# Exploring comparative evaluation of semantic enrichment tools for cultural heritage metadata

Hugo Manguinhas[1], Nuno Freire[1], Antoine Isaac[1,6], Juliane Stiller[2], Valentine Charles[1], Aitor Soroa[3], Rainer Simon[4], Vladimir Alexiev[5]

[1]Europeana Foundation, The Hague, The Netherlands
{hugo.manguinhas,antoine.isaac,valentine.charles}@europeana.eu
nuno.freire@theeuropeanlibrary.org
[2]Humboldt-Universität zu Berlin, Berlin, Germany
{juliane.stiller@ibi.hu-berlin.de}
[3]University of the Basque Country, Bizkaia, Spain
{a.soroa@ehu.eus}
[4]Austrian Institute of Technology, Vienna, Austria
{rainer.simon@ait.ac.at}
[5]Ontotext Corp, Sofia, Bulgaria
{vladimir.alexiev@ontotext.com}
[6]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

**Abstract.** Semantic enrichment of metadata is an important and difficult problem for digital heritage efforts such as Europeana. This paper gives motivations and presents the work of a recently completed Task Force that addressed the topic of evaluation of semantic enrichment. We especially report on the design and the results of a comparative evaluation experiment, where we have assessed the enrichments of seven tools (or configurations thereof) on a sample benchmark dataset from Europeana.

**Keywords:** Semantic Enrichment, Metadata, Cultural Heritage, Evaluation, Europeana

## 1 Introduction

Improving metadata quality is crucial for Europeana, the platform for accessing digitized Cultural Heritage (CH) in Europe, and many projects that have similar goals in more specific areas. A key technique is semantic enrichment, which puts objects in context by linking them to relevant entities (people, places, object types, etc.). In the CH domain, semantic enrichment provides richer context to items and allows systems to add information to existing metadata [2,3]. Systems can indeed later obtain "semantic" descriptions for the related entities when these are published, e.g., using Linked Open Data technology. Semantic enrichment is a useful technique with a variety of applications. For instance, enriching flat documents with instances of structured knowledge can be used in search, where results for named entity queries will include

adfa, p. 1, 2011.

facts about the entities involved [1]. A variety of tools have been recently developed – or adapted from other domains – to enrich objects by exploiting the existing metadata.

Enrichment of CH metadata is however a very difficult task due to: 1) bewildering variety of objects; 2) differing provider practices for providing extra information; 3) data normalization issues; 4) information in a variety of languages, without the appropriate language tags to determine the language [2]. Adding to this difficulty, the importance of these factors may change between one dataset or application and another. Moreover, the various tools and approaches available can of course perform very differently on data with different characteristics. This makes it difficult to identify and apply the right enrichment approach or tool (including using the right parameters when an approach can be tuned). To help practitioners, especially those from the Europeana family of projects, the following R&D questions required specific effort:

- perform concrete evaluations and comparisons of enrichment tools currently developed and used in the Europeana context;
- identify methodologies for evaluating enrichment in CH, specifically in Europeana, by making sure that evaluation methods (i) are realistic wrt. the amount of resources to be employed, (ii) can be applied even when enrichment tools are not trivially comparable (i.e. when they link objects with different target datasets) (iii) facilitate the measure of enrichment progress over time;
- build a reference set of metadata and enrichments that can be used for future comparative evaluations.

To gain insight on these points, a group of experts from the Europeana community have gathered as a Task Force and undertook an evaluation campaign for representative enrichment tools[1]. In this paper we explain the phases of the evaluation, covering methodology, results and analysis. We conclude with a summary of the group's lessons learned and recommendations. We refer the reader interested in more detail to browse the technical report of the Task Force, where advanced explanations on the evaluation method and results can be found [4].

