# Discovery - User scenarios and their metadata requirements (version 3)

*Harvest on a cooperative farm on the puszta, the Hungarian plain*
*Austrian National Library, Public Domain Marked*

*Editors:* **Valentine Charles, Antoine Isaac, Timothy Hill** (Europeana Foundation)
*Contributors*: Members of the **Data Quality Committee**

Europeana Foundation, December 2015

# Scope and Purpose of this Document

The purpose of this document is to capture existing information-access user needs as comprehensively as possible, and identify the metadata requirements for these. Although originally motivated by the necessity to better define user information requirements, the centrality of metadata quality to improving information retrieval means that this document also serves to guide and structure the work of the Data Quality Committee (DQC). As noted in the DQC Mission Statement[1], the focus is on 'stating … known problems we have to fix and identifying the gaps in current recommendations and guidelines' rather than describing innovative or ideal functionality: these scenarios should all in principle be actionable in the near-term, even if they are satisfiable only incrementally or in the medium- to long-term.

The primary audience for this document is members of the DQC; secondarily data providers, Europeana partners, and Europeana staff liaising with these; and thirdly other Europeana staff as appropriate.

Motivations for scenarios should ideally make reference to the user personas outlined in Europeana DSI D6.2 Requirements for europeana.eu, Appendix II[2]. That said, it is recognised that readers and editors of this document will have extensive knowledge of and experience with user needs, and that the personas provided are not exhaustive.

The focus of the scenarios is specifically on metadata. User-interface (UI) and user-experience (UX) issues are out of scope, as is discussion of specific software components and approaches to ameliorate the issue; for example, while automatic translation of queries might help in relation to the 'cross-language recall' scenario, it relates to the question of enhancing data quality only marginally.

[1] http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee//DataQualityCommittee_MissionStatement_032016.pdf
[2] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d6.2-requirementsforeuropeana.pdf

This document reuses the terminology defined in the Europeana Data Model (EDM)[3]. You may want to refer to the model documentation for further details on the classes and properties referred throughout the document.

The DQC has identified for each scenario a list of elements from the Europeana Data Model (EDM) which would support particular desirable but optional functionalities from a specific (set of) usage scenario(s). These "enabling elements" are defined as :

- Elements whose mandatory-ness is 'qualified' by a specific (set of) usage scenario(s);
- Elements that enable functions and improve services in Europeana or from third parties;
- Elements that are highly desirable as they increase user satisfaction.

Our goal is to demonstrate what these elements 'do' in specific situations to motivate providers to provide them. While it is expected that Europeana functions requiring enabling elements won't reach their full potential as well as expected if the enabling elements are missing from CHO metadata descriptions (e.g. cultural objects lacking these elements may not be shown or rank as well in the results on the Europeana portal), their absence won't make a dataset invalid.

## Scenarios

Note that all scenarios below are rated for Desirability and Effort Required on a 5-point scale. With regard to Desirability, '0' would indicate that a feature was not at all desirable, while '5' would indicate a 'must-have'. With regard to 'Effort Required', it is assumed that all data-quality tasks will require significant work, and are difficult to estimate: '0' thus broadly indicates a task that might foreseeably be accomplished within a month, while '5' would be a long-term task that could take a year or more to implement. It is also assumed that data-quality tasks are all ongoing efforts; the number given should thus be a time estimate to achieve the bare minimum useful to an end-user, with the intention that this minimum be improved over time.
Ideally we want all interested DQC members to provide an estimate of the desirability and/or effort required. If you do provide such an estimate, please provide your initials after your estimate to guide future discussion.

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

**Guide to Assessor Initials**

AI - Antoine Isaac
CD - Cécile Devarenne
CM - Cristina Muntean
DA - Dimitra Atsidis
GG - Gretchen Gueguen
HV - Henk Vanstappen
JM - Jef Malliet
TH - Tim Hill
MT - Melissa Terras
VC - Valentine Charles
WB - Werner Bailer

## 1. Basic Retrieval with high precision and recall

Desirability: 5 (TH), 5 (CM), 5 (CD), 5(AI), 5(JM), 5(MT)
Effort Required: 3 (TH), 4 (CM), 3 (CD), 4(AI), 3(JM), 4(MT)

**Scenario**
As a user, I want my search results list to contain all documents relevant to my search, with the most relevant results appearing at the top of the result-page list.

We have identified two sub-scenarios to cover all the aspects of basic retrieval, following the categories defined by Broder[4] and Behnert[5] regarding search queries in the library domain. We therefore distinguish topic (or 'informational') search from known-item (or 'navigational') search.

- Topic (or informational) search: queries executed in this context are more problem-oriented. A user will be browsing Europeana to find resources supporting his specific informational needs. His query will be supported by metadata elements providing topical and contextual information.
- Known item search (navigational): queries executed in this context are concrete. A user will search Europeana for a very specific resource or for "factual" information ( a specific title, an author...). His query will be supported by metadata elements providing accurate and distinct metadata.

In addition to this distinction between topic and known-item search, it is worth noting that queries for Europeana (as with the cultural heritage domain in general) tend focus on named entities, esp. person names and geographic entities. These entities make up almost 50% of the queries, followed closely by the large number of queries looking for broad subject categories such as "art", "theatre" or "landscape".

## 1a. Topic search

**Scenario**

As a user I want the result list to contain several documents relevant to my (topical) information need, with the most relevant results appearing at the top of the result-page list.

**Motivation**

This scenario is obviously fundamental to the Europeana platform: users need to be able to find what they are looking for, based on their informational needs.

[4] Broder, A. (2002). A taxonomy of web search. In ACM Sigir forum (Vol. 36, pp. 3–10). ACM.

[5] Behnert, C. (2016). Evaluation Methods within the LibRank Project (Working Paper). LibRank. Retrieved from http://www.librank.info/wp-content/uploads/2016/07/Working_paper_LibRank201...

**Metadata Analysis**

Simply put, the metadata quality for all records must be as high as possible. This includes such factors as:

- Metadata completeness - all relevant metadata elements should be supplied with values
- Metadata accuracy
    - Metadata values must be correct
    - These values must be appropriate to their metadata elements
- Metadata precision
    - Values should be of an appropriate length
    - Where applicable, they should be taken from a relevant controlled vocabulary
    - Where applicable, they should be (resolvable) URIs

**Proposed Actions**

Ongoing work in the DQC is intended to address the above points. The most immediately-pertinent work is Peter Kiraly's on metadata completeness - but there are few aspects of the DQC that do not touch on this.

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:creator, dc:contributor, dc:description, dc:language, dc:publisher, dcterms:conformsTo, dc:coverage, dc:date, dc:format, dc:source, dc:subject, dc:title, dc:type, dcterms:created, dcterms:issued, dcterms:medium, dcterms:provenance, dcterms:spatial, dcterms:temporal, edm:currentLocation, edm:isRelatedTo, edm:type. |
| ore:Aggregation | edm:data Provider, edm:language |

**Notes**

As the note under 'Proposed Actions' indicates, this is a fundamental - and therefore very broad and generic - scenario. It is intended more as a reference point to tie together existing actions and documents than as a source of distinct actions in its own right.

## 1b. Known-item search

**Scenario**
As a user I want the result list to contain the document I am looking for, with the most relevant results appearing at the top of the result-page list.

**Motivation**
This scenario is obviously fundamental to the platform: users need to be able to find the specific resource they are looking for.

**Metadata Analysis**
Simply put, the metadata quality for all records must be as high as possible, following the factors identified for the previous scenario "topic (or informational) search".

**Proposed Actions**
Ongoing work in the DQC is intended to address the above points. The most immediately-pertinent work is Peter Kiraly's on metadata completeness - but there are few aspects of the DQC that do not touch on this.

