Comparative evaluation of semantic enrichments

29/10/2015

Editors
Antoine Isaac, Hugo Manguinhas, Valentine Charles (Europeana Foundation R&D) and Juliane Stiller

Co-financed by the European Union
Connecting Europe Facility

1. Introduction

In order to gain better insight over the quality of enrichment tools and in particular on the methods and metrics to evaluate them, the Task Force members have undertaken an evaluation campaign. This document explains the phases of the evaluation, covering the methodology used, the results obtained and their analysis. It concludes with a summary of the recommendations and the lessons learned.

2. Dataset chosen for the evaluation

For this evaluation, we need a dataset that represents the diversity of the data in Europeana. For this reason, we looked at The European Library[1] (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 allowing the gathering of an evaluation dataset with metadata of varied countries, languages, as well as heterogeneous metadata practices and formats.

We have selected an evaluation dataset that contains metadata records using the Europeana Data Model[2] (EDM) delivered by TEL to Europeana from all 19 countries[3] that contribute to TEL. We first removed some collections (newspapers) where the metadata had been mostly "artificially" generated during the digitization process (e.g., issues of journals that are described by merely appending the journal's title with the number of the issue). We then selected a maximum of 1000 metadata records for each country. When more than 1000 records were available for a country, we did a random selection. In order to have heterogeneous data within the evaluation dataset, we partitioned these larger datasets in 1000 sequential parts, and blindly selected one record from each partition. In total the evaluation dataset[4] contains 17.300 metadata records in 10 languages[5] (based on the language tags on literals and not on the field edm:language). Figure 1 lists the properties that were found within the evaluation dataset ordered from the most frequent to the least.

[1] http://www.theeuropeanlibrary.org/
[2] http://pro.europeana.eu/edm-documentation
[3] Austria, Belgium, Bulgaria, Czech Republic, Finland, France, Germany, Ireland, Latvia, Luxembourg, Netherlands, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain and United Kingdom
[4] see the "dataset" folder under the Task Force resource archive in the following Assembla space: https://www.assembla.com/spaces/europeana-r-d/documents?folder=58725383
[5] English, German, French, Polish, Portuguese, Serbian, Latin, Italian, Dutch, and Spanish (from the most frequent to the least)

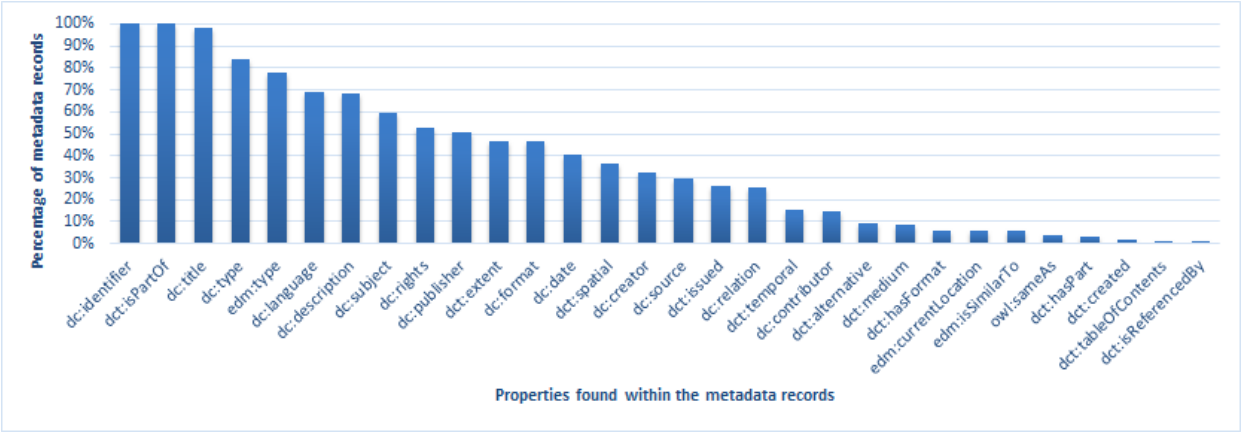

Figure 1: Frequency of properties found within the evaluation dataset.

In spite of our efforts, however, one may say the dataset reflects a specific perspective. The European Library has a strong focus on library and research material, including many scientific articles on mathematics, biology, agriculture, etc. These are particular cultural heritage objects, and general enrichment tools can miss the specific scientific terminology present within the metadata descriptions of the objects. For example an article mentions varieties of wheat called "Pliska, Sreca, Vila, Holly". But Pliska[6] is also the name of both the first capital of the First Bulgarian Empire and a small town in Bulgaria. When a scientist uses known proper names to name newly discovered things, the concept extractor will miss it and return the most common meaning. However, a similar problem would very probably appear for other domains in other collections. A key problem for semantic enrichment in Europeana is how to tackle all the different domains it includes.

Another issue can be illustrated by the following example: the description of an article[7] indicates that a researcher has used ethanol in her study. The enrichment tool recognized “ethanol”, but it is only marginally relevant, as it is a mere technical means and not the general object of the study described in the paper. In the evaluation, we should thus aim at judging the relevance of an enrichment as opposed to a bare “correctness”, which requires specific evaluation methods to be defined and implemented (see the discussion on evaluation criteria the section "Methods and metrics for evaluation" in the Task Force's main report).

3. Enrichment results obtained from the participants

For this evaluation, we called on participants of the Task Force that use and/or develop enrichment tools, to apply them to the evaluation dataset. The following participants have

[6] http://dbpedia.org/page/Pliska
[7] http://data.theeuropeanlibrary.org/BibliographicResource/1000095953660

answered: the **Europeana Foundation (EF)** with the current version of the semantic enrichment framework; **The European Library (TEL)** with the TEL enrichment service; the LoCloud project contributed two tools, the English version of the Background Link service (**BgLinks**) and the Vocabulary Match service (**VocMatch**); the **Pelagios** project used their NER and geo-resolution tool; and **Ontotext** used two different settings of Ontotext's concept extractor.

3.1. Tools and settings used for the evaluation

The **Europeana Foundation,** as part of its aggregation workflow, applies semantic enrichment to EDM properties that may refer to contextual entities such as Places, Agents, Time Spans, and Concepts. This is done by the Semantic Enrichment Framework[8]. Depending on the kind of contextual entity that may be present within a metadata property, it applies different enrichment rules to link to different target datasets. Before the enrichment takes place, the target datasets are identified, carefully selected and the data is mapped into the appropriate classes and properties of EDM. Currently, it uses DBPedia[9] for both agents and concepts, Geonames[10] for places, and Semium Time for time spans. The rules for matching between the source metadata and target vocabulary boil down to strict string matching between the text within the metadata field and the labels defined in the vocabulary, with some pre-processing for some fields.

The European Library applies semantic enrichment as part of its aggregation workflow to metadata properties that may refer to Places and Agents. It also applies different enrichment rules for different target datasets. For places, enrichment is done by an in-house (dictionary-based) Named Entity Recognition process combined with entity resolution and using Geonames as target vocabulary, which were developed in the EuropeanaConnect project[11][12]. When several candidate places are present, the decision process uses a heuristic based mainly on the name match, type of geographic feature, population, and the origin of the metadata record. For agents, enrichment is done by conversion of internal authority files from the data providers into corresponding Semantic Web URIs of the GND (Gemeinsamen Normdatei - Germany) Integrated Authority File[13] using coreference information.

