Report on Enrichment and Evaluation

29/10/2015

Co-financed by the European Union
Connecting Europe Facility

Contributors

- Aitor Soroa (Univ. Basque Country)
- Cécile Devarenne (Europeana Foundation Data Partners services)
- Cesare Concordia (ISTI)
- Daniel Vila Suero (UPM)
- David Haskiya (Europeana Foundation Product Development)
- Dimitris Gavrilis (ATHENA research center)
- Eetu Mäkelä (Aalto)
- Maarten Brinkerink (Netherlands Institute for Sound and Vision)
- Marie-Claire Dangerfield (Europeana Foundation Data Partners services)
- Nikos Simou (National Technical University of Athens)
- Nuno Freire (The European Library)
- Rainer Simon (AIT Austrian Institute of Technology)
- Rodolphe Bailly (Cité de la Musique)
- Roxanne Wyns (LIBIS)
- Ruben Verborgh (ULB)
- Seth van Hooland (ULB)
- Tatjana Gornostaja (Tilde Terminology tilde.com/term)
- Vivien Petras (Humboldt-Universität zu Berlin)
- Vladimir Alexiev (Ontotext)
- Yorgos Mamakis (EF Dev team)

Editors

- Antoine Isaac (Europeana Foundation R&D)
- Hugo Manguinhas (Europeana Foundation R&D)
- Juliane Stiller (Max Planck Institute for the History of Science)
- Valentine Charles (Europeana Foundation R&D)

Executive Summary

This report on Evaluation and Enrichment provides an overview of the different processes in semantic enrichment and offers guidance on how to assess each of these steps to implement a coherent enrichment strategy.
The report begins by introducing the terminology used in the report. While defining the notion of semantic enrichment, the Task Force has identified several other associated notions that are commonly used in the cultural heritage domain when addressing semantic enrichment.
We also provide an overview of the enrichment tools and services developed in the Europeana Network over the past years, reflecting the diversity of processes at hand: tools for manual enrichment and annotation, tools for automatic enrichment and workflow design tools. We also focus on the interoperability issues such as rules for specifying the linking or the format used to describe the enrichment outputs.
As well as looking at the details of the enrichment processes we pick up the work done by the previous Task Force by specifying criteria for selecting and assessing target datasets. These criteria are based on vocabularies and datasets examples relevant to the Cultural Heritage domain. This selection strategy is available in a companion document to this report.
The last component of the enrichment strategy is the evaluation of the enrichment processes. So far, evaluation in this domain has not been much documented even though a lot of work has been done in the field. We have tried to summarise different evaluation methodologies developed in related projects. These methods highlight the different components of the enrichment process that can be subject to evaluation.
In order to validate all the recommendations provided in the previous sections, we have performed a quantitative and qualitative evaluation of seven enrichment services on a same subset of the Europeana dataset. The report of the evaluation is available in a companion document to this report while the main conclusions remain in this report.
This report is a result of an inventory of tools, practices and standards that define the current state of the art for semantic enrichment. The analysis and evaluation work done during the course of the Task Force have allowed us to compile a series of lessons learnt that should be considered for the design and enhancement of enrichment services and their evaluation.

1. Introduction

Enriching metadata has become a way in Cultural Heritage to overcome data quality issues while providing more context and multilingual information. Considered first as an experiment, metadata enrichment has become part of Europeana and its data providers' strategy. Understanding the key aspects of the process has become necessary to provide best practices and guidelines

A first Europeana Task Force attempted to define a strategy for multilingual and semantic enrichment (Stiller, Isaac & Petras (eds.), 2014). While analysing results of semantic enrichments, the Task Force identified several data quality issues that impacted the analysis of the enrichments and the definition of a proper strategy. The Task Force concluded that the quality of the metadata whether it is in the source or target datasets has a clear influence on the quality of the enrichments. Quality issues in the source metadata have been further analysed by the Task Force on metadata quality (Dangerfield et.al, 2015). The Task Force concluded that the enrichments increased quality issues and spread them across datasets and languages. However, the real impact of the enrichment process itself on the results of the enrichments had yet to be investigated.

Since the publication of the Task Force report, several activities were initiated with the goal to better steer the impact of the enrichment process. Europeana, for example, has tried to make the semantic enrichment process more transparent to its data providers by maintaining and enhancing its documentation. It also applied some specific recommendations from the previous Task Force report, i.e. on augmentation of the search index with GEMET concepts related to the ones directly connected to the objects to improve the overall quality of enrichments.
Identifying metadata enrichments issues requires a good knowledge of the source metadata. The increased involvement of Europeana's data providers has also impacted the quality of the enriched metadata. For instance, many data providers have embarked on providing themselves metadata connected to semantic resources available as linked data. Others are moving to automatically enriching their collections. Several projects are also investigating how crowdsourcing could be used to contribute richer, semantic data or identify errors introduced by an automatic process.

This Task Force has set up to make a snapshot of this progress, and further explore one of the progression areas identified by the previous Task Force: the enrichment process.

We have gathered a number of experts among Europeana's Network, whom have developed or applied enrichment services to the metadata.
The report presents the work carried out over seven months in 2015, during which our group has:

- inventorized relevant semantic enrichment work in the Europeana Network in the past years. We have tried to do this in way that we hope will facilitate identification and use of services relevant for the application that need them. This includes general considerations on the diversity of the processes at hand, and a focus on interoperability issues.
- further developed criteria to select datasets for semantic enrichment and illustrate with a thorough analysis of selected examples
- explored methodological issues for the evaluation of semantic enrichment services
- performed a quantitative and qualitative evaluation of seven enrichment services on a same subset of the Europeana dataset, containing 17.300 records.

This report, together with its two companion documents and a data space on the Europeana Assembla, presents all this. We have also included, as a conclusion, a number of lessons learnt we think should be considered for the design and enhancement of enrichment services as well as for their evaluation.

2. Concepts used in this report

Throughout this report, we used various terms to describe people with different functions and activities within the enrichment process. At the beginning of each enrichment process, there is the source data which will be enriched. This data comes from different *data providers*. The agent who is in charge of selecting the different datasets and processes them for enrichment is the *enricher* - the one who handles the process of enrichment. The user of the services made possible or enhanced through enrichment is the *end user*. Sometimes, enrichments can be crowdsourced. In these cases, the volunteers using the crowdsourcing tool and annotating data are the *annotators.* For testing the quality of evaluations, one often needs to create gold standards or annotate the quality of a given automatic enrichment. Persons doing this are called *raters* throughout this document.
Data providers, enrichers, end users, annotators and raters are also the audiences targeted by this report. We collected an enormous amount of information which can be used to define an enrichment strategy and help different groups with its execution.

2.1. Defining semantic metadata enrichment

The Task Force did not only try to gather and evaluate different enrichment tools. It also tried to find a common ground for the definition of enrichments and the different nuances it entails. This section defines some of the most important notions used by the Task Force for defining its objective and work.[1]

In the context of this Task Force, 'enrichment' is always conceived as being applied to the metadata that is exchanged between Europeana, the aggregators and the data providers.

Generally, a metadata enrichment task can be described as a process that improves metadata about an object by adding new statements about the object that this metadata describes. The term 'enrichment' can be used to refer to the process - e.g., the application of an enrichment tool - or its result - the new metadata created at the end of the process. The enrichment strategy refers to all workflows components and the processes which determined these components.

[1] A more general glossary of Europeana terms can be found at http://pro.europeana.eu/glossary

Enrichments can be created manually, semi-automatically or automatically (e.g. by means of information extraction).
Annotations (in the sense of the Open Annotation community[2], among others) can be one form of automatic enrichments. Reciprocally, enrichments can be represented as annotations (see **section 3.1 Techniques and tools for semantic enrichment**).

In previous work[3] we have identified the main component of enrichments:

- **Source:** the source objects whose metadata is being enriched (by extension it will also refer to the metadata set about these objects)
- **Target:** the set of resources used to enrich the source metadata, i.e. the values that will appear at the end of the process as in the new metadata for the source. Targets can be of different types, from simple uncontrolled strings to resources published as linked data (knowledge organization systems, datasets of cultural objects).
- **Rules:** enrichment rules specify how the enrichment between the source and target should be executed. Typically, in the context of automatic enrichment, rules take the form of instructions to create links based on matches between the various string representations attached to the resources in the source and the target.

Enrichments can be distinguished by the nature of their results:

- some enrichments produce *typed links*: the result includes links of a certain type between the resource (source object) described in the metadata and other resources (e.g. a dc:subject statement linking the resource representing a book and a resource representing a concept, or a label for such a concept). In the RDF model, these enrichments correspond to a set of RDF statements.
- other enrichments may result in simpler *tags* such as (non-semantic) string tagging. One only knows that a certain string is attached to the object, without knowing what is the relation between the object and whatever the string may represent.