**Enabling metadata elements or metadata features for this scenario**
While high quality metadata is required for this scenario, we have defined a restricted list of the preferred elements. The enabling elements are:

| | |
|---|---|
| edm:ProvidedCHO | dc:title, dcterms:alternative, dcterms:hasPart, dcterms:isPartOf, edm:currentLocation, dc:contributor, dc:publisher, dc:creator, dc:identifier, owl:sameAs. |

**Notes**
As the note under 'Proposed Actions' indicates, this is a fundamental - and therefore very broad and generic - scenario. It is intended more as a reference point to tie together existing actions and documents than as a source of distinct actions in its own right.

## 2. Cross-language recall

Desirability: 5 (TH), 5 (AI), 5 (CM), 5 (GG), 5 (DA), 5 (CD), 5(JM), 5(MT)
Effort Required: 2 (TH), 3 (AI), 3 (CM), 3 (GG), 3 (DA), 3 (CD), 4(JM), 2(MT)

**Scenario**
As a user, I want to search Europeana collections in the language I am most comfortable in, and feel confident that I will receive relevant results irrespective of language of documents or the particular language label that was indexed.

**Transversal scenario 1: See Annex 1**

**Motivation**
Multilinguality is central to Europeana, in terms of both its collections and its user base. See the Best Practices for Multilingual Access in Digital Libraries White Paper[6], and in addition the various non-English users and collections described in the D6.2 Personas appendix.

**Metadata analysis**
For cross-language recall of EDM Concepts, see the analysis for Browse by subjects and types. For this scenario all contextual entities should be labelled and available in all supported languages.

[6] http://pro.europeana.eu/publication/best-practices-for-multilingual-access

In addition, this scenario requires that all searchable freetext elements be consistently tagged with their language. The language tag should be taken from a controlled list, for instance the ISO 639-1or 632-2[7] list, the IANA language tag registry[8], or the Languages Name Authority List (NAL)[9]

**Proposed actions**

With regard to Contextual entities (e.g Agent, Place, Concept and Timespan linked to objects), care needs to be taken to ensure that the content of these entities is available in all supported languages, either as an intrinsic part of the vocabulary (as with the AAT thesaurus), or through harvesting (Agent names, for example, could in many cases be harvested from the various-language versions of Wikipedia). Vocabularies like GEMET have too specific a scope for broad applicability in our domain.

**Enabling metadata elements or metadata features for this scenario**

In this given scenario metadata values matter more than the chosen elements themselves. The presence of language tags for every metadata values is crucial for this scenario.

We therefore recommend that:

- all EDM elements supporting literals SHOULD be provided with language tags.
- using EDM elements in combination with (i.e. which link to ) a contextual entity with multilingual features is RECOMMENDED.

**Further recommendations**

With regard to free-text elements[10], Europeana should establish best practices for publishing multilingual content as highlighted in the Best Practices for Multilingual Access in Digital Libraries White Paper.

One specific issue is language identification of metadata. To ensure proper language identification, there are two pathways:

[7] https://www.loc.gov/standards/iso639-2/php/code_list.php
[8] http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
[9] https://open-data.europa.eu/en/data/dataset/language
[10] EDM fields for completeness, multilingual saturation and other measures (EDM External with EDM Internal mapping)

- ensuring data providers identify the language of their metadata appropriately with language and script tags as applicable[11]
- application of machine-learning and Natural Language Processing (NLP) techniques for language detection and automatic tagging[12]

**Notes**
To take advantage of improved metadata, the ingestion process and Solr schema will need to be changed to apply language-specific text-analysis processing to our records, such as stemming and stop-word removal.

**Example Implementations**
The European Library (TEL) has developed software for normalizing dc:language that was applied to the library datasets contributing to (TEL). Normalisation software and results can be found on Github at https://github.com/europeana/metis-language-normalization

## 3. Improved language based facets

Desirability: 5 (TH), 5 (AI), 4 (CM), 4 (GG), 4 (DA), 4 (CD), 5(JM), 4(MT)
Effort Required: Assuming the work on improving metadata for Cross-language Recall has been completed, 2 (TH), 2 (AI), 2 (CM), 2 (GG), 2 (DA), 2 (CD), 3(JM), 3(MT)

**Scenario**
We have defined two scenarios for language facets, language of metadata descriptions and language of CHO and Work, depending on whether users wants to find descriptions or content in a specific language. See below.

[11] Even if practical it would be a very slow process as all data partners would have to update and resubmit their datasets to Europeana. Doing that for 50 million records would take a VERY long time.
[12] Preliminary tests on a small dataset indicate that Solr's built-in language detection capacities are 89% accurate across English, French, Spanish, Italian, German, and Polish. https://github.com/europeana/search/tree/master/util/language_detection

**Notes**
For either language facet scenario to work well:

- Languages should be fully and consistently normalised or follow a limited set of ISO-standards.
- It must be crystal-clear whether the language refers to the language of the metadata or the digital representation of the work (the language a book is written in, the language people talk in a video, the language Maria Callas sings in in the music recording, etc.).

The exact number of facets required to represent language adequately remains to be determined. A distinction between language of metadata and language of content is fundamental. Beyond this, additional facets indicating, e.g. subtitle language may be desirable.

**Example Implementations**
Examples are given in the Best Practices for Multilingual Access in Digital Libraries white paper.

## 3a. Improved facets based on language of metadata

**Scenario**
As a user, I want to filter my results to see only cultural heritage objects descriptions in languages I can competently read.

**Motivation**
This is captured in Marcel's need to 'read the metadata in his own language'.

**Metadata analysis**
From a metadata point of view, this scenario requires language tags identifying the language of the free-text values provided in EDM properties.

**Proposed actions**
For the Language of metadata descriptions, see Cross-Language Recall, above.

**Enabling metadata elements or metadata features for this scenario**

For enabling this scenario,

- all the EDM elements supporting literals SHOULD be provided with language tags.
- using EDM elements in combination with (i.e. which link to ) a contextual entity (with link to a multilingual features) is RECOMMENDED.

**Notes**
In addition and only applicable to Europeana, edm:language MAY be used as a fallback option when no multilingual literals are available in the metadata.

The existence of multilingual metadata records in our collection demands that language-tagging occur at the level of the metadata element rather than the document.

## 3b. Improved facets based on language of content

**Scenario and motivation**
As a user, I want to filter my results to see only cultural heritage objects and works in languages I can competently read.

**Metadata analysis**
From a metadata point of view, this scenario requires the dc:language property describing the language of a CHO.

**Proposed actions**
For the Language of metadata descriptions, see Cross-Language Recall, above.

**Enabling metadata elements or metadata features for this scenario**

For enabling this scenario the dc:language element, describing the language of a CHO, MUST be provided when a resource is of edm:type TEXT and SHOULD be provided for this other types (AUDIO, IMAGE, VIDEO, 3D).

We RECOMMEND the use of the ISO 639-2 code[13] for non linguistic content (ZXX). When dc:language is missing for non linguistic content, edm:type will be used to deduce the value of dc:language.

**Further recommendations**

For Language of CHO and Work, we have for now no choice but to rely upon the metadata of our partners: machine-learning and other automated techniques are helpless against the fact that title-strings may be extremely short, may not indicate language at all ('Don Quixote', 'Oliver Twist') or may have a language tag that fails to reflect the language of the work. Furthermore, the language of a title has very weak correlation to the language of the Work (as a functional entity as defined in FRBR[14]):

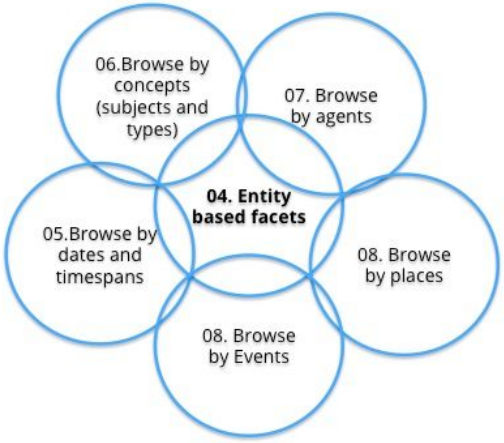
- Consider that many works have multiple titles in many languages
- Most visual works have no language at all (images are a "universal language")
- Some works have several languages (e.g. a video can have subtitles in many languages)

In the future, language identification based on content (text or audio) could be applied. The feasibility of this option increases as more textual content becomes directly accessible.

