Europeana Search Strategy

23/05/2016

Editors
Timothy Hill, David Haskiya, Antoine Isaac, Hugo Manguinhas and Valentine Charles (Europeana Foundation R&D)

Co-financed by the European Union
Connecting Europe Facility

Introduction

Search has long been acknowledged as a difficult area for the Europeana platform.

The problem exists on two levels.

On the one hand, with 48 million+ records in some 30 languages, of extremely varied nature and quality, and a small and specialised user-base, there is simply a problem locating what people are looking for. Multilinguality, heterogeneity, user diversity are all well-recognised challenges in web search, and it cannot be said that Europeana has always met them successfully.

On a deeper level, the very nature of the Europeana project poses distinct - and in some ways unique - challenges of its own. The stated ambition of the Europeana site is to 'transform the world with culture'. And the concept of 'culture' is of course a complex and multilayered one; at the very least, its presence in the mission statement would seem to imply that it is envisaged as something more than a queryable database of artefacts. This was much of the substance of the critiques of Europeana made by Seb Chan and George Oates at the EuropeanaTech 2015 conference: the search box paradigm is too narrow and reductive for Europeana's galleries, libraries, archives, and museum (GLAM) focus. As Oates puts it,

> "When you think about an amazing cultural heritage collection and your first entry point is a search box that is just a tragedy … Things like Google are good when you need to search everywhere for anything, but in our [GLAM] world, we know what sort of content we are looking at, so let's pay a bit of attention to that content and design around it. "[1]

Chan is even blunter: "We haven't come up with a better general interface than search for presenting vast collections, but we know that search is not the best way of doing things. It might be the best way if you know what you are looking for, but it is terrible for everything else".[2]

Yet the solutions advocated and implemented by Chan and Oates - interfaces to relatively small collections lovingly wrought and strongly rooted in the curatorial traditions of a single, physical museum[3] - also do not quite fit the Europeana case. Europeana is something different - much larger, less coherent, and more diverse - than the individual institutions that make it up.

The remainder of this document is an attempt to answer the question of the search and navigation paradigm required for a platform that attempts to represent not just an individual exhibition or organisation, but the GLAM sector as a whole. How does one navigate and discover 'culture', at scale?

[1] Pekel, 2015a.
[2] Pekel, 2015b.
[3] See for instance Oates' work with the Victoria & Albert Museum (http://va.goodformandspectacle.com/) and Chan's with Cooper Hewitt (http://labs.cooperhewitt.org/).

Context

These questions arise in the context of significant technical and structural changes in the Europeana platform.

Of most immediate relevance is the ongoing work on improving known-item search. This is the bedrock of any web-search service: users with a clear idea of what they are looking for need to be able to find it easily and immediately, or the entire platform is next to useless. Accordingly, work in evaluating known-item search and exploration of a variety of techniques for improvement, such as machine-learning approaches, has been ongoing, and a Search Improvement Plan[4] has been put in place outlining the progress to be made in known-item search over the next five months.

Also of immediate relevance is the recent launch of Europeana Collections[5], replacing the Europeana Portal. This new platform in itself leaves search essentially unaltered; the existence of more specialised collections and exhibits, however, can be expected to have an effect on user behavior. In addition, changes in UI and UX will plausibly affect users' interactions with the site. While the cumulative impact of these changes is difficult to predict, one might expect a drop in the number of broad, vaguely-formulated query terms used as initial entry points into the Europeana corpus.

Meanwhile, the platform's functionality is set to expand in the medium term with the further development of the Annotations API[6]. This API will allow users to annotate Europeana records in a variety of ways, including freetext and semantic tagging, as well as object-to-object linking. The Annotations API is intended to be made available for public use by July 2016.

The most significant change to the platform, however, is the creation of the Entity Collection and associated Entity API[7], with first iterations being available in 2016 Q1. While on the one hand this change can be more or less invisible to end-users, with progressive enhancement allowing gradual modification over time, on another it involves a fundamental reconceptualisation of the platform. Currently, the Europeana corpus is fundamentally a document-store searchable by keyword; when the work on 'entityfication' is complete the user will instead be navigating a network of entities defined in large part by their connections to each other. Any medium- to long-term search strategy needs to take into account this underlying change in platform and data paradigm.

