Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy

Marlies Olensky, Juliane Stiller, and Evelyn Dröge

Humboldt-Universität zu Berlin, Berlin School of Library and Information Science
Dorotheenstr. 26, 10117 Berlin, Germany
{marlies.olensky,juliane.stiller,evelyn.droege}@ibi.hu-berlin.de
http://www.ibi.hu-berlin.de

Abstract. Cultural heritage information systems offer access to objects coming from museums, archives and libraries. To enhance retrieval performance and access across languages, metadata is enriched with controlled vocabularies or other datasets with structured information. During this process many pitfalls occur which lead to wrong or poor enrichments thus decreasing the user experience. Taking the use case of Europeana, this paper investigates the extent of enrichment flaws and their causes. A categorization of these deficiencies is proposed as well as a strategy to avoid common enrichment mistakes.

Keywords: Semantic and multilingual enrichment, problem diagnosis, enrichment strategy, Europeana

1 Introduction

When a user in Europeana[1], the single access point to European cultural heritage, searches for *poison* in the collections provided by Swiss institutions, she will find photographs from India and Indian movie covers. The relevance of the retrieved documents to the query is not comprehensible. A deeper investigation reveals that retrieved objects were automatically enriched with the term *poison* and its multilingual equivalents. In Latvian *poison* means *Inde* which is the same keyword the French-speaking domain expert gave the objects to describe its content: India. This striking example shows one of the potential pitfalls in semantic and multilingual enrichments if no strategy is applied.

Semantic and multilingual enrichment of information objects is a process with the goal to enhance the retrieval experience for the user. Digital libraries like Europeana aggregate a vast amount of cultural heritage information objects from different countries and in different languages; semantic and multilingual enrichment of metadata supports disambiguation in such multilingual environments. Synonyms, homonyms and cross-lingual ambiguities are the main reasons for improper search results and consequentially a poor user experience. Enrichment of metadata with structured information resources can support the disambiguation on the one hand and the enhancement of multilingual search results on

[1] http://europeana.eu/

the other hand. However, the question is: what makes enrichments usable and valuable and how can we ensure that enrichments are correct? In this paper, we show the importance of applying a semantic and multilingual enrichment strategy. We identify the influencing factors that lead to successful, correct and in the best case useful enrichments. Europeana serves as use case. From the investigated enrichments we derive a set of factors and rules that should constitute an enrichment strategy which can be applied across domains.

The paper is structured as follows: section 2 elaborates on related work on semantic and multilingual enrichment and its evaluation; section 3 describes the use case Europeana and the applied methodology; section 4 presents the diagnosis of enrichment problems, section 5 derives a generalized strategy from the findings and section 6 concludes the paper.

2 Background

The paper focuses on semantic and multilingual enrichment which can also be referred to as semantic and multilingual tagging [1]. The Europeana Data Enrichment Requirements [2] define data enrichment as the overall process of enrichment, cleaning and normalization of collections with insufficiently rich metadata to be carried out by the data ingestion team. This includes de-duplicating objects across collections, adding string-valued fields to metadata records and linking objects to other internal or external knowledge sources and or to other objects. We define the term semantic and multilingual enrichment as the process of identifying concepts, places, agents and time periods in the metadata of a cultural heritage object (CHO) and linking them to a knowledge resource (such as ontologies, thesauri or other controlled vocabularies) by adding the respective labels and URIs from these vocabularies to the CHO. For example, a CHO might hold the term *London, UK* as a value in its metadata field coverage and enriching this object would mean finding the place *London* in the UK in an appropriate vocabulary (a suitable one would be GeoNames[2]); adding the label / URI of the correct *London* to the metadata would be a semantically correct and valuable enrichment, adding labels in other languages would be a multilingually correct and valuable enrichment.