Typed links enrichments can be further characterized by

- the *type of link*: enrichments can produce equivalence or others semantic relationships (broader/narrower), or any domain relationship (dc:subject), including even vague/unspecified relations like dc:relation
- the *type of source/target*. The linked resources can be of same type. For example, two (EDM) Cultural Heritage Objects, places, concepts… Or they can be of different types, as when an object is linked to a conceptual subject.
-

[2] http://www.openannotation.org/
[3] In a previous Task Force on multilingual and semantic enrichment strategy, several Europeana collections were analyzed, identifying enrichment problems and their causes. The report of the Task Force can be found here: http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/MultilingualSemanticEnrichment//Multilingual%20Semantic%20Enrichment%20report.pdf. Furthermore, a presentation by Antoine Isaac lists the workflow and processes in Europeana with regard to enrichment:http://de.slideshare.net/antoineisaac/enrichment-and-europeana

For example, ***co-referencing*** creates semantic equivalence relationships (using owl:sameAs or skos:exactMatch[4]) between resources of same type, using (semi-)automatic processes (e.g. CultuurLINK[5], SILK[6], LIMES[7])[8] and/or manual processes (e.g. crowdsourcing as in Wikidata Mix-n-Match).[9]

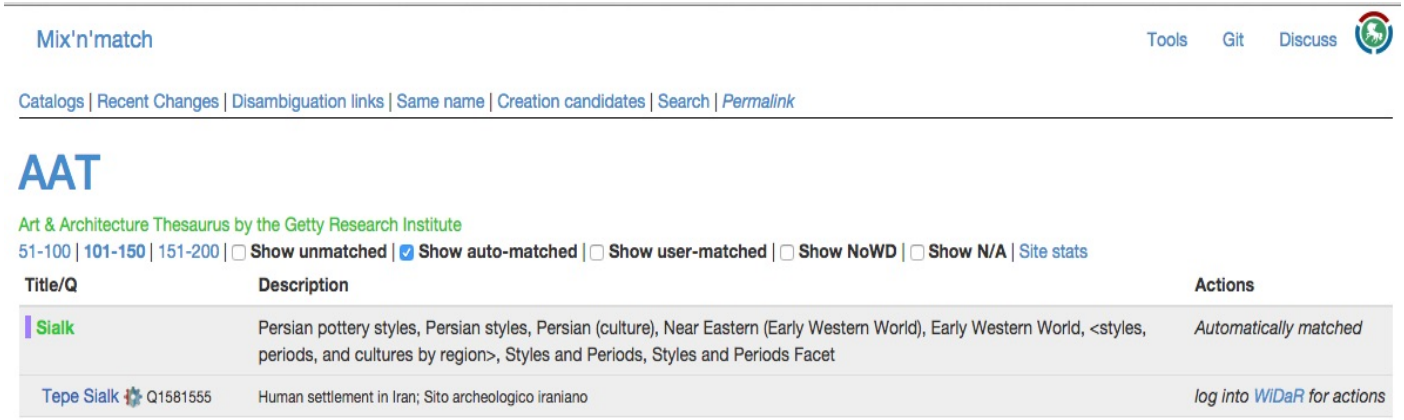

Figure 1: Example of co-referencing between AAT and Wikidata for the concept "Sialk" done in Mix-n-Match[10].

Alignment (or 'matching') of Knowledge Organization Systems (KOS) creates semantic relationships (in SKOS, using the mapping properties skos:exactMatch, skos:broadMatch, skos:narrowMatch, etc) between resources of the same type from two different KOSs.

Contextualization creates typed relationships between resources of different types. For example the Europeana enrichment relates Europeana Cultural Heritage Objects to concepts, places, etc., using the properties in EDM (dc:subject, dc:creator).

[4] http://www.w3.org/2002/07/owl#sameAs or http://www.w3.org/2004/02/skos/core#exactMatch
[5] http://cultuurlink.beeldengeluid.nl
[6] http://silk-framework.com/
[7] http://aksw.org/Projects/LIMES.html
[8] These tools can be used for other enrichment scenarios.
[9] More links at https://twitter.com/hashtag/coreferencing
[10] https://tools.wmflabs.org/mix-n-match/

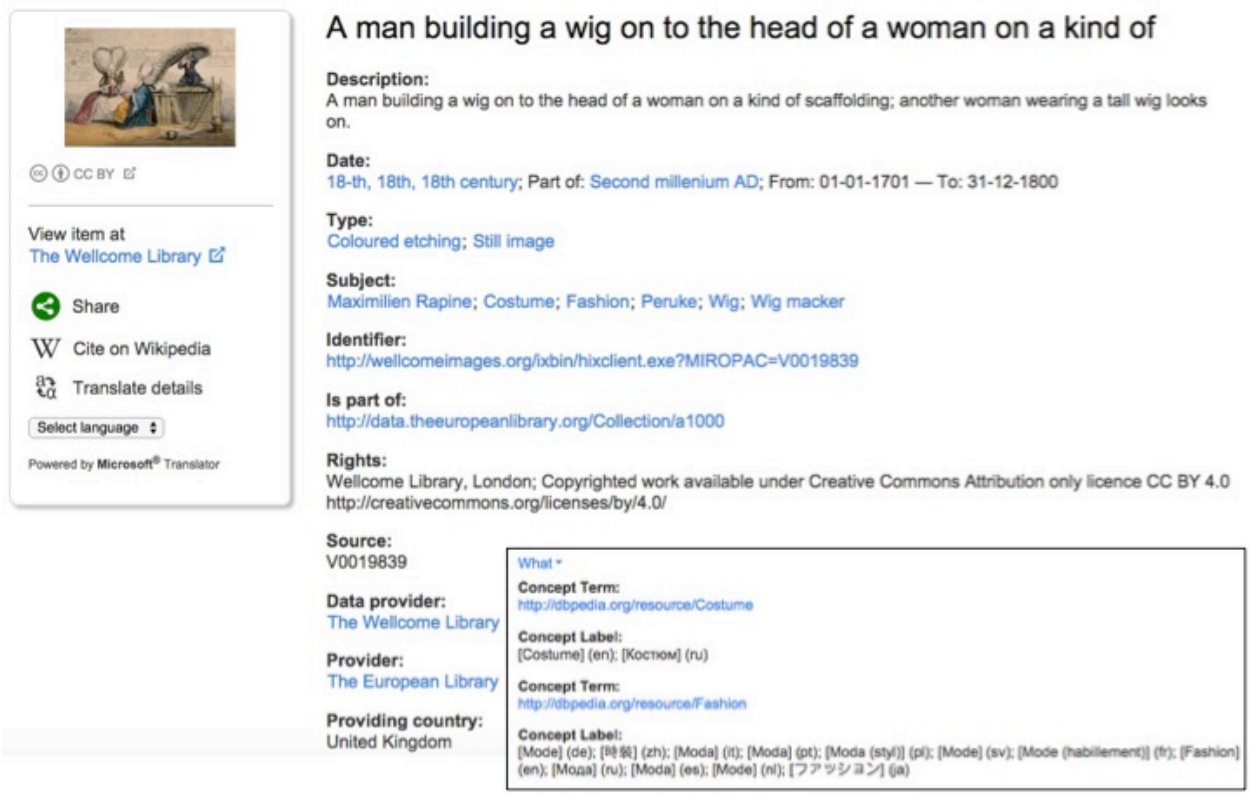

Figure 2: Example of Europeana semantic enrichment (or contextualisation) with concepts terms from DBpedia. A man building a wig on to the head of a woman on a kind of scaffolding; another woman wearing a tall wig looks on, Wellcome Trust: http://www.europeana.eu/portal/record/9200105/BibliographicResource_3000006114081.html

An enrichment may materialize links that were already implicitly present in the data, as in the case of metadata 'massaging' (e.g. by means of advanced mapping or cleaning tools such as OpenRefine[11]) where the (string) label of a concept used in an object's metadata is replaced by the identifier of this concept in its source vocabulary. It may add "new" metadata, as in the case of post-mapping enrichment (as in Europeana) where a link is created to a target resource, which was not referred to in the original metadata, even implicitly. While the former kind of enrichment can be seen as a (sophisticated) case of metadata mapping, the latter can be seen as a case of information extraction (albeit a possibly manual one), and is more error-prone.

Finally, the term 'enrichment' may denote a broader process, encompassing some of the preparation steps for the semantic enrichment task proper, as well as some of the metadata management tasks that follow it. Such a broader process has been described in (Stiller et al, 2014):

[11] http://openrefine.org/

- **Analysis**: the pre-enrichment phase focuses on the analysis of the metadata fields in the original source descriptions, the selection of potential target(s) and the creation of rules to enrich the source with the target. The choice of enrichment tools can also be also part of this analysis activity, and they can have a great influence on the choice of targets and the definition of rules.
- **Linking**: (automatically) applying the rules to connect the source resources to the target ones.
- **Augmentation**: adding more values from the target data to the original object metadata, after the basic enrichment data has been produced. When a cultural heritage object is enriched with a concept from a knowledge organization system, this step could include in the search index for this object all the multilingual labels of the concept. It could also include data about broader or narrower concepts (as shown in Figure 2).

This report will mostly focus on tools, methods and workflows for semantic enrichment.

2.2. Other relevant notions

Metadata mapping

- An expression of rules to convert structured data from one format or model to another such as EDM. (from the Europeana glossary[12])

Knowledge Organization System

- A list (vocabulary) of concepts, terms (labels) for them, scope notes (definitions) and relations between them (hierarchical and associative). E.g. thesauri like GEMET, Getty AAT, or authority files.

Authority List, Authority File

- A list of persons, families, their names, and relations between them. Eg Getty ULAN, LCNAF. Sometimes authority files are extended to include any kind of entry, e.g. German Integrated Authority File (GND) or Virtual Integrated Authority File (VIAF)

Gazetteer

- A knowledge organization system (thesaurus, authority file) of place names, e.g. GeoNames, Getty TGN

[12] http://pro.europeana.eu/get-involved/projects/glossary

Contextual Entity

- In Europeana, any resource (e.g. an edm:Agent representing a creator) that provides contextual information for a cultural heritage object (edm:ProvidedCHO)

Named Entity Recognition (NER, or NERD)

- An automatic enrichment task that seeks to locate elements within text such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Typically the outcome of this task is an annotated block of text that highlights the names of entities. In the context of metadata enrichment, named entity recognition is typically applied to textual data elements, such as titles and descriptions. Often NER is complemented by a Disambiguation phase (between several candidate entities), leading to the acronym NERD.