Challenges

The uniqueness, value, and challenge of the Europeana platform arises, from a search and information-architecture perspective, from its existence in a state of creative tension among three sets of opposed poles.

[4] Hill, Isaac, and Charles 2015.
[5] http://www.europeana.eu/portal/
[6] https://docs.google.com/document/d/1R2_RDuP5E2n0SBnZOqvjedqmKVoBavo6wtS0pN1UJtA/edit#heading=h.cqat0tgajpb2
[7] https://docs.google.com/document/d/16Rw_qlSpINxztGpI5sM6NcXhnQeZkW3hRA1IYjFtE1w/edit#heading=h.l2fg46yn5tej

First, as noted above, there is the tension between the intense focus on individual objects traditionally characteristic of GLAM institutions, and Europeana's status as a truly mass aggregator. Researchers in the humanities, artists, and designers such as Oates and Chan, have typically taken small and coherent collections as the focus of their attention, with an emphasis on deep and immersive engagement with particular works of art encountered in their distinctive cultural, historical, and aesthetic context. By contrast, the nascent fields of data science and the digital humanities tend to focus instead on trends and patterns that emerge only when phenomena are considered at scale. Maintaining the GLAM sector's traditional richness and intense curation while at the same time realising the Big Data potential of Europeana's large scale is accordingly central to any successful search strategy.

A second tension arises from consideration of the needs and expectations of users. Reviewing the personas developed for the D6.2 Requirements for Europeana.eu[8] deliverable - and in particular those of the 'culture vultures' - one finds two somewhat-opposed requirements. On the one hand, they need to be able to find what they are looking for, and navigate through clear and well-structured data.[9] On the other hand, they also come to Europeana looking for 'inspiration'.[10] - that is to say, for something new and unexpected that points them towards possibilities they had previously been unaware of; what, in the formal literature of user experience and search design, is sometimes referred to as 'serendipity search'.[11] Europeana's users need the platform to be structured and predictable - but not entirely so.

Another tension within user needs is that between their requirements in the context of Europeana and expectations they bring with them from their experience of other applications - including, but not limited to, Google. If we analyse the D6.2 personas in terms of the influential taxonomy of search modes[12] developed by Russell-Rose, Lamantia and Burrell, we find their needs cluster around the 'Comprehend', 'Explore', and 'Synthesize' behaviours.[13] - or, in terms of a slightly different analytical framework, we find that they are engaged in 'sensemaking'..[14] This is an extremely distinctive profile. The majority of search applications are focused on relatively concrete, targeted, and better-understood tasks such as 'Locate', 'Verify' and 'Analyze'. In practical terms, then, there is a relatively low ceiling to how much we can unhesitatingly adopt or adapt approaches taken from other platforms. Correspondingly, we can anticipate that users' mode of interaction with Europeana will be in the first instance to some extent maladaptive, in the sense that any learned behaviours are unlikely to in fact match their information need.

8

https://asset1.basecamp.com/1768384/projects/2913457/attachments/184610002/11f0de4445ca921f264a947455de6efe0010/original/D6.2_Requirements_europeanaeu.pdf

9 See e.g., Marcel, who wants to 'filter search results in clear, understandable ways'; Linda, who wants to filter images by 'color table, color name & paint name'; Paul's need to 'search for content under a clear structure', and Marie, who 'would like to filter by dates, seasons, designers, or materials'.

10 See the personas for Linda, James, and Marie.

11 On serendipitous search and 'not knowing what you need to know', see Russell-Rose and Tate 2013, 76-79; further Spencer 2006.

12 https://isquared.wordpress.com/2011/11/02/a-taxonomy-of-enterprise-search-and-discovery/

13 Of course, this is not to say that Europeana users do not engage in the other modes identified. But this is where the preponderance of activity is.

14 The term is a common one in the search design literature; its contemporary sense originates from work at Xerox-PARC. See Russell, Stefik, Pirolli, and Card (1993).