Semantic enrichment experiments have been carried out in the Europeana-Connect project[3] where the Free University of Amsterdam (VUA) used their Amalgame tool[4] to enrich metadata values by mapping them to existing vocabularies. The Amalgame tool is basically a vocabulary alignment tool; to use it as enrichment tool they created a temporary vocabulary from the metadata values and in a second step mapped this vocabulary to existing ones. In principle, the alignment and the enrichment processes are quite similar to each other, as they both involve a matching process, and therefore might use similar quality evaluation methods. Tordai et al. [3] checked all alignments manually in order

[2] http://www.geonames.org/
[3] http://www.europeanaconnect.eu
[4] http://semanticweb.cs.vu.nl/amalgame/

to evaluate their quality. As a manual check is not feasible for large vocabularies, they developed a disambiguation technique to improve the precision of alignments where the parent and/or the child match of the respective term is taken into account. Even though this will increase the quality, it still does not provide a method to evaluate the quality of the alignments. To align vocabularies semantically, the project EuropeanaConnect [4] identifies six characteristics that influence the mapping of vocabularies and that also need to be considered during the enrichment process: lexical variance of the labels, use of preferred/alternative labels, number of labels, use of diacritics, nature of hierarchy and multilinguality. Furthermore, it points out that thesauri or controlled vocabularies for alignments should be chosen according to their institutional and collection adequacy, in terms of scope and uptake [5]. With regard to vocabulary evaluation, a lot of research has been conducted recently, with the main focus to find categories which allow for comparison of knowledge organization systems or other controlled vocabularies. Approaches of Vrandecic [6] in measuring ontology quality or of Mader [7] for choosing SKOS quality criteria are more elaborated as both have additionally identified evaluation criteria regarding the completeness or consistency of vocabularies, among others. Still, even if the result of the evaluation suggests that one vocabulary suits best for the enrichment task, this may not be the case in a specific context. If, for example, a vocabulary is too general it may not be as appropriate as a vocabulary that is less linked to other vocabularies but more precise than the first one.

3 Use Case - Europeana

Europeana is a single access point to digitized cultural heritage coming from libraries, archives, museums and audio-visual archives. Currently, Europeana provides access to over 23.5 million objects (images, textual objects, sound and audiovisual files). More than 2,200 institutions based in 33 different countries contributed to the aggregated content representing the diverse and heterogeneous cultural objects of Europe. This poses a challenge as each record has two multilingual dimensions: the language of the object and the language of the metadata, both not necessarily matching. The goal of Europeana is to provide access to this material in different languages and to unlock the cultural heritage. A means to reach this objective is the semantic and multilingual enrichment of Europeana's content carried out by the Europeana Office. Table 1 shows the enriched metadata fields and the datasets[5] used for the enrichment. All of them are linked open data resources which can be either described as controlled vocabularies or datasets representing structured information (e.g. DBpedia). At the time of writing this paper over 16 million records were enriched with either one or more of these labels.

[5] Two of the datasets, DBpedia and GeoNames, were analyzed by [8] with the qSKOS tool: DBpedia concepts are never documented, 77,062 concepts (∼10%) have no associative or hierarchical relationships and 3,058 concepts (∼0.4%) are not labeled. GeoNames concepts have no semantic relations at all. Both vocabularies are nevertheless used for enrichments.

[6] `http://www.eionet.europa.eu/gemet/`

Table 1. Controlled vocabularies and datasets with structured information used to enrich Europeana's metadata fields

Vocabulary	**Tag type**	**Enriched metadata fields**
GEMET Thesaurus[6]	Concept	dc:subject dc:type dcterms:alternative
DBpedia[7]	Agent	dc:contributor dc:creator
Semium Time Ontology[8]	Period	dc:date dc:coverage dcterms:temporal
GeoNames[9]	Place	dc:coverage dcterms:spatial

The AnnoCultor Tagger has been used to enrich objects in Europeana[10]. In terms of quality control, enrichments were applied to certain sets of metadata fields to avoid mislabeling. For example, a geographic location occurring as a subject keyword was not enriched with GeoNames. Furthermore, the tagging tool only applied the European subset of cities in GeoNames to avoid ambiguous matches with cities outside of Europe. In general, the enrichment rules are not documented but can be extracted from the actual source code[11]. Although Europeana requirements [1, 2] point out the need to evaluate the enrichment results before they are included in the Europeana metadata base, this requirement was disregarded during the enrichment process.