## 2 Related work

For information access systems, several initiatives exist for conducted structured evaluation experiments, e.g. CLEF[2]. In the Semantic Web community, the Ontology Alignment Initiative assesses ontology alignment systems[3]. For enriching data, no well-established benchmark exists, although the number of tools and enriched datasets is growing constantly. A notable exception in the Linked Data community is GERBIL [5], a framework for benchmarking systems on various annotation tasks, including named entity recognition and named entity linking. GERBIL enables implementers to compare results of tools on the same datasets, in a principled, reproducible way. Us-

[1] http://pro.europeana.eu/taskforce/evaluation-and-enrichments
[2] http://www.clef-initiative.eu/
[3] http://oaei.ontologymatching.org/

ing GERBIL, Usbeck et al. compared more than 15 systems on 20 different datasets. GERBIL can be used with systems and datasets from any domain. However, these datasets do not include multilingual CH metadata. Evaluating enrichment tools against them would not bring the insight needed to answer our research questions[4].

Within the digital CH domain, numerous studies have evaluated automatic semantic enrichments. The DM2E[5] project performed sample evaluation of alignment of (local) places and agents to DBpedia where 150 random agents and 150 random places were selected. The sample was based on the amount of agents/places each collection contains. The results showed that 18% of the agents and 60% of places are linked. From these, 83% of the agent links and 85% of the place links are good.

The Paths Project[6] developed functionalities for information access in large-scale digital libraries, focusing on metadata enrichment to let users better discover and explore CH material. Evaluation of the Paths prototype focused on assessments with focus groups within laboratory settings. Enrichments were tested indirectly following a methodology of Interactive Information Retrieval in a laboratory setting: users performed tasks and logs, screen recordings and observer notes were collected [6].

In [3], the authors explore the feasibility of linking CH items to Wikipedia articles. They develop a small dataset comprising 400 objects from Europeana and manually link them to Wikipedia whenever there exists an article that exactly describes the same object as the CH object. This dataset is then used to evaluate two systems, which yielded relatively poor performances.

OpenRefine[7] has also been used to perform evaluations. One was an evaluation of structured field reconciliation[8]. The other was an evaluation of named-entity recognition on unstructured fields (performed with OpenRefine and a plugin): Both have been evaluated on concrete datasets with a manual validation.

Stiller et al. in [2] 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 were subjective as the choice of queries for the evaluation influenced the results. Nevertheless, if a representative query sample is chosen the approach can give insights about the likelihood of a user encountering beneficial enrichments or incorrect ones. A previous evaluation of the enrichment in Europeana qualitatively assessed 200 enrichments for the four different types: time, persons, location and concepts [7].

As enrichments impact the search performance and are often implemented to improve search across several languages, all evaluations targeting the search performance are also relevant as enrichment evaluations. Within the cultural heritage domain, search evaluation was performed [8] as well as retrieval experiments bases on data from cultural heritage portals [9]. Additionally, user-centric studies aimed at improving usability of these services [10].

---

4 In fact a possible outcome of our work could be to contribute datasets and gold standards to GERBIL so as to make it a more suitable platform for future evaluations in our domain.

5 http://dm2e.eu/

6 http://www.paths-project.eu/

7 http://openrefine.org/

8 http://freeyourmetadata.org/publications/freeyourmetadata.pdf

Although not specific to cultural heritage, the evaluation studies conducted in the area of ontology alignment are relevant, as CH makes extensive use of ontologies in the data, and many enrichment tools enrich data by targeting ontologies. Work in this area has found similar difficulties in evaluation as our case, such as defining a gold standard and reaching a high level of rater agreement [11].

## 3 Evaluation setting

The Task Force conducted the evaluation of selected enrichments tools fulfilling the following steps which are more detailed in Fig. 1: select a sample dataset for enrichment, apply different tools to enrich the sample dataset, manually create reference annotated corpus and compare automatic annotations with the manual ones to determine performance of enrichment tools.