Strategy

As the points listed in the Challenges section indicate, the Europeana platform is in its search requirements a special case - distinctive, and possibly unique. The best approach to its search, navigation, and information architecture is accordingly an open research question.

What follows, then, is not a search strategy in the sense of a set of defined steps to be taken with known results obtained. It is instead a research programme, aimed at the question of the underlying search structure to be implemented rather than optimising a known approach.

Evaluation

Fundamental to pursuit of a research programme is defining a means of evaluating it. Because Europeana's profile is so unusual, measurements widely employed as success metrics in search - for example, precision vs. recall or nDCG ratings - are not necessarily applicable to the platform, and certainly not in isolation from other measures.

We are accordingly pursuing a partnership with Professor Paul Clough and the University of Sheffield to define a framework for evaluating Europeana's search architecture. Professor Clough and his students have worked extensively on search evaluation[15] in digital heritage and other areas of the humanities, and have collaborated with Europeana in the past[16]. The current round of collaboration is projected to last 6 - 8 person-months, allocated over a period of 14 - 16 calendar months.[17]

Approaches

If the Europeana platform's search structure is an area for ongoing research, then options should not be foreclosed, and explored as they seem necessary or required. That said, three areas can readily be identified as useful for immediate focus.

Surfacing Users' Mental Models

As noted above, one requirement arising consistently from the D6.2 personas is a need for a clear information structure. This is consistent with their engagement in 'sensemaking' activities, a significant component of which is the user's discovery and refinement of his or her own mental models - often referred to in the literature as 'internal schemas' - for the domain being explored, and their subsequent comparison with, and elucidation as, shared, communicative structures - referred to as 'external schemas'.[18]

The difficulty, of course, is that we have at best only a very coarse-grained understanding of what our users' mental models are, and therefore of what schemas they would find most useful in navigating and understanding Europeana's content. There is no schema for 'culture' - and while

[15] http://www.sheffield.ac.uk/is/staff/cloughallpubs
[16] http://pro.europeana.eu/project/paths
[17] Though note that this time estimate also potentially includes work on Named-Entity Recognition (NER) using the University's General Architecture for Text Engineering (GATE) framework.
[18] See generally Russell, Stefik, Pirolli, and Card (1993); also Russell-Rose and Tate (2013), 34-6.

individual disciplines in some cases may have articulated well-developed taxonomies for their subjects, these relatively well-defined structures often differ from country to country and tradition to tradition, or are incompatible in whole or in part with each other.[19] The appeal of new fields such as the digital humanities, furthermore, lies to a large extent in their tendency to cut through previously-established disciplinary boundaries, and in this context much of Europeana's value lies in its ability to collocate items normally kept physically and institutionally separate.

Given that we can rely only minimally upon disciplinary or other pre-established schemas to structure our content, then, how do we discover the internal schemas our users are bringing to bear upon our content?

There are several resources at our disposal.

- *Query logs.* These record the search terms people are using, as well as the results they eventually click on. These logs, and in particular query reformulations, are currently the primary evidence we have for the mental categories users employ when attempting to navigate our content

- *Annotations.* The tags and relations added by users to Europeana contents will provide a more direct insight into how they understand our content and its structure than the query logs.

- *User Sets.* The Europeana Sets API[20] allows users to save and label their own searches. The searches and sets thus created have the potential to form a valuable guide to user mental categories.

- *Formal/disciplinary ontologies and models.* For reasons noted above, we cannot simply apply existing schemas to our data and expect our users to comprehend and/or benefit from them. That said, such schemas are typically grounded in a deep understanding and appreciation of the domain(s) they cover, have evolved to avoid many of the pitfalls encountered by domain practitioners, and furthermore may form the basis of communication among students and experts in that domain. As such, they should be used to guide and inform experimentation and development.

In addition to the above datasets, which are either pre-existing or arise naturally as a result of user interaction with the site, a very direct means of enquiry into user mental models might be provided by experimental clustering; that is to say, Solr's native clustering functionality could be used to generate dynamic facets based on the terms present in returned documents, and a clicklog of these would then indicate clearly which terms were considered central to the domain in question. Testing would be required, however, to determine whether such dynamic facets might adversely affect users' experience of the rest of the site.