To get an overview of the areas of concern for semantic and multilingual enrichment, a purposeful sample of 200 records enriched with controlled vocabularies was pulled from Europeana. The goal was not the selection of a statistical representative sample but the aggregation of insightful and diverse enrichments across providers, languages and metadata fields. For each of the four tag types, 50 metadata records were analyzed and the enrichment process reproduced. Of value here are the so-called information-rich cases offering insights into the pitfalls which can occur during enrichments [9, p. 230]. The analysis was performed with focus on executed enrichments and missed ones were touched peripherally. Deducing causes for missed enrichments is mostly impossible and reasons can be

[7] http://dbpedia.org/

[8] http://semium.org/time.html

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

[10] A thorough explanation of the enrichment process can be found here: http://europeanalabs.eu/wiki/EDMPrototypingTask21Annocultor which is a copy of the following blog post: http://borys.name/blog/semantic_tagging_of_europeana_data.html

[11] http://europeanalabs.eu/browser/europeana/trunk/tools/trunk/annocultor/src/main/java/eu/annocultor/converters/solr/BuiltinSolrDocumentTagger.java

multifaceted. Therefore, we refrained from a deeper analysis, acknowledging that omitted enrichments can decrease retrieval performance and user experience.

4 Enrichments - Problem Diagnosis

In this section, different reasons for the error-proneness of enrichments[12] in the presented use case Europeana will be listed, grouped in categories and described.

4.1 Incorrect Metadata

When an object is enriched, it can introduce semantic errors, simply because its metadata is incorrect. This includes mapping errors at ingestion time, i.e. mapping provider metadata fields to wrong Europeana metadata fields, typographical mistakes that were made at indexing time or in the worst case wrong metadata assigned at indexing time. Irrespective of the reason for incorrect metadata, the insufficient metadata quality is the basis for wrong, and in most cases absurd enrichments and can also lead to omission of potential enrichments. A measure to avoid these enrichments is to have a data cleaning process installed at ingestion time. This corresponds to the functional requirements of data enrichment where the need for data cleaning is emphasized [2].

4.2 Inconsistent Structure of Metadata

Related to incorrect metadata is the inconsistent structure of metadata in Europeana, which causes major problems at enrichment time. The following three aspects of inconsistent metadata structure, again, correspond to the functional requirements of data enrichment which state the need for data normalization [2].

Inconsistent name structure. We found incorrect enrichments caused by the names of creators and contributors not being structured as last / middle / first names or identified as named entities. For example, the tagging tool enriched any value in a name field with a matching agent in DBpedia. Therefore, the *[Copy of request and confirmation of special dispensation granted to the friars of the Irish Franciscan province in 1663.]*[13] by *Bongiorno, Michelangelo, Fr* and *Docherty, Anthony, Fr* was enriched with the "wrong" *Michelangelo*[14] *(Buonarrotti)*. Defining a consistent structure for names would increase the enrichment precision of agents enormously. The structure could follow common bibliographic conventions, like *Last_name, first_name middle_name*. First and middle names could optionally be abbreviated by the respective initial(s). Multiple agents should be distinguished by a semicolon. As Europeana does not have such a structure implemented, the safest way to enrich agents would be to use exact matches only, which would lead to a decrease in the amount of enrichments

[12] Europeana enrichments can be found in the grey box as *Auto-generated tags* in the full view.
[13] http://europeana.eu/portal/record/09714/B179B7E51E87F4EA7CE5E1472AABD19F60252AB4.html
[14] http://dbpedia.org/page/Michelangelo

but also to an increase in quality.

