'm just browsing your collections"), or
- *ambiguous or unclear* (e.g., "I'm an Opera lover").

The results are shown in Figure 3.

[14] The full results are reported in: https://docs.google.com/document/d/16geDwFkuIxKIKtc6MWkdkRZo__7Qg8CipRhLPkZn1-U/

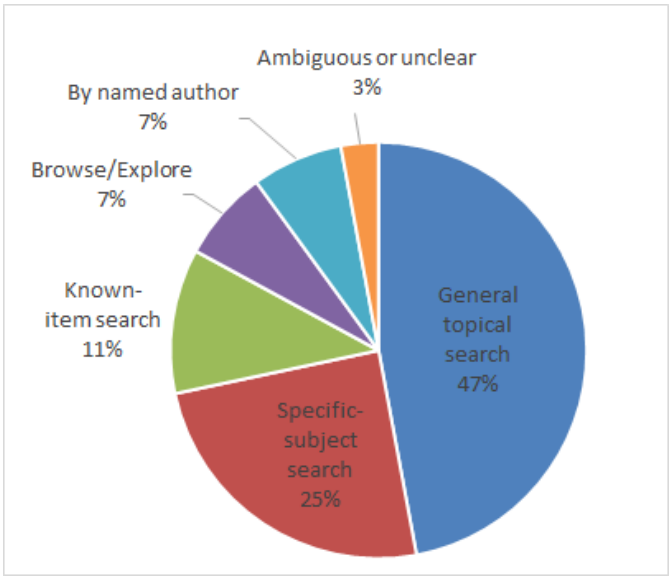

Figure 3. Types of searches in Europeana

We also analysed and categorised the contents of search tasks (Q4 answers) into different mode/facets. The mode/facet analysis helps to provide insights into the subject or content of search tasks. As shown in Figure 4, the most frequent mode/facet is *general object/thing* (71 occurrences), followed by *specific location* (42 occurrences). Search tasks comprise an average of 1.53 modes/facets (min=1, max=5). The most common combinations are "*Creator* + *Specific object/thing*", such as "I want to find some information about a painting of Willem van de Velde, 'Het kanonschot'" (9 occurrences), and "*Creator* + *General object/thing*" (8 occurrences), e.g., "I am looking for artworks by Leonardo da Vinci". We also find that users often use the *Medium* mode/facet to refine the search, e.g., 'images of Stuttgart' and 'I am looking for photographs of The Trachian tomb near to village of Mezek, Bulgaria'[15].

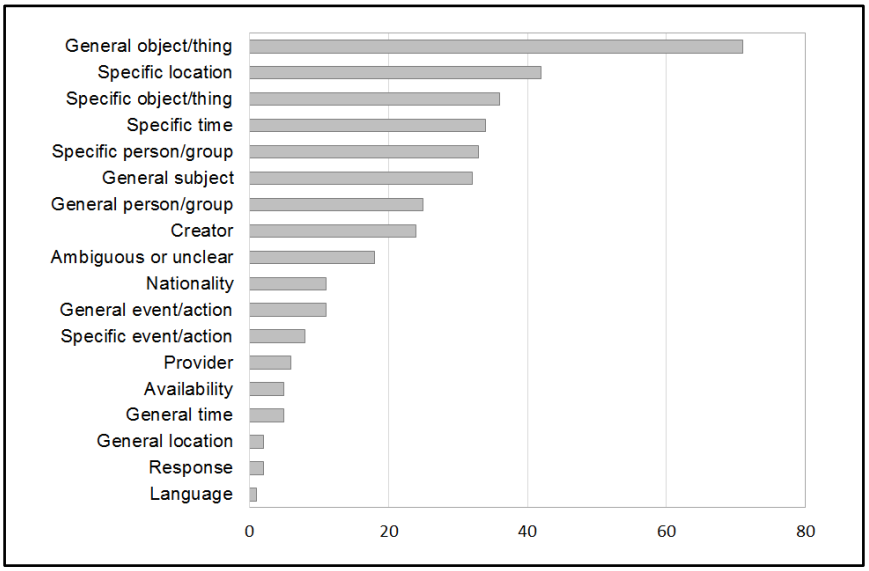

Figure 4. Frequency of occurrences of each mode/facet in the search request

[15] However, due to the lack of details provided in some users' answers (e.g. some users specified that they were searching for "trees", "dental tools", etc.), we were not able to accurately identify the occurrences of Medium.

We also found that the majority of users (37.1%) were searching Europeana with the intention of using the information found to create a new work, e.g. “to write a book”, “to prepare an exhibition”, “to use images for a presentation”, and “to find additional material for my PhD-thesis.” Around a quarter of the search tasks (27.5%) were being conducted for personal interest, such as “to enrich my personal archive”, and “inspiration and general interest”. Professional activities accounted for 20.8% of search tasks, e.g. “it’s my job”, “to check whether the information was correct”. Finally, 7.9% of search tasks were categorised as teaching, e.g., “to illustrate a university lecture”. Such insights helps us better understand our users and could be used to complement existing artefacts, such as the personas. We could also use the findings to derive groups of queries that could be used for system-level evaluation to represent specific tasks or scenarios. The findings also help in identifying potential indicators of success (e.g. for different user groups, search tasks or purposes).

Further Actions
Further work will be conducted on analysing user’s searching behaviours (within and between sessions) to better understand ‘what’ people search for, and ‘how’, as more data is collected.

5.2. Evaluation of Individual Search Components

A site such as Europeana 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. A process of *formative evaluation*[16] will be necessary to design and optimise each of these components individually. We identified two approaches for evaluating individual search components: (i) *system-testing evaluation* in which the performance of components is assessed in isolation (either using a test collection or behavioural data extracted from query logs), and (ii) *user-testing evaluation* where user feedback is gathered and used to assess their performance. This section describes the methods to evaluate these components and preliminary results.

5.2.1. Components: System-Testing

Objective	To design an evaluation framework for search components
Approach	- Literature review for component-level testing of search systems - Create a comprehensive listing of Europeana components eligible for testing - Provide methods, criteria and measures for search components - Determine which behavioural data (extracted from query logs) can be used to assist in component testing - Comparison of Europeana components and other CH systems
Success criteria	This task should result in the following outputs: 1. A table describing all components of Europeana and guidelines on

[16] Formative evaluation includes activities that occur during the development of systems compared with summative that evaluate the system, normally at the end of a development cycle. This is a key aspect for an *agile* development methodology.

	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.

We propose that the iterative development (incl alpha testing) of individual components in Europeana should be undertaken in isolation using specific testing resources (e.g., test collections, corpora) and methods, in order to identify the best techniques and settings[17]. When the components are integrated into the operational system, the Europeana Search Logging framework should be used to gather usage data on components (a potential indicator of success) in combination for other sources, such as explicit user feedback (see Section 5.2.2).

5.2.1.1. Europeana search and discovery components and metrics

Our main effort for component testing has been to catalogue the wide variety of search and discovery features offered within Europeana to form an inventory of components. Components include features provided by the user interface (e.g., search box, filters), back-end components (e.g., indexing and ranking algorithms) and wider system features (e.g., blogs, curated content). For each component, a number of metrics that can be used to evaluate it is also listed. The table describing all components of Europeana, their metrics, and the guidelines to testing is available here: https://docs.google.com/spreadsheets/d/1RAv1oZ3rVyfKC7bnwW29RYUWOsy-rnDZE6j0d2Yup5k/

5.2.1.2. Proposed framework for evaluating search components

A challenge with proposing some overarching framework for evaluating components is that numerous approaches are possible and the selection of candidate methods often depends on many factors, such as experimental goals, whether the component can be surfaced for gathering user feedback, resources available, skills and expertise of the researcher, etc. The effectiveness of components can be measured *directly,* for example the accuracy of different autocomplete algorithms, as well as *indirectly*, i.e. the effects of different algorithms on retrieval effectiveness (assuming the performance of the component as an independent variable can be isolated). In the case of testing with an operational system (i.e., *online* evaluation), behavioural signals (e.g., click positions, pageviews, query reformulations, etc.) can be used to evaluate and optimize the performance of components in-situ, for example the use of click data to infer relevance for tuning the relevance ranking algorithm.

[17] This may be similar to the approach taken in the PATHS project where the functionality and effectiveness of components were carried out independently and separate from the integrated system (see, http://ir.shef.ac.uk/cloughie/papers/Clough_FIRE2014.pdf)

Often general resources, such as test collections and corpora, can be utilised in the early stages of component development, typically within an *offline*[18] evaluation setting. For example, different stemming or indexing algorithms could be explored using TREC and CLEF test collections. However, ideally Europeana would develop its own sets of resources to make evaluation results more predictive of what could be expected in operation. To facilitate the system-testing of components the following resources can be utilised: (i) a custom-built interaction logging framework to track user's searching (and browsing) activities and use of components; and (ii) the creation of test collection resources. To date we have focused efforts in DSI-2 on the Europeana search logging framework.

Europeana search logging framework

DSI-2 saw several improvements in Europeana's logging infrastructure. First, we have undertaken infrastructural work to improve the stability of logging, and to ensure that additional information - such as user id - is reliably recorded. Second, we have written a number of scripts to convert these logs into a variety of formats (for example, Apache-formatted logs as required by our LTR framework, along with human-readable forms) and to apply simple heuristics assessing the nature of any given logged interaction in cases where this might otherwise appear ambiguous. As a result, we are now able to collect information on the user's searching sessions, such as the query terms used, the number of search results found, the rank position of clicked results, and which URLs were clicked, in a manner that is both human-readable and machine-processable. The information collected is summarised in Figure 5 and an extract of the log shown in Figure 6.

```
SearchInteraction [timedate] [session id] [query] [filter] [number of results
     found]
RankedRetrieveRecordInteraction      [timedate] [sessionId] [query] [filter] [URL
     of clicked item] [rank of clicked Item] [number of results found]
RefreshOrPaginationInteraction   [timedate] [session id] [query] [filter] [number
     of results found]
CollectionFilterAdditionInteraction     [timedate] [session id] [query] [filter]
     [number of results found]
CollectionFilterRemovalInteraction      [timedate] [session id] [query] [filter]
     [number of results found]
ArbitraryAccessInteraction     [timedate]  [session id]  [URL requested]
UnknownInteraction     [timedate]  [session id]  [message]
...
```

Figure 5. Information recorded in the log

```
SearchInteraction   2017-07-14T12:08:51.886Z     b009d2c5bbc44fe0e8b3c547174f90a5
     Juan Gris     {'TYPE': ['IMAGE']}     607
RankedRetrieveRecordInteraction     2017-07-14T12:09:07.498Z
      b009d2c5bbc44fe0e8b3c547174f90a5     Juan Gris     {'TYPE': ['IMAGE']}
      /9200376/BibliographicResource_3000136360436     4     607
RankedRetrieveRecordInteraction  2017-07-14T12:09:28.249Z
      b009d2c5bbc44fe0e8b3c547174f90a5 Juan Gris     {'TYPE': ['IMAGE']}
      /9200365/BibliographicResource_1000055123313   13     607
...
```

Figure 6. Query log extract

[18] By *offline* we mean testing that is carried out on components and system, external to the operational system.