A proper evaluation requires a dataset representative of the multilingual and cross-domain diversity of Europeana's metadata. We selected data from The European Library[9] (TEL). TEL is the biggest data aggregator for Europeana; it has the widest coverage of countries and languages. For each of the 19 countries, we randomly selected a maximum of 1000 records. In order to have varied data in 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 contains 17,300 metadata records in 10 languages (see more info, incl. the full list of countries and languages, in our extended document and archive [12]). TEL records are expressed in the Europeana Data Model[10]. Fig. 2 lists the properties used in the evaluation dataset.

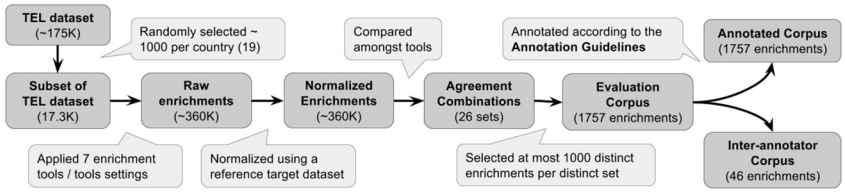


**Fig. 1.** Evaluation workflow.

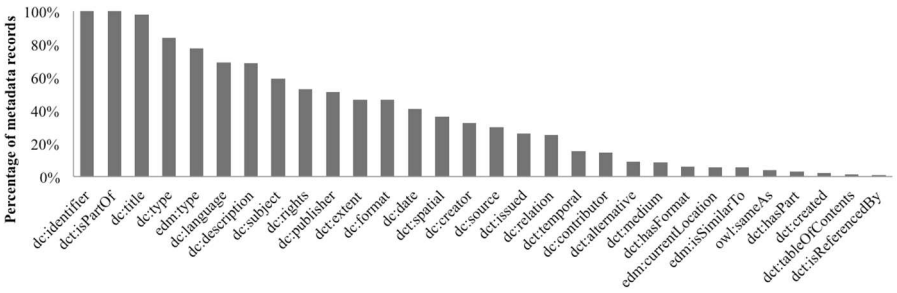


**Fig. 2.** Frequency of properties found within the evaluation dataset.

[9] http://www.theeuropeanlibrary.org/

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

Albeit coming from libraries, our dataset is quite heterogeneous, as TEL includes books, prints, and maps. However, at a later stage of the evaluation, we detected a bias towards scientific materials, which as key TEL resources are more frequently represented than in Europeana. Still, these documents belong to varied domains of science: mathematics, biology, agriculture, etc.

## 4 Enrichment results obtained from the participants

Within our Task Force, the following participants have applied their tools to enrich the evaluation dataset: the Europeana Foundation (EF); TEL; the LoCloud project[11] (with two different tools); the Pelagios project[12] and Ontotext[13] (with two different settings for determining language of metadata). Table 1 lists the tools, methods and target datasets each participant used. Participants sent their enrichment results using an agreed format containing (i) the identifier of the enriched object; (ii) the enriched property (e.g., dcterms:spatial); (iii) the identifier of the target entity (e.g., a DBpedia URI); (iv) the enriched literal (word or expression) where the entity was identified. A total of about 360K enrichments were obtained for the 7 different tools or tool settings. Fig. 3 and 4 shows respectively the number of enrichments and the coverage of the evaluation dataset's records for each tool. More statistics and tool information can be found in our extended document and archive [12].

## 5 Creation of the reference annotated corpus

Building a complete "gold standard" of correct enrichments for every object, as done in related work, is not feasible for us: the amount of objects and the variety of tools and targets are just too large. We instead tried to build a reference dataset starting from the enrichments themselves, reflecting their diversity *and* their commonalities.

The variety of target datasets hides cases where tools agree on the semantic level, i.e., they point to semantically equivalent resources in different datasets. We have "normalized" the targets of enrichments into a reference target dataset using existing coreference links (i.e., owl:sameAs or skos:exactMatch) between original targets, so that original enrichments can be "reinterpreted" as linking to resources from the reference dataset. We selected GeoNames (for places) and DBpedia (for other resources) as reference datasets, as they benefit from the highest overlap across the output of all tools. It was possible to normalize 62.16% of the results this way.