[19] Archaeology, for example, is one of the most fertile of the humanities disciplines in its creation of formal ontologies, both digital and non-digital. This, however, has created its own challenges, and it appears that an excess of structure may create at least as many problems as not enough; see for example the presentation by Isaksen, L., Martinez, K., and Earl, G. (2009) on "Archaeology, Formality, and the CIDOC CRM".

[20] https://docs.google.com/document/d/15nqqs7M9V25iku9NsiEfXl-vJpWh31LmyaYVA__cvho/edit#heading=h.7hc21jtki1ok

Preliminary data inspection will need to be done to determine what information can be derived from each of these sources, and how this data could best be exploited in our search functionality. Possibilities include making user annotations directly searchable, machine learning approaches blending user data into relevance rankings, and record enrichment based on user feedback, among others; and of course there is no reason for all of these approaches to be mutually exclusive. Care will in addition need to be taken regarding biases introduced by each of these sources, so that, e.g., reliance upon sources such as user annotations and sets does not slant the findings towards expert users. However, it is idle to speculate further in the absence of further evidence from the datasets listed above.

User Interface and User Experience Changes

As the discussion of internal and external user schemas indicates, search and sensemaking functionality is at least as much about user cognition as it is about inherent characteristics of the data itself. User Interface (UI) and User Experience (UX) design accordingly affect search significantly, and experimentation should be done here to determine what UI/UX factors serve to promote clarity and understanding for users.

As elsewhere, it would not do to be over-prescriptive here. The following areas, however, suggest themselves as promising areas for further investigation, in roughly ascending order of importance:

1. *Link Mode.* That is to say, whether links should open in the same tab, a new tab, or a new window.

2. *Breadcrumb and breadbox building.* There are a number of ways of indicating a user's navigational history. These should be experimented with.

3. *Hit-highlighting on the Search Results Page* (SERP). What fields to favour, how much context to show, and how best to indicate hits visually all need to be investigated.

4. *Result clustering*. This falls into three categories.

    i. *Media-type clustering*. Most web-search engines cluster results by media-type (video, images, text, etc.) and present the results in tabbed form. We should investigate how useful this is for Europeana.

    ii. *Parent/child clustering*. Many object groupings have a natural parent/child structure - for example audio CDs and their constituent tracks. Failure to cluster these under a single heading means our SERPs are often flooded with dozens of closely-related items that might well be better presented together.

    iii. *Semantic clustering*. Europeana SERPS often have a problem with term polysemy: 'Turkey' as a bird and as a country; 'Athena' as either a goddess or a city; 'Freud' as a therapist, painter, or photographer. Clustering these results by their sense would almost certainly make our SERPs more coherent and understandable to users.

While not wishing to prejudge the work with Professor Clough on search evaluation, A/B comparison probably provides us with a straightforward means of testing all of the above approaches.

Diversification of Search Results

A range of considerations - most notably, term polysemy, knowledge-domain representation, and the need to avoid flooding the SERP with near-identical items - point toward the desirability of maintaining a degree of diversity in our search results. This is, however, a delicate balancing act, as we also need to ensure that the items most relevant to a given search query are presented earlier in the result-set. Research accordingly needs to be performed into the degree of diversity that ought to be maintained, the means by which this is to be achieved, and how this should be presented to the end user.

Investigation should initially focus in the first instance on semantic clustering (see User Interface and User Experience Changes, point 4c, above) as a simple means of allowing categorisation within the SERP. It is, however, conceivable that other approaches show themselves viable or desirable - for example, ensuring diversity of media-type or provider through crafting of a custom Solr plugin.

Constituting Knowledge Domains Through Object Linking

Through organisation in accordance with user schemas and appropriate UI and UX design choices, we should be able to ensure the platform's ease of use and clarity for users. This, however, leaves the problem of inspiration and serendipity - the capacity for users to find records they need, but didn't know they needed.

Serendipity search is a famously paradoxical problem, and difficult to design for.[21] The only acceptably unobtrusive means of integrating it with web-search, however, appears to be through document-to-document linking:[22] by following associational chains of related items, users can arrive at points connected to their original search that were nevertheless not captured by it.