[8] https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y
[9] http://wiki.dbpedia.org/
[10] http://www.geonames.org/
[11] 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
[12] Charles, V., Freire, N., Antoine, I., 2014, 'Links, languages and semantics: linked data approaches in The European Library and Europeana', in 'Linked Data in Libraries: Let's make it happen!' IFLA 2014 Satellite Meeting on Linked Data in Libraries.
[13] http://www.dnb.de/EN/Standardisierung/GND/gnd.html

The Background Link (**BgLink**) service is one of the cloud services developed within the LoCloud project. It analyzes the textual elements of metadata (title, description, subject, etc), and links named references within the data to DBpedia resources. It uses DBpedia Spotlight[14] v0.6 as a backbone (statistical back-end). DBpedia Spotlight is an annotation tool that links mentions in a text to DBpedia. It follows the standard NERD steps, namely detecting the mentions occurring in text, and disambiguating each one by assigning a unique DBpedia concept. Unlike NERD systems, where only named entities are considered, Dbpedia Spotlight tries to disambiguate any mention occurring in the text. The statistical back-end is a supervised approach that learns how to disambiguate textual mentions using a bag-of-words type of features, and the model is trained from Wikipedia. BgLink uses the FSAspotter[15] (a finite state automata) for mention detection, and a bayesian model for disambiguation (instead of tf-idf scores[16]). An important feature is that it enriches exclusively the most relevant terms in the text. Note that this means that it does not disambiguate all the terms it finds, and that the disambiguated terms are often more ambiguous than the average. The service is deployed in two versions, one for English[17] and another for Spanish[18]. Both BgLinks and the VocMatch service described below are further documented on the LoCloud site[19].

The Vocabulary Matching (**VocMatch**) service is another cloud service developed under the LoCloud project to automatically assign relevant to items concepts and terms from selected SKOS vocabularies. Those are developed on top of the TemaTres tool[20], an open source server to manage and exploit vocabularies - thesauri, taxonomies and formal representations of knowledge. The version 2.0 of TemaTres has been integrated with the LoCloud test lab where it can be accessed via its portal[21] or the test platform of the microservices[22]. The SKOS vocabularies currently used by VocMatch can be accessed at the documentation page[23] of the project or browsed directly from its portal.

The **Pelagios** 3 project[24] enriches text and map image objects with gazetteer URIs. The workflow is semi-automatic and supported by the Recogito open source tool[25]. It differs whether it is applied on plain text or an image, but in general is composed of NERD and user

[14] http://spotlight.dbpedia.org/
[15] Available as part of DBpedia Spotlight bundle, available at: https://github.com/dbpedia-spotlight
[16] https://en.wikipedia.org/wiki/Tf%E2%80%93idf
[17] http://test183.ait.co.at/rest/bglink
[18] http://lc013.ait.co.at/rest/bglink
[19] http://support.locloud.eu/Metadata%20enrichment%20API%20technical%20documentation
[20] http://www.vocabularyserver.com/
[21] http://test113.ait.co.at/tematres/unesco/index.php
[22] http://lc004.ait.co.at:8080/portal/site/wp3
[23] http://vocabulary.locloud.eu/?p=36
[24] http://pelagios-project.blogspot.nl/2013/09/pelagios-3-overview.html
[25] http://pelagios.org/recogito

interfaces for verification and correction. The NERD recognition step is implemented using the Stanford NLP Toolkit[26] while disambiguation uses gazetteers. Different gazetteers can be plugged on the tools, which can therefore enrich against multiple aligned gazetteers such as Pleiades (see companion document on "Selecting target datasets for semantic enrichment"), the Digital Atlas of the Roman Empire[27] and the Archaeological Atlas of Antiquity[28].

The **Ontotext** Semantic Platform consists of a suite for text mining and semantic annotation, data integration tools to transform data into RDF and tools for semantic curation. As part of this suite, Ontotext offers a tool for Semantic Enrichment[29], which as other evaluated tools, follows a standard NERD approach. It applies a text processing pipeline that uses GATE[30]. The pipeline includes various Natural Language Processing (NLP) components: tokenization, part-of-speech (POS) tagging and chunking. Rules are written using GATE's Java Annotation Patterns Engine (JAPE)[31]. It uses as target dataset an in-house repository called MediaGraph[32], which integrates data from both DBpedia and Wikidata. The tool can enrich terms in English against any kind of concepts, but is limited to Persons and Places for terms in other languages.

The following table explains the settings that each tool applied for this evaluation:

Tool	Metadata fields	Type of Entity	Target Vocabulary	Rules or methods applied
EF	dc:subject, dc:type, dct:spatial, dc:coverage, dc:contributor, dc:creator, dc:date, edm:year, dct:temporal	Places, Agents, Concepts, Time Spans	**Agents:** Selection of DBpedia; **Concepts:** Selection of DBpedia and GEMET; **Places:** Selection of Geonames for places within Europe; **Time Spans:** SemiumTime.	Strict string matching between the text within the metadata field and the preferred plus alternative labels of the target vocabulary. Some pre-processing of the labels and text may occur depending on the kind of entity.
TEL	dc:subject, dct:spatial, dc:creator, dc:contributor	Places, Agents	**Places:** Geonames; **Agents:** GND Integrated Authority File.	**Places**: NERD using heuristics based mainly on the name match, type of geographic feature, population, and the origin of the metadata record. **Agents**: conversion of internal authority files from the data providers into GND URIs using

[26] http://nlp.stanford.edu/software/corenlp.shtml
[27] http://darmc.harvard.edu/icb/icb.do
[28] http://www.vici.org/
[29] http://ontotext.com/products/ontotext-semantic-platform/semantic-enrichment-and-text-mining/
[30] https://gate.ac.uk/
[31] https://gate.ac.uk/sale/tao/splitch8.html
[32] http://mediagraph.ontotext.com/

				coreferencing.
BgLinks	All fields	All types	English DBpedia	Based on DBpedia Spotlight 0.6. **Note**: only English terms are annotated since the statistical model was trained for English.
VocMatch	All fields	Concepts	About 30 vocabularies[33] curated in the LoCloud project. **Note:** Vocabulary terms are referred by means of an internal Locloud URI.	String match between the source text and the preferred and alternative labels of the target vocabulary.
Pelagios	dc:title, dc:description, dc:source, dc:publisher, dct:spatial and dc:subject.	Places	Wikidata was chosen since it contains contemporary places, since the other supported Gazetteers were found to be too focused on the ancient world.	The metadata field was subject of either: (a) NER and then geo-resolution, or (b) the entire field was used as a search term for geo-resolution, without prior NER. For the title, description, source and publisher (a) was applied and (b) for the remainder.
Ontotext v1 & v2	All fields	All types[34] for English, Persons and Places for other languages.	DBpedia & Wikidata, integrated as a dataset called "MediaGraph".	**Version 1:** used the record language for determining the language of text. **Version 2:** used the language tag of individual literals.

3.2. Data received

Each participant was asked to send the results from their enrichment tools in a basic predetermined format containing:

- the identifier of the metadata record (URI of the edm:ProvidedCHO);
- the qualified name of the property (e.g., dcterms:spatial);
- the identifier of the entity (e.g., URI of a DBPedia resource);
- a floating point value for the confidence, from 0 (certainty) to 1 (certain), or empty if the confidence is unknown (eventually this value was not taken into account in the evaluation as only Ontotext and BgLinks were able to output it);
- the source of the enrichment, i.e. the literal (word or expression) where the entity was identified.

[33] http://vocabulary.locloud.eu/?page_id=2
[34] The results from Ontotext contained an extra column to indicate the type of entity that was emitted.