Inconsistent date structure. Our investigation showed that the date enrichment with the time vocabulary caused the least problems. However, we did find objects that were not enriched or not fully enriched due to inconsistent date structure. Dates and time periods can have different formats (numeric, numeric and literal or literal characters only). The inconsistent date structure is similar to the structure of names. A standardized format for dates, e.g. YYYY-MM-DD, should be used. Also, a clear structure for dates BC and AD as well as time durations need to be agreed on. Multiple values in one field must be clearly indicated. One interesting example, where no time labels were enriched although the object holds a valid date and historical period, is *Fragment eines ionisches Kapitells*[15] with the date *285 - 280 v. Chr.* and the time period *Hellenistisch*. A correct and valid enrichment would have added the label for the first millenium BC[16]. An additional benefit would be the label for the *hellenistic period*[17], if German labels were available in the Time Vocabulary.

Inconsistent field structure / refinements. In the specification for the Europeana Semantic Elements [10], the current metadata model in Europeana, the field *dc:coverage* should be used to describe "the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant". It therefore comprises a temporal or a spatial aspect and should be refined to *dcterms:temporal* or *dcterms:spatial* where applicable. We found objects that held the values about their temporal and spatial coverage in the same field (*dc:coverage*) instead of splitting these data to the correct refined elements. These inconsistencies can lead again to missing out on potential enrichments, as additional values in the same field are disregarded by the enrichment tool. Additionally, it is important to get the structure of the fields and their refinements straight, in order to choose the correct fields to enrich with a certain vocabulary. For example, our investigation showed that the video *Akten werden hinausgeworfen*[18] holds *Wien* and *20. Jahrhundert* in the field *dc:coverage*. Had this coverage been distinguished into the temporal and the spatial aspect, the tagging tool could have identified *20. Jahrhundert* (20th century) as time period as well as *Wien* as place and enriched them with the respective labels. Here, a consistency check at ingestion time should be carried out in order to ensure that metadata is accurately refined and represented in appropriate granularity.

[15] http://europeana.eu/portal/record/15502/AEED91B8CF6FCF1D3C81EB71E471108BD82D83F6.html
[16] http://semium.org/time/BC1xxx
[17] http://semium.org/time/greek_hellenistic
[18] http://europeana.eu/portal/record/00901/57525CB2B138706A9094714E76C38D7C2B41FF5D.html

4.3 Context Disregarded

Another reason for erroneous enrichments is the syntactically correct but semantically incorrect matching of labels. The video with the title *Renault*[19] holds a contributor *Daniel Richter*. In this case, *Daniel Richter* is a French trade unionist and not the German artist who was found as matching DBpedia label for *Daniel Richter*[20]. Another example is the incorrect matching of places that exist in more than one country, e.g. *Córdoba* in Spain and Argentina or *Guadalajara* in Spain and Mexico. As noted in the description of the use case, Europeana intentionally only included the European subset of GeoNames in order to avoid mismatches with ambiguous places outside of Europe. Yet, we found two objects[21], where exactly this restriction caused incorrect enrichments. All three examples prove that if the enrichment tool had considered the context of objects, i.e. other metadata and broader or narrower labels, in the matching processes, the persons or places could have been disambiguated and correct enrichments could have been made.

4.4 Choice of Enrichment Fields

The decision on the enrichment fields and the corresponding vocabularies depends on quality control and on considerations what value a vocabulary can add to a certain metadata field. It is debatable if *dc:type* is a good choice for concept enrichment, as *dc:type* does not describe the concept an object is about. We found objects that are of type *book, photo, video, map, patent,* etc. and were enriched with the respective labels from the GEMET thesaurus. These enrichments add multilingual labels and therefore enhance the multilingual retrieval experience for the user. Yet, they do not add value in terms of semantics. Therefore, these enrichments optimize recall but also create a lot of noise.