[13] https://www.loc.gov/standards/iso639-2/php/code_list.php
[14] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

# 4. Entity-based facets

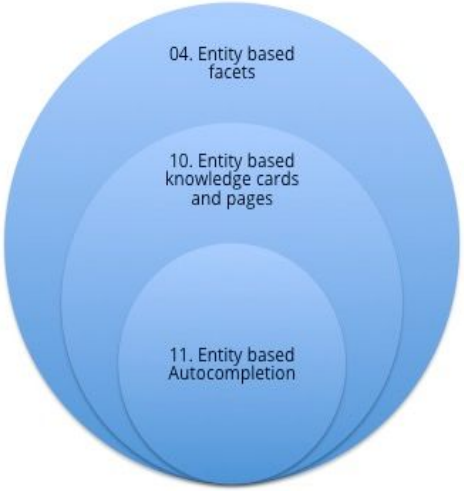
Desirability: 4 (TH), 4 (AI), 4 (CM), 4 (GG), 4 (WB), 4 (DA), 4 (CD), 5(JM), 4(MT)
Effort Required: 4 (TH), 4 (AI), 4 (CM), 4 (GG), 4 (WB), 4 (DA), 4 (CD),4(JM), 4(MT)




**Note on scenario scope**

This scenario stands at the intersection of other scenarios mentioned in this document. It focuses on the presence of entities linked to a given CHO. The main requirement is the presence of relationships between a CHO and one or more entities represented as a fully-fledged resource (Agents, Places, Concepts, etc.).
Scenarios 5, 6, 7, 8 and 9 require metadata information that would enable a browsing on specific entities. This required information might consist of literals (place names, a date, a creator name), or of links to specific contextual entities. Because of the latter there are intersections between the different scenarios

While scenario 04 requires the presence of entities (a URI to an entity), scenarios 10 and 11 require more data about these specific entities. Preferred labels are currently the core enabler of scenario 11 while the list of metadata enabling elements for scenario 10 will grow according to end-users requirements.



**Scenario**
As a user, I want to be able to filter results by the people, places, and subjects associated with CHOs. I may also want to search for a more general (broader) concept or place, and find objects related (indexed with) all more specific (narrower) concepts or places - i.e., hierarchical query expansion.

**Transversal scenarios 1 and 2:** **See Annex 1**

**Motivation**
From the personas: Marcel 'wants to filter results in clear, understandable ways'; Marion 'is not a searcher, she is a browser'; Paul 'wants to search content under a clear structure'; Marie 'wants to filter search results to find exactly what she is looking for', etc.

Other search examples based on semantic hierarchies relevant for this scenario:

- Searching for dc:type Painting should find objects indexed with the narrower concept Watercolor;
- Searching for dc:subject Mammal or Domestic Animal should find objects indexed with the narrower concepts Pig or Horse;
- Searching for "place of creation" Netherlands should find objects indexed with any place in the Netherlands;
- Searching for dct:spatial Netherlands **maybe** should find objects indexed with any place in the Netherlands.

Faceting examples:

- It is very useful to see search results faceted by eg "place of creation" in a hierarchy, starting from the continents (and oceans), then proceeding to countries, etc
- It is also useful to see objects by Nationality of creator. If we view Nationality as a group, this can be considered as a 2-level hierarchy of Agents.

This report from the Europeana Creative project provides motivations for this scenario. For instance 153 names for Lucas Cranach[15] were identified across 7 data sources, of which VIAF and Wikidata/DBpedia are the most important.
See Name Data Sources for Semantic Enrichment. Alexiev, V. Technical Report, Part of Deliverable D2.4, Europeana Creative project, February 2015[16].

**Metadata analysis**

In each case the underlying requirement is that the relevant EDM elements for objects MUST be populated with URIs from controlled vocabularies rather than free-text. These URIs need to be related, at a minimum, to a label for each of the supported languages. The target vocabularies should also be hierarchically structured.

Searching using semantic hierarchies depends on:

1. Having thesauri with appropriate hierarchies (e.g. Geonames has gn:parentFeature, TGN has gvp:broaderPartitive and skos:broader, AAT has gvp:broaderGeneric and skos:broader)
2. Objects using such thesauri, or semantic enrichment being able to link (coreference) object metadata to such thesauri

**Proposed actions**

Metadata support for this scenario can be supplied through Europeana's metadata-enrichment activities[17]. In addition Europeana will encourage its data providers to provide more links to entities.

**Notes**

Note that the four Contextual classes described below can be approached independently.

[15] http://vladimiralexiev.github.io/CH-names/README.html#sec-3-1
[16] http://vladimiralexiev.github.io/CH-names/README.html
[17] https://pro.europeana.eu/page/europeana-semantic-enrichment

Initial inspection of our query logs in 2015 seemed to indicate that, if importance is assessed in terms of query frequency, Concepts are the most problematic class and least valuable; Places the most valuable, but problematic; and Agents less problematic (even though spelling variants can represent a challenge), but less frequently queried for than Places. More recent query logs will have to be analysed to see if this prioritisation continues to hold true.

Note further that the requirements for subsumption searches assume hierarchically-structure controlled vocabularies where relevant (i.e., in the cases of geographical areas and concept taxonomies).

**Enabling metadata elements or metadata features for this scenario**

This scenario relies on the EDM elements that can accommodate an entity and contain a link to an entity. It excludes any elements used with literal values. Any linked entity SHOULD also have at least one Preferred Label with a language tag. The enabling elements are:

| | |
|---|---|
| edm:ProvidedCHO | dc:contributor, dc:creator, dc:format, dc:publisher, dc:subject, dc:type, dcterms:medium, dcterms:spatial, edm:currentLocation |
| edm:Place | skos:prefLabel |
| edm:Agent | skos:prefLabel |
| skos:Concept | skos:prefLabel |

**Example Implementations**

Entity-based facets:

- Europeana Food and Drink Semantic Demonstrator (Places and Food&Drink topics), http://efd.ontotext.com/app
- Prosopography of Anglo-Saxon England: http://www.pase.ac.uk/jsp/index.jsp
- German Digital Library: www.deutsche-digitale-bibliothek.de; sample search: http://bit.ly/1tsKVty

Facets using semantic hierarchies:

- The Europeana Food and Drink Semantic Demonstrator provides hierarchical facets for Places and Food&Drink topics: http://efd.ontotext.com/app/ but not search
- ResearchSpace semantic search (based on CIDOC CRM fundamental relations) had hierarchical search but not facets.

# 5. Browse by date or timespan based facets

Desirability: 4 (TH), 5 (AI), 3 (CM), 4 (HV), 5 (GG), 5 (WB), 3 (DA), 3 (CD), 4(JM), 4(MT)
Effort Required: 4 (TH), 3 (AI), 3 (CM), 3 (HV), 3 (GG), 4 (WB), 4 (DA), 4 (CD), 4(JM), 4(MT)

**Scenario**
As a user, I want to be able to filter my results by a variety of dates or timespans. For CHOs, this primarily consists of: the date of creation or production (as captured in the dcterms:created element); of publication (dcterms:issued); and dates and timespans that form the subject of a CHO (dc:subject).