A focus on document-linking sits well with the development of the Entity Collection and Entity API - that is to say, the transformation of the Europeana data architecture into an inherently-linked object graph. This, however, leaves open three questions:

1. What kinds of links will prove useful to our users? For example, should they be untyped or typed? If typed, what types should be provided?

2. How are such links between entities to be generated?

3. How should such links be integrated with other search functionality? Should they be searchable in themselves?

[21] Russell-Rose and Tate describe it as the 'holy grail of the search experience' (Russell-Rose and Tate 2013, 77-8). Since Eli Pariser's 2011 book The Filter Bubble, it has also been something of a cultural and political issue; see further Morozov, E. (2013), 140-180, for an overview and a scathing critique of Google's views on engineering this.
[22] Russell-Rose and Tate (2013), 78, citing Spencer (2006).

With regard to the first question, the only sensible way to proceed would seem to be to define the kinds of links possible as broadly as possible, and then use clickstream data to eliminate those which appear useless or harmful: in the first instance, both typed and untyped links should be available, with as many individual types as is convenient defined. Sources for these definitions could be various, derived from a review of existing schemas, from discussion with domain experts, by harvesting tags from the Annotations API, and so forth. For example, the kinds of relations useful for art-history records might at first be determined by reviewing RKDArtists[23] and other sites; CIDOC-CRM and other vocabularies; discussing the options with our Art History Digital Collections Manager; and analysing our logs. This would yield a rough and ready ontology that could then be winnowed back through assessment of what links are actually followed by users.

The task of populating such links might similarly be handled in a variety of ways. Some relations could be populated through a process of inferencing: for example, if two Agents are listed as creators of an artistic work, a relationship of (at least) foaf:knows might be deduced to exist between them. Relations between entities might also be harvested from existing sources, such as DBPedia, Wikidata and, most recently, the Getty Institute's Cultural Objects Name Authority[24] (CONA). And finally, if appropriate UI/UX support is given - for example, population of a dropdown with appropriate relations whenever a user opts to assert that two entities are connected in some way - relations could be defined using the Annotations API. Assuming some degree of uptake and a smooth user experience towards making such assertions, our object graph could be richly connected with contextual relations.

The final question, of how such relationships should in turn affect search functionality, can only be answered once we can in fact draw upon such relationships. It should be noted, however, that the impact can only be positive. Every important web-search application of the past decade and a half employs the number and nature of inter-document links as a significant parameter in its relevance algorithms. The importance of this parameter for the Europeana platform, with its relatively small and extremely diverse user-base, will almost certainly be immense. In other words, a focus on context, serendipity, and navigation over the blank face of the search box will end up also improving the utility of that search box immensely.

Conclusion

There is a way in which the mutual reinforcement of known-item and exploratory, sensemaking search should be unsurprising. Objects only make sense in context; and that context in turn consists of other objects. This interdependence between entity and environment is well known - and it is an aspect of experience to which humanities scholars, artists, and other agents in the field of 'culture' are particularly sensitive. In anthropologist Clifford Geertz's phrase, understanding culture, or a culture, is an act of 'thick description' that encompasses both that which one seeks to understand and the context that makes it comprehensible.[25] And it is presumably the absence or thinness of the context provided by an empty search box and an endlessly-mutable SERP that critics such as Chan, Oates, and the proponents of the traditional

[23] http://english.rkd.nl/Databases/databases-old/RKDartists
[24] http://getty.edu/research/tools/vocabularies/cona/index.html
[25] See Geertz 1973, "Thick Description: Toward an Interpretive Theory of Culture"; for the Europeana context in particular see Gradmann 2010, in particular sections 2 and 5.

physical museum are thinking of when they complain about the 'illegibility' of the search interface, or that metadata does not 'tell stories'.[26]

But Europeana is, as an aggregator, in a unique position. It lacks the capacity for bespoke curation readily provided by both the physical and digital infrastructure of some well-funded physical institutions. On the other hand, with its corpus of 48 million records and growing, it is capable of providing more depth of understanding, through provision of greater context, than any other single organisation. It is a question of ensuring that this context is visible, available, and meaningful.