We propose to use the search log data for component testing, such as tracking and quantifying the usage of components; as implicit forms of relevance judgment for developing evaluation resources (e.g. test collections) and deriving traditional forms of IR metrics (e.g., nDCG); and deriving metrics for search analytics, such as reformulation rate, number of queries per session, number of clicks per query, failed searches, bounce rates, etc. The proposed measures for evaluating search and discovery components are reported in the KPI document[19].

An example of utilising query logs for component evaluation is Santo et al. (2015) who compare different approaches for auto-complete. A prefix of the user query is used as input to the auto-completion algorithm to be evaluated, while the final query that users issued is used as the correct answer and create auto-completion relevance judgments. Different auto-completion tools were tested with different lengths of prefixes (2, 4, 6, 8 and 10 characters) to represent users in different stages of typing the query and *effectiveness* of the ranked suggestions measured using Mean Reciprocal Rank (MRR). Query logs can be used in this way to develop test resources without the need of gathering further feedback from users. A variation on this approach has been used for basic relevance evaluation with regard to the Entity Collection's autosuggest functionality (see above, Section 4.2.1: Manual (Coarse) Tuning).

Further Actions

Further work will involve a more systematic review of past evaluations involving testing components of an IR system. Examples of evaluations carried out on system components similar to those implemented in Europeana will be reviewed. Further work will also be carried out to analyse data collected using the Europeana Search Logging framework, such as patterns of user's searching behaviour and sets of queries that can be used for testing (e.g., 'popular' queries; samples over the long tail, etc.). We also envisage the creation of resources (e.g., test collections) for offline testing would be beneficial, although we are aware of the effort involved and the limitations of using a static document collection.

In addition, further work on improving the logging infrastructure should ideally be undertaken. Analysis of our logs currently shows minimal activity on the Europeana Collections home page for long periods. These prolonged periods of apparent stasis probably arise as a result of users exploring content, such as Europeana Blogs and Galleries, which is left unlogged by the current framework, and work is underway to integrate these into our standard logging stack.

Additional improvements which might also be undertaken would be to improve Europeana's tracking of user IP addresses, which would allow us to identify user search sessions more effectively, and to find some means of measuring user dwell time (that is to say, the amount of time spent examining a given page) precisely.

5.2.2. Components: User-Testing

Objective	To evaluate the contribution each component makes to users' search

[19] https://docs.google.com/document/d/16TKUfpZVM7m3SXjgfPD1_9Z2QvScxrJ8MIpdGHbCgb4/

	experience and search process overall
Approach	To carry out user feedback on the relevance of individual components in their search activities, using SASI or a similar protocol.
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. The primary purpose of this is to assist in refining the recommendation. Secondarily, however, it may serve as a benchmark for future evaluations.

Components of a search application can be evaluated independently and in isolation (as described in Section 5.2.1), which is typically performed offline during their initial development and without user involvement (i.e., in a more system-oriented manner).

However, evaluation of components can also involve human input (real users or test subjects) in various forms (i.e., user-oriented evaluation) where feedback can be gathered on the usefulness of, and users satisfaction with, components. We envisage that in the future, the creation of a specific framework for conducting multiple forms of component testing would be highly beneficial, along with further investigation of online testing.

5.2.2.1. User-testing of components in isolation

Various approaches can be used to test components of Europeana (in isolation) with users. This could include offline methods and testing with custom prototype systems and *online* methods (Hoffmann et al., 2016) where changes could be made to the operational system and changes in user's search behaviour used to quantify success (e.g., A/B testing, interleaving, etc.). Configuring and running such tests could be possible using commercial search analytics software, such as Google Analytics or Adobe Analytics. Offline methods of user evaluation could involve in-lab studies where user's task and environment are controlled, or controlled task field studies, where the tasks is controlled but experiments can be conducted in the participant's own environment, i.e. remotely. An example of a novel evaluation setup to gather user feedback is the experiment conducted by Karl Pineau to gather the relevance of 'similar items' from Europeana Collections using a Chatbot[20]. Alternatively, SPIRE[21], a tool for developing user test/evaluation, can be used as a tool for testing component prototypes.

5.2.2.2. SASI-style evaluation of 'whole page' relevance

An approach that has recently gained interest in the IR community is the notion of assessing 'whole page' relevance. This provides another perspective on user-testing whereby the relevance of components on a search interface are assessed with respect to a user's query or search task. An example of a 'whole page' relevance framework implementation is the

[20] Results of the Chatbot study reported here: https://docs.google.com/document/d/1Cbjf2i2rtFaPK1AB7xpQ49YxrTyc0Xp2k8db3fwtfG8/
[21] https://promise.sheffield.ac.uk/spire

School Assignment Satisfaction Index (SASI)[22]. In this approach, it is possible for users to judge the relevance and quality of individual components (e.g., ads, filters, search results) shown in response to a query. We have not been able to implement a SASI-style test harness as planned, but still envisage that this would provide a useful addition to the results of online testing. An initial test of the feasibility of gathering component-level was included in the pop-up survey (described in Section 5.1.3) where instead of rating components for relevance given a query, we rated components for perceived usefulness for supporting a specific search task (described next).

5.2.2.3. User-testing of components in-situ

User feedback can also be gathered on components in Europeana during surveys carried out more generally. For example, as a part of the pop-up survey (described in Section 5.1.3), specifically Q8-Q10 (see Appendix 1), we asked respondents to rate their perceived importance of components or features provided by Europeana in supporting their current search task. In order to balance participant's time and effort against the need to capture sufficient detail about their search activities and needs, only nine features were assessed in the survey as they were considered to be the most important features in users' search and discovery tasks. Q8 and Q9 aimed to gather feedback from users on the importance of Europeana features (shown in Table 2) in helping them carry out their tasks. We asked users to indicate the importance score in a scale of 1-10 (1=not important; 10=extremely important; N/A if not applicable for their task being carried out).

Table 2. Europeana features

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

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

The pop-up survey data enabled us to identify the perceived importance of individual components for supporting different type of search tasks (shown in Figure 7). The availability of high quality images (F2), detailed information about an item (F3), availability of information in the user's own languages (F4), links to download an object (F5) and the availability of freely reusable contents (F6) was rated the highest by users who carried out a known-item search. F6 was also rated highly for three tasks: known-item search, general topical search, and browse/explore, but significantly lower for search by named author and specific subject search. Having links to categorised collections (F9) were not identified to be as important on any tasks (average score below 7), except for the "Browse/Explore" activities (7.53). The availability of information in the user's' own languages (F4) is the least important feature, regardless of the tasks.

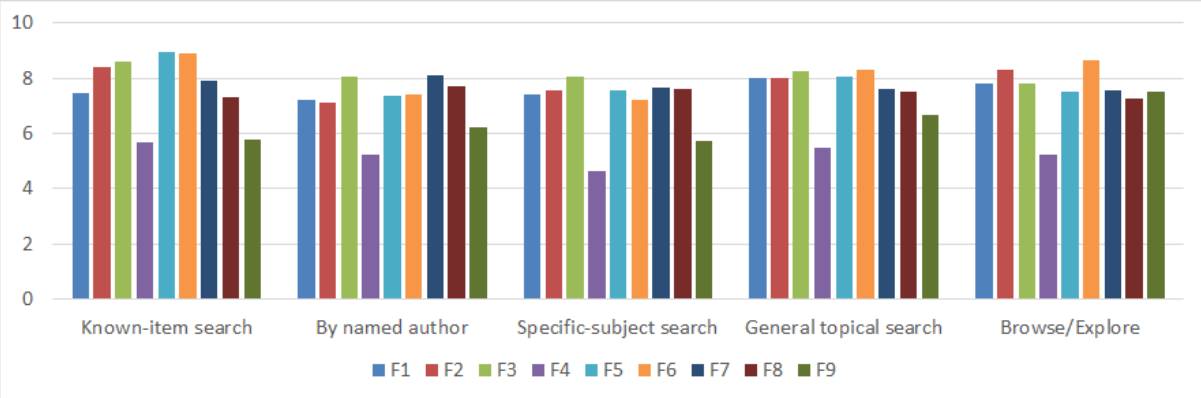

Figure 7. Importance of features for each search task category

We also investigated the feasibility of using user surveys to gather user's feedback on their satisfaction with individual search components/features more generally. This was collected as a part of the task-based evaluation (further described in Section 5.3) in which participants were asked to rate their satisfaction (using a 5-point Likert Scale; 1=very dissatisfied, and 5=very satisfied) on aspects of search in Europeana, including the following:

- Quality of the search (component):
  - Relevance of search results
  - Relevant search results ranked highly
  - Diversity and variety of search results
  - Completeness of search results
  - Response time during interaction
- Degree of support for searching:
  - Access to online help documentation
  - The level of description about objects
  - Support provided for browsing and exploration
  - Availability of links to categorised collections (e.g. Art Nouveau posters, Irish folk music, Fashion illustrations)
  - Availability of options to refine your search (e.g. filters)
  - Availability of links to download an object
  - Access to content you can freely re-use
  - Availability of links to an object provider's site