Other dates of interest might include date of modification; of discovery; use; or change of custody.

**Transversal scenario 2:** **See Annex 1**

**Motivation**
From the personas: Marie wants to filter by date; Paul's historical focus also strongly implies a need for reliable date data.

**Metadata analysis**
For date facets that work well the following guidelines should be followed:

- **Normalisation:** Dates should be fully and consistently normalised to follow standards such as the XSD date-time data types. For instance, dates expressed in styles like "490 avant J.C" that are inherently language dependent should be avoided, as these are very difficult to normalise (e.g. this should be represented as "-0490"^^xsd:gYear)
- **Specialisation**: Use specialisations (e.g. dcterms:created) of dc:date whenever possible.

- Specialisations of dc:date element might be required to capture more specific types of dates (date of manufacture, date of performance, etc.) - as in the case of the EDM profile developed for the Europeana Sounds project: http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/EDMSound//TF_Report_EDM_Profile_Sound_301214.pdf. However these specialisations should be aligned with a finer-grained version of this user scenario.
- An implementation of Event could also improve the representation of dates but wouldn't solve the problem on its own. A specific date element would need to be aligned with specific types of events.
-

- **Alignment:** Many time elements contain periods or events information, which could be mapped to dates if we have the right background knowledge. ARIADNE has for example published its period collection in collaboration with PeriodO, which means permanent URIs are available like https://test.perio.do/#/p/Canonical/periodCollections/p0qhb66/. Clarifying the distinction between dcterms:temporal and dc:date might help data partners during the mapping exercise.

**Enabling metadata elements or metadata features for this scenario**

The presence of dates as numeric values is the enabler of this scenario. All the EDM elements representing dates MUST have numeric values. If these elements refer to a timespan, it MUST have a beginning and an end date. Dates related to an Agents must also have numeric values where present.

The enabling elements are:

| | |
|---|---|
| edm:ProvidedCHO | dcterms:created, dcterms:issued, dcterms:temporal, edm:hasMet (with a time span)). |
| edm:Timespan | edm:begin and edm:end |
| edm:Agent | edm:begin and edm:end |

NB: 'with a time span' indicates that the element can be used with an explicit reference to an instance of the edm:TimeSpan class or with a simple string (literal) that makes an implicit reference to a time span. A literal making an implicit reference could be used by Europeana or others to derive an explicit reference to an instance of edm:TimeSpan, as part of a metadata enrichment process.

**Further recommendations**

More user research will be required to evaluate what users are interested in when using a date facet. The most common user assumptions of dates is that it represents the date an object was first created in a generic sense - i.e., the classic "How old is it? When is it from?" questions. In the case of a book "when was it written or first published?". In additions, specialist users might be interested in more specialised dates (events) in the life of the object (e.g. where it was in a place, who was with it, what was happening...)

We could use heuristics and normalise existing date elements similarly to what Europeana Fashion or DPLA are doing:

- In Europeana Fashion, normalisation of date strings (even the very messy ones) is done with simple regular expressions (See Example implementations)
- This is the code for DPLA's date parsing:https://github.com/dpla/KriKri/blob/develop/lib/krikri/enrichments/parse_date.rb Obviously, this works with standard English terms like "circa" "after" "BC" etc, and is therefore inherently limited in the European context. It is included here out of interest, and as a possible basis for future development.

Further investigations will be also required to analyse how semantically-distinct and fine-grained the date elements in the current Europeana dataset are. A similar analysis has for instance been undertaken as part of the work on identifying the “20th Black Hole”[18]. More information of this analysis is available at https://github.com/hugomanguinhas/europeana_experiments/tree/master/blackhole

**Notes**

**Display of dates**

The same actions would also open up for a better display of Europeana items on interactive timelines. Such timelines could also benefit from the additional labelling detail provided by vocabularies such as Semium Time or Perio.do.

[18] http://pro.europeana.eu/blogpost/the-missing-decades-the-20th-century-black-hole-in-europeana

**Managing date uncertainties**
This scenario also requires better management of timespans for CHOs created before the year 1 of our modern calendar (BC dates).

Uncertainties in historic dates are modeled in CIDOC CRM with 4 date points (crm:P82a_begin_of_the_begin, crm:P81a_end_of_the_begin, crm:P81b_begin_of_the_end, crm:P82b_end_of_the_end), and in Wikidata are modeled with a "date precision" attribute (eg days vs centuries). The same approach could be adopted in EDM if needed.

For visualisation purposes capturing uncertainty is also useful. For example, a creation date "19th century" should be visualised differently than a period that is known to be exactly 1.1.1800-31.12.1899.

**Normalisation of dates**
The problem with normalization of dates is that it will often result in the loss of semantics in the metadata. This is why museums often use both a text element ("ca. 1955"), along with a normalized start and end date ("1950";"1959"). This differentiation could be adopted in EDM. Europeana Fashion is currently managing its dates this way (see Example Implementations).

**Support of periods**
Additional considerations include: periodisations ('Medieval', 'Renaissance', 'post-war'); overlap of schools and styles with periodisation ('Baroque', 'Gothic'); alternative calendars (Julian; Islamic; trans-March).

A Period (in the sense of crm:E4_Period) is much more than a time interval. It is a set of coherent cultural-social phenomena that can have multiple spatio-temporal extents, and researchers will never agree on the exact extents. Two researchers may agree that an object is from the Neolithic, while disagreeing on the exact extent of the Neolithic. Period can be captured in dcterms:temporal, and can be used to derive a date-span heuristically; but reducing the Period to a date-span is inappropriate.

**Representation of dates as part of an event**
It is also possible to envisage the representation of roles for timespans in this scenario. Timespans can be broken down into such categories as 'date of creation', 'date of manufacture', 'date of finding'.... In this case timespans would often refer to an event in the CHO history which should be defined as part of a list of event types.

The CHO and the event would be connected either by definition of the specific subproperties (e.g. createdAt, occuredAt) or by using an explicit link such as the Event class. For backward compatibility we would recommend to connect the CHO to a given Timespan entity using for instance dcterms:created and to an Event that refers to the same place using edm:occuredAt.

**Example Implementations**
Modemuze (https://www.modemuze.nl/collecties) has a filter on date ranges. For this filter, all dates are normalised so values match either the YYYY or the YYYY/YYYY format. This normalization is done in Mint, using a regular expression. The regex extract the first and (if present) last string of 4 digits from a text string. So e.g. "ca. 1960" becomes "1960", "1/2/1965" is returned as "1965" and a long string such as "between ca. 1800 and 1899 (?)" is returned as "1800/1899". This is ideal for searching and filtering purposes, but the exact dates and nuances ("ca.") are lost. To allow for a sufficiently normalised representation of the dates, a differentiation is made between dc:date (a text representation) and dcterms:created for the normalized value. dcterms:created is used for searching and filtering, but only dc:date is visible in the record detail view. A similar filter will be implemented on the Europeana Fashion portal.

DPLA does this kind of normalization for dates already. They then feed into a timeline view: http://dp.la/timeline. A date range filter is also available in the search results: http://dp.la/search?q=walt+whitman&utf8=%E2%9C%93 (4th facet)

# 6. Browse by subjects and types

Desirability: 5 (TH), 5 (AI), 4 (CM), 5 (GG), 5 (WB), 5 (DA), 5 (CD), 5(JM), 5(MT)
Effort Required: 3 (TH), 4 (AI), 4 (CM), 4 (GG), 4 (WB), 4 (DA), 4 (CD), 4(JM), 3(MT)

Concepts in the broad sense, which would include also e.g. genres of art and music and resource types.