Crosswalk

- Establishing correspondences between two metadata schemas or ontologies, for the purpose of mapping data from one schema to another. Equivalent to 'metadata mapping', different from 'alignment of KOS' (see above).

3. Existing practices for tools and semantic enrichment services

It is nearly impossible to write an inventory of tools that can help enriching data. This is a very active field of development in the database and semantic web communities, for example. Many vendors (e.g. PoolParty[13]) or open-source projects (e.g. OpenRefine) have developed tools that have semantic enrichment as a core or auxiliary function. Rather than giving a complete overview of the state-of-the-art, the Task Force has preferred to focus on (i) an inventory of enrichment efforts and tools initiated in the context of the Europeana Network, which is presented in the table of projects in **Appendix A**; (ii) a general characterization of the methods and rules used by enrichment tools, which is presented in this section.

3.1. Techniques and tools for semantic enrichment

3.1.1. Tools for manual enrichment/annotation

The first approach to semantic enrichment, and arguably the more natural one in a domain where human cataloguers have described cultural objects for decades, is to let people perform the task of linking these objects to the resources made available thanks to new (linked data) technology. In parallel to *crowdsourcing* efforts that seek to empower

[13] https://www.poolparty.biz/

more people to create new data about objects, there are also efforts that seek to guide such efforts into creating finer-grained metadata, shifting the focus from basic tags to a richer form by linking to contextual resources.

This section further elaborates on such manual enrichment and annotation tools. Often these tools are either used for crowdsourcing annotations for large corpora, or give individuals the possibility to add, correct or edit existing enrichments. Depending on the digital objects enriched/annotated, one can distinguish between annotation tools for texts, images, metadata, audiovisual content, or multimedia items such as websites. Each of these content types presents different challenges for technical tools.

Images

One of the challenges in annotating images is to determine image sections the enrichment refers to. One example is Annotorious[14] which is an image annotation tool for the web. A very specialized tool is Pelagios' Recogito[15] annotation tool which allows annotators to annotate places in maps and other geospatial documents.

Audiovisual content

Annotating audiovisual content is challenging, as the annotation needs to carry the information of the particular frame or point in the time dimension of the object which was annotated. One example is Waisda?[16], which is a video annotation tool for the crowd which allows free tagging with optional matching against controlled-vocabularies.

Web resources

Many annotation tools let end-users annotate web content. For example, Pundit[17] is based on linked open data technology, and lets user annotate web content creating semantic annotations as RDF triples. Similarly, The Annotator[18] allows users to annotate web content - annotations are shareable and linked to their respective creators generating links between resources (webpages), annotations and annotators.

Other scenarios - citizen science and research, crowdsourcing in culture

More and more crowdsourcing tools are used to let the public generate knowledge on a large scale as input for science and research. These so-called citizen science projects are very successful in helping scholars to generate massive amounts of structure (and often annotated) data. The well-known platform Zooniverse[19] is one of them - here users can voluntarily assist research by tagging, classifying or enriching objects from various disciplines. The source code is also available on Github. Crowdcrafting[20] on the other hand

[14] https://annotorious.github.io/
[15] https://github.com/pelagios/recogito
[16] https://github.com/beeldengeluid/waisda
[17] http://thepund.it/
[18] http://annotatorjs.org/
[19] https://www.zooniverse.org/ and source code available at https://github.com/zooniverse
[20] http://crowdcrafting.org/

is a free and open source alternative to existing citizen science platforms. It is built with the software Pybossa[21] which is a framework for handling crowdsourcing projects and in collaboration with the Open Knowledge Foundation[22] and Citizen Cyberscience Centre[23]. Micropasts'[24] projects help gather data from the crowd about our human history (Note that the source code of Micropasts reuses a lot of the PyBossa code).

Historypin[25] is a web platform that enables the public to share their memories in form of collections of photographs, sound and stories mapping them to places and enabling others to explore local history.

Similar to the projects presented here, cultural heritage institutions themselves are moving in the direction of letting the end-user annotate or tag digital objects to improve metadata. The JocondeLab[26], for example, is a semantic tagging prototype. The Rijksmuseum also offers a semantic annotation tool. With Accurator[27], end-users can tag parts of the digitized objects with information about the depicted object.

Annotation support in Europeana projects

The Annotations API developed by Austrian Institute of Technology (AIT) for Europeana Sounds[28] is a backend component that will be implemented by reusing and extending the User Generated Content service developed within the scope of Europeana Creative[29] project. It implements the basic functionality for managing web annotations related to items in the Europeana metadata aggregation, based on a REST API and a JSON serialization of annotations.

The REST API acts as the interface of the annotation backend, encapsulating a Java API that implements the storage and retrieval of annotations. The current implementation, described in the Europeana Creative D2.2 *Services and Messaging API*[30], supports the creation of annotations on Europeana objects or web resources (images and other media files) using simple text comments, tags or semantic tags.

In Europeana Sounds support for the new W3C Web Annotation model[31] and protocol[32] will be implemented, and more annotation and body target types will be supported[33] (see

[21] http://pybossa.com/
[22] https://okfn.org/
[23] http://www.citizencyberscience.net/
[24] http://crowdsourced.micropasts.org/ and source code available at https://github.com/MicroPasts
[25] https://www.historypin.org/
[26] http://jocondelab.iri-research.org/jocondelab
[27] http://accurator.nl/
[28] http://www.europeanasounds.eu/
[29] http://pro.europeana.eu/structure/europeana-creative
[30] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Creative/Deliverables/eCreative_D2.2_NTUA_v1.0.pdf
[31] http://www.w3.org/TR/annotation-model/
[32] http://www.w3.org/TR/annotation-protocol/
[33] More details on this work in http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Delivera

also **section 3.2 Formats and specifications for semantic enrichment data** on the use of the Web Annotation data model).

3.1.2. Automatic enrichment processes and tools

Manual enrichment processes are of course limited in their coverage. The amount of objects that can be enriched in datasets like the ones of Europeana or other aggregators is too high for the human resources available. Semantic enrichment may be in the future included in the regular cataloguing processes, but this wouldn't solve the problem of how to enrich legacy metadata. Also, one could get input from the crowd but crowdsourcing efforts are often limited to quite specific collections, and with a specific objective in mind (e.g., focusing on persons, places, one specific data field...).

On the other hand, new technology also comes with the means to automate enrichment processes, at least partially. Very often the original (source) metadata or the content itself (text, audio, maps...) contain mentions of the concepts, places and other contextual resources that are in the target datasets. This allows one to create tools that exploit such traces to create semantic links between source and target resources.

As said earlier, it is impossible to present readers with a complete comprehensive report on all tools and techniques. For approaches based on metadata alone, enrichment can use Named Entity Recognition (NER) techniques that are essentially string-based, statistical analysis on the datasets to match, comparison of common information attached to the resources in source and target datasets, employing third-party contextual resources, or even machine-learning techniques where data elements are used in an enrichment process that seeks to reproduce the positive examples listed in a learning set.

We can only give readers hints of the various techniques employed, through some observations in this section, and pointing to more references for the enrichment tools we have evaluated as part of our comparative evaluation, which are included in the companion document to this report (see Task Force page). We also list some relevant APIs in Appendix C.

Linking rules for enrichment

Many enrichment services are based on rather simple *rules* that specify how to recognize when a resource from the source dataset should be connected to a resource from the target dataset, by comparing the lexical or numerical attributes of these resources.

For example the MORe tool used in the LoCloud project[34] allows enrichers to create simple rules in the form: if the subject of an object contains X or some keyword attached

bles/EuropeanaSounds-D2.4-Crowdsourcing-infrastructure-V1-assessment-and-reccommendations-v1.0.pdf

[34] http://www.locloud.eu/Resources/MORe-repository

to it contains Y then add the subject K, where K is taken from a collections of subject terms that refer to a specific theme.

The DM2E project[35] has used the Silk[36] linked data framework to align places and agents from local collections to DBpedia. The process is based on one rule (expressed below in the Silk syntax) that states two resources should be linked when the Jaro-Winkler distance[37] between their (lower-cased) labels (values of skos:prefLabel or rdfs:label in their metadata) is below the 0.1 threshold:

```
<LinkageRule>
  <Compare id="unnamed_11" required="false" metric="jaroWinkler" threshold="0.1">
    <TransformInput id="unnamed_13" function="lowerCase">
      <Input path="?a/skos:prefLabel"/>
    </TransformInput>
    <TransformInput id="unnamed_14" function="lowerCase">
      <Input path="?b/rdfs:label"/>
    </TransformInput>
  </Compare>
</LinkageRule>
```

This rule has been established in a conservative way, after tests by the enrichers. The quality of enrichment results depends of course on the way the target dataset has been selected (see Section 3). But even with a good target, defining rules require very good knowledge of the data at hand, both source and target. And even then, carefully created rules may fall short in tackling some hard issues. In the case of DM2E, performance was judged fine for place enrichment, but there are still problems with agents (persons and organizations).

The first difficulty is in handling cases of ambiguity: for a given resource in the source dataset, several resources in the target could comply with the selection criteria defined in the matching rule. This especially happens for approaches that use in natural language processing (NLP) techniques, essentially based on the labels of the resources to link. Very often, especially for persons, a name can be ambiguous, i.e. there are several persons with the same name in the dataset. In which case, determining the right candidate require the matching strategy (or rule) to take into account more contextual data about the resource, such as dates of birth and dates, associated places, etc.