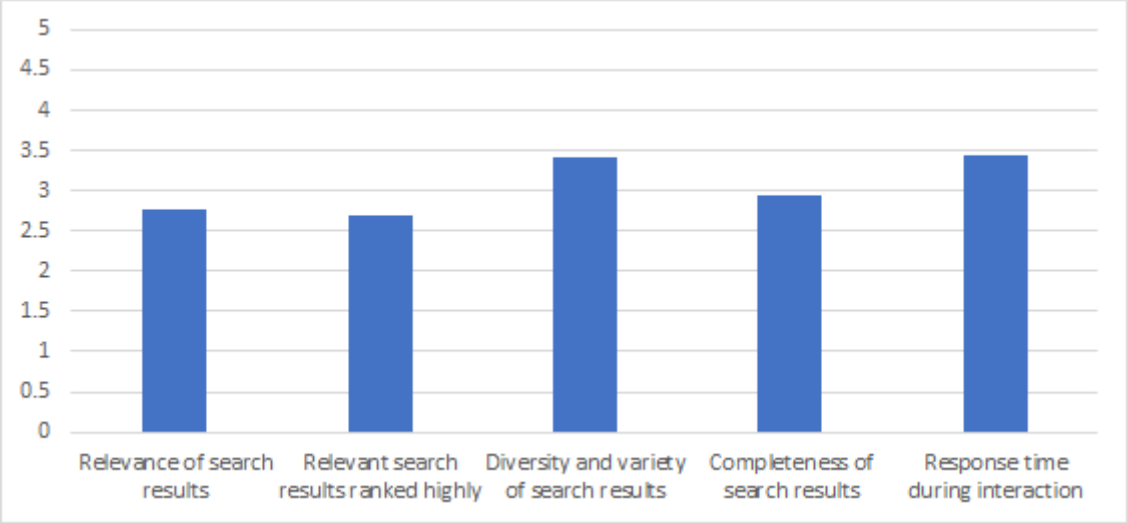

Figure 8. User assessment on the quality of search

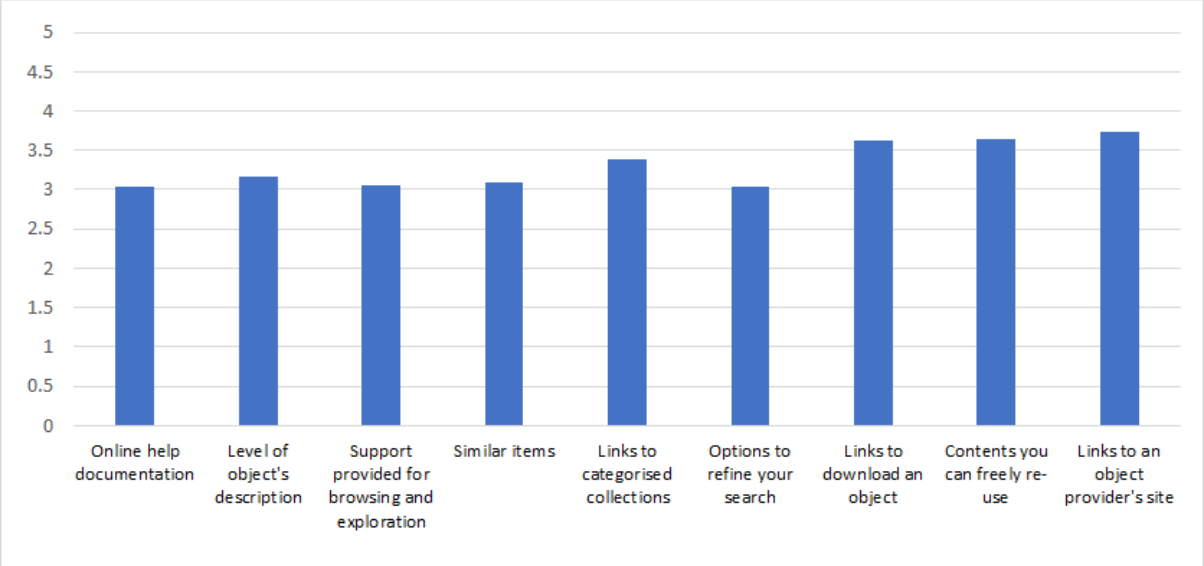

Figure 9. User assessment on the degree of support for searching

5.3. Task-Based Evaluation of Europeana Search and Discovery

Objective	To propose an evaluation framework for Europeana search and discovery
Approach	- To carry out an initial benchmarking and exploratory evaluation based on a simulated work task - To create a number of simulated work tasks - To carry out a lab-based observation and evaluation
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.

	3. Technical and other recommendations regarding the application of this framework in the absence of a usability lab (e.g, a remote-testing framework).

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 and completing their tasks. In this task we sought to evaluate search performance of Europeana in a more holistic manner and in the context of supporting users with search and discovery activities using a task-based evaluation approach. Task-based evaluation is common in Interactive IR methods (Kelly, 2009) where user's interaction with a search system is investigated and evaluated. In this approach to evaluation the focus is user-centric and on whether people can use the system to retrieve relevant information for their needs or tasks. Accordingly the proposed evaluation framework takes a user-oriented perspective on evaluation and involves suggestions for Key Performance Indicators to aid search evaluation. This aspect of the work aims to go beyond evaluating search *outputs* and also evaluate aspects of the search *process*[23]. The task-based framework proposed in this section can be viewed as another source of evidence to inform evaluation more broadly (as highlighted by the conceptual model in Figure 2).

5.3.1. Key Performance Indicators (KPIs) for Search

To aid evaluation activities a set of KPIs have been identified to measure search quality[24]. Europeana currently has one KPI related to search: *"Percentage of all user searches matched to an entity from the Entity Database: 30%"*. This KPI measures the coverage of Europeana collection and the suitability of its content with regards to user searches. In this work we have defined additional KPIs for evaluating Europeana, both for individual components - specifically those that directly relate to the search process - and Europeana as a whole.

5.3.2. Task-based Evaluation Framework

To evaluate how well Europeana supports users during their searching activities we investigated the use of a task-based approach that allows two modes of execution: (i) controlled lab-based observation and evaluation, and (ii) remote testing.

5.3.2.1. Search tasks

To carry out task-based evaluation appropriate search (or work) tasks must first be identified and specified. To facilitate this, we used findings from the pop-up survey which suggested that Europeana users carried out a wide range of search activities on a daily basis. The support provided by Europeana for these different search tasks may vary and needs to be evaluated. To achieve this, we propose the use of a method for a controlled task field study

[23] This distinction between evaluating search output and process is discussed in White (2016) who identifies evaluation measures based on: (i) the search *process* (e.g. learning, user effort, cognitive load, enjoyment, frustration and engagement); and (ii) the search *outcomes* obtained after the search process is completed (e.g., relevance-based metrics, novelty, diversity, search success, satisfaction, etc.).

[24] https://docs.google.com/document/d/16TKUfpZVM7m3SXjgfPD1_9Z2QvScxrJ8MIpdGHbCgb4/

that incorporates three different types of search tasks (derived from the findings described in Section 5.1.3.1):

1. *Task 1: Specific item / information task (5 minutes)*

> "Please use the Europeana site to list the names of three paintings by Juan Gris (a Spanish painter)"

2. *Task 2: Specific-subject search (5-10 minutes)*

> "Imagine that you have been asked to create a presentation about a European city or country that you have visited before."
>
> You may prefer to illustrate your presentation with old/new images, videos, or other media of your choice. Your presentation may include interesting places to visit, illustrate the history of the city/country, etc. Please find at least 3 different items.

3. *Task 3: Browsing/exploring task (10 minutes)*

> "Please imagine that you are on your lunch break. You would like to spend 10 minutes to either search Europeana for any topic that you are interested in, or to simply explore the items provided by Europeana. After your time has expired, please write a short description of what you learned from your browsing/exploring session."
>
> Note: If you are from a cultural institution, please do not explore your own collection.

5.3.2.2. Experimental design and protocol

Commonly in an IIR study, multiple (alternative) components or systems are tested to allow for comparative evaluation. An appropriate experimental design is used that reduces learning and order effects (e.g., the use of task-system counter balancing and rotation). In work undertaken to date we have evaluated Europeana as an entire system on its own with different tasks to explore the use of the SPIRE system[25] as a tool/framework for conducting user-testing and to provide an initial benchmark for future comparisons. We also wanted to confirm that the user interactions captured using the Europeana Search Logging framework could be mapped to participants in the user testing to provide measures of interaction to evaluate search performance.

The protocol used for user-testing based on a controlled task field study is shown in Figure 10. (The full list of questions in this task is shown in Appendix 2.) Firstly, participants are given an introduction about the evaluation task and asked to provide their consent about the use of their results. A registration page is then shown, where participants are asked to enter their participant ID into the Europeana site, to enable their query logs to be identified and

[25] https://promise.sheffield.ac.uk/spire

mapped into their feedback in the evaluation system. After registering, participants are asked to answer a pre-questionnaire, containing questions about their demographics and their last visit to Europeana.

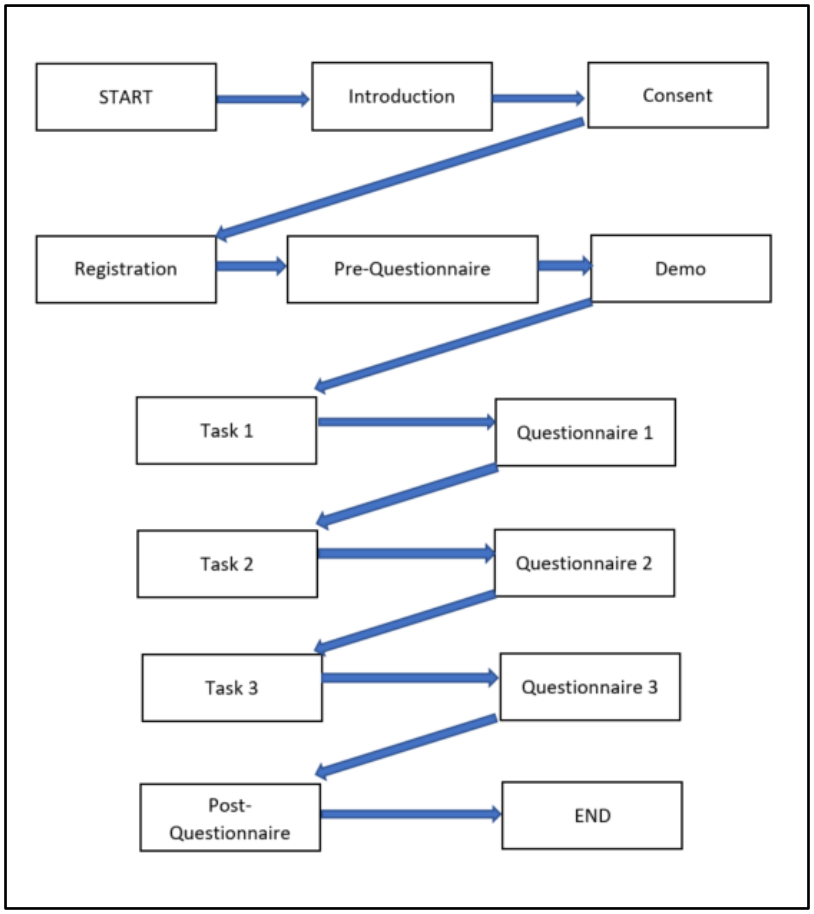

Figure 10. Task-based evaluation protocol

Once completing the pre-questionnaire they were shown the “Demo” page that briefly describes the different methods to access Europeana collections and items. They were also given some time to familiarise themselves with the system, if needed, and then proceeded to complete three evaluation tasks. Following each task, participants were asked to complete a short questionnaire to indicate their familiarity with the topic of the search task, how easy the tasks were to complete, the usefulness of information found, and the rate of success of Europeana in helping them complete the task. They were also asked if they had any comments about using Europeana for the particular task.