**Scenario**
As a user I want to be able to browse a controlled index (or visualised browse entry points) of concepts from a controlled vocabulary or a list of subjects established by a curator (e.g. for a virtual exhibition) represented in the Europeana corpus.
As a user, I want to refine my results by concepts the CHO is *about* (as captured in the dc:subject element), and concepts the CHO *instantiates* (i.e., dc:type).

**Transversal scenarios 1 and 2: See Annex 1**

**Motivation**
This satisfies the general need expressed in the personas for 'findability'. See in particular Paul's need to 'search for content under a clear structure'.

**Metadata analysis**
This scenario requires consistent use of term-based subject and resource type classifications. In Europeana's case the terms also must have multilingual labels.

Developing this would be possible if:

All Europeana partners used the same SKOS-compliant terms for subjects and resource types and supplied URIs for it. The terms would need to have labels in all official languages of the EU and ideally also some regional European languages. BUT this is not the case and so we need to take specific actions to be able to support browse scenarios within an acceptable time frame.

It is important to note that it will be difficult to find vocabularies covering the complete range of concepts mentioned in Europeana's data. A mix approach including domain-specific vocabularies such as AAT and more general datasets such as Wikidata might help.

**Proposed actions**
In order to begin supporting this scenario we need to develop the Europeana Entity Collection[19] of aligned terms drawn from DBpedia, Wikidata and specialised vocabularies like Getty AAT, and then use them for semantic enrichment of dc:subject and dc:type.

[19] https://docs.google.com/document/d/16Lcuddgw7fNV0EQ7gnIJW5-C76z4HUIvRA8aunY13U4/

The achievability of this scenario would be improved if dc:subject and/or dc:type elements could be made mandatory. Without values in these elements Europeana has nothing to base its semantic enrichment on.[20] The DQC will discuss this recommendation and in the process might need to refine the definitions of the dc:subject and dc:type elements.

We also encourage the provision of terms based on selected vocabularies that are aligned with the Europeana Entity Collection.

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:subject (with a concept), dc:format, dc:type, dcterms:medium |
| skos:Concept | skos:prefLabel, skos:broader, skos:narrower, skos:exactMatch, skos:closeMatch, skos:related |

NB: 'with a concept'' indicates that the element can be used with an explicit reference to an instance of the skos:Concept class or with a simple string (literal) that makes an implicit reference to a concept. A literal making an implicit reference could be used by Europeana or others to derive an explicit reference to an instance of skos:Concept, as part of a metadata enrichment process.

**Example Implementations**

- Wellcome Library prototype explorer by Good, Form & Spectacle: http://scale.whatsinthelibrary.com/?_ga=1.236184904.890182881.1447844692
- British Museum Twoway.st by Good, Form & Spectacle: http://twoway.st/
- Walters Art Museum: http://art.thewalters.org/
- Europeana 1914-1918: http://europeana1914-1918.eu/en/explore
- LoCloud Vocabulary Service: http://support.locloud.eu/LoCloud%20Vocabulary%20Microservice and http://vocabulary.locloud.eu/

[20] Look at e.g. this otherwise excellent music manuscript object. As it has no dc:type or dc:subject value classifying that it is indeed a music manuscript Europeana semantic enrichment can't connect it to a vocabulary term.

# 7. Browse by Agents

Desirability: 4 (TH), 4 (AI), 3 (CM), 3 (GG), 4 (WB), 3 (DA), 4 (CD), 4(JM), 5(MT)
Effort Required: 3 (TH) 4 (AI), 3 (CM), 4 (GG), 3 (WB), 3 (DA), 4 (CD), 4(JM), 4(MT)

**Scenario**
As a user I want to be able to browse an index (or visualised browse entry points) of people and organisations represented in the Europeana corpus to access Europeana objects, filtered by the role of that person in relation to their associated CHOs. For instance, I want to be able to filter by whether a person is the subject (“aboutness”) of a CHO, or its creator (author, engraver, printer etc.).

We consider time-based filtering used on Agent lists a case of data and timespan based facets (which has edm:Agent/begin and edm:Agent/end as enabling elements).

**Transversal scenario 2:** **See Annex 1**

**Metadata analysis**
This scenario requires consistent use of term-based authorities.

This scenario will benefit from the presence of elements such as dc:subject, dc:creator, dcterms:contributor or other simple properties.

Further work on the representation of role information may be also required. It is for instance possible to envisage the representation of roles as part of an Event. e.g. for a portrait the production event typically has two agents participating, one in the role of “painter”, one in the role of “sitter”; in an acquisition event, by contrast, you have a 'seller' and a 'buyer'.

**Proposed actions**

In order to begin supporting this scenario we will develop the Europeana Entity Collection of aligned terms drawn from DBpedia, Wikidata and specialised authorities like VIAF and use them for semantic enrichment of the dc:creator and dc:contributor elements.

The achievability of this scenario would be improved if

- either dc:creator or dc:contributor would be made gradually mandatory (to give Europeana's semantic enrichment something to work on to connect the value to an authority term). Note that this action would need to be investigated considering domain specificities (e.g. lot of museums objects have for instance no creator).
- more recommendations were provided on how to code this element for a CHO when:
    - No information is provided on the creator of CHO. Since it is not possible to connect dc:creator to a person in an authority list, one suggestion would be to omit the element or leave it empty.
    - Information is known about the creator but not enough to properly identify the person. In this case the person can be added as 'unidentified' to an authority list as recommended by Getty ULAN[21]. Note that these persons often have been given a name in ULAN and are available as resource for semantic linking. If it is known that the creator is "Anonymous", then this value should be added in the metadata.
- the provision of terms based on select agent authorities that are aligned with the Europeana Entity Collection is encouraged.

The normalisation of dates, mostly date of birth and date of death (as mentioned in scenario 3.1) could also improve the quality of Agent information.

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:creator, dcterms:contributor, dc:publisher, dc:subject (with an agent), edm:hasMet (with an agent) |
| edm:Agent | skos:prefLabel, skos:altLabel, rdaGr2:professionOrOccupation |

[21] http://www.getty.edu/research/tools/vocabularies/ulan/about.html

NB1: 'with an agent' indicates that the element can be used with an explicit reference to an instance of the edm:Agent class or with a simple string (literal) that makes an implicit reference to an agent. A literal making an implicit reference could be used by Europeana or others to derive an explicit reference to an instance of edm:Agent, as part of a metadata enrichment process.

NB2: skos:altLabel is defined here as an enabling element to capture information about pseudonyms and name changes but not for translations. Language variants should be captured in the skos:prefLabel field, and tagged with the appropriate language tag.

**Notes**
To support Agents as *creators* of resources, we need to parse the dc:creator and dc:contributor elements - presumably populated exclusively by Agents. Supporting Agents as the subjects of records is a little trickier, as we'll need to determine whether the entity indicated in the dc:subject element is an Agent or not.

For sources of person names, see Name Data Sources for Semantic Enrichment. Alexiev, V. Technical Report, Part of Deliverable D2.4, Europeana Creative project, February 2015. Eg there are 153 names for Lucas Cranach across 7 data sources, of which VIAF and Wikidata/DBpedia are the most important.

**Example Implementations**
- BBC Your Paintings (now on Art UK): https://artuk.org/discover/artists
- RKD Explore RKD Artists: https://rkd.nl/nl/explore/artists

# 8. Browse by Places

Desirability: 2 (TH) 2 (AI) 2 (CM), 1 (GG), 3 (WB), 2 (DA), 2 (CD), 4(JM), 4(MT)
Effort Required: 5 (TH) 4(AI) 5 (CM), 4 (GG), 4 (WB), 4 (DA), 4 (CD), 4(JM), 3(MT)

**Scenario**
As a user, I want to search and explore topics geographically - being able both to retrieve geographical information about my result set, and to use geographical information in searching the collections.