To build the corpus to be manually annotated, we compared normalized enrichments sets to identify the overlap between tools (i.e., enrichments with the same source object, target resource and enriched property) and sets specific to one tool. This gave 26 different sets that reflect the agreement combinations between tools (for more details see [12]). For each set, we randomly selected at most 100 enrichments (if

[11] http://www.locloud.eu/
[12] http://pelagios-project.blogspot.nl/
[13] http://ontotext.com/

the set contained less than 100, all enrichments were selected) resulting in a total of 1757 distinct enrichments.

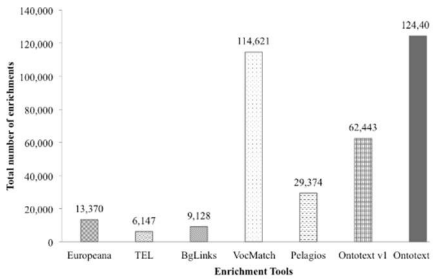


**Fig. 3.** Number of enrichments by tool.

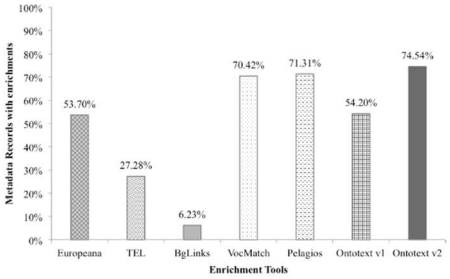


**Fig. 4.** Dataset records enriched, by tool.

**Table 1.** Overview of the tools evaluated.

| Part. | Tool | Entity types | Target datasets | Methods |
|---|---|---|---|---|
| EF | Europeana Semantic Enrichment Framework[14] | Places, Agents, Time spans, Concepts | DBPedia[15] (Agents, Concepts), GeoNames[16] (Places), Semium Time (Time spans) | Rule based tool, string normalization and matching. |
| TEL | In-house dev. | Places, Agents | GeoNames, GemeinsamenNormdatei | NERD, heuristic-based (Places); coref. information [13,14] (Agents) |
| Lo-Cloud[17] | Background Link (BgLink), DBpedia Spotlight 0.6[18] | Wide range of entities and concepts | DBpedia | NERD; supervised statistical methods (English) |
| | VocMatch service and TemaTres[19] | Concepts | Several thesauri and taxonomies[20] | SKOS vocs, automatic term assignment |
| Pelagios | Recogito[21] | Places | Pleiades, Digital Atlas of the Roman Empire[22], Archaeological Atlas of Antiquity[23] | NERD; user verification and correction |
| Ontotext | v1: Ontotext Sem. Platform, GATE[24] | Concepts (English only), Persons, Places | MediaGraph (a custom KB including DBpedia and Wikidata) | NERD, rule based and machine learning |
| | v2: same v1 | | | |

14 http://pro.europeana.eu/page/europeana-semantic-enrichment
15 http://wiki.dbpedia.org/
16 http://www.geonames.org/
17 http://support.locloud.eu/Metadata%20enrichment%20API%20technical%20documentation
18 http://spotlight.dbpedia.org/
19 http://www.vocabularyserver.com/
20 http://vocabulary.locloud.eu/?p=36
21 http://pelagios.org/recogito
22 http://darmc.harvard.edu/icb/icb.do
23 http://www.vici.org/
24 https://gate.ac.uk/

This approach helps us to identify when a similar logic is shared across tools (such as using the same rule or same part of the target dataset). However it negatively affects the evaluation of some tools. VocMatch for instance used target datasets that have no co-reference links the reference datasets. Therefore it didn't share enrichment with any other tool, and its results were underrepresented in our corpus.