We chose CSV as format as it was considered to be the easiest to generate and to process. Below is an example of an enrichment from one of the participants:

```
http://data.theeuropeanlibrary.org/BibliographicResource/2000085482942;dcterms:spatial
;http://dbpedia.org/resource/Prague;0.9;Praha
```

A total of about 360k enrichments were obtained from 7 different tools or tool settings. Figure 2 shows an indication of the metadata record coverage for each tool and Figure 3 shows the distribution of enrichments across tools. Note while interpreting Figure 3, that tools such as VocMatch, Ontotext and BgLinks output an enrichment for each time a term is present in the source metadata field which helps explain the high number of enrichments. Additionally, and since VocMatch enriches against multiple vocabularies, it also outputs one enrichment for each term found across all target vocabularies, unlike other tools that always return the best candidate.

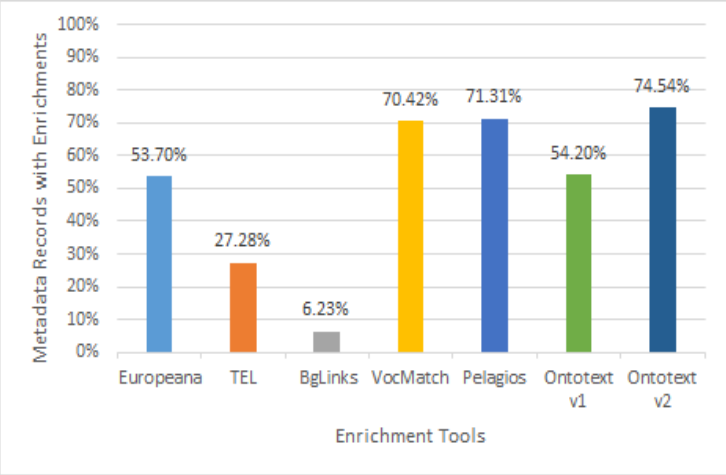

Figure 2: Frequency of enriched records for each enrichment tool.

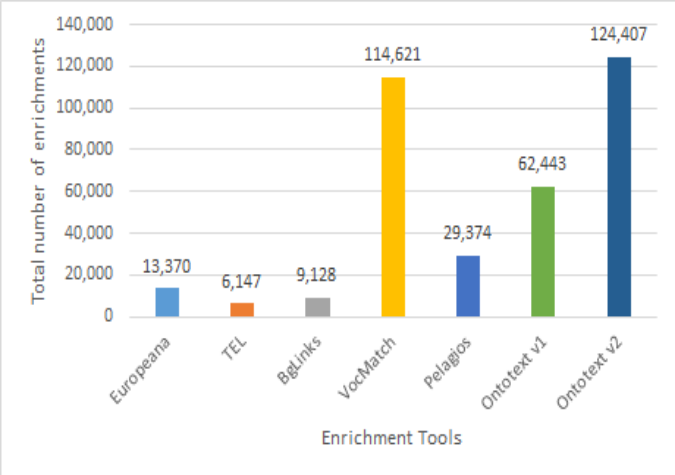

Figure 3: Number of enrichments by enrichment tool.

A preliminary analysis of the enrichments confirmed that the tools enriching against different target datasets, as seen in Figure 4. A closer look, shows that only two of the datasets (DBpedia and Geonames) were re-used by more than one tool (note that only Ontotext tools produce MediaGraph enrichments). In Appendix A, a table with the distribution of enrichments based on their source is presented, while a more complete table combining both distributions can be access on the Assembla page of the Task Force[35].

[35] See folder "enrichments/stats" on the Assembla page of the Task Force: https://www.assembla.com/spaces/europeana-r-d/documents?folder=58725383.

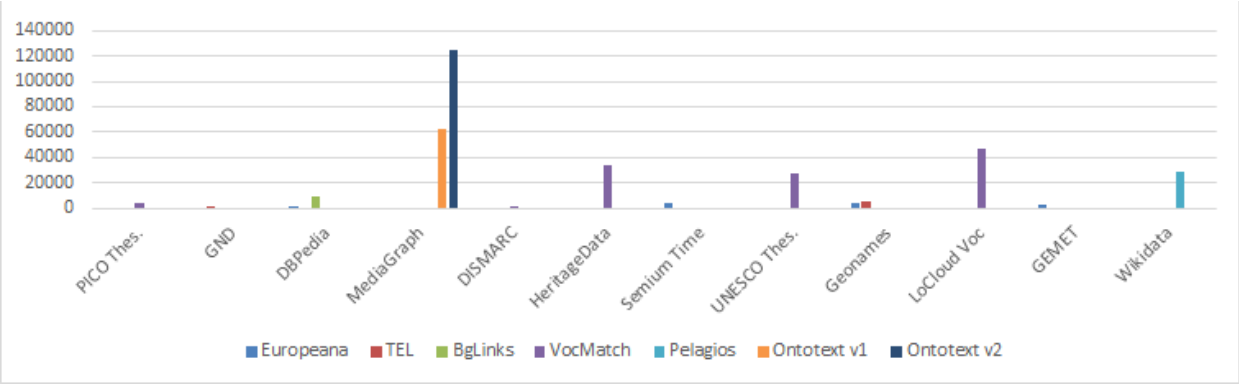

Figure 4: Overview of the target datasets that were referred by the enrichments.

4. Creating the annotated corpus

Assessing the performance of enrichment tools often involves comparing their results with a "gold standard", i.e. a reference set of correct object annotations. In our case creating a gold standard would require too much effort. Our evaluation dataset is big and enrichment tools use different target datasets. One would have to look at every object and create enrichments to each target dataset. Instead of creating a true gold standard, we have tried to focus on assessing and comparing the enrichments as they are produced by the tools, after a "normalization" step where we recognize enrichments to different targets that have the same semantics.

NB: as explained in the subsections below, our choice has important methodological drawbacks. We felt however it was preferable to get at least some comparative assessment of the enrichment tools than none. We also hope the lessons learned in this process will help design more accurate and fair evaluations in the future.

4.1. "Normalization" of the enrichments

Variation on the target vocabulary brings yet another level of complexity for a comparative evaluation, as it hides cases where tools actually agree on the semantic level, i.e. they enrich with the same contextual resource (concept, place, person, time period). We have tried to palliate this by "normalizing" the target of enrichment links into a common (canonical) dataset. The idea is to take advantage of existing coreference links between the original target dataset and this common dataset, so that original enrichment results are "re-interpreted" as referring to a resource within the common target dataset.

The vocabularies selected as common datasets were Geonames for places and DBpedia for the remaining resources. These two datasets were chosen as they were the ones that

benefited from the highest agreement level. Even though we were able to automatically exploit coreference links for a significant portion of the results, it was not always possible due to the lack of coreference links for some target vocabularies. The most prominent case was VocMatch that linked to the PICO thesaurus, the LoCloud vocabulary, the UNESCO thesaurus and HeritageData, which do not have any coreference links.

The following enrichment result sets were "normalized": for **Pelagios**, we were able to exploit the "wikidata:P1566c"[36] links from Wikidata to Geonames, while for DBpedia we exploited the owl:sameAs links from DBpedia to Wikidata; for **Ontotext** we were able to exploit the "tgsi:exactMatch" links to DBpedia from the MediaGraph data. We realized that both Ontotext and BgLinks could have been coreferenced for places, but this was only later in the evaluation and would have required us to redo too much work.