Finally, after completing all three tasks and the short questionnaires, participants were asked to complete a post-questionnaire to summarise their experiences with Europeana. Firstly, were asked to rate the overall success of Europeana in helping them complete their tasks, and to rate their satisfaction with Europeana for various components/features provided to support them. The post-questionnaire also asked users to specify how well Europeana supported different types of tasks and also asked for suggestions on how Europeana search and discovery could be improved. Finally, users were asked if they would recommend Europeana to their friends/colleagues using a 5-point Likert Scale (1=not at all, 5=definitely).

Overall, the experiment took around 45 minutes to complete.

We used SPIRE to develop the task-based evaluation. A pilot task was run in a lab-based testing to gather feedback on the evaluation framework and make amendments to the protocol. The lab-based testing was carried out at the University of Sheffield and involved 4 participants (3 students and 1 member of staff at the Information School, University of Sheffield). The evaluation framework used the stages shown in Figure 11, with the addition of a 10-minute post-questionnaire interview. The results gathered from the pilot evaluation task were used to improve the clarity of the tasks and questions in the remote evaluation task. The final tasks and questionnaire proposed in this framework are shown in Appendix 2.

5.3.2.3. Results of task-based user study

The final task-based evaluation was run in SPIRE in a remote evaluation for the period of 2.5 weeks (14th July 2017 - 31st July 2017) involving three different user groups: EuropeanaTech members, Europeana Improvers members, and the volunteers mailing list containing staff and students at the University of Sheffield. Different user groups were selected to investigate how Europeana support users from various backgrounds, and also to examine if participants' behaviours and responses varied across different groups. A total of 51 participants completed the task-based evaluation in full (26 users from EuropeanaTech members, 6 users from Europeana Improvers, and 19 Sheffield Volunteers). Almost half of the participants (47%) had English as their first language, followed by German (14%), Italian (10%), Dutch (6%) and Danish (4%). Overall, 29% of the participants visited Europeana for the first time when they participated in the evaluation task, 22% visiting a few times a month, 18% less than once a month, and 16% visiting once a week and 16% visiting a few times a week. None of the participating users visited Europeana every day. Over one-third identified themselves to be cultural heritage professional (37%), whilst the others were academics (22%), cultural heritage enthusiast (16%), students (8%), teacher (2%) and others (16%).

After completing each task, users were asked to give feedback on their experiences on four aspects (shown in Table 3).

Table 3. Average score (mean) of post-task questionnaire (N=51)

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

At the end of the evaluation, users were asked to specify the overall success of Europeana in helping them fulfill their information needs. An average score (mean) of 3.21 (out of 5) was achieved across all users. Similar scores were found in the EuropeanaTech and Europeana Improvers group (3.35 and 3.33, respectively), and a slightly lower average score (3) in the

Sheffield Volunteers group. We also found that users with higher experience of Europeana (i.e. those who visited Europeana more often) rated a higher success rate than those with less frequent visits or first time visitors (shown in Figure 11). This may imply that users with different degrees of experience on Europeana may have different success criteria. Further work is needed to understand these further in order to improve Europeana success rates for both new users and experienced users.

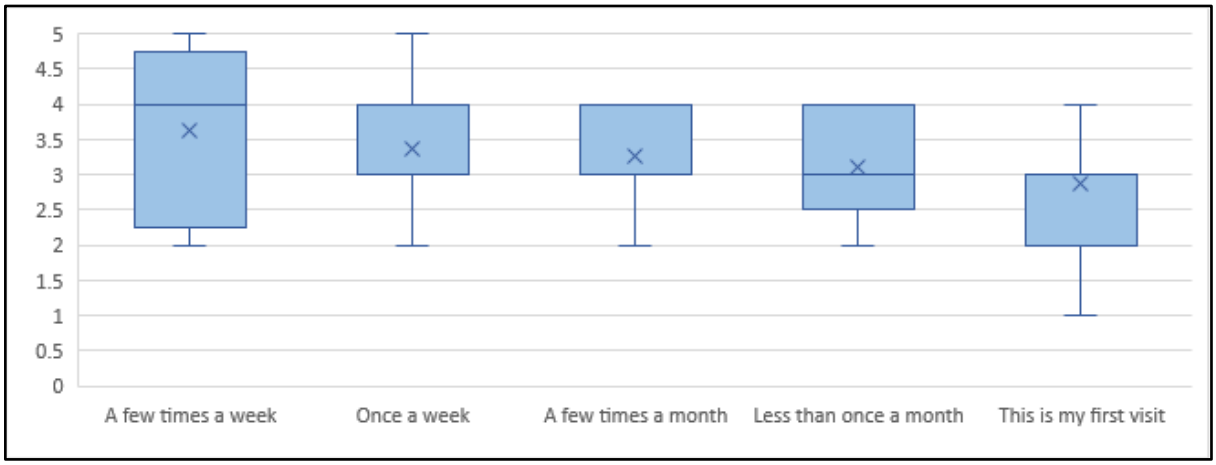

Figure 11. Europeana success rate for users with different visit frequency

Participants were also asked to specify how well Europeana supports different tasks (shown in Table 4) with support for browsing/exploring tasks rated most highly. Europeana is shown to support browsing/exploring task the most (mean=3.42), while searching by specific subject (e.g., named person, place or location) or general subject (e.g., medieval illuminations) were scored the lowest.

Table 4. Average score (mean) across all users for how well Europeana supports different search tasks (N=51)

ID	Task	Mean	St. Dev.
Task 1	Searching for specific known items or finding facts	3.20	1.04
Task 2	Searching for information by specific authors	3.23	1.03
Task 3	Searching for information on named person, place or location (e.g., images of Stuttgart)	3.06	1.02
Task 4	Searching for information on general subjects (e.g., medieval illuminations)	3.07	1.03
Task 5	Browsing or exploring contents with no specific goal in mind	3.42	1.15

In the future, the SPIRE system can be used for future study, such as evaluating different versions of Europeana, evaluation of Europeana components, etc. We describe methods to create an evaluation using SPIRE in Appendix 3. A document describing the full results from the task-based evaluation is available in:

https://docs.google.com/document/d/1z8ITlN6H5Y9Ft9F2R_XqvZOQ6ysyBQFd6_ptsaCJW6w/

Future Actions
Results from the task-based evaluation, including analysis of the log data, need to be analysed in-depth and will be reported in a publication. We also plan to identify behavioural signals that correlate with users' task success. We would recommend the creation of a set of SPIRE templates for conducting various types of experiment with Europeana that could be used for various user studies and user-oriented evaluations involving IIR methods.

6. Other Work Ongoing

6.1 Infrastructural/Architectural Work

Rationalisation of our datastore schema This proved to be both a more arduous but also more valuable exercise than anticipated. Removing extraneous fields from the schema.xml configuration required extensive analysis of our search logs and field usage, in alignment with the Data Quality Committee's extensive work to identify enabling elements for supported search scenarios.[26] Once a reindex was completed with the altered schema, however, the size of Europeana's search index had shrunk from 435GB to 74GB - a reduction of over 80%. The immediate effect of so significant a reduction is improved system stability and performance. In addition, however, this increased headroom opens up the potential for technical improvements. A number of Europeana search components - for example, the choice of search handler and the manner in which Similar Items are retrieved - have been strongly constrained by the fragility of the Solr server and the fact that it has usually been working near the limits of its capacity. A new, and probably more optimistic, assessment of our technical options in future needs to be made accordingly.

Investigation of ElasticSearch as an alternative to Solr Initial investigation of this question appeared to indicate that, while the differences between these two Lucene-based technologies are minor, ElasticSearch is targeting the web-analytics and logging domains and is therefore unlikely to provide a good fit with Europeana's search needs in future.

In addition, much of the motivation for investigating ElasticSearch has been the apparent fragility and unreliability of our existing Solr configuration. The recent radical reduction in the Solr index size, however (see above, *Rationalisation of our datastore schema*), makes this question less urgent.

6.2 Image and Audio Similarity Search

Work with the Austrian Institute of Technology on Image Similarity Search has been planned in Milestone 6.1: Advanced Image Discovery Plan[27] and implementation is going to be reported in Deliverable 6.1, to be submitted at the same time as this deliverable. Initial work

[26] https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/
[27] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms6.1-advanced-image-discovery-development-plan.pdf

has also been carried out preparatory to embarking on Audio Similarity work in subsequent DSIs.

6.3 Horizon Scanning

Of the technologies listed in the original DSI-2 Search Improvement Plan (Mediachain, Persistent Reproducible Identifiers, Commons Machinery, and Pastec), Persistent Reproducible Identifiers and Pastec should continue to be monitored. Initially, the chief application of Mediachain in the Cultural Heritage domain was, essentially, Digital Rights Management. This is of little concern for an organisation dedicated to rights-free redistribution such as Europeana. Furthermore, Mediachain has now been bought by Spotify, which jeopardizes the future of the open source software built by Mediachain Labs. Commons Machinery has been discontinued, with some of the ideas and efforts being now gathered with Persistent Reproducible Identifiers.

6.4 Work Carried Over From DSI1 MS30[28] and MS31[29]

Document the search mechanisms employed by Europeana. Documentation work has continued steadily over the course of DSI-2. [30]The recent changes to the Solr schema mean that updating Europeana's public-facing API documentation is now an urgent priority.