Next, 16 members of our Task Force manually assessed the correctness of the 1757 enrichments annotating the corpus accordingly. We prepared a first version of annotation guidelines looking at the extrinsic (user-focused "usefulness") and intrinsic (system-focus) value of an enrichment. The extrinsic criteria assessed the informational value and specificity of an enrichment. Testing these criteria with three raters on six enrichments per set revealed that the extrinsic category was very subjective and applying it would have required onsite training of the raters. Due to constraints in time and resources, the extrinsic criteria were dropped and the definitive version contained three intrinsic categories for assessment: **semantic correctness** (correct, incorrect, uncertain), if the enrichment is appropriate; **completeness of name match**, if a whole phrase/named entity from a metadata field was enriched or only parts of it; **completeness of concept match**, if the target resource is at the same level of conceptual abstraction as the named entity/phrase in the metadata field being enriched.

To assess the reliability of the annotations, we measured inter-rater agreement on the "semantically correct" assessment for 46 enrichments that were assigned to all 16 raters – resulting in 736 annotations. We selected the enrichments manually to make sure the low sampling rate would not result in missing interesting cases and losing variety of enrichments. Agreement was measured using the Fleiss Kappa [15], calculated with parameters N=46, n=16, k=3. Inter-rater agreement is 0.329, i.e., "fair agreement" under the typical Kappa value interpretation, although the observed percentage agreement was high (79.9%). One reason for this is that the ratings were not evenly distributed between the different categories as most of the enrichments were considered to be correct, so the prevalence of correct ratings was very high. We therefore also report on the free-marginal multirater Kappa [16], which is 0.698, an agreement we considered satisfactory.

## 6 Analysis of enrichment results

The results of enrichment tools were compared against the manually annotated corpus, adapting Information Retrieval's common precision (fraction of enrichments that were judged to be correct over all the enrichments found by a tool) and recall (correct enrichments found by a tool against all the correct enrichments that could have been found) measures. We chose to compute our measures for two 'aggregates' of the three correctness criteria above: **relaxed**, where all enrichments annotated as semantically correct are considered "true" regardless of their completeness; and **strict**, considering as "true" the semantically correct enrichments with a full name and concept completeness. Enrichments for which raters were unsure were ignored in the calculations.

The fact that we could not identify all possible enrichments for the evaluation set lead us to apply **pooled recall** [17], in the total amount of correct enrichments is re-

placed by the union of all correct enrichments identified by all tools. As mentioned in section 5, however, some tools' results are under-represented in the corpus. This especially impacts the pooled recall. To take this into account in our analysis, we computed the **maximum pooled recall**, i.e. pooled recall assuming that all enrichments from a tool are correct and applying the 'strict' approach as it gives an upper bound for this measure (NB: for precision and pooled recall, "true" depends on the choice of 'strict' or 'relaxed', while for max pooled recall we use only 'strict'):

$$\text{Pooled Recall} = \frac{\{\#\text{ "true" enrichments of a tool}\}}{\{\#\text{ "true" enrichments of all tools}\}} \quad (1)$$

$$\text{Max Pooled Recall} = \frac{\{\#\text{ enrichments of a tool}\}}{/\{\#\text{ "true" enrichments of all tools}\}} \quad (2)$$

In general one must keep in mind the coverage of enrichments (Fig. 2 and 3) when analysing the results of our evaluation. For example, from Ontotext v2's relaxed precision (92.4%) and its total amount of enrichments (124,407), we can extrapolate that this tool probably produces over 100K correct enrichments, which is a good indicator of its performance in the absence of recall based on a complete gold standard.

Results of the evaluation are presented in Table 2. A first look at these, in particular the strict precision, shows a divide between two groups: EF and TEL (group A), and BgLink, Pelagios, VocMatch and Ontotext (group B). Tools in group A enrich records based only on metadata fields which typically contain (semi-) structured information (e.g., dc:creator) while tools in group B enrich using fields with any sort of textual description (e.g., dc:description). In semi-structured metadata fields, the difficulty of identifying the right named reference is lower since 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) often obey 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).