Lessons Learned

Being able to *fully* compare enrichment services requires that they use comparable target datasets. Unless they can be configured or re-developed to do this, evaluators should explore exploiting coreferences between the target datasets. For this purpose, some of the criteria for selecting target datasets (see companion document on "Selecting target datasets for semantic enrichment") will be even more important. A dataset with coreference links to other datasets is a key asset. Dereferenceable URIs also ease the process of accessing and exploiting coreference data. Conceptual coverage and granularity should facilitate the coreferencing of as many (right) resources as possible (e.g., the DBpedia resource for Paris is coreferenced to two Geonames resources, for respectively the capital and the city).

Note that we do not recommend that targets should only be (mappable to) DBpedia or Geonames. The report on the past EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy has already found that Europeana needs a diversity of enrichments to match the needs of a diverse community of data providers and aggregators. Enrichments to "local" vocabularies remain relevant, especially when they are technically and conceptually compatible with the later establishment of a rich semantic network across institutions, languages and domains, as is the case for the Heritage Data[37] linked data vocabularies used by VocMatch. Enrichments with such targets should be assessed according to the specific applications or institutional contexts that motivate them, even though it will be of course much harder then to compare services.

[36] https://www.wikidata.org/wiki/Property:P1566
[37] http://www.heritagedata.org/

4.2. Building the sample corpus

The next step was the sampling of the evaluation dataset. The enrichments (after normalization) were compared across tools to recognize sets shared by several tools and sets specific to one tool (two enrichments are considered equal if they have the same source and same target). The purpose was to make sure we evaluate enrichments that represent appropriately the variety of enrichments produced by the evaluated services, in all their differences and commonalities. This would also help us to identify similarities that could indicate a same logic being shared across enrichment services (such as using the same rule or same specific data from the target dataset). This resulted in a total of 26 different sets (shown in Table 1) that reflect the agreement combinations between the tools. For each set, at most 100 enrichments were randomly selected (if the set contained less than 100, all enrichments were selected) to be part of the sample corpus, for a total of 1757 distinct enrichments. Note that the sets are not necessarily disjoint, as they were built based on different "agreement configurations", not following a clustering process. However, we have removed from the evaluation the sets that happened to include exactly the same enrichments (for example, #8 to #11). As can be observed in Table 1, VocMatch does not share any enrichment with any other tool as there are no coreference links between its targets and other services.

Table 1: Sets obtained from the comparison made between the enrichments from each tool.

Sample	Europeana	TEL	BgLinks	VocMatch	Pelagios	Ontotext v1	Ontotext v2	Set Size
#02	A		A			A	A	12
#03			A		A	A	A	10
#04			A			A	A	6082
#05	A	A			A			905
#06	A					A	A	68
#07					A	A	A	22
#12						A	A	62330
#13			A			A		6142
#14			A				A	6085
#15	A	A						3101
#16		A			A			1283
#17	A						A	1085
#18	A				A			1053
#19					A		A	831
#22			A		A			16
#23	A		A					15
#24							N/A	118257
#25				N/A				114621
#26						N/A		56293
#27					N/A			28447
#28	N/A							12397
#29		N/A						5242
#30			N/A					3046
	N/A	Enrichments produced by one service only (for the given column)						
	A	Enrichments shared by several tools (across the selected columns)						

Lessons Learned

The number of sampled enrichments is not even across tools (see Table 5 below). Some systems are over-sampled (Ontotext, Europeana and Pelagios) and others down-sampled. There is a specific bias against the tool that does not share results with others, VocMatch,

which reinforces the bias from the "normalization" step. Although it does not change the individual assessment of each tool, this heavily impacts the *comparative* aspect of our evaluation, as will be shown later. For future evaluations, a more balanced selection could be applied. Computing pooled recall (see Section 6) could use different weights for the results coming from the different sets, so as to reflect the (assumed) quantity of correct enrichments from each tool. Entirely different selection approaches could also be followed, such as sampling based on the distribution of enrichments over the target datasets, or the properties of the source dataset. One could also only sample over objects that were enriched by all tools. This could however introduce another bias, as these objects' metadata may have very specific features.

4.3. Annotating

To guide raters with the evaluation task, a set of guidelines was created. The first version of the guidelines contained only three evaluation criteria: semantic correctness, general completeness (combining both name and concept completeness), and informational value. This first version was then tested by three participants, starting with the first 6 rows of the annotated corpus, to assess whether the criteria were relevant and easy to apply. This resulted in a revision of the guidelines by both changing some of the criteria and adding clear examples on how enrichments should be annotated. In particular, the completeness criteria was split into name and concept completeness. Informational value was theoretically relevant, but was found to be too subjective and would have required to further train the raters, taking more time and effort we could afford. The final version of the guidelines contains three criteria (Semantical correctness, Completeness of name match and Completeness of concept match) and can be seen in Appendix B.

To perform the annotation tasks, two online spreadsheets were created, one for each corpus. For this evaluation, online spreadsheets were considered to be: (a) easiest to use; (b) simplest to implement; (c) easily customizable/adaptable; (d) with the ability to collaboratively work with other raters. Both spreadsheets contained the following columns: a link to the record (URI), its title, the property which was the source of the enrichment together with its value; the portion of the text that was considered by the tool; link to the entity (URI); three columns reflecting the three criteria; and a column for the rater to place his comments.

Since the main evaluation corpus was built from a selection of the sets, the set organization was preserved in the spreadsheet as separate sheets. Generally, one sheet was assigned to each rater. For the second corpus, the spreadsheet had one sheet for each rater with an exact copy of the corpus. After both spreadsheets had been created, 16 participants were

asked to annotate them following the guidelines defined within the Task Force. This task was completed in two weeks.

Lessons Learned

Further, even though the online spreadsheet met most of our initial evaluation requirements, it was not perfect. When performing the annotation task, most of raters missed either one or more enrichments, made annotations using assessment codes (see Appendix B) unexpected for the column at hand. In some situations, it may have been because the raters were not careful enough when making the annotations but it was also difficult to spot when an annotation was missing or made in an incorrect way. These situations were only spotted at later processing time, requiring some raters to come back to the sheet to fully complete their tasks.
To prevent this, we recommend to use a specialized environment that would fulfill the following additional requirements:

- Support a sort of **validation mechanism** to avoid raters making syntactically incorrect annotations on enrichments;
- Display information on the **progress of the annotation task** so that raters know when the task is finished; more specifically count and highlight the annotations that are missing with direct links to them so that raters can more quickly act on them.

The raters also found it difficult to identify the exact portion of the source metadata that was being enriched. This happened because of the inability in online spreadsheet to highlight specific pieces of text within a cell, which could have made it easier for the rater to spot the enrichment. We thus recommend that such highlight is present in the annotation environment, in a similar fashion as displayed for example in text annotation tools.

Finally, for some enrichments the information displayed in the spreadsheet was not enough for the rater to make an immediate judgement of the enrichment, in particular with regards to the target entity for which only the link was displayed. The raters often felt the need to access the source and/or target resources at their origin (website) to have a better view of the data. However, in some situations, this was still not enough as web pages often do not display all the data used by the enrichment tools, but only the most relevant data for a general web user. To make the assessment task more efficient, the Task Force recommends that both the metadata from the source and target should be displayed in a side by side manner with the source text highlighted (as explained in the previous paragraph) so that raters have all the data needed to assess the enrichment and make the right judgement.

4.4. Inter-Rater Agreement

As part of a sound evaluation methodology, it is essential to measure the level of agreement among raters when applying the criteria defined in the guidelines to annotate an enrichment, such is often referred to as **inter-rater agreement**. This agreement reflects the reliability of the evaluation results.

A second corpus was built to evaluate the agreement. This was done by selecting two enrichments from each set from the annotated corpus, resulting in a total of 52 enrichments to be annotated by each participant (see Appendix A). The selection was done manually so that it would contain the cases that may cause less agreement, but also represent best the variety of enrichments in the results.