**Transversal scenario 2:** **See Annex 1**

**Motivation**
This need is not well-captured in the user personas - but searches by geographical name (towns, cities, provinces and administrative areas, and countries) are the single most common category in our query logs.

**Metadata Analysis**
At a minimum, all metadata elements with Place values need to be populated with URIs rather than strings, and these URIs associated with long/lat values.

Beyond this, however, we confront significant semantic problems in relation both to our existing metadata, and to user expectations.

With regard to our existing data, the only geographical element we have with an explicit and clear semantics is edm:currentLocation. Beyond this, the chief distinctively spatial elements are dc:coverage and dcterms:spatial, the latter of which is simply a specialisation of the former. These elements are only usable from a browse perspective if they are populated in a sufficiently consistent way that their semantics are clear, and an assessment needs to be made of this.

In addition, as noticed by DPLA in their previous work, developing map browsing will require the normalisation of metadata, which is itself a significant challenge. Data partners might record hierarchically-related place terms in individual elements, resulting in different sets of coordinates for each level in the hierarchy (e.g. "United States" "Pennsylvania" "Pittsburgh"). Coordinates for larger regions are furthermore frequently set by default to their geographic center, often nowhere near the coordinates of more specific elements of the hierarchy.

In sum, map browsing requires clean and sophisticated data that might be difficult to create.

With regard to users, it is not at all clear what they are hoping to find when they submit a geographic query, and it is likely that their needs are diverse; archaeologists, for example, are probably trying to search by the geographical provenance of archaeological

artifacts, while tourists by contrast hope to find cultural institutions located in the area. It may accordingly be necessary to create qualifiers or specialisations for the various geographic elements.

Notwithstanding the above, most users would probably benefit considerably from linking data-provider institutions to geographic locations. User research needs to be conducted into what exactly people are looking for when they perform geographical searches. Right now their search intention (in the sense of what records they are expecting to see returned) is totally opaque.

**Proposed Actions**
An assessment of the completeness of Place data (in particular long/lat information) needs to be undertaken.

An evaluation of the semantics of dc:coverage and dcterms:spatial in our collections, and the consistency of their application, is also required. In the event that the semantics are unclear, we need to develop heuristics for clarifying them.

Entityfication of data-provider institutions should be undertaken, and the resulting entities further enriched with geographic data.

Ultimately some kind of validation needs to be put in place for this kind of data. For instance, we have many many many records from the Bodleian listing their location simultaneously as Lond (a village in Pakistan) and as being within the UK.

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:subject (with a place), dcterms:spatial, edm:currentLocation, edm:hasMet (with a place) |
| edm:Place | wgs84_pos:lat, wgs84_pos:long, dcterms:isPartOf,skos:prefLabel, skos:altLabel |

NB1: 'with a place' indicates that the element can be used with an explicit reference to an instance of the edm:Place class or with a simple string (literal) that makes an implicit reference to a place. A literal making an implicit reference could be used by Europeana or others to derive an explicit reference to an instance of edm:Place, as part of a metadata enrichment process.

NB2: skos:altLabel is defined here as an enabling element to capture information about local or historical variations of a place name but not for translations. Languages tags should be used to capture language variants.

**Notes**

It is also possible to envisage the representation of roles for places in this scenario. Place can be broken down into such categories as 'original location', 'place of manufacture', 'place of origin'.... Note, however, that the EDM does not cater for all of these, and might need to be modified to support them. In this case places would often refer to an event in the CHO history which should be defined as part of a list of event types.
The CHO and the event would be connected either by definition of the specific subproperties (e.g. currentLocation, happenedAt or by using an explicit link such as the Event class. For backward compatibility we would recommend to connect the CHO to a given Place entity using for instance dcterms:spatial and to an Event that refers to the same place using edm:happenedAt.

**Example Implementations**

- Prosopography of Medieval Scotland map interface: http://db.poms.ac.uk/map/browse/
- Peripleo (Pelagios): http://pelagios-project.blogspot.co.uk/2015/07/peripleo-sneak-preview.html
- The Google Field Trip app might be more relevant
- DPLA Map view http://dp.la/map
- eCultureMap: http://eculturemap.eculturelab.eu/eCulture14m/map.html
- Ajapaik.ee

# 9. Browse by Event

*Disclaimer: this scenario is still being discussed within the DQC and will be refined at a later stage.*
Desirability: 3(MT)
Effort Required: 5(MT)

**Scenario**
As a user do I want to:

- get a meaningful representation of the information available about a cultural heritage object- information which is understandable, and suitable for research.
- be able to browse an index (or visualised browse entry points) of events represented in the Europeana corpus.
- be able to filter results by the type of event associated (and not only historical events but also events relevant to the history of a CHO such as survey and fieldwork events)

As a user, I want to refine my results by

- CHO is about (dc:subject)
  (e.g. return all CHOs about the Storming of the Bastille)
- CHO was present at an event which is identified by edm:Event
  (e.g. return all CHOs present at the Storming of the Bastille)

**Motivation**
This scenario addresses the search behaviour of the following personas: Marcel 'wants to filter search results in clear understandable ways', Marie 'wants to filter search results to find exactly what she is looking for', Amy 'wants to find stories in WW1, especially those related to her family and places in her memory', may also support Marion in browsing by place and time.

**Metadata analysis**
The underlying requirement for this scenario is that the relevant EDM Event class is implemented.

In order to ingest data, either an equivalent to event needs to exist in the source data, or appropriate qualified times and locations need to exist, that can be used to unambiguously create an event.
While it may in many cases not be possible to create events from place and date properties only, the inverse is possible, i.e. data providers using Event may omit other place/time properties, as they could be reconstructed if needed for backward compatibility.

This scenario will be also improved if an Event type authority list is created and consistently used.

**Proposed actions**
This scenario would require in the first place an evaluation of the Event class and the properties defined for this class in the EDM definitions. Mapping exercises with a sample of providers would enable us to check whether the basic requirements for this scenario are met by the current proposal. For instance, current metadata from related types of date or place could be analysed and modelled according to the Event structure.

A list of relevant event types and associate types for dates and places will need to be collected and ideally represented as a linked data vocabulary. In addition, this scenario would be improved if URIs to controlled vocabulary are used for critical events such as battles, war, etc.

**Example Implementations**
- Fasti Online has implemented a database of excavations since 2000 in the Classical world: http://www.fastionline.org/excavation/, databases of archaeological conservation and archaeological survey are in development: http://www.fastionline.org/
- Historic England has maintained an excavation index since 1978: http://archaeologydataservice.ac.uk/archives/view/304/
- The EEXCESS project has started to support edm:Event in the mappings for some data providers, see eg https://github.com/EEXCESS/recommender/blob/knowDev/modules/partners/kimportal/src/main/resources/mapperObject.xsl
- German Digital Library has the “nucleus”, see examples from the Basecamp thread:
  https://www.deutsche-digitale-bibliothek.de/item/SQZTNMRYPP5647VYR7TULLB3HX34FL7Q
  https://www.deutsche-digitale-bibliothek.de/item/URQDUNLEOWNQ4BDF5APNECAEKLEPE2ZQ?lang=en

https://www.deutsche-digitale-bibliothek.de/item/4E2RRHLM2735ADDHVLN4ACWMIFSMXXI6?lang=en
https://www.deutsche-digitale-bibliothek.de/item/ZKUBVGY42Y5V7WPEGSGXJFEWOVFJO4RY?lang=en
https://www.deutsche-digitale-bibliothek.de/item/HGJ34PCPQXT3JHK2B6MBZQLX7C546ANC
https://www.deutsche-digitale-bibliothek.de/item/xml/HGJ34PCPQXT3JHK2B6MBZQLX7C546ANC

- Example why the grouping of information is necessary and specialisations /sub-properties for agent, date, place can't do the job: https://www.deutsche-digitale-bibliothek.de/item/SEIQHBPT6NGFSHKWMOMR73AKXQEKUZL2

**Notes**
The scenario overlaps partly with the various Entity Browse scenarios described below: see Browse by Agents, Browse by date or timespan based facets, Browse by Places, and Browse by subjects and types.