This is, of course, a challenge. The corpus is large, the user-expectation high, and the approach somewhat experimental. But it is precisely the research-oriented aspect of the work that makes it valuable: Europeana represents a unique opportunity to do something new. Users who enter the site expecting it to behave like Google - to fire off a query and get back a single hit that provides an unequivocal answer - will usually be disappointed, because the platform is in fact not a web-search engine; while users who enter it expecting it to be like going to, say, the Victoria and Albert Museum will be disappointed, because we are not in fact a single, physical institution. With the right data structure, Europeana will be something new, and different: a means not just for finding particular cultural artefacts, or even of exploring 'culture' - but a new way of engaging with and understanding culture. Or of extending and repurposing it.

The strategy outlined above can be seen as, in a sense, a collaboration between Europeana and its users: ultimately, it is they whom we are relying upon to provide structure and connectivity to our content. As such, it is a long-term strategy: the resource will grow richer over time, and needs some minimum amount of user-driven data to bootstrap itself effectively. Considered in another way, however, its implementation focus is short- and medium-term: we need first to determine what kind of scaffolding users require in navigating our content, and then to provide it. Once this structure is in place, 'search' is no longer simply a service provided by the Europeana platform. It is a kind of dialogue by and for users,[27] both with each other and with our content - allowing them to participate in culture, and, through participating, help to create it.

[26] 'Illegibility' is Oates, in the interview cited above; the observation on metadata's lack of narrative is from the interview with Chan.
[27] On the concept of search as a dialogic process, see Russell-Rose and Tate (2013), 47-65.

References

(Geertz 1973) Geertz, Clifford: “Thick Description: Toward an Interpretive Theory of Culture”, in *The Interpretation of Cultures: Selected Essays* (Basic Books), 3-33.

(Gradmann 2010) Gradmann, Stefan: *Knowledge = Information in Context. On the Importance of Semantic Contextualisation in Europeana*. Europeana White Paper 1. Retrieved 31.03.2016 from http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20White%20Paper%201.pdf.

(Haskiya, van Duijn, and Gardien 2015) D. Haskiya, M. van Duijn, R. Gardien. *Core Service Platform: Requirements for the Platform Backend*. Europeana Deliverable D6.1. Retrieved 31.03.2016 from http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d6.1-requirementsfortheplatformbackend.docx.pdf.

(Hill, Isaac, and Charles 2015) T. Hill, A. Isaac, V. Charles. Search Improvement Plan. Europeana Milestone MS30. Retrieved 31.03.2016 from http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms30-search-improvement-plan.pdf.

(Morozov 2013) Morozov, Evgeny:*To Save Everything, Click Here: The Folly of Technological Solutionism* (Public Affairs).

(Pariser 2011) Pariser, Eli: *The Filter Bubble: What the Internet is Hiding From You* (Penguin Press).

(Pekel 2015a) Pekel, Joris: “Discovery without search: an interview with George Oates”. *Europeana Pro Blog*. Retrieved 31.03.2016 from http://pro.europeana.eu/blogpost/discovery-without-search-an-interview-with-george-oates.

(Pekel 2016b) Pekel, Joris: “Making a museum part of the internet: an interview with Seb Chan” *Europeana Pro Blog*. Retrieved 31.03.2016 from http://pro.europeana.eu/blogpost/making-a-museum-part-of-the-internet-an-interview-with-seb-chan.

(Russell, Stefik, Pirolli, and Card 1993) D. M. Russell, M. J. Stefik, M.J., P. Pirolli and S. K. Card: “The cost structure of sensemaking”, *Proceedings of the SIGCHI conference on human factors in computing systems CHI 93,* 93, 269-276. Retrieved 31.03.2106 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.3353&rep=rep1&type=pdf.

(Russell-Rose and Tate 2013) T. Russell-Rose. and T. Tate: *Designing the Search Experience: The Information Architecture of Discovery* (Elsevier).

(Spencer 2006) Spencer, Donna (2006): “Four modes of seeking information and how to design for them’, *Boxes and Arrows*. Retrieved 31.03.2016 from http://boxesandarrows.com/four-modes-of-seeking-information-and-how-to-design-for-them/.