In the literature, measuring agreement is often done by computing a **kappa** coefficient, such as Cohen's one[38]. If the kappa value is low, then the evaluation is questionable since assessments depend very much on the individual rater. Besides attaching a confidence level to the evaluation, the inter-rater agreement is also useful for determining if a particular criteria is appropriate for assessing an enrichment or if the guidelines need to be revised either by adding or removing some criteria or adding examples that could clarify situations of possible disagreement. If raters do not agree, either the criteria is inappropriate or the raters need to be re-trained.

Calculating Kappa

We chose the Fleiss Kappa[39], an extension to Cohen's kappa for a number of raters higher than two. As for Cohen's kappa, no weighting of the different criteria is used and the judgement categories ('correct', 'incorrect', 'uncertain') are considered to be unordered.

Only the values from the "semantically correct" criteria were used for measuring the agreement. Extending it to the other two criteria would only have an impact on the difference between relaxed and strict measurements for precision.
Table 2 shows the Fleiss Kappa calculation for each enrichment in the corpus for the number of n=16 raters, N=46 enrichments (subjects) and k=3 categories. The numbers within the subjects column represent references to enrichment within each sample in the form "<sample_number>.<index>". The column 'Categories' shows how many raters chose a given

[38] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20:37-46, 1960.
[39] Fleiss JL. Statistical methods for rates and proportions second edition. Wiley Series in probability and mathematical statistics. Chapter 13 p. 212-236

code. The values for the Pi and Kappa in the table are color coded according to the interpretation from Table 3.

Table 2: Fleiss Kappa calculation table with (N=46,n=16,k=3) for the inter-rater agreement for the annotated corpus.

Subjects	Categories			Subjects	Categories			Subjects	Categories			Subjects	Categories		
	C	I	U		C	I	U		C	I	U		C	I	U
#02.04	16	0	0	#12.05	16	0	0	#18.19	14	1	1	#26.77	16	0	0
#02.05	16	0	0	#12.06	5	11	0	#18.02	15	1	0	#26.99	16	0	0
#03.10	16	0	0	#13.02	13	3	0	#19.02	16	0	0	#27.25	3	13	0
#03.05	15	0	1	#13.03	16	0	0	#19.56	15	1	0	#27.03	8	7	1
#04.18	16	0	0	#14.06	16	0	0	#22.16	16	0	0	#28.15	6	9	1
#04.50	15	1	0	#14.07	14	1	1	#22.03	16	0	0	#28.02	15	1	0
#05.10	14	1	1	#15.60	13	1	2	#23.15	16	0	0	#29.27	15	1	0
#05.68	13	1	2	#15.61	15	1	0	#23.07	14	2	0	#29.03	6	5	5
#06.36	16	0	0	#16.14	15	1	0	#24.74	9	6	1	#30.32	12	4	0
#06.06	16	0	0	#16.37	15	1	0	#24.77	13	2	1	#30.04	6	10	0
#07.14	13	1	2	#17.03	16	0	0	#25.11	8	6	2	**Total**	607	101	28
#07.19	15	0	1	#17.07	16	0	0	#25.05	1	9	6				

Table 3: Interpretation table for the Kappa value (from Wikipedia[40]).

kappa	Interpretation	kappa	Interpretation
< 0	Poor agreement	0.41 – 0.60	Moderate agreement
0.01 – 0.20	Slight agreement	0.61 – 0.80	Substantial agreement
0.21 – 0.40	Fair agreement	0.81 – 1.00	Almost perfect agreement

Interpretation

The Fleiss kappa calculation for the inter-rater agreement is 0.329. According to Table 3 the obtained value represent a fair agreement among raters.

The enrichments on which raters disagreed most were #25.11 and #29.03, but also #24.74, #27.03 and #28.15[41]. Looking in particular to these 5 enrichments we can identify two

[40] https://en.wikipedia.org/wiki/Fleiss%27_kappa
[41] See file "evaluation/agreement/inter_rater_corpus.csv" under the Task Force's archive: https://www.assembla.com/spaces/europeana-r-d/documents?folder=58725383

different patterns. A first pattern found in #29.03 and #27.03 shows that some raters missed to recognize that the enrichment was not correct. This may have happened because the rater have overlooked it or were not experienced enough with the particular topic of the enrichment. A second pattern emerges from #25.11, #24.74 and #28.15, where the disagreement may have come from misunderstanding the guidelines regarding partial matches either at the name or concept categories, which led some raters to annotate them as incorrect, some as correct and others to simply not know how to annotate.

Even though the agreement was slightly lower than ideal, the evaluation was still considered satisfactory by the Task Force based on the following reasons:

- According to the Fleiss kappa method, the number of categories and subjects affect the magnitude of the value. The number of subjects was significantly high, which has more (negative) impact on the kappa value in case of disagreement.
- The enrichments annotated as unsure were not taken into account for measuring the performance of the tools. Even though it has impact on the size of the sample, it does not directly penalize the performance assessment of the tool.

Lessons Learned

For future evaluations using an annotated corpus, the Task Force recommends:

- Further analysis of non-trivial evaluation cases, in particular the ones that were the least agreed upon. The guidelines provided here should be improved with examples that could help clarify their use;
- The raters should be trained before the evaluation until they reach a higher inter-rater agreement (e.g., 0.81 or higher);

5. Analysis of enrichment results

The results from the enrichment tools were compared with the annotated corpus, adapting Information Retrieval's common **precision** and **recall** measures. For enrichments, precision measures the fraction of enrichments that were judged to be correct over all the enrichments found by a tool; recall ideally measures the enrichments judged correct against all the correct enrichments that could have been found. Simply put, high precision means that a tool returned substantially more correct enrichments than incorrect ones; high recall means that a tool returned most of the correct enrichments. Figure 5 shows a visual representation of these two metrics.

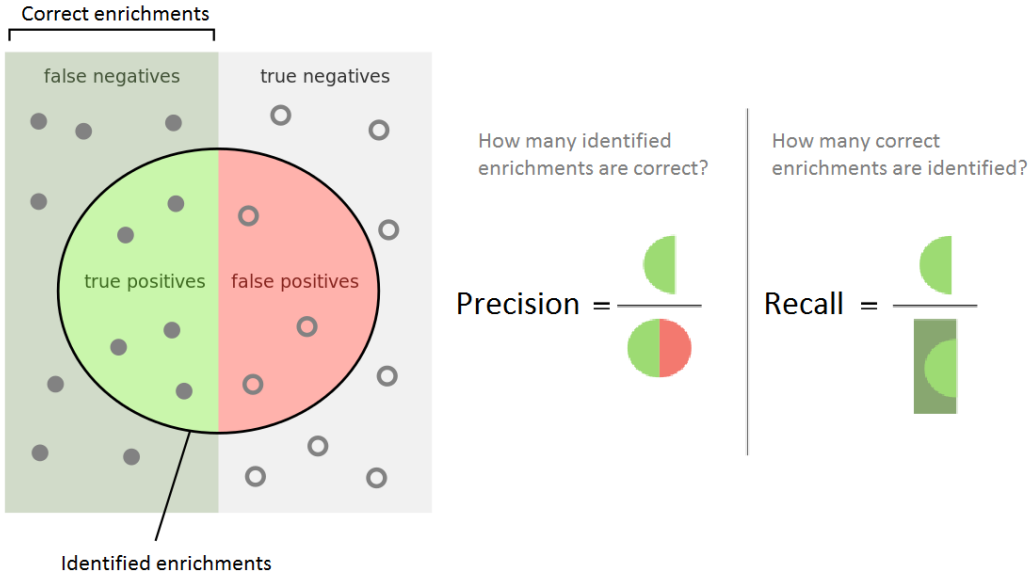

Figure 5: Diagram explaining the precision and recall notions adapted to semantic enrichment (adapted from Wikipedia[42]).