Some measures can be applied to limit the possibilities of ambiguity, some of them also influence the choice of the target dataset (see Section 3). For example, make sure that labels are matched when they are from the same language, to avoid multilingual ambiguities. Or try to have source and target dataset cover the same 'intellectual domain'. In fact the more that can be done to restrict the size of the target dataset, the higher the precision can be expected to be. It may also be interesting to try to "boost the relevance"

[35] http://dm2e.eu
[36] http://silk-framework.com/
[37] The Jaro-Winkler distance is a measure for the (dis)similarity between two strings. Cf https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

of a particular part of the target dataset, for example a set of resources focusing on a restricted geographic or temporal coverage. Namely, to give priority to candidate matches that come from this part, as opposed to resources coming from other parts.

The second difficulty is in the definition of the weights assigned to the various data element comparisons and checks used in the linking process. This is a place where machine learning techniques can be used, to determined the ideal combination of weights and checks, based on an existing set of good matches (most often produced manually). Examples can be found for places (Freire et al., 2011) for persons (Freire et al., 2008), for work titles in library catalogues (Freire and Juffinger, 2011), etc.

Iterative experiments can also be very helpful, i.e., when the enricher herself tries a configuration of weights, runs the enrichment process, assess the results and depending on their quality, possibly adapts the weights and perform new tests, leading to a new assessment, etc., until a satisfactory level of quality is reached.

3.1.3. Designing semantic enrichment workflows

A first degree of freedom in the design of an enrichment process is of course the ability to choose the most relevant tool or service among the various ones available - including tools that would allow enricher or end-users to manually enrich a dataset or validate an existing enrichment. One has to pick one that fit well the case at hand: source, target, and application scenario[38].

A second degree of freedom is the parameterization of the enrichment method employed. While some semantic enrichment tools function as black boxes, many of them offer possibilities for tuning the process. Offering parameters that the enricher can tune has very recently appeared as a key requirement, since monolithic procedures cannot fit a wide variety of sources and targets. Some parameters can be relatively obvious, depending on the matching processes. For example, it is easy to define a varying threshold for fuzzy string-based similarity measures: when the similarity between the labels of the source resource and the target resource is above this threshold, the match gives raise to an actual enrichment. When it is below, no new link is created. Some other tuning methods require more effort to be employed in the most appropriate way, such as choosing a right learning set for a machine learning-based tool.

Some tools empower even more the enricher by allowing her to choreography different enrichment techniques, where different entity matching modules are applied in different *steps* of a *workflow*[39].

[38] The quality of an enrichment partly depends on the application that will employ it. One's good enrichment may be another's (relatively) poor enrichment.
[39] This approach somehow reproduces at the level of a general enrichment process the sort of "piping" of different natural language processing tools (part-of-speech tagger, stemmer, etc), which are employed for NER.

For example, the MORe aggregation toolkit developed and used in the LoCloud project makes available various enrichment services, which the enricher can combine using an enrichment plan editor.

In any case, for choosing an enrichment technique, let alone, properly tuning it, one requires good knowledge of the working of this technique, as well as the characteristics of the source and target datasets it will be applied to.

A first solution is to make available documentation that is appropriate to compare the various options available - as opposed to purely technical documentation on the various techniques in isolation. For example, in the case of geo-coding, several services are available (DAI gazetteer[40], Geonames-based, Pelagios). Although these services are quite different and serve different purposes, this is not always easy to grasp. The MORe team has started to build a simple web guide in order to state the differences, as part of MORe's enrichment plans editor.

Another option is to offer agile toolkits in which the enricher can easily test different techniques or parameterization thereof, quickly assess the results, and change the method or the parameters when the results are not appropriate. An example of such tool is the CultuurLINK service, which follows the Amalgame framework[41] developed in the EuropeanaConnect project (Wielemaker et al., 2011). While these tools were developed for alignment of vocabularies, their philosophy can be applied to general semantic enrichment problems. The principle is to "pipe" individual (tuned) alignment techniques onto each other. Different modules can be applied to different parts or the source or target dataset, offering the flexibility to apply one technique on the parts of the datasets where it will perform best, and to use different techniques to complement each other: a module can be applied on the part of the data on which another module has failed to provide (good) results. The enricher can evaluate the results of any step, directly re-run this step with different parameters, as well as the steps that build on the results of this step.

[40] http://gazetteer.dainst.org/
[41] http://semanticweb.cs.vu.nl/amalgame/

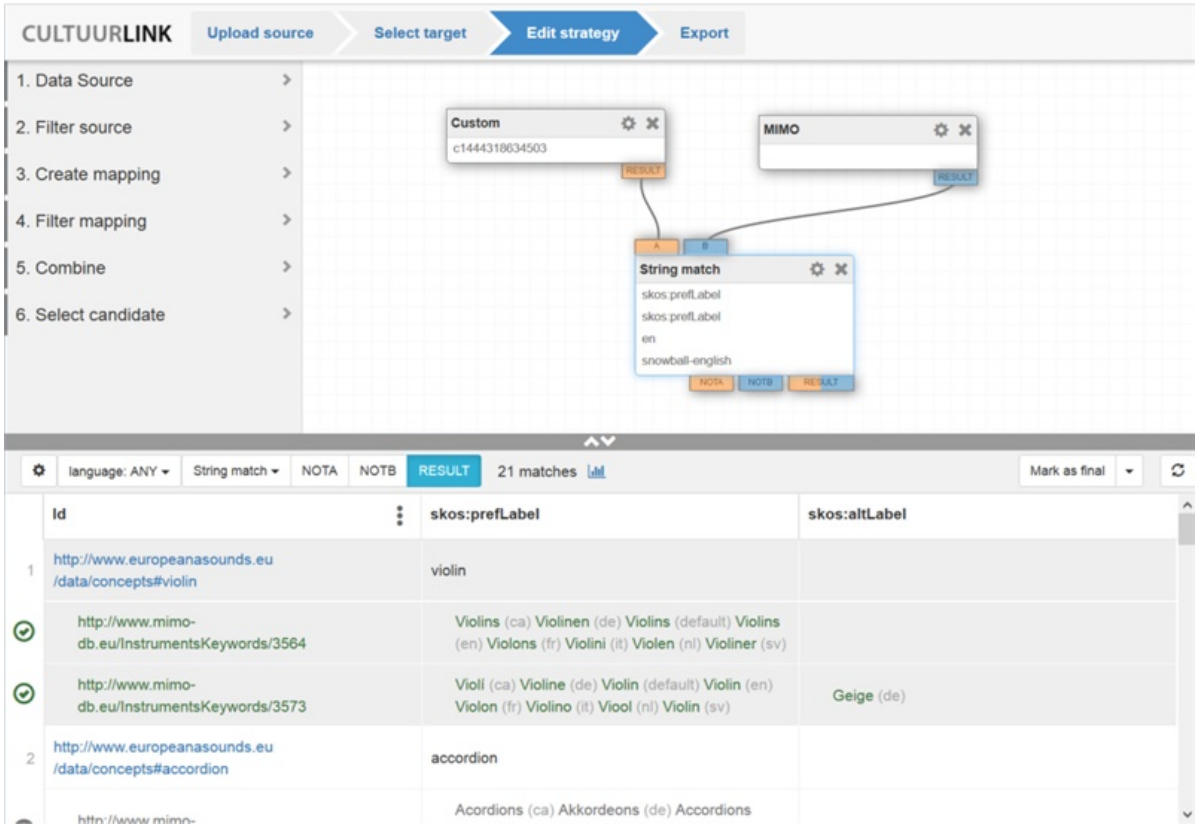

***Figure 3**: Using the CultuurLINK alignment workflow tool.*

3.2. Formats and specifications for semantic enrichment data

Semantic enrichment services should provide appropriate representations of enrichments. Ideally, these should follow similar patterns and standards. They should also publish metadata about the enrichment, e.g., information about how the enrichment was provided (provenance), and the confidence on its correctness. This information would enable data consumers, such as Europeana and aggregators, to automatically process individual enrichments, using only those with the required characteristics (including performance against quality indicators) for a particular purpose.

EDOAL (Expressive and Declarative Ontology Alignment Language)

The EDOAL[42] alignment format is an example from the related area of ontology alignment. In EDOAL, both the method used for alignment, and its measured confidence, can be specified. EDOAL is however a fairly generic framework. Its simplicity helped build a very active Ontology Matching[43] community in the Semantic Web research area. Its genericity is also a great asset there: whether an alignment is produced using a simple string-based matching or a combination of dozens of different techniques does not make a big

[42] http://alignapi.gforge.inria.fr/edoal.html
[43] http://ontologymatching.org

difference. However, it does not specify a lot of the information features one could need to have a clearer picture of an alignment characteristics, and hence its potential for re-use, even for enriching object metadata. As it is based on RDF, it is in principle extensible with specific classes and properties that would represent provenance and trust better, as soon as a community agrees on such constructs.

ITS (Internationalization Tag Set)

ITS is a W3C-based effort to provide vocabularies to support internationalization work, e.g. translation or localization. ITS can be used to represent information about an enrichment process, with a strong focus on natural language processing tools (like named entity recognition) and their output. ITS 2.0 tools annotation[44] enables to present information about a tool used in this process, and ITS 2.0 text analysis[45] and machine translation[46] provide attributes to represent confidence in the results for such processes. Note that in general, automatically generated confidence values are tool specific, hence in best practices one should require to have confidence information only if the tool is known too.

Other relevant information in ITS mark-ups includes terminology references. When a term from a reference terminology resource is recognized, it is possible to use the ITS terminology[47] specification to reflect this. One example of how digital (XML and HTML) content can be enriched with terminological data is "Terminology Annotation Showcase[48]" from the FREME[49] project. Note that if a controlled vocabulary is made available using Linked Data receipes, i.e. if it is possible to get machine-readable version about a term by just looking up its web identifier (URI) then the need for embedding extra terminological information becomes less important. But many terminologies are not yet published this way.