Search in Annotations. Work on assessing the potential contribution of user-generated content to enhancing search is dependent upon the integration of the Annotations API with the Collections site. No date has yet been set for this integration.

Translation in Search. Although the Translation API remains available for interested developers, our efforts are currently concentrated on improving linguistic coverage of the Entity Collection. We have liaised with the CEF AT (eTranslation building block DSI) on this topic, communicating what relevant requirements Europeana could have for their language technology this DSI is developing. Future reintegration of the Translation API into Europeana Collections in its current form appears unlikely. There might be however indirect application of translation in Search, if it is applied to translate the object metadata that Europeana's search index is built on.

Language Detection. While this option should continue to be explored in relation to METIS development, discussions with experts from the British Library (among the only ones we know to have experimented with that technology in our sector) indicate that implementation of a Language Detection framework may provide less benefit than previously hoped: the preponderance of named-entity queries (which tend to share identical or closely-similar labels across languages) in our logs, combined with the somewhat-lowered search precision language-analysis procedures such as stemming tend to introduce, works to lower the

[28] DSI1 MS30. *Search Improvement Plan*, available at http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/euro peana-dsi-ms30-search-improvement-plan.pdf

[29] DSI MS31. *Report on the improvement of search*, available at http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms-31-report-on-the-improvement-of-search.pdf

[30] The most up-to-date (though still partial) documentation can be found at: https://europeanadev.assembla.com/spaces/europeana-r-d/wiki/Search_Infrastructure

perceived value of language-specific processing for end-users. Further testing will need to be performed, then, to determine how desirable this feature may in fact be.

Search taking into account hierarchical objects. The status here is as described in DSI1 MS31. It should be noted that if METIS brings with it the possibility of upgrading our Solr infrastructure, our technical options here might be expanded.

Configure Solr text analysis chain in a language-specific way. As with language detection (see above), discussion with experts at the British Library would seem to indicate that from the user perspective the gains to be had here are marginal. The British Library's use case is, however, quite different from Europeana's, and further testing will need to be performed to determine how useful this functionality is when weighed against the technical overhead of implementation.

Explore ways to mine and exploit query reformulation strategies from users. Little progress has been made here beyond confirmation that the existing logging framework allows the use of various different strategies for discovering user reformulations in our logs.

Metadata translation experiments. Options continue to be explored regarding the multilingual enrichment of the Entity Collection. Preliminary discussions with CEF AT (eTranslation building block DSI) appeared to indicate that the service was an unlikely candidate for use specifically with reference to metadata - though it might prove useful for other aspects of the Europeana Collections site.

Europeana data in Peripleo. A full description of this work can be found in Deliverable 6.4: Pilot for Time-and-Space Discovery.[31]

[31] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d6.4-pilot-for-time-and-place-discovery.pdf

References

Clough, P., Hill, T., Paramita, M. L, and Goodale, P. (2017). Europeana: What Users Search For and Why. In: *Proceedings of TPDL 2017*.

Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: a ten-year update. *Journal of management information systems*, 19(4), 9-30.

Di Santo, G., McCreadie, R., Macdonald, C., and Ounis, I. (2015). Comparing approaches for query autocompletion. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Dumais, S., Jeffries R., Russell D.M., Tang D., and Teevan J. (2014). Understanding User Behavior Through Log Data and Analysis. In: *Olson J., Kellogg W. (eds) Ways of Knowing in HCI*. Springer, New York, NY.

Hofmann, K., Li, L., and Radlinski, F. (2016). Online Evaluation for Information Retrieval. *Foundations and Trends® in Information Retrieval*, 10(1), 1-117. http://dx.doi.org/10.1561/1500000051

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2), 1-224.

White, R. W. (2016). *Interactions with search systems*. Cambridge University Press.

Appendix 1: Pop-up Survey Questions

About You

1. How often do you visit Europeana.eu? [Select one]
   - Every day
   - At least once a week
   - At least once a month
   - Less than once a month
   - This is my first visit

2. How would you identify yourself? [Select one]
   - Cultural heritage enthusiast (e.g., hobbyist, genealogist, amateur historian)
   - Student (e.g., college, university, further education)
   - Academic (e.g., lecturer, professor, post doc researcher, academic support)
   - Teacher (e.g., primary and secondary teaching)
   - Cultural heritage professional (e.g., curator, historian, archivist)
   - Other: _____

3. How did you get to Europeana today? [Select one]
   - Via a link from a search engine (e.g., Google)
   - Via a link from social media (e.g., Facebook)
   - I knew about the site already so came directly here
   - Via a link from teaching resources
   - Other: _____

For your current activity in Europeana.eu

4. What are you currently looking for in Europeana.eu? *(e.g., "I want to find an image of the Mona Lisa", "I'm trying to explore what's available in Europeana on World War I", "I am looking for photographs of Sheffield in the 1980s", "I am looking for artwork by Leonardo Da Vinci", "Don't know / nothing specific")*

5. Why are you looking for this information? *(e.g., "To create a presentation for my student class", "To write an article", "To help plan a visit to Turin and want to know about artworks to visit whilst there", "To learn about the history of English folk music", "General interest / no specific reason")*

6. After finding this information, you will:
   - Look for more information on the same topic using Europeana
   - Look for more information using other resources
   - Browse Europeana (e.g., look for other interesting things)
   - Have completed everything I need to do
   - Other: _____

7. How would you rate your level of subject knowledge for your current activity?

Level of subject knowledge	N/A	No knowledge 1 2 3 4 5 6 7 8 9 10 Extensive knowledge

8. For your current activity, please rate the importance of the following functionality:

Features to allow you to refine your search (e.g. filters)	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Availability of high-quality images	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Detailed information about an object (e.g. provenance, provider institution, etc.)	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
The availability of information about an object in your own language	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Links to download an object	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Access to content you can freely re-use	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Links to an object provider's site (e.g. library, archive, etc.)	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Links to similar items	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important
Links to categorised collections (e.g. Art Nouveau posters, Irish folk music, Fashion	N/A	Not important 1 2 3 4 5 6 7 8 9 10 Very important

illustrations)		

9. What other features could be added / improved to help you complete your current activity?

10. If you have any other comments about Europeana:

Appendix 2: Task-Based Evaluation

This appendix only shows the questionnaire. The full interface for the task-based evaluation (which includes the introduction, consent, registration and demo pages) is shown in: http://paramita.staff.shef.ac.uk/europeana/Europeana%20Evaluation%20Interface.pdf

Pre-Questionnaire

Before starting with the task, we would appreciate it if you could answer the following questionnaire about your background.

Part I: About you

1. What is your gender?*
   - ☐ Female
   - ☐ Male
   - ☐ Prefer not to say
   - ☐ Other: _____

2. Which category below includes your age?*
   - ☐ Under 15
   - ☐ 15-18
   - ☐ 19-24
   - ☐ 25-34
   - ☐ 35-44
   - ☐ 45-54
   - ☐ 55-64
   - ☐ 65+
   - ☐ Prefer not to say

3. Which category below best represents your highest level of education?*
   - ☐ Some high school
   - ☐ High school graduate
   - ☐ College graduate
   - ☐ Trade/technical/vocational
   - ☐ Undergraduate degree
   - ☐ Pre-Questionnaire
   - ☐ Postgraduate degree
   - ☐ PhD
   - ☐ Other: _____

4. What is your native language?*
   - ☐ Dutch
   - ☐ English
   - ☐ French
   - ☐ German
   - ☐ Italian

□ Spanish
□ Other: _____

5. How often do you visit Europeana?*
□ Every day
□ A few times a week
□ Once a week
□ A few times a month
□ Less than once a month
□ This is my first visit

6. How would you identify yourself?*
□ Cultural heritage enthusiast (e.g., hobbyist, genealogist, amateur historian)
□ Student (e.g., college, university, further education)
□ Academic (e.g., lecturer, professor, post doc researcher, academic support)
□ Teacher (e.g., primary and secondary teaching)
□ Cultural heritage professional (e.g., curator, historian, archivist)
□ Other: _____

Part II: Based on the last time you used Europeana

7. What did you look for the last time you visited Europeana?* *(e.g., "I wanted to find an image of the Mona Lisa", "I tried to explore what's available in Europeana on World War I", "I was looking for photographs of Sheffield in the 1980s", "I was looking for artwork by Leonardo Da Vinci", "Nothing specific")*

8. Why did you look for this information?* *(e.g., "To create a presentation for my student class", "To write an article", "To help plan a visit to Turin and want to know about artworks to visit whilst there", "To learn about the history of English folk music", "General interest / no specific reason")*

9. What did you do after finding this information?*
□ I looked for more information on the same topic using Europeana
□ I looked for more information using other resources
□ I browsed Europeana (e.g., looked for other interesting things)
□ I had completed everything I needed to do
□ Other: _____

10. How would you rate your level of subject knowledge for that activity?*
□ N/A □ 1 (No knowledge) □ 2 □ 3 □ 4 □ 5 (Extensive knowledge)

11. Overall, how useful was the information you found in completing your task?*

□ N/A □ 1 (Not at all useful) □ 2 □ 3 □ 4 □ 5 (Very useful)

12. How successful was Europeana in helping you carry out the task?*

□ N/A □ 1 (Not at all successful) □ 2 □ 3 □ 4 □ 5 (Very successful)

13. Which features of Europeana were most important to you in completing your task?

Features to allow you to refine your search (e.g. filters)	N/A	Not at all important 1 2 3 4 5 Very important
Availability of high-quality images	N/A	Not at all important 1 2 3 4 5 Very important
Detailed information about an object (e.g. provenance, provider institution, etc.)	N/A	Not at all important 1 2 3 4 5 Very important
The availability of information about an object in your own language	N/A	Not at all important 1 2 3 4 5 Very important
Links to download an object	N/A	Not at all important 1 2 3 4 5 Very important
Access to content you can freely re-use	N/A	Not at all important 1 2 3 4 5 Very important
Links to an object provider's site (e.g. library, archive, etc.)	N/A	Not at all important 1 2 3 4 5 Very important
Links to similar items	N/A	Not at all important 1 2 3 4 5 Very important
Links to categorised collections (e.g. Art Nouveau posters, Irish folk music, Fashion illustrations)	N/A	Not at all important 1 2 3 4 5 Very important

14. What other features could be added / improved to help you complete that activity?*

Task 1: Specific item / information task (5 minutes)

You will now be given 5 minutes to carry out the task below. Please read it carefully.