The correctness of an enrichment was assessed along three different criteria (see Section 5.3). We chose to compute the measures in two ways: **relaxed**, for which we considered as "true" all enrichments that were annotated as semantically correct regardless of their completeness (i.e. name or concept match); and **strict,** considering as "true" only enrichments that were annotated as being semantically correct and with a full name and concept completeness. Enrichments for which the rater was unsure were ignored in the calculations.

Formula 1: Calculation for precision for a specific tool (note that the meaning of "true" varies depending on the strict vs relaxed approach):

Precision:

$$\frac{\{\, All\ "true"\ enrichments\ of\ a\ tool \,\}}{\{\, All\ enrichments\ of\ a\ tool \,\}}$$

Table 4: Summary of the results for precision obtained for this evaluation.

Tools	Relaxed	Strict	*Diff.*
Europeana	0.985	0.965	*0.020*
TEL	0.982	0.982	*0.000*
BgLinks	0.888	0.574	*0.314*
Pelagios	0.854	0.820	*0.034*
VocMatch	0.774	0.312	*0.462*
Ontotext v1	0.842	0.505	*0.337*
Ontotext v2	0.924	0.632	*0.292*

[42] https://en.wikipedia.org/wiki/Precision_and_recall

An important issue, the task of identifying all possible enrichments for a given set of objects would have required too much effort for the Task Force members. We therefore chose to apply **pooled recall**[43], in which recall is measured by considering the total amount of correct enrichments as the union of all correct enrichments identified by all tools. To have an estimate of what could be the **maximum pooled recall** for each tool, we applied the pooled recall formula assuming that all the enrichments from a tool would be correct and applying the strict precision approach as it gives us an upper bound for this measure.

Formula 2: Calculation for pooled recall for a specific tool (note that the meaning of "true" varies depending on the strict vs relaxed approach):

Pooled Recall:

$$\frac{\{\ All\ "true"\ enrichments\ of\ a\ tool\ \}}{\{\ All\ "true"\ enrichments\ of\ all\ tools\ \}}$$

Formula 3: Calculation for the maximum pooled recall for each specific tool (note that for this formula the meaning of "true" is obtained from applying the strict approach only):

Max Pooled Recall:

$$\frac{\{\ All\ enrichments\ of\ a\ tool\ \}}{\{\ All\ "true"\ enrichments\ of\ all\ tools\ \}}$$

Table 5: Results for pooled recall and f-measure obtained for this evaluation.

Tools	Annotated Enrichments	Max Pooled Recall	Pooled Recall			F-measure		
			Relaxed	Strict	*Diff.*	Relaxed	Strict	*Diff.*
Europeana	550 (31.3%)	0.458	0.355	0.432	*-0.077*	0.522	0.597	*-0.075*
TEL	391 (22.3%)	0.325	0.254	0.315	*-0.061*	0.404	0.477	*-0.073*
BgLinks	427 (24.3%)	0.355	0.249	0.200	*0.049*	0.389	0.296	*0.093*
Pelagios	502 (28.6%)	0.418	0.286	0.340	*-0.054*	0.428	0.481	*-0.053*
VocMatch	100 (05.7%)	0.083	0.048	0.024	*0.024*	0.091	0.045	*0.046*
Ontotext v1	489 (27.8%)	0.407	0.272	0.202	*0.070*	0.411	0.289	*0.122*
Ontotext v2	682 (38.8%)	0.567	0.418	0.354	*0.064*	0.576	0.454	*0.122*

As mentioned above, the unbalance in the sample selection impacts the *comparative* aspect of our evaluation. This especially concerns the "pooled" recall. Because they were made for different application scenarios and use different targets, the results of some tools have been under-represented (see Sections 5.1 and 5.2). We urge the reader to keep in mind the general coverage of enrichments (as reported in Figure 2 and Figure 3) when looking at the figures for pooled recall. For example, extrapolating Ontotext v2's relaxed precision (92.4%) and its total amount of enrichments (124,407), we can infer that this service probably produces above 100K correct enrichments, which is also an interesting indicator of its performance in the absence of recall based on a complete gold standard.

[43] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

5.1. Overview

A quick look at the results for precision, in particular the *strict* one, shows a divide between two groups: on one side Europeana and TEL (group A), and on the other the BgLink, Pelagios, VocMatch and Ontotext (group B). The difference between these two groups is that both EF and TEL limit their enrichments to metadata fields which typically contain (semi-) structured information (e.g., dc:creator) while others also apply enrichment to metadata fields containing any sort of textual description, from short (dc:title) to long (dc:description). In semi-structured metadata fields, the complexity of identifying the right named reference is much lower due to the fact that these fields tend to: (a) contain only one named reference, or several entities with clear delimiters (author names within a dc:creator field are often delimited by a semicolon); (b) obey to a normalized format or cataloguing practice (e.g., dates with a standardized representation); (c) contain references to entities whose type is known in advance (e.g., dcterms:spatial should refer to places and not persons).

5.2. Individual enrichment services

Europeana Foundation

The EF enrichment tool ranks first on relaxed precision. Besides the fact that it focuses enrichment only to semi-structured fields, the tool benefits from enriching only against a specific selection of the target vocabularies (made prior to enrichment) which reduces the chance of picking incorrect enrichments because of ambiguous labels (cf. Section on techniques and tools in the main Task Force report). However, the results for EF drop to the second place for strict precision.

The problem is that in case of ambiguity, the tool cannot select the right entity. A typical example is references to places that may correspond to different levels of administrative division with the same name. In such cases, the tool always picks the same entity (based on the order in which it was loaded in the tool), which can result in partial (and approximate) enrichments. For example, the named reference "London" is always enriched with the "City of London"[44] and not the capital city[45] (a case of enriching with a narrower place instead of a broader); "Madrid" is enriched with the province[46] and not the capital city[47]; and "Bratislava" is enriched with the region[48] instead of the capital city[49]. Besides the granularity issue, the

[44] http://sws.geonames.org/2643744/
[45] http://sws.geonames.org/2643743/
[46] http://sws.geonames.org/6355233/
[47] http://sws.geonames.org/3117735/
[48] http://sws.geonames.org/3343955/

tool is also not able to take into account the time dimension when selecting the entity. This is particularly relevant for place entities whose boundaries change over time. For example, some objects from the 18th century with the named reference "Germania" are enriched with "Federal Republic of Germany". This can be seen as an avoidable side effect of using Geonames which mostly contains contemporary places. Still, one could argue that the service could recognize the time deviation and not output the enrichment. Another time-related issue, the service does not correlate the time of the publication with a creator's life span. For example (also found in Ontotext results), a book with a named reference of "Joaquim Costa" was enriched with a Person[50] that was born after the book had been published.

Finally, the results confirm previous findings[51] that some incorrect enrichments could be avoided if the language of the metadata was taken into account. The named reference "minister" is enriched with the GEMET concept entity for "Ministry"[52] due to the fact that in Romanian the same name is used to refer to the Ministry.