Finally, ITS, as a result of a community focusing on internationalization, has also metadata categories that could be relevant for Europeana's own multilingual efforts, e.g. language information, translation provenance and localization quality issues.

NIF (NLP Interchange Format)

Some of the ITS categories have been re-used in the context of NIF[50] (formerly NLP2RDF), which is an RDF format for representing the results of NLP analysis, including:

- word/phrase/sentence demarcation
- parsing (including dependency and deep parsing) using OLIA[51] and constituent ontologies

[44] http://www.w3.org/TR/its20/#its-tool-annotation
[45] http://www.w3.org/TR/its20/#textanalysis
[46] http://www.w3.org/TR/its20/#mtconfidence
[47] http://www.w3.org/TR/its20/#terminology
[48] http://taws.tilde.com/
[49] http://www.freme-project.eu/
[50] http://persistence.uni-leipzig.org/nlp2rdf/
[51] http://purl.org/olia/

- NER based on the ITS2 ontology.
- sentiment and opinion analysis, using MARL[52]

NIF has the potential to act as an umbrella for exchange (in RDF) of the results of many existing (XML-based) NLP formats, such as GATE[53] and FOLIA[54].

NIF relies on mixing different vocabularies (e.g. the Provenance ontology[55]) to meet its requirements and other NLP frameworks can build on it in turn. A major point of NIF is to enable easy lightweight integration of different NLP tools through NIF adaptors. See NIF Combinator[56] and GERBIL[57] (easy integration of 9 NLP tools). As another example, the Named Entity Recognition and Disambiguation (NERD[58]) ontology and API combine the NIF framework with a set of classes for types of recognized entities (e.g. nerd:Organization) that ties together the classifications coming from several different NER tools. NIF can point to them using itsrdf:taClassRef, or to individuals (which is richer info) using itsrdf:taIdentRef.

Links and marked-up examples (including NER provenance info about tool and confidence) are available online[59]. NIF supports 3 profiles, listed in order of increasing representation cost (i.e., triples per enrichment):

- *Simple*, for making enrichment statements directly over words/phrases
- *FISE/Stanbol*, for accommodating several annotations per phrase, with different confidence, eventually by different tools
- *Open Annotation*, for most general annotations

There is also a NIF tutorial[60] from the FREME project, in which further development of NIF is ongoing.

Using the Web Annotation Data Model for representing manual enrichments

W3C's Web Annotation data model, based on previous work from the Open Annotation community, is model is a natural fit with manually produced enrichments, as it provides the basic building blocks to represent annotations: the *target* of an annotation (what is being annotated), its *body* (the annotation text, or the URI of a linked data resource, for semantic annotations) as well as basic provenance metadata.

The Europeana Creative and Europeana Sounds projects have started to use this model as a basis for the annotation services mentioned above, as can be seen for the Annotations API[61].

[52] http://purl.org/marl/ns
[53] https://gate.ac.uk/sale/tao/splitch5.html#x8-1070005.5
[54] http://proycon.github.io/folia/
[55] http://www.w3.org/TR/prov-o/
[56] http://demo.nlp2rdf.aksw.org
[57] http://aksw.org/Projects/GERBIL.html
[58] http://nerd.eurecom.fr/
[59] http://vladimiralexiev.github.io/Multisensor/20141008-Linguistic-LD
[60] http://www.slideshare.net/m1ci/nif-tutorial

Using the Web Annotation Data Model for representing automatic enrichments

The Europeana related PATHS project[62] proposed to create a profile/extension of the Web Annotation data model to represent automatic enrichments as annotations[63].

It mostly consists of guidelines on how to represent additional information about enrichment, which are usually not the focus on manual annotations, such as confidence and the target vocabulary being used.

More investigation needs to be done on which model would be best in the context of Europeana and its Network for simple exchange of enrichment information, and the design of interoperable enrichment tools and their clients. With a focus on Linked Data technologies, it is advised to have a close look at the models presented in this section, which are based on RDF, chiefly Web Annotation or EDOAL for a basic layer, and NIF for the tools that specifically use text processing techniques.
The Fusepool[64] Annotation Model, related to the Apache Stanbol[65] project, could provide a useful inspiration, as it combines Web Annotation with NIF to meet the requirements of different usage scenarios such as language annotation, entity detection / linking and topic classification[66].

Finally, Europeana's existing EDM proxies (from the OAI-ORE model) can be used to represent enrichments. Europeana's own automatic enrichment are represented by attaching to a Europeana-specific proxy the URIs from the target datasets, using the same properties as the original fields these are derived from. In in the context of general RDF-based frameworks, named graphs (a.k.a quadruples) could be used to the same effect, possibly adding more provenance information (potentially using one of the models mentioned above).

4. Selecting target datasets for semantic enrichment

Sourcing relevant datasets for semantic enrichment was an issue already identified by the EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy (Stiller, Isaac & Petras (eds.), 2014). Their report had listed some example vocabularies. We have sought to give more methodological guidance on this aspect. We recommend to follow several steps to select potential enrichment targets. A proper evaluation and selection of enrichment targets is key to improve the quality of enrichments.

[61] http://www.slideshare.net/gsergiu79/europeana-creative-annotation-api
[62] http://www.paths-project.eu/
[63] http://www.paths-project.eu/eng/content/download/5342/40580/version/6/file/edmPATHS.pdf
[64] http://p3.fusepool.eu/
[65] https://stanbol.apache.org/
[66] https://github.com/fusepoolP3/overall-architecture/blob/master/wp3/fp-anno-model/fp-anno-model.md

- *Analyse the source data*: a good knowledge of the source data in terms of topic coverage, gaps, quality issues is necessary before selecting an enrichment target.
- *Identify the enrichment requirements*: before performing an enrichment, the enricher should have already define the expected results. For instance an enrichment could be performed to improve the overall quality of a dataset. In this case the quality issues to be fixed should be identified before performing the enrichment.
- *Find datasets available on the Web*. The Task Force recommends to select datasets available on the Web. Several inventories are available to help enrichers to source enrichment targets.
- *Select the enrichment targets*. Before selecting a target, the enricher will have to evaluate potential targets. The Task Force has identified a series of criteria that can be used by an enricher to evaluate targets against his source data.
    - Availability and Access: The Task Force recommends to select targets available on the Web and compliant with the Linked Data recipes. These targets should re-usable under an open licence.
    - Granularity and Coverage. The enricher should select targets that have the same coverage than the source data or that can complement the source data.
    - Quality. The enricher should pay attention to the quality of the target in terms of semantic and data modelling.
    - Connectivity. The Task Force recommends to select target with incoming and outgoing links to other targets.
    - Size
    -
- *Test the selected target on a sample of source data*. Once the target is selected, it should be tested on a sample of data before being applied to the whole dataset. A test will allow to verify whether the target really covers the source data or whether it doesn't introduce semantic ambiguities.

In some cases, it might not be possible for an enricher to select an existing target. The enricher will have therefore to build its own target.

The dataset selection strategy is available as a separate document on the Task Force page[67].

5. Methods and metrics for evaluation

Enrichment tools and enrichment workflows need to be evaluated during their development and employment to ensure they are suitably adjusted to the domain and requirements of the underlying data avoiding decrease in user satisfaction. Enrichment

[67] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments

tools should be evaluated on their performance and suitability for their domain and envisioned goal.

5.1. Scope of evaluation

One can distinguish between two different types of evaluations which are characterised by different methods and results:

Intrinsic Evaluation

Here, the main focus of the evaluation is the enriched metadata and the corresponding links itself. This type of evaluation will be mainly executed by domain experts (often the data providers) who know the data basing their judgment on the criterion "correct" or "incorrect". The enrichment is evaluated based on its semantic correctness and the evaluation is technology orientated trying to create an experimental setting where only the enrichment as such is evaluated. This can also be further elaborated by evaluating the scope of the match between the source metadata and the target vocabulary or considering related entities, such as a city and the region it is located in. Evaluation criteria could focus on close matches as defined in SKOS (skos:closeMatch) for example when entities evolve over time - especially true for geographic entities which can change due to change in jurisdiction . An even more advanced criteria could be to assess the enrichment depending on the field and/or the types of entities that is used in one enrichment.

Extrinsic evaluation

Extrinsic evaluation considers the context of the enrichment, the enrichment workflows itself and other external factors which shape the outcome of the enrichment process. So it is more targeted on defining the impact of the applied enrichments on the end-user experience and performances. The goal of such evaluations can be manifold and the extrinsic evaluation considers the context and goal of an enrichment and its impact on search performance, browsing functionalities and user experience. Often this type of "end-to-end"-evaluation considers the whole range of impact enrichment can have various fields within a system.

Search Performance

To evaluate the search performance is a classical area of Information Retrieval with various methods applied. The field moved from classical information retrieval evaluation in laboratory settings (often referred to as the Cranfield paradigm) to a more user-oriented approach with more diverse evaluation methods.[68]

Another area to explore is the impact of enrichments on browsing capabilities. The impact on retrieval is quite straightforward whereas browsing capabilities due to enrichment is probably tricky to assess.

[68] A good overview on the evaluation of interaction information retrieval is given by (Kelly, 2009).

User Experience

These methods are targeted in defining the user satisfaction with given results or the experience of a given portal as such. Here, one can think of A/B-testing where users decide which result set or object description (with or without enrichment) they prefer. Also, all methods of usability research can be applied here either through quantitative methods such as log files or qualitative methods such as surveys. Both have advantages and limitations and choosing the right method for evaluation depends on the information one wants to obtain. For all these methods it needs to be noted that the results might be influenced by more factors than just the enrichments. For example, a survey could be asking the users if they can access the content using their native language, if the response rate is high, it could mean that the multilingual enrichments are qualitatively high but it could also mean that the user did only encounter good examples.