# 10. Entity-based knowledge cards and pages

Desirability: 4 (TH) 4 (AI) 4 (CM), 4 (GG), 3 (DA), 4 (CD), 4(JM), 2(MT)
Effort Required: Assuming work for Browse by subjects and types is complete, 1 (TH), 1 (AI), 1 (CM), 1 (GG), 1 (DA), 1 (CD), 3(JM), 1(MT)

**Scenario**
As a user, I want to see important information about persons, places, and subjects summarised in a single card or page.

**Motivation**
Entityfication-related stories are all grounded in the needs for (a) findability and (b) inspiration repeatedly expressed in the user personas.

For the relationship between 'inspiration' and 'entitification', see further the Search Strategy[22].

**Metadata analysis**

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

For knowledge cards representing EDM Concepts, see the analysis for Browse by subjects and types, below. The logic can be extended also to e.g. places and periods.

**Proposed actions**
Same as for Browse by subjects and types and Browse by Agents.

**Enabling metadata elements or metadata features for this scenario**
The enabling elements for this scenario are all the elements present in EDM[23] to describe contextual resources. The more information about a contextual resource is provided the richest the information for the user will be. New elements could be included in EDM if motivated by user requirements for a given scenario.

**Notes**
The generation of knowledge cards might be difficult in few cases (lack of metadata but also presence of approximations, ambiguities, homonyms). From a visualisation perspective, it might be difficult to capture all the information related to an entity in a clear and succinct card. For instance, some entities like Agents might have lot of related CHOs in Europeana, while some entities like places (a huge site like Pompeii or a large battlefield like Waterloo) might be too complex. Concepts might be also difficult to "visualise" unless they are based on some illustrated thesaurus.
*In v1 of this document there were separate headings for Entity Cards and Entity Pages. The two concepts, however, are identical in their metadata requirements and differ only in terms of UI/UX factors. Accordingly they have been merged in v2.*

**Example Implementations**
- Google
- Deutsche Digitale Bibliothek: https://www.deutsche-digitale-bibliothek.de/ search e.g. for Leibniz (https://www.deutsche-digitale-bibliothek.de/entity/118571249 )
- Person pages from data.bnf.fr http://data.bnf.fr/11907966/victor_hugo/
- Datos BNE ES: http://datos.bne.es/
- RKD provides a list of their entity types at https://rkd.nl/nl/explore.

[23] https://github.com/europeana/corelib/wiki/EDMObjectTemplatesProviders

**Examples for Europeana**

- Europeana mock-ups
  - http://demo.deanbirkett.name/400PDV/#p=home
  - Wireframe/mockup of a Bach knowledge card in a Europeana search result (list style): https://invis.io/YQ2G1HE4V
  - Irish Traditional Music Archive card in a Europeana search result (list style): https://invis.io/323301P5S

# 11. Entity-based autocompletion

Desirability: 4 (CM), 3 (GG), 4 (TH), 3 (CD), 4(JM), 5 (AI), 4(MT)
Effort Required: 4 (CM), 4 (GG), 1 (TH), 4 (CD), 3(JM), 4 (AI), 3(MT)

**Scenario**

As a user, I want the search interface to suggest entities to me as I type, ranked in order of relevance and accounting as much as possible for spelling errors, variations or synonyms. It should also show brief info about the entity (see scenario 9 Entity cards), to let me select the exact entity that I am interested in.

**Motivation**

For the motivation of entityfication scenarios, see above under Scenarios 4 (Entity-Based Facets) and 10 (Entity-based Knowledge Cards and Pages).

**Metadata Analysis**

Given that the autocomplete will necessarily operate over whatever contents of the Entity Collection there are, metadata requirements in an absolute sense are low: the Entity Collection must be populated by entities consisting of at least a canonical URI, associated labels (preferred and alternative), and some kind of relevance-boost factor; and the document store must have been enriched with corresponding canonical URIs.

The relevance factor right now is calculated as Wikipedia PageRank * Europeana term incidence, with some smoothing applied across the collection.

**Proposed Actions and Enabling metadata elements or metadata features for this scenario**
The key enabler of this scenario is the presence of an entity linked to a CHO and the presence of at least one preferred label.
The current implementation of the Entity-based Autocomplete[24] only requires the presence an entity and its preferred label. Preferred label SHOULD be multilingual.
The current implementation doesn't take into account alternative labels. This implementation could change in the future if alternative labels prove to be useful. While they can be interesting to capture variations of a given name for Agents and Places, altLabels for Concepts (corresponding to non-prefered terms in thesauri) could bring significant noise in the search results.

**Example Implementations**
- Google Knowledge Graph Search Widget: https://developers.google.com/knowledge-graph/how-tos/search-widget
- Multimedian eCulture demo (was used for the Europeana Semantic Search Thought Lab[25])
- Wikipedia has excellent auto-complete when searching for pages

# 12. Categorised similar items

Desirability: 4 (TH), 4 (AI), 4 (CM), 3 (GG), 4 (WB), 3 (DA), 3 (CD), 2(JM), 4(MT)
Effort Required: 3 (TH), 5 for categorized links and 2 for uncategorized (AI), 3 (CM), 3 (GG), 5 (WB), 3 (DA), 3 (CD), 4(JM), 3(MT)

**Scenario**
As a user, I want to navigate from the item I'm viewing to other, similar items, in a manner that is clear and meaningful to me.

[24] As of July 2017
[25] https://pro.europeana.eu/data/searchengineeuropeana . See also *Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator*, Schreiber et al., Web Semantics: Science, Services and Agents on the World Wide Web, 6(4), 2008. https://doi.org/10.1016/j.websem.2008.08.001

**Motivation**

This scenario answers to the personas' needs for inspiration and serendipity: James's desire to 'find inspirational things easily' and Marie's to 'see what jumps out at her'.

It is mostly about work-based relationships (between derived works). The relationships links supporting the scenario are complementary from the ones derived from other contextual entities (same place, same author…)

**Metadata Analysis**

The metadata requirements for this will depend on the categorisations used for similarity. Such categorisation was for example established for the innovation pilot between OCLC and Europeana working on hierarchical clustering. The clusters automatically obtained based on metadata similarities were of very diverse nature and required a manual categorisation of the type of relationships and similarities found in the different clusters[26].

Broadly speaking, however, we can probably expect them to be met by work undertaken to support other scenarios..

In the simplest case, if the similarity categories correspond to our standard metadata elements ('Of the same type', 'On the same subject', 'By the same creator'), quality improvements in these elements will obviously facilitate our 'More Like This' functionality.

Other approaches might rely on term-based comparison of different metadata elements ('with a similar description', 'also with keyword x'), geographic proximity ('within 10 miles of ….'), or some other similarity metric ('using the same colours …'). In each case, however, we'd be relying on data or metadata provided as a result of some other effort.

Ranking and diversity will also be crucial: how to rank items that match with different subsets of the elements, and how to ensure that more similar items come first, without presenting e.g. all the different copies of the same book that only have some different metadata elements

**Proposed Actions**

[26] http://pro.europeana.eu/blogpost/hierarchical-clustering-making-sense-of-europeana-data

Research needs to be undertaken into what notions of similarity can be supported using Solr's built-in similarity functionality, and into what points of similarity end-users find useful.
This scenario may require the modelling of any types of relationships between CHOs such as hierarchical relations (e.g., the relationship of an album to its tracks).