The European Library

The TEL results show the same precision regardless of whether a relaxed or strict metrics is applied, meaning that the TEL enrichment tool is primarily aiming at high precision. Besides limiting the enrichments to metadata fields that are typically (semi-) structured, it features a disambiguation mechanism to pick the entity most likely to be the one being referred, based on its description. In particular, for places it uses the classification of the place (e.g., the 'feature type' in Geonames) or demographic information (also used by Pelagios) as indicators for the relevance of an entity. However, like EF, it does not take into account the time frame of the object when selecting a place. An object from the 16th century has been enriched with "Russian Federation". Again, the use of a gazetteer containing names of contemporary places does not help the tool to selecting the right place. In addition, results show that disambiguation could have performed better had it considered language. The named reference "Romãe" referring to the latin name for Rome[53] in Italy has been instead linked to a small village in Portugal[54].

---

[49] http://sws.geonames.org/3060972/
[50] http://dbpedia.org/resource/Joaquim_Costa
[51] 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
[52] http://www.eionet.europa.eu/gemet/concept/5299
[53] http://sws.geonames.org/3169070/
[54] http://sws.geonames.org/2735029/

BgLinks

In group B, BgLinks appears just behind Ontotext v2 and in front of Ontotext v1. Table 1 shows that these tools are the ones that share the biggest number of enrichments (except Ontotext v1 and v2, of course) which can help explain the proximity in their performance (both relaxed and strict).

A more detailed look show that a particular challenge for BgLinks is to enrich acronyms. Very few of these are correct. For example the acronym for "intracranial pressure"(ICP) has been enriched with an hip-hop group[55] and the acronym for "angiotenzin-converting enzyme" (ACE) has been enriched with "Accumulated Cyclone Energy"[56]. On the other hand, BgLinks performs significantly better in determining the right references within the text to enrich, compared to Ontotext. Some exceptions were still found, such as the term "permeability", used to characterize a soil, being enriched with the same term as used in electromagnetism[57]. Additionally, BgLinks is also successful at enriching more complex named references such as the term "C. acutatum" being rightly enriched with "Colletotrichum acutatum"[58] in DBpedia and the term "active metabolites" with "Produg"[59]. This feature is a result of applying more relaxed approaches to name matching.

An aspect that explains in part the difference between the results for relaxed and strict is that many partial enrichments were produced for terms that denote entities without a full semantic equivalent in the target dataset. A typical example, the term "Archive of Oncology" which may lead to enrichments with the entities in DBpedia for "Oncology" and "Archive" as there is no DBpedia resource that combines both. Other examples are "lithogenic bile", for which only bile is enriched, and "mass concentration" that is linked to a general "concentration" notion that combines mass, molar, number, and volume concentrations. This issue appears for all tools but is particularly found in group B, as such references are more common in long text descriptions than in normalized or structured fields.

Pelagios

Pelagios has the best strict precision in group B, and is slightly under BgLinks for relaxed precision. The fact that Pelagios is specialized for place name enrichments certainly helped achieving this result. The target vocabulary used is smaller and more specialized than the broader datasets used by other tools, which also makes it able to apply place-specific heuristics. This can explain why in terms of deviation between relaxed and strict precision it

[55] http://dbpedia.org/resource/Insane_Clown_Posse
[56] http://dbpedia.org/resource/Accumulated_cyclone_energy
[57] http://dbpedia.org/resource/Permeability_(electromagnetism)
[58] http://dbpedia.org/resource/Colletotrichum_acutatum
[59] http://dbpedia.org/resource/Prodrug

performs similarly to TEL and EF, which apply rules and target datasets depending on the type of the entity expected to be found in certain fields.

A closer look shows that the most common reasons for incorrect or partial enrichments are related to Pelagios' issues with disambiguating between target entities. It has a disambiguation mechanism in place, but the Wikidata target vocabulary does not yet provide the necessary demographic information that Pelagios (like TEL) uses as one of the indicators of the relevance of an entity. For example, "Siberia" is enriched with a place in California[60], "Paris" with a place in the US[61] and "France" with a crater on the moon[62]. Additionally, Wikidata contains a very wide range of geographical entities from administrative divisions (countries or cities) to any sort of location or physical body, e.g., café, statue, public building, monument.

The lack of proper disambiguation made it possible to enrich a named reference for "Poland" with a fire department in the U.S.[63] and a "Postcards" subject with the Postcards memorial in NYC. Pelagios also applies a fuzzy matching between the named reference and the labels of the target entity, which leads to enrichments for all sorts of nouns, such as "people" with Peoples[64], a place in U.S., or "men" with Menz[65], a village in Germany. Additionally, even though Pelagios aims at enriching old place names, it had some issues determining if an entity actually corresponds to the time frame of the description. A reference for the pannonian basin[66] is enriched with the roman province related to it[67] (an enrichment also made by Ontotext). The disambiguation problems were not significantly felt in the overall performance since only a small amount of the enrichments evaluated were referring to text fields (about 20% of the total number of enrichments against 50% that was found in average for the other group B tools, see Appendix A for the complete distribution of enrichments per property).

VocMatch

The tool that displayed the poorest performance was VocMatch. The fact that it was exceptionally difficult for the rater to identify the actual portion of the text that served as indicator for the enrichment made it hard to assess its correctness. In particular for the

[60] http://sws.geonames.org/5395524/
[61] http://sws.geonames.org/4402452/
[62] http://dbpedia.org/resource/Franck_(crater)
[63] http://sws.geonames.org/4263238/
[64] http://sws.geonames.org/4303909/
[65] http://www.wikidata.org/entity/Q1767603
[66] http://dbpedia.org/resource/Pannonian_Basin
[67] http://dbpedia.org/resource/Pannonia_(Roman_province)

pooled recall, we were unable to reconcile its results with the results from other tools because VocMatch uses specialized vocabularies not used by the others and no coreference information was available to link them.

A more detailed look at the results shows that the reason for some incorrect enrichments are because it matches against all terms available as target vocabularies, with no disambiguation. An example is the word "still" as part of the term "still image" present within a dc:type property being enriched with the term "distillery"[68] of the Wales Monument Type Thesaurus, or the named reference ("Włodzimierz Press") for a Person being partially enriched with the term "Press"[69] of the UNESCO Thesaurus. However, this approach is effective when applied to semi-structured properties such as dc:subject or dc:type, which becomes more evident when comparing the results from VocMatch with EF which also applies the same methods as VocMatch but only to semi-structured fields. In fact, complementary investigations show that using only semi-structured fields the tool reaches 86.7% relaxed precision.

Ontotext v1 and v2

The comparison of the results from the two Ontotext versions show a difference in performance between them, from which one can deduce that language played an important role in the methods that it applied. Additionally, close to 100% of the enrichment identified in version 1 were also detected in version 2. Version 1 discarded about half of the enrichment that was identified in version 2 but still reducing its performance.

A closer look shows that a great amount of enrichments were identified for non-named references like verbs (think, conduct, caused), adverbs (viz.), adjectives (valid, inadequate, randomly, red, inferior), abbreviations (Mrs), simple nouns (purpose, predictor, stone, left), etc., which do not really contribute to improving the object description. They can even lead to wrong enrichments. For example, the word "old" in "16 years old" has been enriched with the term referring to an old vineyard[70] on DBpedia, or the word "leis" in Portuguese (which mean "laws") being enriched with DBpedia term for Low Energy Ion Scattering[71], a method used in chemistry. Such enrichments were mostly found within text fields but also within semi-structured fields. As other tools, it is not capable of determining the correct time frame of the enrichment (see discussion on Pelagios and Europeana).