Relevance

The relevance criteria is a tricky one. It tries to assess how relevant is a given enrichment to the object. That can be considered with regard to the other enrichments a digital object might have or considering the weight the enrichments has when for example it comes to push the enriched object in the search results.

5.2. Approaches and Tools

Gold standard based evaluation

This is often used for measuring the performance of a given algorithm for example for ranking. In the particular case of information retrieval, the relevant documents for a given information need are identified and aggregated to a gold standard. The algorithm can then be tested by analysing the overlap between the found documents and the gold standard. Transferring this to other applications, a gold standard would always display the ideal result a given algorithm should deliver. In most cases, the gold standard is developed manually and can be applied in intrinsic and extrinsic evaluations looking how well an algorithm performed a given task.

Tools for evaluation

As mentioned before, Information Retrieval has a long tradition of gold standard evaluation in fixed setting. There are tools which support the evaluation workflow and are able to compute the relevant measures. One of these tools is Direct[69] which was developed during the EU-funded project Promise[70] and allows researchers to manage all workflows of an evaluation campaign and produce the relevant metrics making the research data shareable and reusable (Ferro, 2011). As our use case is different from the

[69] http://direct.dei.unipd.it/
[70] http://www.promise-noe.eu/

ones typical in IR, the adaption of the tool to Task Force's requirements would have been too difficult.

Another tool is GERBIL[71] which is a community driven evaluation framework for entity annotation. It addresses some common problems of annotation tool evaluation, e.g. it provides not only persistent IDs and a repository for archiving results and settings but also aggregates the measures from different experiments to assess certain tools (Usbeck, Ricardo, et al., 2015). GERBIL uses NIF as the format for the output of the enrichment tool and need to be fed with an already existing gold standard which should be also in NIF.

The Ontology Matching community has produced a tool similar to GERBIL: the SEAL platform, which is used to run the yearly Ontology alignment Evaluation Initiative (OAEI)[72] campaigns. GERBIL seems closer to the needs of NER-like enrichment, as the NIF format allow more detail than the OAEI format[73] (which is specific to ontology alignment, see **section 3.2 Formats and specifications for semantic enrichment data**).

Finally, and while it is dedicated to evaluating co-referencing tasks only, the Mix'n Match tool (see Figure 1) could be a useful inspiration. It lets users confirm in a very quick and easy way matches between Wikidata and other resources, prior to loading these co-reference links into Wikidata.

Comparing semantic enrichment tools

Some evaluation might target different tools and their performance. In the case of comparing different enrichment tools with each other, one needs to determine the basis of agreement between the tools. This is important for determining the overlap of enrichments found by the used enrichment tools and interpreting the results.

There are several possible levels of agreement:

- *super-basic*: whether the enrichments link an object to entities of the same type
- *direct:* whether the enrichments link an object to the same entity identifier (URI)
- *direct, with co-reference*: whether the enrichments link an object to the same entity, considering that different URIs appearing in different alignments may co-refer to the same real-world entity (thus they are in "semantic agreement") or not
- *independent on type of link*: whether the enrichments link an object to the same entity, but ignoring the type of link (e.g., one relates an object to a concept using dc:type, the other using dc:subject)
- *other option*: finding related entities, e.g., a company and its headquarters. The problem is to find appropriate data to compute such softer agreement

For our comparative evaluation, we have opted for computing agreement taking into account the type of enrichment link and co-reference relationships.

[71] http://gerbil.aksw.org/gerbil
[72] http://oaei.ontologymatching.org/2014/seals-eval.html
[73] http://alignapi.gforge.inria.fr/edoal.html

Annotating Enrichments

If one tries to classify the results delivered by an enrichment process or algorithm, a good solution is to annotate a representative sample of the enriched corpus. The criteria used for the evaluation can be of intrinsic (system-focused) or extrinsic (user-focused) nature. The Task Force came up with several categories which can be used to assess the quality of a given enrichment. Whereas the intrinsic criteria seem to be easier to assess, extrinsic criteria depend on the goal of the enrichment assessing the relevance of the enrichment with regard to search performance, user experience or by weighting in the other created enrichments. The following table gives an overview on the different criteria within the intrinsic and extrinsic category of the evaluation of enrichment.

Category	Criterion	Description	CH-Domain example
Intrinsic	Semantic correctness	Is the enrichment semantically correct or not?	
	Completeness of name match	Was the whole phrase/named entity enriched or only parts of it?	
	Completeness of concept match	Whether the matched concept is at the same level of conceptual abstraction as the named entity/phrase. Since sometimes the exact concept is not available in the target vocabulary, a narrower or broader concept may be used in the enrichment. This is also true for a geographical region where the concept identified describes a smaller entity (narrower code), whereas the “broader” code refers to a bigger geographical region. The category would also cover change of geographic locations over time. For these cases, it is especially important to train the enricher thoroughly.	
Extrinsic	Informational value of enrichment	Enrichment targets content descriptive elements such as the author or the subject keywords, formal elements as rights statements, type of the resource and provenance; Does the enrichment focus on content describing keywords or on keywords describing rather formal aspects?	"book" or "documentation" would be formal, “Shakespeare” would be content; another example would be the enrichment of a place where a book’s publisher is located would be a formal characteristic whereas the enrichment of a place which is described in the book would be a content characteristic.
	Specificity of an enrichment	This category refers to the coverage of an enrichment considering all other possible	For example, enriching an article describing an

		enrichments.	scientific experiment with "measurement" would be considered a random enrichment because just one word out of many more was picked. A targeted enrichment would be one which described the source well, e.g. enriching the Novel Romeo and Juliet with "Shakespeare".

5.3. Past evaluations of semantic enrichments in the cultural heritage domain

Stiller et al. (2014) evaluated enrichments in Europeana looking at the intrinsic relationship between enrichments and objects but also taking extrinsic factors such as queries into account. The results for the extrinsic evaluation are subjective and based on the queries chosen for the evaluation. Nevertheless if a representative query sample is chosen it can give insights about how often user encounter potentially beneficial enrichments or incorrect ones. The evaluation was based on experiments done during the project Europeana v2.0[74]. The evaluation can be found in the section (4.3) of the report D7.8: Final report on Innovative Multilingual Information Access[75].

A previous evaluation of the enrichment in Europeana qualitatively assessed 200 enrichment for the four different types: time, persons, location and concepts (Olensky, Stiller, and Dröge, 2012).

As mentioned before, enrichments impact the search performance and are often executed to improve search across several languages. Having this in mind, all evaluation targeting the search performance are also relevant as enrichment evaluations. An overview of the search performances in Europeana can be found in **Appendix B**.

For the datasets used for enrichment there should be clear quality guidelines. In EuropeanaConnect, a document on evaluation criteria for language resources[76] was

[74] http://pro.europeana.eu/project/europeana-v20

[75] http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/Readings/D7.8%20Final%20Report%20Innovative%20Multilingual%20Access.pdf

[76] http://www.europeanaconnect.eu/documents/06_%20Europeana_Language_Resources_Evaluation_Criteria.pdf

produced. Coming up with criteria like this could be feasible for resources for enrichments, too.

OpenRefine was also used to perform evaluation. One was an evaluation of structured field reconciliation which can be read upon here: http://freeyourmetadata.org/publications/freeyourmetadata.pdf. The other is an evaluation of named-entity recognition[77] on unstructured fields (performed with OpenRefine and a plugin): Both have been evaluated on concrete datasets with a manual validation.

DM2E[78] has done a sample evaluation of alignment of (local) places and agents to DBpedia. For that, 150 random agents and 150 random places were taken. The random sample was based on the amount of agents/places each collection contains, e.g. of all agents across all collections 9% come from the KPE collection and the random sample also contains 9% KPE agents. The results show that 18% agents and 60% places are linked. From this, 83% agent links are good and 85% place links are good.
The 60% just comes from the fact that of all places in all collections, a link to DBpedia (can be either correct or incorrect) could be found for 60%, for the remaining 40%, the automatic tool cannot detect a link at all (caused by the linkage rule focusing on precision). The evaluation did not consider how many links are actually possible or whether there is no linking resource at all.

The Paths Project developed functionalities for information access in large-scale digital libraries. They focused on metadata enrichment as one way to improve user experience letting user discover and explore cultural heritage material. The evaluation of the prototype centered around assessments with focus groups and within laboratory settings. They also developed a methodology for semantic enrichments[79] (Paths Project, 2013). The enrichments itself were tested indirectly by choosing a methodology of Interactive Information Retrieval (IIR) in a laboratory setting - user performed tasks and log data, screen recordings and observer notes were collected.[80]

[77] http://freeyourmetadata.org/publications/named-entity-recognition.pdf
[78] http://dm2e.eu/
[79] Semantic enrichment of cultural heritage content in PATHS - *Mark Stevenson and Arantxa Otegi with Eneko Agirre, Nikos Aletras, Paul Clough, Samuel Fernando and Aitor Soroa*
[80] D5.2 Evaluation of the second PATHS prototype - *Jillian Griffiths, Sheena Bassett and Paula Goodale with Rodrigo Agerri, George Chryssoichoidis, Kate Fernie and Jen Smith*

6. Comparative evaluation of semantic enrichments

In order to gain better insight over the quality of enrichment tools and, in particular, on evaluation methods and metrics, the Task Force members have undertaken an evaluation campaign. The complete process that was applied, including an analysis of its results, has been described in detail in a separate document that is available on the Task Force page[81].