**Group A.** The tools from **EF** and **TEL** rank first and second on relaxed and strict precision. Besides the fact that they focused enrichments mainly to semi-structured fields, they benefit 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 [4]). EF results drop to second place for strict precision since in case of ambiguity, the tool cannot select the right entity. A typical example is place references that may correspond to different levels of administrative division with the same name. TEL 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 classification (e.g., 'feature type' in GeoNames) or demographic information as indicators for the relevance of an entity. Both tools do not take into account the historical dimensions of the object when selecting a geographical entity. For example, some objects from the 18th century with the named reference "Germania" are enriched with "Federal Republic of Germany" in EF. This can be seen as an avoidable side effect of using GeoNames, which mostly contains contemporary places. Finally, the results

confirm previous findings that some incorrect enrichments could be avoided if the language of the metadata was taken into account [2].

**Table 2.** Precision, Pooled recall and F-measure results.

| Tools | Annotated Enrichments (% of full corpus) | Precision | | Max Pooled Recall | Estimated Recall | | Estimated F-measure | |
|---|---|---|---|---|---|---|---|---|
| | | Relax. | Strict | | Relax. | Strict | Relax. | Strict |
| EF | 550 (31.3%) | 0.985 | 0.965 | 0.458 | 0.355 | 0.432 | 0.522 | 0.597 |
| TEL | 391 (22.3%) | 0.982 | 0.982 | 0.325 | 0.254 | 0.315 | 0.404 | 0.477 |
| BgLinks | 427 (24.3%) | 0.888 | 0.574 | 0.355 | 0.249 | 0.200 | 0.389 | 0.296 |
| Pelagios | 502 (28.6%) | 0.854 | 0.820 | 0.418 | 0.286 | 0.340 | 0.428 | 0.481 |
| VocMatch | 100 (05.7%) | 0.774 | 0.312 | 0.083 | 0.048 | 0.024 | 0.091 | 0.045 |
| Ontotext v1 | 489 (27.8%) | 0.842 | 0.505 | 0.407 | 0.272 | 0.202 | 0.411 | 0.289 |
| Ontotext v2 | 682 (38.8%) | 0.924 | 0.632 | 0.567 | 0.418 | 0.354 | 0.576 | 0.454 |

**Group B**. **Pelagios** has the best strict precision in group B, and its relaxed precision is slightly below BgLinks'. The fact that Pelagios is specialized for place name enrichments certainly helped achieving this. Its target vocabulary is smaller and more specialized than the datasets used by other tools, which makes it able to apply place-specific heuristics. This can explain why in terms of deviation between relaxed and strict precision it performs similarly to TEL and EF, which apply rules and target datasets that depend on the type of the entity expected to be found in certain fields. The most common reasons for incorrect or partial enrichments in Pelagios are related to issues with disambiguating between target entities. It does disambiguation, but the Wikidata target vocabulary that it exploits does not yet provide the necessary demographic information that it (as TEL) uses as indicator for the relevance of an entity. For example, "Siberia" is enriched with a place in California[25]. Pelagios also applies fuzzy matching between the named reference and the labels of the target entity, which leads to enrichments across different types of nouns, such as "people" with Peoples[26], a place in the U.S. Additionally, even though Pelagios aims at enriching old place names, it had issues determining whether an entity actually corresponds to the time frame of the description. The disambiguation problems did not significantly impact 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, to be compared with an average of 50% for other tools in group B[27]).

The two **Ontotext** versions perform differently, due to the fact that v1 applied enrichment only to objects with dc:language "en" and uses NLP methods for English as an attempt to increase precision. As a matter of fact nearly 100% of v1 enrichments were also detected by v2 but v1 discarded about half of the ones detected in v2. Yet

[25] http://sws.geonames.org/5395524/
[26] http://sws.geonames.org/4303909/
[27] see Appendix A of [12] for the complete distribution of enrichments per property.

performance was reduced overall since dc:language gives the language of the object not that of metadata. A great amount of enrichments were identified for non-named references like verbs (e.g. think), adverbs (e.g. viz.), adjectives (e.g. valid), abbreviations (Mrs), simple nouns (e.g. purpose), etc., which do not really contribute to improving the description of objects and sometimes lead to wrong enrichments. For the remainder of the enrichments, Ontotext shows a good performance.