4.5 Non-domain Specific Vocabulary

Choosing the right enrichment vocabulary is not a trivial task. Especially across domains, terms occur to be ambiguous and the problem rises exponentially in a multilingual environment. For example, in German the term for *print* is *Druck*. In physical science, *Druck* also means *pressure* and is therefore one of the many homonyms in the German language. In Europeana, this ambiguity leads to poor enrichments as many records are indexed with the term *Druck* and then wrongly enriched with the term *pressure* in the GEMET thesaurus[22]. Domain-specific vocabulary introduces certain implications even if the term as such is not ambiguous. An example is the term *paper* which, in cultural heritage, is a type of material used for printing and drawing. In environmental science, *paper* is

[19] http://europeana.eu/portal/record/04802/F51D452365426ECD303C40F87134A383B91D89C3.html

[20] http://dbpedia.org/page/Daniel_Richter

[21] http://europeana.eu/portal/record/10102/BA5342F824A2CF7EAD1F7130FC5EDFFFBB2BD2E2.html, http://europeana.eu/portal/record/00901/7D1F2919B80CE8BF070CE1695BF304473FE07419.html

[22] http://europeana.eu/portal/record/92060/2B66D3FACA9A0047916E51E0C0556BECF9259142.html

mainly understood to be an industrial product with the emphasis on the production of this resource[23]. Enrichment flaws like this can be avoided by choosing a domain-related vocabulary.

4.6 Named Entity Treatment

Named entities always require special treatment as they carry particular characteristics such as being predominantly language-agnostic or at least require specific translations. Therefore, in retrieval and natural language processing, the first step is to identify these named entities. In the cultural heritage domain, named entities relate to geographic locations, names or time periods but also work titles of books or performances. The dimension of named entities in this domain needs to be considered to avoid deficient enrichments.

4.7 Cross-lingual Ambiguity

When dealing with cross-lingual collections and records, the issue of multilingual ambiguity needs to be addressed. With a growing number of languages, the potential for having the same term with totally different meanings in different languages rises. This is a pitfall for enrichments which do not acknowledge the language of the metadata. Terms which are the same across languages but with completely different meaning are sometimes referred to as "false friends" in language learning and this term is very suited to be applied here. One example are German records dealing with *power* (in German: *Strom*) erroneously enriched with the term *tree*[24]. The explanation is the Czech word for *tree*: *strom*. In German, this term means *power*, the enrichment presumed that *strom* is a Czech word meaning *tree*. This example might appear like a one-off but in a portal with records in more than 23 different languages, this is an area of concern. Avoiding this means to identify the language of the metadata and map only terms with the appropriate language.

4.8 Weighting of Enrichments

It is obvious that the enrichment of terms makes the associated documents much more retrievable across languages. An enriched term has a lot of influence on the retrievability of documents. If an object has many keywords, choosing only one of them for enrichment can be counter-productive. One example is the enrichment of the word *history* for a record which has very specific keywords attached to it in Estonian and its English translations[25]. It is disputable whether such an enrichment is useful. In total, almost 80,000 records[26] were enriched with *history* and its translation equivalents. This is adding to the pool of records which are

[23] http://www.eionet.europa.eu/gemet/concept?cp=6023&langcode=en&ns=1
[24] http://europeana.eu/portal/record/92063/B1CD66B8D6FB2FF6CC33B0279C81571572F2F90B.html
[25] http://europeana.eu/portal/record/92067/28296EA118D9DF7E307F3B51E3C552F5A2D3E1F1.html
[26] http://europeana.eu/portal/search.html?query=enrichment_concept_label:histoire

retrieved as they have *history* somewhere in their metadata. In general, every record in Europeana is related to history. In particular, if *history* is only one aspect of the resource, the danger is to decrease precision in the search results and thus create noise. Nevertheless, such an enrichment might still be relevant in minor languages with few objects. Anyway, enrichments should be weighted according to their significance for the record.

4.9 Workflow

Most of the items listed above are also of concern with regard to the enrichment workflow. The workflow summarizes the rules and strategies in place to balance out poor metadata quality and vocabulary restrictions. Additionally, the choice of the mapping or enrichment tool is crucial as it should be able to handle special cases.