For this evaluation, we looked at metadata aggregated by The European Library[82] (TEL), as it is the biggest aggregator for Europeana and has the widest coverage in terms of countries and languages among its data providers. Therefore, TEL provides a good representativity of the diversity of data in Europeana. An evaluation dataset was selected containing 17.300 metadata records from all 19 countries[83] that contribute to TEL. The dataset was provided in the Europeana Data Model[84] (EDM) to some participants of the Task Force that use and/or develop enrichment tools. A total of about 360k enrichments were obtained from 7 different tools or tool settings from which 1757 enrichments were sampled and manually annotated by the Task Force members, making up the annotated corpus used for the evaluation.
In this process, we have learned a number of lessons that made us change our original plans, or would need to be considered for future experiments:

- **Select a representative dataset for your evaluation:** Make sure your corpus sufficiently gathers the diversity of your source data, covering aspects such as language diversity, spatial dispersion, as well as distinct subjects and domains.
- **Building a gold standard is ideal but not always possible:** Apply a manual strategy to build a reference set of correct alignments if you have sufficient time and human resources to commit to it, otherwise go for a semi-automatic strategy by selecting and assessing the enrichments identified by the tools under evaluation- or even by other enrichment tools results. The tradeoff is that the latter option will not provide absolute recall figures.
- **Consider using the semantics of target datasets for evaluation:** Some target datasets may be connected together by coreference links. These links may be used (e.g., in a process that "normalizes" the enrichments) to get a better view on how enrichment compare across tools, or to reuse a gold standard coming from another evaluation.

[81] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments
[82] http://www.theeuropeanlibrary.org/
[83] Austria, Belgium, Bulgaria, Czech Republic, Finland, France, Germany, Ireland, Latvia, Luxembourg, Netherlands, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain and United Kingdom
[84] http://pro.europeana.eu/edm-documentation

- **Try to keep balance between tools in comparative evaluations:** Some of the above strategies are likely to result in a bias against some tools. Make sure such bias is recognized and if possible properly related to the concerns that motivated your evaluation strategy.
- **Write clear guidelines on how to annotate the corpus:** guidelines should be simple enough for raters to understand but still detain the necessary information to make the right judgement. Consider having examples for the cases that may raise the most doubt. Consider testing your guidelines with the raters before and if necessary train them.
- **Use the right tool for annotating your corpus:** Choose or develop a tool that can best help raters efficiently and effectively perform their task. It should fulfill the following requirements: display the necessary information; respects the guidelines that were defined; and guide raters through their task.

The annotated corpus was then analysed by generally following the metrics from Information Retrieval, i.e., precision and recall. The analysis of the results shows a "relaxed" precision ranging from 77.4% to 98.5%. If a more stricter metrics is applied for precision then it significantly drops for most of the tools (from 31.2% to 98.2%). Regarding recall, we applied a pooled recall measure which ranged for the relaxed metrics between 4.8% to 41.8%, while for the strict it ranged between 2.4% to 43.2%. More details can be consulted on the separate document about evaluation.

After measuring and analyzing the results, the Task Force makes the following recommendations in order to improve the general quality of enrichments:

- Consider applying different enrichment methods and techniques depending on the (kind of) property subjected to enrichment; not only considering whether it is a semi-structured or textual description field, but also whether it is a field that generally contains references for locations/places, persons or time periods.
- For most if not all application cases in the Europeana context, concepts so general as "general period" do not bring any value as enrichment targets. It could be relevant to include additional logic to the enrichment rules so that they are not used to enrich objects.
- Enrichment tools seeking matches on parts of a field's textual content may result in too general enrichments or even meaningless ones if they miss to recognize compound expressions[85]. This especially hurts when the target datasets include

[85] This is the case for example of enrichment that recognize http://dbpedia.org/resource/Cf. or the general concept of Library for specific (named) libraries.

resources of a very general nature, which are less relevant for the application needs.

- Apply a strong resolution and disambiguation mechanism that considers the accuracy of the name reference in a record together with the relevance of the entity both in general (looking as its data properties) and in particular, i.e., within the context it is being referred (which requires determining the correct context of its use). For example, we observed that most enrichment tools could be improved if they determine the temporal scope of the records and compare it to the temporal scope of the enriched entities.
- Quality issues originated in the mapping process had been already identified as a great obstacle to get good enrichments, in the 2014 report of the Task Force on Multilingual and Semantic Enrichment Strategy. Our evaluation has confirmed it. Semantic enrichment rules aimed at specific metadata fields (e.g., spatial coverage of an object) should be designed and applied carefully to the source datasets, in case the fields would be populated with values that result from wrong mappings (e.g., publication places).

7. Conclusions

The report of the EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy concluded that there isn't one solution fitting all cases, and that the Europeana Network should facilitate the emergence of target datasets that fit enrichment needs and enrichment workflows that successfully exploit them. The work of our Task Force expands and refines this general recommendation. The implementation of a successful enrichment strategy makes it necessary to plan efforts and choose components of the enrichment workflows wisely. Workflows need to be created, which utilise enrichment tools and/or services to accomplish certain enrichment tasks. The following list is a digest of the more fine-grained recommendations of the previous sections. It details ten steps for developing and maintaining a successful enrichment strategy:

1. **Define your enrichment goals (annotation guidelines) that will guide your enrichment strategy.**

At the beginning of these reflections you should define the enrichment goals. They define the “competence questions” that you want enrichment to facilitate, they are driving the decision for each component of the enrichment process, and determine the criteria for evaluation.

2. **Choose the right components for your enrichment workflow: enrichment solution and target datasets.**

Choosing the right enrichment service is crucial to achieve the goals set by the enrichment strategy. For example, the chosen enrichment services should provide an interface so that different enrichment steps can be piped one after the other. Ideally,
each enrichment service or tool recognizes a specific metadata schema (schema/semantics-aware) and can apply different enrichment rules depending on the metadata fields subjected to enrichment, but still keep in mind that wrong mappings or partial match strategies may undermine them (see Section 6 for more detailed recommendations). In general, one should aim at keeping the enrichment service simple and manageable. The selection of the target dataset that the enrichment service will use is also an important criteria for obtaining good quality enrichment. The selected target needs to cover the goals identified as part of the enrichment strategy while remaining close enough to the source's metadata (e.g., with comparable terminology, as enrichment tools always exploit some form of NLP).

3. **Define the enrichment workflow.**

It is essential to set the enrichment workflow and make sure which actions will be taken in which order. For example, enrichers need to apply enrichment services in the most suitable order, like firstly, normalising the value of a dcterms:spatial field and then apply geo-coding. One should avoid to pipeline two services or tools that are doing the same thing (e.g. two geocoding services). In this step, the rules for the enrichment need to be defined and documented. Consider including logic to the enrichment workflow so that less valuable enrichment targets are not used to enrich objects (see Section 6 for more detail).

4. **Make sure your enricher has sufficient knowledge.**

Enrichers that are called upon to design a enrichment workflow must have a clear understanding of how each enrichment tool works and which parts of the data each enrichment service affects. The training should also cover the assessment of results and potential problems which might occur, e.g. the ambiguity of place names.

5. **Test your enrichment workflow.**

Once the initial enrichment workflow is set, it needs to be tested to see if adjustments need to be made. Tools that allow short feedback loops are priceless in this respect.

6. **Assess the quality of your enrichment and have an evaluation strategy.**

This goes hand in hand with the desired goals of the enrichment. These goals are likely to define what a high quality enrichment is for your strategy. Only if you know what the best possible outcome of the process is, you can evaluate how well the process worked. You should have methods in place which determine how well the tool(s) you used and the enrichment strategy worked. Here, you should not only limit the evaluation to intrinsic evaluation but also use extrinsic evaluation methods which focus more on the user satisfaction.

7. **Choose the right evaluation method.**

One point is to choose an evaluation method that suits the goal of the enrichment. If you enrich for better search performance, you should consider an evaluation of the search performance. Additionally, the chosen method should be sound, resulting in statistically strong and representative evidence: the evaluation dataset should represent the diversity of your data, and balance between tools should be ensured while performing comparative evaluations. In case of creating gold standards or annotated samples, it is strongly advised to compute inter-rater agreements but also remember to make clear guidelines for raters and train them if necessary (see Section 6 for more detailed recommendations). Using an appropriate (collaborative) annotation tool will help, but not many tools can currently be applied out-of-the-box.

8. **Apply user-initiated enrichment workflows.**

User initiated enrichment workflows should be applied to specific datasets (e.g. datasets that are coming from a specific source or system). Usually such workflows aim at fixing/enrichment flaws that are system specific. In order for the user to design an enrichment workflow that targets specific flaws, these flaws must be first identified. Statistics/Quality on ingested datasets is a great source of information.

9. **Document your enrichment process and learnings.**

Since datasets are in constant change and evolution, the lessons learned and evaluation conclusions from the enrichment process, will prove a valuable contribution for future enrichment strategies and maintenance of an enrichment process.

10. **Monitor your enrichment process and re-assess.**

Components of the enrichment process are often subject to change over time. The source data might have been expanded either by adding new digital objects or by added more information to existing objects. Also new terms might have been added to the target datasets. There might also be changes in the workflow itself as new sophisticated rules might have been applied. For all these cases, it is necessary to re-asses the process and adjust it accordingly**.**

References

(Dangerfield et.al, 2015) Dangerfield, MC. et. al (2015): Report and Recommendations from the Task Force on Metadata Quality. Retrieved October 19, 2015 from http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata Quality Report.pdf

(Ferro, 2011) Ferro, Nicola. "DIRECT: the First Prototype of the PROMISE Evaluation Infrastructure for Information Retrieval Experimental Evaluation." *ERCIM News* 86 (2011): 54-55.