**BgLinks** appears just below Ontotext v2 and above Ontotext v1. These tools are the ones that share the biggest number of enrichments, which partly explains the proximity in their performance. A closer look shows that enriching acronyms is a particular challenge for BgLinks. Very few of these were correct. BgLinks performs significantly better in determining the right references within the text to enrich, compared to Ontotext and is also successful at enriching more complex named references. This comes from 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 are produced for terms that denote entities without an exact semantic equivalent in the target dataset. This is an issue for all tools, but is particularly found in group B, as references to such entities are more common in long text descriptions than in normalized or structured fields.

**VocMatch** had the lowest performance. The fact that it was exceptionally difficult for raters to identify the actual portion of the text that served as clue for the enrichment made it hard to assess its correctness. As already hinted, VocMatch's pooled recall is impacted by its use of specialized vocabularies not used by the others, and for which no coreference links were available to reconcile them with other enrichments. A closer look shows that some incorrect enrichments come from matching against all terms available in the target vocabularies, without disambiguation. An example is the word "still" as part of the term "still image". This approach is much more effective when applied to semi-structured fields like dc:subject or dc:type; this is quite visible when comparing VocMatch and EF, which applies the same methods as VocMatch but only to semi-structured fields. Subsequent investigations have shown that using only semi-structured fields, VocMatch reaches 86.7% relaxed precision.

## 7 Conclusion

Our experiment is the result of an effort to gather representatives of several projects over a couple of months. While we have stumbled over issues in the evaluation procedure that in retrospect seemed obvious, the exercise has proven to be fruitful for exchanging practitioners' perspectives on the assessment of enrichment tools. Our work is important for users from the CH communities and/or owner of digital library applications using generic frameworks like GERBIL, which offers many options but little domain guidance and may lack test datasets that fit specific cases. As a matter of fact, we find it useful to articulate and share the following recommendations regarding the evaluation process:

1. **Select a dataset for your evaluation that represents the diversity of your (meta)data:** Covering language diversity, spatial dispersion, subjects and domains.

2. **Building a gold standard is ideal but not always possible:** Build a reference set of correct alignments manually if you have sufficient time and human resources, otherwise annotate the enrichments identified by the tool being evaluated or other enrichment tools. The trade-off is that the latter option does not allow one to obtain absolute recall figures.
3. **Consider using the semantics of target datasets:** When datasets are connected by coreference links, these may be used in a process that "normalizes" enrichments to get a more precise view on how they compare across tools, or to reuse a gold standard from another evaluation.
4. **Try to keep balance between evaluated tools:** Some of the corpus creation strategies can result in a bias against some tools. Make sure bias is recognized and properly related to your evaluation strategy's motivations.
5. **Give clear guidelines on how to annotate the corpus:** Guidelines should be simple but still complete enough for raters to deliver appropriate judgements. Consider having examples for the cases that may raise the most doubt. Try testing the guidelines with raters early in the process.
6. **Use the right tool for annotating the corpus:** Choose or develop a tool that displays all the necessary information; respects your guidelines and guides raters to efficiently and effectively perform their task.

In addition, we want to share recommendations for enrichment tools:

1. **Consider applying different techniques depending on the values used with the enriched property**; e.g., semi-structured or textual descriptions, or values generally containing references for places, persons or time periods.
2. **Matches on parts of a field's textual content may result in too general or even meaningless enrichments** if they fail to recognize compound expressions. This especially hurts when the target datasets include very general resources that are less relevant for the application needs.
3. **Apply strong disambiguation mechanism** that considers the accuracy of a name reference together with the relevance of the entity in general (looking as its data properties) and in particular, i.e., within the context of reference and use. For example, we observed that most tools would benefit from identifying and comparing the temporal scope of both records and candidate target entities.
4. For most if not all cases in the Europeana context, **concepts so broad as "general period" do not bring any value as enrichment targets**. Additional logic should be added to the enrichment rules so that they are not used to enrich objects.
5. Our evaluation has confirmed that **quality issues originating in the metadata mapping process are a great obstacle to get good enrichments** [2]. Enrichment rules designed to work on specific metadata fields (e.g., spatial coverage of an object) should be applied carefully when these fields can be populated with values that result from wrong mappings (e.g., publication places).