5 Framework of Strategies for Semantic and Multilingual Enrichments

By generalizing the findings from our case study, we found that the consequences of these problem areas are always the same: enrichments are semantically or multilingually wrong, objects have not been enriched with the most useful labels or objects were not enriched at all. The areas of concern that influence an enrichment strategy can be divided into three different levels: metadata, vocabulary and workflow (Table 2).

On the metadata level, the quality and structure of the underlying metadata is crucial. When deciding on an enrichment strategy, one needs to be aware of the metadata quality. A data cleaning and standardization process should be applied at ingestion time and ideally, metadata quality is assessed and measured by a score. Afterwards, a minimum level of quality can be defined and only records above this score will be enriched. In the standardization process, syntactic rules on how to format values within metadata fields are defined, thus ensuring a common structure.

On the vocabulary level, an enrichment strategy needs to specify what collections to enrich by what vocabularies. In the cultural heritage domain, you will hardly find a thesaurus or controlled vocabulary that can be applied for any collection available. Yet, in order to make most of the enrichments, one has to ensure that the right vocabulary is chosen for the right purpose. The pros and cons of selecting a domain-specific vocabulary versus a non-domain specific one must be weighted. A non-domain specific vocabulary might be available in more languages with a broader coverage; however, it probably will hold more ambiguous terms. The choice of the vocabulary also influences the enrichment workflow.

On the workflow level, several aspects need to be taken into account. The semantics of metadata fields as well as the semantics of the actual values should be considered for enrichment. For example, a birth date or place of birth could

Table 2. Framework of strategies for semantic and multilingual enrichments

Level	**Areas of concern**	**Strategic execution**
Metadata	Metadata quality	Quality score for metadata, no enrichments below the score, data cleaning process
Metadata	Structure of metadata	Data normalization e.g. surname forename, rules for syntax, validate fields against a schema, consistency check for field refinements
Vocabulary	Choice of vocabulary	Choose domain-specific vocabulary or a subset of a vocabulary, exclusion of parts of the vocabulary
Vocabulary	Scope of enrichment	Choose fields to be enriched with a specific vocabulary or even limit enrichment to subsets or specific collections
Workflow	Semantics	Disambiguate metadata values and use context
Workflow	Named entities	Apply automatic named entity recognition
Workflow	Cross-lingual ambiguities	Metadata records and enrichment term need to have the same language
Workflow	Weighting of enrichments	If multiple values in one metadata field are enriched, they should be weighted according to their relevance
Workflow	Matching rules	Use exact matches, include variants from the controlled vocabulary, rule on how to enrich multiple values in a field
Workflow	Quality assurance	Quality checks (automatically or manually) before enrichments go live
Workflow	Quality assessment	Assess the scope of the enrichments with regard to their occurrence in search results

be leveraged to identify the correct match for agents carrying the same name. Applying automatic named entity recognition, especially for agents and titles, will avoid enriching named entities with wrong tags which match only parts of the term. To avoid cross-lingual ambiguities, one should only allow enrichments of objects where the language of the metadata and the enriching labels are the same. If a metadata field holds multiple values, a weighting of these must be carried out according to their relevance for the object.

Decisions made on the metadata and the vocabulary level influence the enrichment workflow. The better the underlying metadata quality is and the better it is standardized, the less strict the matching rules can be. For example, the better structured the metadata is, the less important it is to use exact matches; the less structured the metadata is, the more important it is to include (spelling) variants from the vocabulary. Furthermore, the choice of vocabularies and the limitations one sets to the fields / objects / collections to be enriched influence the grade of complexity of the matching rules. For example, explicit rules must define how to enrich multiple values in a field. Applied quality assurance (manual, automatic or semi-automatic in order to check whether the enrichments are correct) can also influence the matching rules. It is obvious that without any quality checks, the matching rules must be as conservative as possible. This implies decreasing the number of enrichments, but at the same time increasing the quality of the actual enrichments. Finally, the scope of the enrichments with regards to their occurrence in search results should be assessed to know what the influence of erroneous enrichments might be for the user.