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:source, dc:relation, dcterms:isPartOf, dcterms:hasPart, dcterms:isReferencedBy, dcterms:isVersionOf, edm:incorporates, edm:isNetxtInSequence, edm:isSimilarTo, edm:isSuccessorOf, edm:isDerivativeOf, edm:isRepresentationOf, edm:hasVersion, dcterms:hasFormat, dcterms:isFormatOf. |

**Notes**
The most readily supported similarity categories (those corresponding to standard metadata elements) might also be the least useful, particularly if they're already present as facets.

There is considerable overlap between this and Scenario 13. The differences are:

- This Scenario concerns individual items; Scenario 13 is about search result lists.
- For Scenario 13, all results will be related to the original search term. For this Scenario, they need not be related (similarity might be calculated entirely on the basis of other terms)
- At a technical level, implementation will be very different (in Solr, for example, Scenario 10 is probably a job for the MoreLikeThis handler, and Scenario 13 would use carrot2)
- This scenario would exploit quality data about relations between objects.

**Example Implementations**

- Europeana mock-up proposing an implementation of similar items: https://invis.io/GJ632VI7M
- Data.bnf.fr (http://data.bnf.fr/ ) creates pages for Works to connect with the galaxy of documents related to the same work. Note that such linking implies that the Work level exists in the metadata either by creating it while cataloguing or by automatically extracting it from legacy data.

# 13. Diversification of results

Desirability: 5 (TH), 5 (AI), 5 (CD), 4(JM), 1(MT)
Effort Required: 4 (TH), 4 (AI), 4 (CD), 4(JM), 5(MT)

**Scenario**
As a user, I want the first page of my search results to represent as wide a range as possible of both (a) the various meanings of my search term(s) and (b) the kinds of CHO related to them. For example, a search for 'Freud' should include results from both Sigmund and Lucian, and a search for 'Sigmund Freud' should include images, videos, and texts early in the result set.

**Motivation**
This scenario is relevant to:

- polysemous or ambiguous queries, where the number of relevant results is high, but the user needs to run through a large number of results to grasp the multiple query meanings. For example the query "venus" has several meanings: planet, goddess, could be related to pictures, sculptures or different historical periods or artistic movements (difficult to be illustrated in a single result page or to be explored if the user is not guided).
- when the result list contains redundant entries. For example for the query "india" on the first page we have 8 sound items belonging to the same album. They should be either:
    - clustered in a single entry
    - near-duplicates should be pushed away from the top of the list
    - near-duplicates should be "hidden" behind a similar results link (expandable).

**Metadata Analysis**
The metadata requirements for this scenario are largely identical to those for Scenario 1 (Basic Retrieval with High Precision and Recall).

In addition, precisely-defined relationships (and specified using relevant properties) between objects (e.g. parent-child, original-translation, work-expression) is extremely useful for clustering.

**Proposed Actions**

- Define list of the most significant metadata elements to diversify results on**.**
- Create a dissimilarity measure between relevant results (list-wise view of metadata)
- Cluster similar items
- Near-duplicate detection

**Enabling metadata elements or metadata features for this scenario**

| | |
|---|---|
| edm:ProvidedCHO | dc:type, edm:type, dc:format, dcterms:medium, dc:date, dcterms:created, dcterms:issued, dc:language, edm:currentLocation, dcterms:spatial, dcterms:temporal |
| ore:Aggregation | edm:provider, edm:dataProvider, edm:language |

**Notes**

This clustering scenario (13) is the other side of the similarity scenario (12). Both cases are based on bags of terms which lead to the notion of similarity. The Europeana "More Like this" functionality based on Solr similarity is different in that the terms extracted from the document are used launch a new query while for clustering each bag of terms returned for a query are analysed and grouped with similar bags.

Solr's result-clustering functionality is provided by the Clustering Plugin[27], and this represents one possible path forward. Another option would be simply to diversify based on the values within identified metadata elements: for instance, differentiation based on media-type can be done based on the values in the TYPE facet element. edm:isRelatedTo, edm:isSimilarTo, dcterms:isPartOf, dcterms:hasPart, and similar elements, may also provide a basis for such clustering.

[27] https://cwiki.apache.org/confluence/display/solr/Result+Clustering

# Annex 1: Transversal scenarios

'Transversal scenarios' are cross-cutting concerns that, while not standing as scenarios in their own right, affect several of the scenarios listed above.

**Transversal scenario 1 : Synonymisation**

Named Entities typically don't have a 'language' as such - they may need to be transliterated or transposed into another language, but not 'translated' as such. Scenarios primarily concerned with named entities accordingly need not so much language tagging or translation, as a synonyms file/authority list. Synonymisation may also be relevant with other controlled vocabularies (for e.g. subject, roles) and for free-text elements.

**Transversal scenario 2: Event**

Events are a complex topic - in part because they may play a number of different roles.

First, events may describe well-known historical events ('World War I', 'the storming of the Bastille'). In this sense they are useful as subjects, categories, or topics in their own right.

Secondly, events may act as a kind of 'ontological glue' used in event-based modelling. In this case, their chief role is to link together other kinds of entities. For instance, a painting might have a Creation event (linking together a painter and a canvas and a sitter and the same canvas), a Transaction event (linking the canvas to a buyer and a seller), an Exhibition event (linking the canvas to an institution), and so forth.

In the latter case, how we handle and model Events will potentially affect all scenarios involving entities, and the Data Quality Committee has formed a special task force to look at the question of Event-centric modelling.

Scenario 9 Browse by Event describes the possibility for a user to browse by Event. The underlying requirement for this scenario is that an entity of type Event exists.

In addition, a user might want to search for a particular CHO in association with people, places, concepts or time periods themselves part of an Event.

As a user, I want to be able to filter by whether a person is the subject ('aboutness') of a CHO, or took part in an event of the object's history.
I also want to search for a place and/or date that relate to a certain event in the history of the object(e.g. created on the island Crete or. found in Crete or held in a collection there).
And finally I want to search for pairs of place and date (e.g. find objects related to a particular event in 1789 in Paris, as distinct from searches for objects related to the date 1789 and/or Paris.)
As a user, I therefore want to refine my results by

- CHO is about (dc:subject)
- [CHO was present at an event which had participant edm:Agent]
- [CHO was present at an event which occurred at edm:Place]
    - Creation/Production (dc:creator)
    - Publication (dc:publisher)
    - Modification
    - Finding
    - Use
    - Custody (edm:currentLocation)

This transversal scenario raises requirements that are not directly impacting scenario 9 Browse by Event but that are required at the level of the entity specific scenarios: Browse by date or timespan based facets, Browse by subjects and types; Browse by Agents; Browse by Places.

## Document history

| Version | Editor | Date | Comments |
|---|---|---|---|
| v1.0 | Timothy Hill, David Haskiya (Europeana Foundation) | 03/12/2015 | First version of the scenarios, commented by the DQC |
| v2.0 | Timothy Hill, Valentine Charles and Antoine Isaac (EF) | 01/07/2016 | Definition of a stable list of scenarios after input from the Data Quality Committee |
| v3.0 | Valentine Charles (EF) | 11/07/2017 | Addition of the enabling elements defined by the Data Quality Committee for each scenario.<br>Refinement of scenarios based on the discussion on enabling elements. Main changes are:<br>- The addition of the definition for enabling elements in the introduction;<br>- the split of scenario 1 Basic Retrieval with high precision and recall in two new scenarios<br>Topic search (or informational) and 1b. Known-item search (navigational);<br>- the split of scenario 3 Improved language based facets in two new scenarios Improved facets based on metadata language and Improved content language based facets. |
| v3.1 | Antoine Isaac (EF) | 10/04/2018 | Updates to enabling elements - making reference to specific types of contextual entities in the scenarios for browsing by dates, subject, agents and places |

**Co-financed by the European Union**
Connecting Europe Facility

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