(Freire et al. 2008) N. Freire, J. Borbinha, B. Martins. *Consolidation of References to Persons in Bibliographic Databases*. ICADL 2008 – The 11th International Conference on Asian Digital Libraries, 2008. http://dx.doi.org/10.1007/978-3-540-89533-6_26

(Freire et al. 2011) N. Freire, J. Borbinha, P. Calado, B. Martins. *A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records*. ACM/IEEE Joint Conference on Digital Libraries, 2011. http://dx.doi.org/10.1145/1998076.1998140

(Freire and Juffinger 2011) N. Freire, A. Juffinger. *Using National Bibliographies for Rights Clearance*. ACM/IEEE Joint Conference on Digital Libraries, 2011. http://dx.doi.org/10.1145/1998076.1998109

(Kelly, 2009) Diane Kelly (2009): Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1—2 (January 2009), 1-224. DOI=10.1561/1500000012 http://dx.doi.org/10.1561/1500000012

(Olensky, Stiller, and Dröge, 2012) Olensky, M, Stiller, J, and Dröge, E. (2012): Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. In: Metadata and Semantics Research, ed. by Dodero, J., Palomo-, Duarte, M., Karampiperis, P., Springer, Berlin.

(Stiller, Isaac & Petras (eds.), 2014) (Stiller, Isaac &Petras (eds.) (2014): EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report. Retrieved October 19, 2015 from http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/MultilingualSemanticEnrichment//Multilingual%20Semantic%20Enrichment%20report.pdf

(Stiller et al., 2014) J Stiller, V Petras, M Gäde, A Isaac: Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. p. 238-247

(Usbeck, Ricardo, et al., 2015) Usbeck, Ricardo, et al. "GERBIL: General Entity Annotator Benchmarking Framework." *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.available at: http://svn.aksw.org/papers/2015/WWW_GERBIL/public.pdf http://aksw.org/Projects/GERBIL.html

(Wielemaker et al., 2011) J. Wielemaker, V. de Boer, A. Isaac, J. van Ossenbruggen, M. Hildebrand, G. Schreiber, S. Hennicke. *Semantic workflow tool available*. EuropeanaConnect Deliverable D1.3.1. October 2011. http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaConnect/Deliverables/ECONNECT-D1.3.1-Semantic%20Workflow%20Automation%20Method%20Implementation.pdf

Appendix

A. Inventory of enrichment efforts and tools initiated in the context of the Europeana Network

The Appendix A exists a separate document available on the Task Force page.

B. Past evaluations in the Europeana community

Information Retrieval evaluation campaigns with Europeana Data

- several CLEF campaigns
- Humboldt 2014 translation evaluation in context of the project GALATEAS
- CHiC - CH in CLEF
    - 2013: http://link.springer.com/chapter/10.1007%2F978-3-642-40802-1_23
    - 2012: http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-CHiC-PetrasEt2012.pdf
    - 2011: http://www.promise-noe.eu/chic-2011/home

Evaluation of Enrichments

- MTSR2012 - Poisonous India or The Importance of a Semantic and Multilingual Enrichment Strategy
    - http://link.springer.com/chapter/10.1007%2F978-3-642-35233-1_25
- MTSR2014 - A Framework for the Evaluation of Automatic Metadata Enrichments
    - http://link.springer.com/chapter/10.1007/978-3-319-13674-5_23
- EuroMed2014 - Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences
    - http://link.springer.com/chapter/10.1007%2F978-3-319-13695-0_23
- EuropeanaTech: Semantic Enrichment Evaluation
    - http://pro.europeana.eu/documents/468623/8b75b054-712e-432b-a0f7-761898e6f6

Usability Evaluation

- Europeana Interface & Interaction Evaluation
    - Evaluation Report of the Usability of the Europeana Website http://pro.europeana.eu/c/document_library/get_file?uuid=ae1d74de-29c1-463c-887e-a6bc6ee0ed7a&groupId=10602
    - User Centric Evaluation of the Europeana Digital Library http://link.springer.com/chapter/10.1007%2F978-3-642-13654-2_19

    - Multilingual Access to Digital Libraries: The Europeana Use Case http://www.degruyter.com/dg/viewarticle/j$002fiwp.2013.64.issue-2-3$002fiwp-2013-0014$002fiwp-2013-0014.xml
    - Cross-lingual information retrieval and semantic interoperability for cultural heritage repositories http://www.aclweb.org/anthology/R13-1063
- General
    - Evaluating Cultural Heritage Information Access Systems
        - http://link.springer.com/chapter/10.1007/978-3-642-54347-0_2
    - Building for Success? Evaluating Digital Libraries in the Cultural Heritage Domain
        - http://www.igi-global.com/chapter/building-success-evaluating-digital-libraries/74327

Evaluation of query translation

- Evaluation of a 250 query corpus in English, French and German performed within the Galateas project:
    - Documentation of Creation of Gold Standard from Europeana Query Corpus
    - Query corpus in English: file:English_corpus_Europeana.xml
    - Query corpus in French: file:French_corpus_Europeana.xml
    - Query corpus in German: file:German_corpus_Europeana.xml
    - Evaluation of Query Translation in Europeana: file:Auswertung_evaluation.pdf
    - D7.4 - Final Evaluation of Query Translation: file:GALATEAS_D7_4.pdf
- Evaluation using the Portal (done using the same corpus as in the Galateas Project):
    - Basecamp thread
    - Result of the 250 query translations from Europeana
    - Result of the 50 query translations from Europeana

Evaluation of search

- WebWeaving's "Europeana vs. Google evaluation" by Dirk-Willem van Gulik, Ardy Siegert (Antoine will fill later)
- A Use Case Framework for Information Access Evaluation
    - http://link.springer.com/chapter/10.1007%2F978-3-319-12511-4_2
    - Private copy at Dropbox

Europeana Logfiles data and work that uses them

- List of crawlers: file:crawlers.txt

- Europeana Logs page at Labs
- Projects that have used Europeana log files:
    - The CLEF initiative
    - The PATHS project: see more documentation of the project
    - ASSETS - Advanced Service Search and Enhancing Technological Solutions: see the related report
        - D2.2.1 Specification of post querying processing functionalities
    - EuropeanaConnect for analysing users' attitudes and needs and opens new ways of discovering cultural heritage in Europeana: see the results
    - Galateas project
- Information on Europeana logging
    - Log actions: file:Europeana_Logs_Actions2010_2011.pdf
    - Session clickstreams: file:Europeana_Session_ClickStream.pdf
- Europeana 2012-2013: usage and performance update
    - http://ciber-research.eu/download/20130623-Europeana_2013_usage_and_performance_update.pdf
- IliX2014 - Multilingual Interface Preferences
    - http://dl.acm.org/citation.cfm?id=2637002.2637030
    - Private copy at Dropbox
- Maria Gäde's PhD Thesis - Country and language level differences in multilingual digital libraries
    - http://edoc.hu-berlin.de/docviews/abstract.php?id=40595
    - http://edoc.hu-berlin.de/dissertationen/gaede-maria-2014-02-05/PDF/gaede.pdf
- Improving Europeana Search Experience Using Query Logs
    - http://link.springer.com/chapter/10.1007%2F978-3-642-24469-8_39
    - http://miles.isti.cnr.it/~nardini/wp-content/uploads/2011/06/tpdl2011.pdf
- Which Log for Which Information? Gathering Multilingual Data from Different Log File Types
    - http://link.springer.com/chapter/10.1007%2F978-3-642-15998-5_9
    - http://www.europeanaconnect.eu/documents/Gaede_Petras_Stiller2010.pdf

Metadata Exchange

- Implementing Enhanced OAI-PMH Requirements for Europeana
    - http://link.springer.com/chapter/10.1007%2F978-3-642-24469-8_40

Recommendation systems

- Implementing Recommendations in the PATHS System
    - http://link.springer.com/chapter/10.1007%2F978-3-319-08425-1_17

C. Relevant APIs for enrichment

In the course of the task force work, members submitted pointers to relevant APIs. These are not only providing semantic enrichment services, but all of them are relevant in the sense that could be used as part of an enrichment workflow.

LoCloud APIs

The LoCloud microservices APIs can be found at and they can be found at http://support.locloud.eu/tiki-index.php?page=Technical%20Documentation

Part of the MORe API, the MORe enrichment API allows to apply an enrichment plan on a dataset: http://support.locloud.eu/MORE%20Technical%20Documentation

Terminology APIs

Tilde's "Terminology Annotation Showcase" uses the standard W3C ITS 2.0 (see Section 2.2)

http://taws.tilde.com/

Tilde also provide a generic Terminology API[86], being advanced within the EU project FREME:

https://term.tilde.com/

NIF enrichment APIs

NIF offers a simple API for NIF tool invocation, both command-line and as REST service: http://persistence.uni-leipzig.org/nlp2rdf/specification/api.html.
The NIF Combinator can combine NLP components by calling them in order and combining the results.
http://svn.aksw.org/papers/2012/NIF/EKAW_DEMO/public_preprint.pdf
There are other ways of composing NLP processing pipelines using specific software libraries:

- GATE Creole: https://gate.ac.uk/sale/tao/splitch4.html
- IBM UIMA: http://www-01.ibm.com/software/ecm/content-analytics/uima.html
- Stanbol Enhancement engine:
    - REST API: http://stanbol.apache.org/docs/trunk/components/enhancer/enhancerrest.html
    - Developing and composing: http://blog.iks-project.eu/getting-started-with-apache-stanbol-enhancement-engine/

[86] Full description available at https://term.tilde.com/Content/api_spec.pdf