"Please use the Europeana site to list the names of three paintings by Juan Gris (a Spanish painter)"

Please enter the names of the paintings in the form below:

- Painting 1: _____
- Painting 2: _____
- Painting 3: _____

Task 1: Questionnaire

1. How familiar are you with the given topic?*
   □ 1 (Not at all familiar) □ 2 □ 3 □ 4 □ 5 (Very familiar)

2. How easy was the task to complete?*
   □ 1 (Very difficult) □ 2 □ 3 □ 4 □ 5 (Very easy)

3. Overall, how useful was the information you found in completing your task?*
   □ 1 (Not at all useful) □ 2 □ 3 □ 4 □ 5 (Very useful)

4. How would you rate the success of Europeana in helping you complete this task?*
   □ 1 (Not at all successful) □ 2 □ 3 □ 4 □ 5 (Very successful)

5. Do you have any further comments about using Europeana for this task?

Task 2: Specific-subject search

Please spend 5-10 minutes to carry out the task below:

"Imagine that you have been asked to create a presentation about a European city or country that you have visited before." You may prefer to illustrate your presentation with old/new images, videos, or other media of your choice. Your presentation may include interesting places to visit, illustrate the history of the city/country, etc. Please find at least 3 different items.

Please add the URLs of your chosen items (3 or more) that you would like to add into your presentation into the text below.

Task 2: Questionnaire

1. How familiar are you with your chosen topic?*
   □ 1 (Not at all familiar) □ 2 □ 3 □ 4 □ 5 (Very familiar)

2. How easy was the task to complete?*
   □ 1 (Very difficult) □ 2 □ 3 □ 4 □ 5 (Very easy)

3. Overall, how useful was the information you found in completing your task?*
   □ 1 (Not at all useful) □ 2 □ 3 □ 4 □ 5 (Very useful)

4. How would you rate the success of Europeana in helping you complete this task?*
   □ 1 (Not at all successful) □ 2 □ 3 □ 4 □ 5 (Very successful)

5. Do you have any further comments about using Europeana for this task?

Task 3: Browsing/exploring Task

You will now be given 10 minutes to carry out the third task. Please read it carefully.

"Please imagine that you are on your lunch break. You would like to spend 10 minutes to either search Europeana for any topic that you are interested in, or to simply explore the items provided by Europeana. After your time has expired, please write a short description of what you learned from your browsing/exploring session." *Note: If you are from a cultural institution, please do not explore your own collection.*

Please describe what you searched for in this session and if you learned any new / interesting information in your exploration task.

Task 3: Questionnaire

1. How familiar are you with the topic you chose to explore?*
   □ 1 (Not at all familiar) □ 2 □ 3 □ 4 □ 5 (Very familiar)

2. How easy was the task to complete?*
   □ 1 (Very difficult) □ 2 □ 3 □ 4 □ 5 (Very easy)

3. Overall, how useful was the information you found in completing your task?*
   □ 1 (Not at all useful) □ 2 □ 3 □ 4 □ 5 (Very useful)

4. How would you rate the success of Europeana in helping you complete this task?*
   □ 1 (Not at all successful) □ 2 □ 3 □ 4 □ 5 (Very successful)

5. Do you have any further comments about using Europeana for this task?

Post-Questionnaire

Congratulations, you have finished all of your tasks.

Finally, we would appreciate it if you could answer the following questions regarding your experience of Europeana and the evaluation tasks:

1. How would you rate the overall success of Europeana in helping you to fulfil your information needs or complete your tasks?*

□ 1 (Not at all successful) □ 2 □ 3 □ 4 □ 5 (Very successful)

2. Please rate your satisfaction with Europeana on the following:*

Relevance of search results	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Relevant search results ranked highly	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Diversity and variety of search results	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Completeness of search results	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Response time during interaction	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Overall ease of use of the site	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Access to online help documentation	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
The level of description about objects	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Support provided for browsing and exploration	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Availability of links to similar items	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Availability of links to categorised collections (e.g., Art Nouveau posters, Irish folk music, Fashion illustrations)	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Availability of options to refine your search (e.g., filters)	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Availability of links to download an object	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Access to content you can freely reuse	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied

Availability of links to an object provider's site	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
The ability of the site to provide me with inspiration	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
The enjoyment I get from interacting with the site	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied

3. How well does Europeana support the following types of tasks?*

Searching for specific known items or finding facts	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Searching for information by specific authors	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Searching for information on named person, place or location (e.g., images of Stuttgart)	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Searching for information on general subjects (e.g., medieval illuminations)	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied
Browsing or exploring contents with no specific goal in mind	N/A	Very dissatisfied 1 2 3 4 5 Very satisfied

4. In your opinion, how could search and discovery on Europeana be improved?*

5. Would you recommend Europeana to your friends and colleagues?*

□ 1 (Not at all) □ 2 □ 3 □ 4 □ 5 (Definitely)

Appendix 3: SPIRE

The SPIRE system can be accessed using the following URL: https://promise.sheffield.ac.uk/spire/. First, you need to click “Register” to create an account in SPIRE. Once registered, you will be able to create an experiment by clicking the icon on the top-right of the site (Figure 7 below). You will be asked to specify a name for your experiment (Figure 8) and click “Create”.

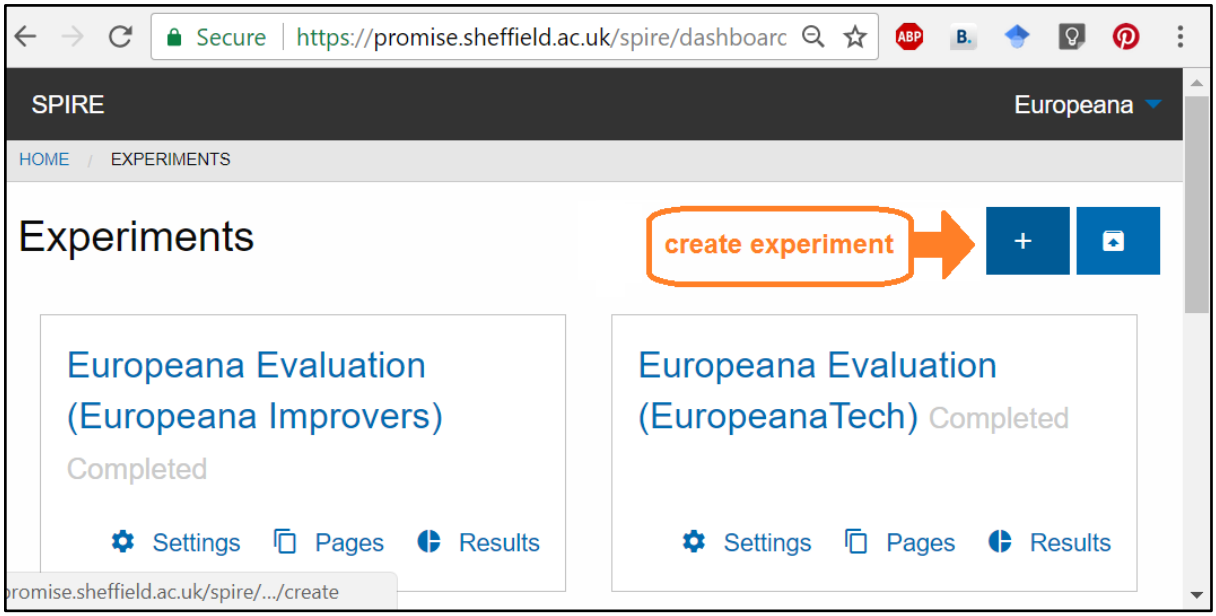

Figure 7. Create experiment

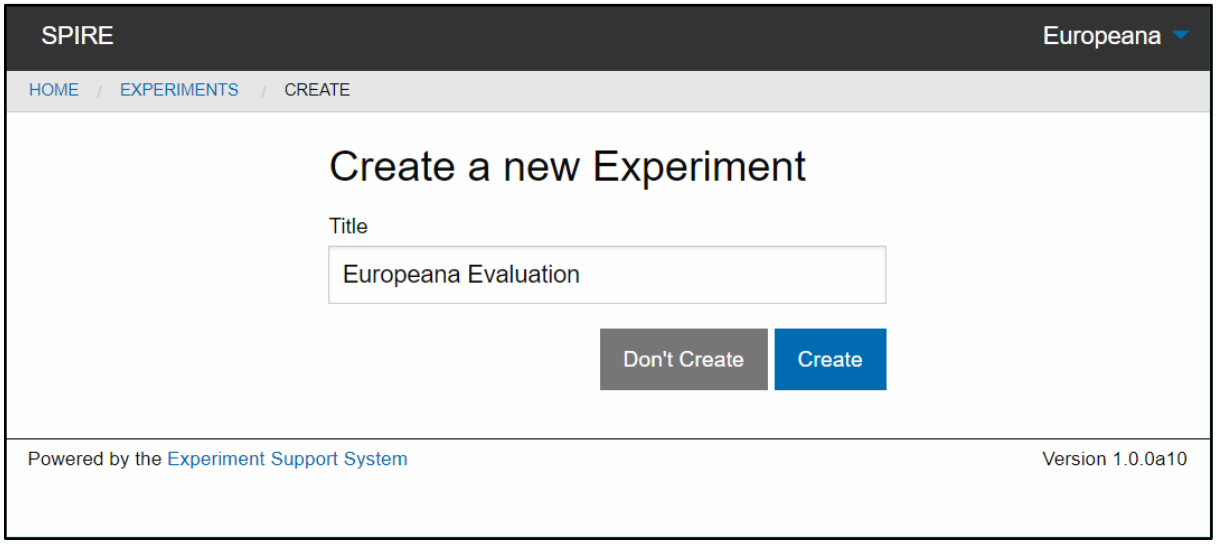

Figure 8. Name the new experiment

Once the experiment is created, you can start creating pages for the evaluation task by going to the “All Pages” menu (Figure 9) and create a new page (Figure 10).