6 Conclusion

When implementing a strategy for semantic and multilingual enrichments, one needs to be aware of the different aspects which impact the quality of the enrichment result. The development of such a strategy implies determining deficiencies in the metadata quality. In addition, certain circumstances, such as access restrictions, can limit the vocabulary choices. To keep the impact of these two factors small and redeem certain shortcomings, workflow and enrichment tools need to be developed. The quality of the metadata and the adequacy of the vocabulary on the one hand and the elaborateness of the workflow and enrichment rules on the other hand tend to be inversely correlated. The more precise and targeted the enrichment rules are, the less impact the flaws in metadata quality and the vocabulary choice have. Thus, lack of quality in the records and the vocabulary can be balanced out with a reasonable workflow strategy and enrichment rules.

In future work, it needs to be determined to which degree the different areas of concern influence the enrichment and consequently the retrieval results. Recall and precision as the common measures of retrieval effectiveness are means to determine the impact of enrichments. Poor enrichments will influence both figures negatively. Either relevant documents cannot be identified anymore among the enlarged pool of retrieved records or none of the retrieved documents are rel-

evant, both resulting in a bad user experience. Furthermore, poor enrichments impact search results in different degrees. A relevant document hidden among less significant ones is much to a provider's regret but might not attract the user's attention in a negative way. Whereas, if an inappropriate document is found based on a semantically wrong enrichment, the mistake is more severe which leads to consequences that are counterproductive to the goal of a cultural institution to carefully curate cultural heritage content. It is beneficial to set priorities in the enrichment strategy to ensure the impact of poor enrichments is as small as possible.

To measure the visibility and impact of poor enrichments not only their number is crucial but additionally, the frequency of the documents occurring in the search results based on these deficient enrichments needs to be included. An enrichment approach based on the quantity of enriched terms is like shooting oneself in the foot, as it increases the potential impact of poor enrichments. To avoid this, quality should outweigh the quantity in an enrichment strategy and an assessment of the process is inevitable. In severe cases, an omitted enrichment can be the better choice.

Acknowledgements. We are very grateful to Antoine Isaac for his feedback. This research was partly financed by the projects Europeana v2.0[27] and DM2E[28] under the ICT PSP Work Programme.

References

1. Isaac, A.: EDM Prototyping, 2.1. Enrichment of EDM data. (2011) http://www.europeanalabs.eu/wiki/EDMPrototypingTask21
2. Isaac, A.: Functional Requirements: Data Enrichment. (2010) http://europeanalabs.eu/wiki/SpecificationsDanubeRequirementsEDMDataEnrichment
3. Tordai, A., van Ossenbruggen, J., Schreiber, G.: Combining Vocabulary Alignment Techniques. In: K-CAP '09, Proceedings of the 5th Int. Conf. on Knowledge Capture, pp. 25–32, ACM, New York (2009)
4. EuropeanaConnect: Milestone 1.2.1: Specification of preferred terms identification methodology. (Internal document). (2010)
5. EuropeanaConnect: D2.3.1 Multilingual mapping of schemes and vocabularies. (Internal document). (2010)
6. Vrandecic, D.: Ontology Evaluation. KIT, Karlsruhe (2010) http://digbib.ubka.uni-karlsruhe.de/volltexte/1000018419
7. Mader, C.: Quality Assurance in Collaboratively Created Web Vocabularies. In: PhD symposium of ESWC2012, Heraklion, Greece (2012)
8. Mader, C., Haslhofer, B., Isaac, A.: Finding quality issues in SKOS vocabularies. In: TPDL 2012, International Conference on Theory and Practice of Digital Libraries. (in press) (2012)
9. Patton, M.Q.: Qualitative Research and Evaluation Methods. Sage Publications (2002)
10. Europeana: ESE specifications 3.4.1. (2012) http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57

[27] http://pro.europeana.eu/web/europeana-v2.0
[28] http://dm2e.eu/