[68] http://purl.org/heritagedata/schemes/10/concepts/69125
[69] http://skos.um.es/unescothes/C03123
[70] http://dbpedia.org/resource/Old_vine
[71] http://dbpedia.org/resource/Low-energy_ion_scattering

For the remainder of the enrichments, Ontotext displays a good performance. In particular, it shows a significantly better performance than BgLinks on enriching acronyms. Only a couple of exceptions were found, among which the "leis" example above, ECCO European CanCer Organisation enriched with a shoe manufacturer[72] and "WI" (referring to the term "Whitening Index") being enriched with Wisconsin[73] (a US state) (both examples were also found by BgLinks).

6. Conclusion

The Task Force members have learned a number of lessons that made us change our original plans, or would need to be considered for future experiments. These were described along this document and are assembled in Section 7.1. With regards to the enrichment tools that were analysed as part of this evaluation campaign, after measuring and analyzing their results, the Task Force has made a list of recommendations in order to improve the general quality of your enrichments which are described in Section 7.2.

6.1. Summary of the Lessons Learned on Evaluation

- **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 alignment if you have sufficient time and human resources to commit to it, otherwise go for a semi-automatic strategy by selecting the enrichments identified by the tool under evaluation or other enrichment tools. The tradeoff is that the latter option does not allow one to obtain 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 more precise view on how enrichment compare across tools, or to reuse a gold standard coming from another evaluation.
- **Try to keep balance between tools in comparative evaluations:** Some of the corpus creation strategies mentioned above are likely to result in a bias against some tools. Make sure such bias is recognized and if possible properly connected to the concern that motivated your evaluation strategy.
- **Make clear guidelines on how to annotate the corpus:** make guidelines that are both simple enough for raters to understand but still detain the necessary

[72] http://dbpedia.org/resource/ECCO
[73] http://dbpedia.org/resource/Wisconsin

information to make the right judgement. Consider having examples for the cases that may raise the most doubt. Consider testing your raters with the guidelines 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 the rater through its task.

6.2. Recommendations for Enrichment Tools

- Consider applying different 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.
- 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[74]. This especially hurts when the target datasets include 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 together with the relevance of the entity in general (looking as its data properties) and in particular, i.e., within the context it is being referred (this implies determining the correct context of its use). For example, one of our observations was 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.
- 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.
- Quality issues originated as part of the mapping process had been already identified as a great obstacle to get enrichments of good quality, in the 2014 report of the Task Force on Multilingual and Semantic Enrichment Strategy. Our evaluation has confirmed it. Semantic enrichment rules crafted to work on specific metadata fields (e.g., for spatial coverage of an object) should be designed and applied carefully to source datasets, in case these fields could be populated with values that result from wrong mappings (e.g. publication places)

[74] 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.

Appendix

A. Distribution across tools of the enrichments based on their source property

Property	Europeana	TEL	BgLinks	VocMatch	Pelagios	Ontotext v1	Ontotext v2	Total
dc:contributor	563	333	51	97	0	802	1899	3745
dc:creator	158	780	161	303	0	1714	2491	5607
dc:date	3221	0	0	823	0	85	373	4502
dc:description	0	0	5901	30277	3228	31742	43979	115127
dc:format	0	0	25	645	0	305	2958	3933
dc:language	0	0	0	16	0	0	0	16
dc:publisher	0	0	322	2375	1152	2959	5301	12109
dc:relation	0	0	3	1110	0	118	2511	3742
dc:rights	0	0	11	23363	0	0	0	23374
dc:source	0	0	12	4441	503	3196	5754	13906
dc:subject	3095	19	1068	18358	15228	7510	13913	59191
dc:title	0	0	1447	12113	2923	8542	19425	44450
dc:type	1087	0	0	6672	0	1020	7293	16072
dcterms:alternative	0	0	15	605	0	454	1460	2534
dcterms:created	0	0	0	0	0	1	12	13
dcterms:extent	0	0	0	160	0	144	657	961
dcterms:hasFormat	0	0	0	0	0	1	1003	1004
dcterms:hasPart	0	0	0	44	0	421	1684	2149
dcterms:isPartOf	0	0	1	32	0	0	0	33
dcterms:isReferencedBy	0	0	0	144	0	128	519	791
dcterms:issued	0	0	1	12	0	40	159	212
dcterms:medium	0	0	0	8106	0	56	1065	9227
dcterms:spatial	4686	5015	88	4733	6340	2913	3297	27072
dcterms:tableOfContents	0	0	22	152	0	163	581	918
dcterms:temporal	560	0	0	24	0	24	148	756
edm:type	0	0	0	16	0	105	7924	8045
owl:sameAs	0	0	0	0	0	0	1	1
Total	13370	6147	9128	114621	29374	62443	124407	359490

B. Guidelines used for the manual evaluation of enrichments

Please follow these guidelines when annotating the gold standard for enrichments.

Each annotation should be evaluated regarding the following categories. Each category has a separate column in the spreadsheets and one should use one of the codes to annotate each category.

To consider while annotating:

- If an enrichment is incorrect (category: semantic correctness), there is no need to fill in the other categories.
- If a better matching concept can be identified, you can leave the URI in the comments column.
- there is the possibility of using "Unsure" for cases where you cannot decide or have no time to investigate in more detail.

Category	Annotation	Description
Semantic correctness	C = Correct I = Incorrect U = Unsure	Is the enrichment semantically correct or not?
Completeness of name match	F = Full match P = Partial match U = Unsure	Was the whole phrase/named entity enriched or only parts of it?
Completeness of concept match	F = Full match B = Broader than N = Narrower than U = Unsure	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. Use B when the concept identified (i.e. target) is broader than the intended concept. Use N, when it is narrower. 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.

<u>Some Examples:</u>

1. Considering a data field with the phrase "Département de Paris":

   a. If the word “Paris” is enriched with Paris as the city / capital of France (http://www.geonames.org/2988507), would result in the following annotations: Semantic correctness: **C**, Completeness of name match: **P**, Completeness of concept match: **F**.
   b. If the complete phrase (that is “Departement de Paris”) is enriched with Paris as the city / capital of France (http://www.geonames.org/2988507), would result in the following annotations: semantic correctness: **C**, Completeness of name match: **F**, Completeness of concept match: **N**.
   c. If either the word “Paris” or the complete phrase is enriched with Paris as the city in the USA (http://sws.geonames.org/4402452/), would result in the following annotations: semantic correctness: **I**, the other two categories do not need to be filled anymore.
   d. If the complete phrase is enriched with Paris, the second-order administrative division (http://www.geonames.org/2968815/), would result in the following annotations: semantic correctness: **C**, Completeness of name match: **F**, Completeness of concept match: **F**.
2. Considering a data field with the phrase “The Jackson Family”:
   a. If the word “Jackson” is enriched with the person “Michael Jackson” (http://dbpedia.org/resource/Michael_Jackson), it would result in the following annotations: semantic correctness: **C**, Completeness of name match: **P**, Completeness of concept match: **N**.
   b. If the complete phrase is enriched with the family (http://live.dbpedia.org/resource/Jackson_family), it would result in the following annotations: semantic correctness: **C**, Completeness of name match: **F**, Completeness of concept match: **F**.
3. Considering a data field with the phrase “Bonnet fils (Avignon)”, if the word “Bonnet” is enriched with Bonnet as a form of hat (http://purl.org/heritagedata/schemes/mda_obj/concepts/96582), would result in the following annotations: semantic correctness: **I**, the other two categories do not need to be filled anymore.
4. Considering a data field with the phrase “Bibliotheque nationale de France”, the word “France” is enriched with the country (http://www.geonames.org/3017382/), it would

result in the following annotations: semantic correctness: **C**, Completeness of name match: **P**, Completeness of concept match: **F**.