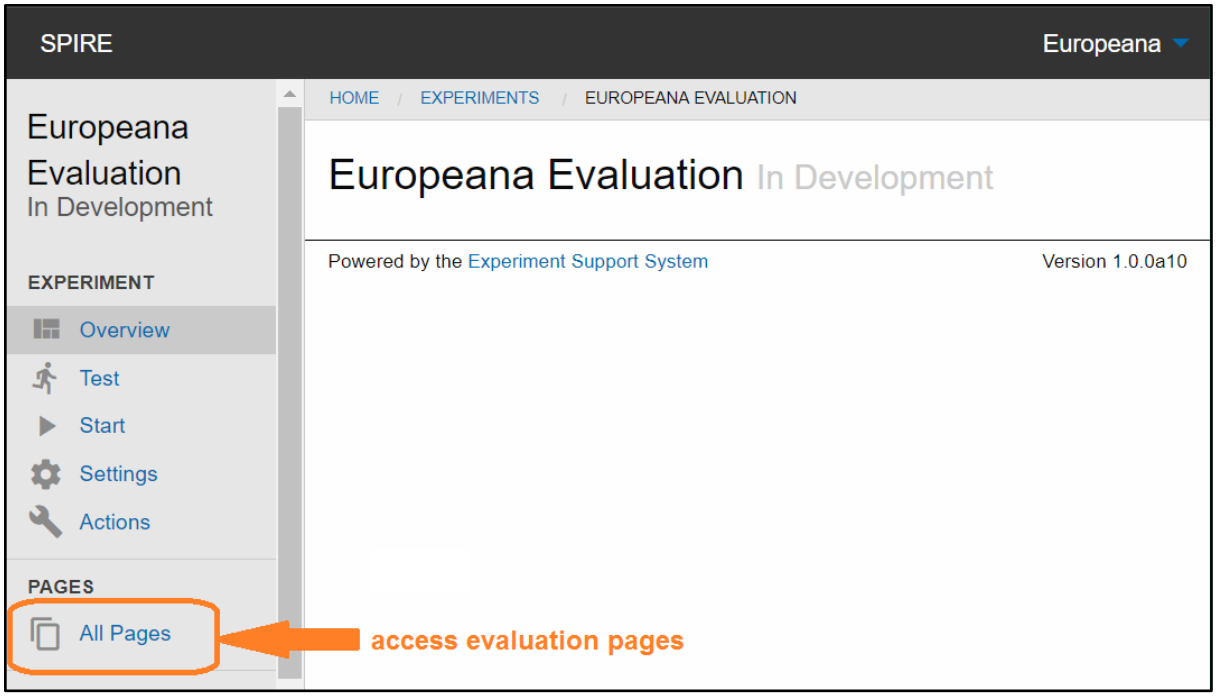

Figure 9. Access evaluation pages

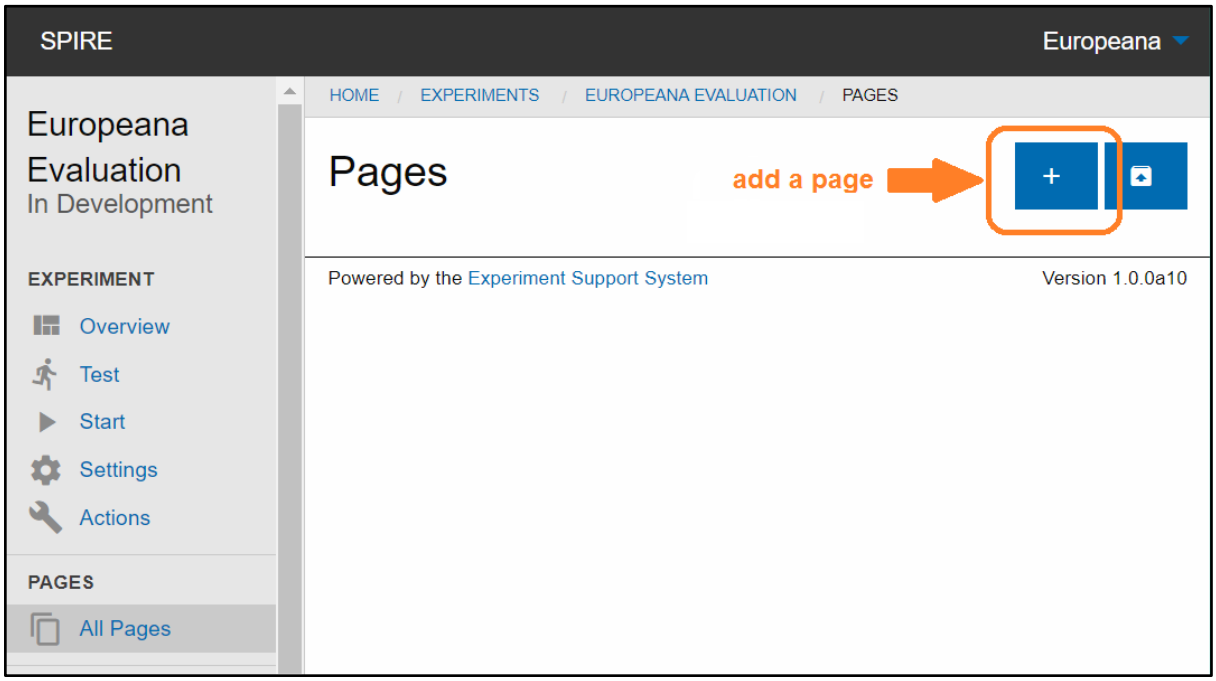

Figure 10. Add an evaluation page

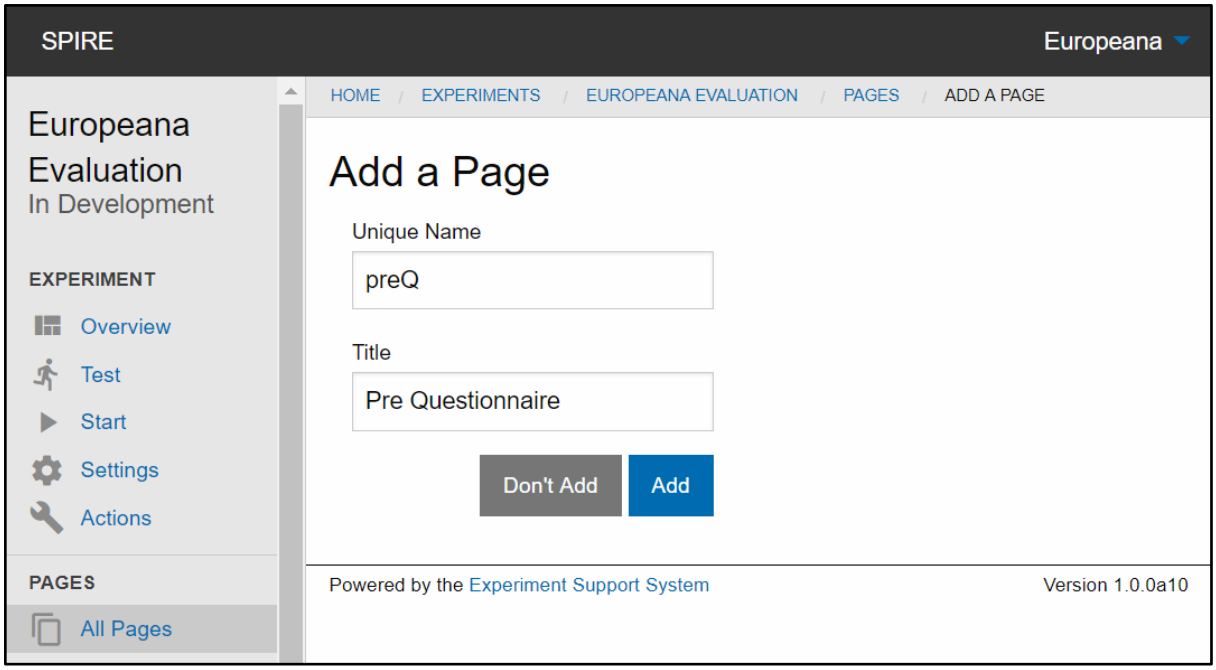

Figure 11. Add the page name

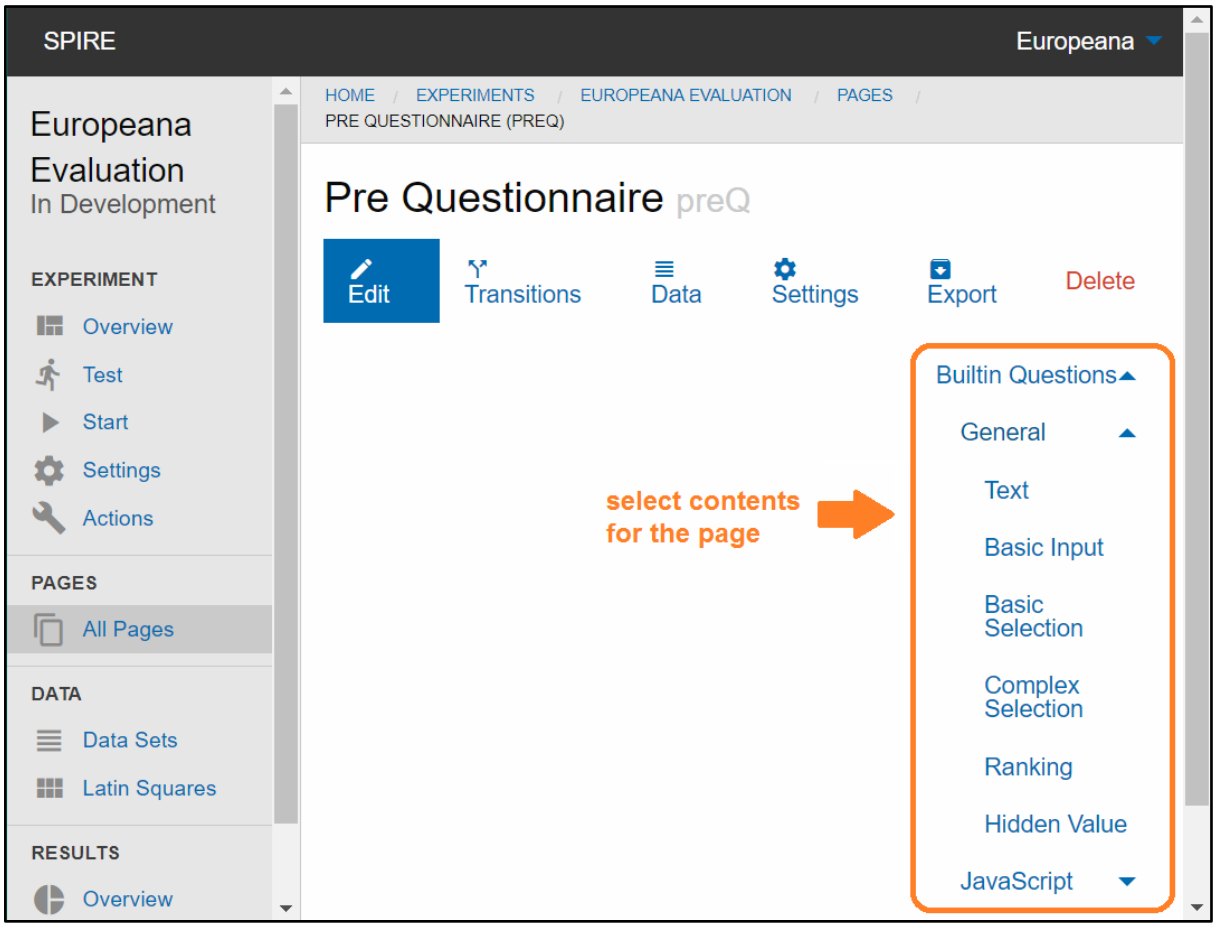

Figure 12. Adding contents into the page

Once the page name is specified, you will be able to add the contents of the page by selecting one of the options (shown on the right hand column) of Figure 12:

- Text: This is to add an HTML code into the site, such as to add an introduction paragraph, a link to external site (e.g. Europeana), etc.
- Basic input: This is to gather an open feedback from users (e.g. “What are you currently looking for in Europeana?”)
- Basic selection: This is to add a multiple option question into the page (e.g. “How often do you visit Europeana?” with optional answers provided)
- Complex selection: This can be used to create a matrix question (e.g. please see Q8 in Appendix 1)
- Ranking: This is to ask users to rank some items
- Hidden value: This is to save any hidden value into the dataset during the evaluation (e.g. timestamp, user ID, etc.)

You will also be able to add any JavaScript code (e.g. to create a timer for the task), if needed.

For example, if the “Basic Selection” option is chosen, you will be asked to specify a unique name for the question, the question title (that will be shown to user), and the optional answers (shown in Figure 13). Once they have been filled, the question can be saved using the green tickmark on the top right page. It will be added to the page automatically.

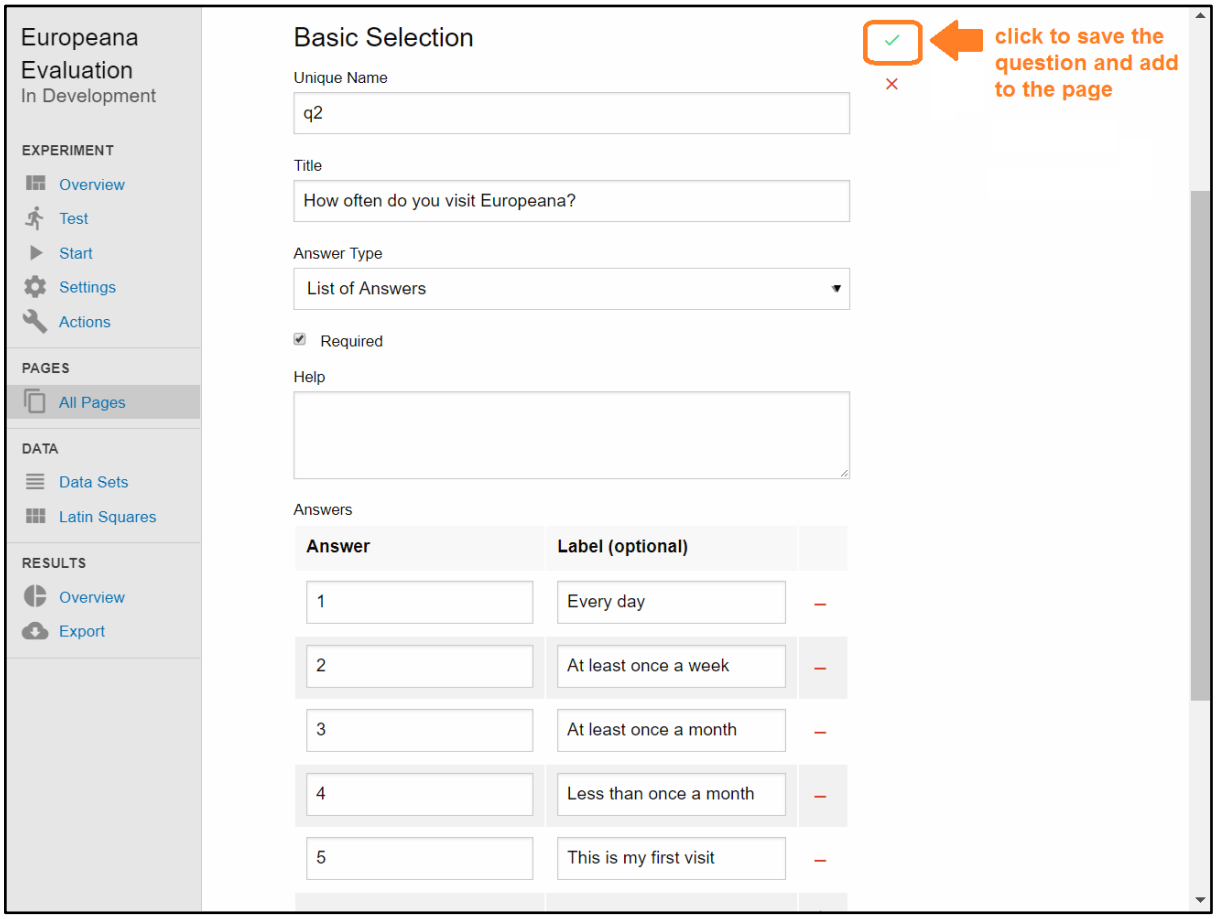

Figure 13. Basic selection questions

You can add as many pages as needed into the evaluation site. Once all the pages are created, you can add transitions between the pages using the “Transition” menu (shown in Figure 14).

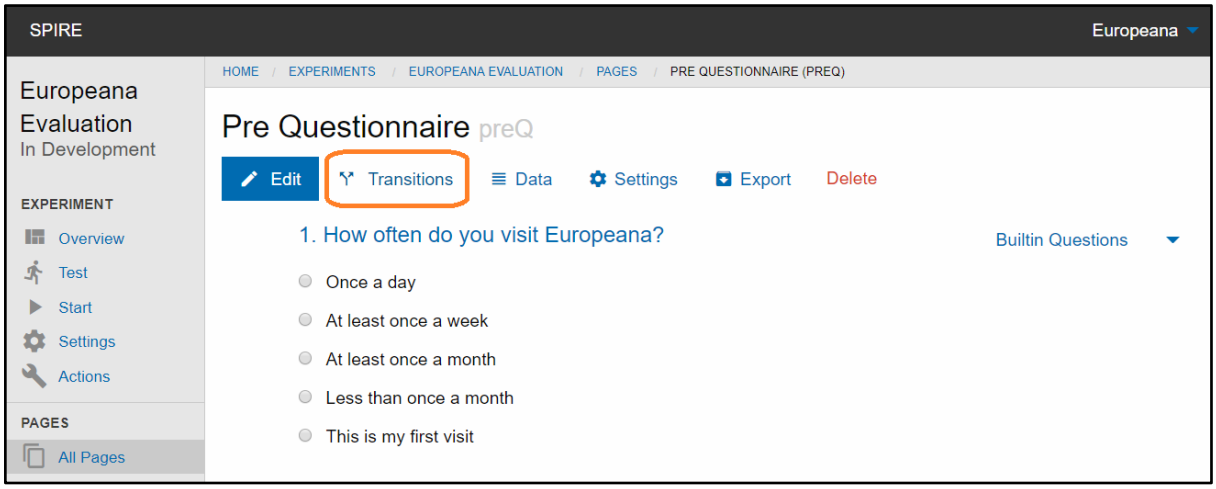

Figure 14. Transitions menu

For example, if you want users to start the task once they complete the pre-questionnaire page, you can click “Add a Transition”, and select “Task 1” as the target page (Figure 15).

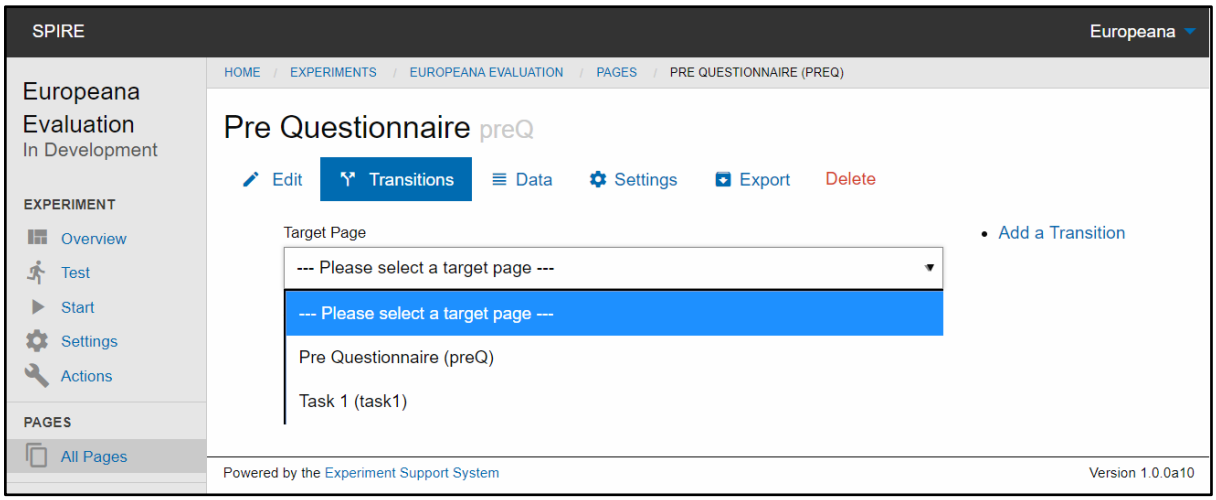

Figure 15. Adding a transition to another page

Once all the pages and transitions have been added, you will be able to test your system using the “Test” link. Experiment can be started by clicking “Start”. You will be able to get a summary of your results in the “Overview” link (this includes the number of completed data, abandoned results, etc.), and finally, the results can be downloaded using the “Export” menu. The results are produced in a CSV format.