We hope that future evaluations of enrichment tools can benefit from our experience, as they are essential to improve the design of the tools themselves, and help make the applications built on these, such as Europeana, deliver better performance.

## References


1. Bunescu, R., Paşca, M. Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 9–16. (2006)
2. Stiller, J., Petras, V., Gäde, M., Isaac, A.: Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. In: Proceedings of the 5th International Conference on Cultural Heritage (EuroMed). (2014)
3. Agirre, E., Barrena, A., Lopez de Lacalle, O., Soroa A., Fernando S., Stevenson, M.: Matching Cultural Heritage items to Wikipedia. In: Proc. LREC 2012. Istanbul, Turkey. (2012)
4. Isaac, A., Manguinhas, H., Stiller, J., Charles, V.: Report on Enrichment and Evaluation. The Hague, Netherlands (2015), http://pro.europeana.eu/taskforce/evaluation-and-enrichments.
5. Usbeck, R., Röder, M., Ngonga Ngomo, A., et al.: GERBIL: General Entity Annotator Benchmarking Framework. In: Proc. 24th WWW conference. ACM (2015). http://doi.acm.org/10.1145/2736277.2741626
6. Griffiths, J., Basset, S., Goodale, P., et al.: Evaluation of the second PATHS prototype. Technical report, Paths Project (2014)
7. Olensky, M., Stiller, J., Dröge, E.: Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. In: Metadata and Semantics Research. Springer, Berlin (2012)
8. Monti, J., Monteleone, M., Buono, M., et al: Cross-Lingual Information Retrieval and Semantic Interoperability for Cultural Heritage Repositories. In: RANLP 2013. Hissar, Bulgaria (2013). http://aclweb.org/anthology/R/R13/R13-1063.pdf
9. Petras, V., Bogers, T., Toms, E., et al: Cultural Heritage in CLEF (CHiC) 2013. In: 4th International Conference of the CLEF Initiative. Valencia, Spain (2013). http://dx.doi.org/10.1007/978-3-642-40802-1_23
10. Dobreva, M., Chowdhury, S.: A user-centric evaluation of the Europeana digital library. In: The Role of Digital Libraries in a Time of Global Change. Springer, Berlin (2010).
11. Tordai, A., van Ossenburg, J., Schreiber, G., et al.: Let's agree to disagree: on the evaluation of vocabulary alignment. In: Proc. 6th K-CAP. ACM (2011)
12. Isaac, A., Manguinhas, H., Charles, V., Stiller, J., et al: Comparative evaluation of semantic enrichments. Technical report (2015). Report available at http://pro.europeana.eu/taskforce/evaluation-and-enrichments. Data archive available at: https://www.assembla.com/spaces/europeana-r-d/documents?folder=58725383
13. Freire, N., Borbinha, J., Calado, P., Martins, B.: 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
14. Charles, V., Freire, N., Antoine, I.: 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 (2014).
15. Fleiss, J.L.:Statistical methods for rates and proportions second edition. In: Wiley Series in probability and mathematical statistics. Chapter 13 p. 212-236.
16. Randolph, J.: Free-Marginal Multirater Kappa (multirater κ free): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Joensuu University Learning and Instruction Symposium 2005*, Joensuu, Finland, October 14-15th, 2005. http://eric.ed.gov/?id=ED490661
17. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Cambridge University Press (2008). http://nlp.stanford.edu/IR-book/pdf/08eval.pdf