DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 4

Semantic Enrichment of a Multilingual Archive with Linked Open Data

Max De Wilde <madewild_at_ulb_dot_ac_dot_be>, Université libre de Bruxelles(ULB), Information Science Department
Simon Hengchen <simon_dot_hengchen_at_helsinki_dot_fi>, University of Helsinki (UH)

Abstract

This paper introduces MERCKX, a Multilingual Entity/Resource Combiner & Knowledge eXtractor. A case study involving the semantic enrichment of a multilingual archive is presented with the aim of assessing the relevance of natural language processing techniques such as named-entity recognition and entity linking for cultural heritage material. In order to improve the indexing of historical collections, we map entities to the Linked Open Data cloud using a language-independent method. Our evaluation shows that MERCKX outperforms similar tools on the task of place disambiguation and linking, achieving over 80% precision despite lower recall scores. These results are encouraging for small and medium-size cultural institutions since they demonstrate that semantic enrichment can be achieved with limited resources.

1. Introduction

Libraries, Archives and Museums (LAM) are increasingly faced with budget cuts, compelling them to review their traditional ways to exploit their collections. Short-term results are often expected by funding bodies, and cultural heritage institutions are therefore encouraged to gain more value out of their data by linking them to existing knowledge bases. In this context, semantic enrichment techniques such as named-entity recognition (NER) and entity linking (EL) have attracted attention since they allow these institutions to enrich their collections semantically with few resources. For LAM, the perspective of freely reusing existing knowledge to map their collections to the Web represents a great opportunity. 1

Using digital methods also allows researchers to tackle larger datasets. While larger corpora do not induce better research *per se*, they allow to have a bird’s eye view on the context of a particular project. In this paper, we evaluate the relevance of NER and EL for an archive of OCRised Belgian periodicals. We focus on historical locations because our analysis, detailed in Section 3, shows that users are mainly interested in them. However, the methodology described in this paper could be extended to other types of entities present in knowledge bases. 2

The remainder of this paper is structured as follows: Section 2 discusses related work, sections 3 and 4 present our case-study and workflow respectively, Section 5 discusses the results obtained and Section 6 concludes the paper whilst providing new research tracks. 3

2. Related Work

Named-entity recognition and other information extraction techniques such as entity linking have been increasingly adopted by DH practitioners, since they help small institutions to enrich their collections with semantic information[1]. According to Blanke and Kristel, “semantically enriched library and archive federations have recently become an important part of research in digital libraries and archives” [Blanke and Kristel 2013]. This is illustrated by such projects as EHRI[2], an 8-million euros EU project focusing on Holocaust research, and CENDARI[3], a European Commission-funded project aiming to integrate digital resources for medieval and WWI history. Similarily, the Europeana Newspapers project[4] has been developing NER tools[5] specifically to process historical newspaper collections. 4

The growing of the Linked Open Data (LOD) cloud and the availability of free online tools have facilitated the access to information extraction for librarians, archivists and collections managers that are not IT experts but are eager to experiment with new technologies. The LOD Around The Clock project of the European Commission, for instance, was started to “help institutions and individuals in publishing and consuming quality Linked Data on the Web”.[6] Its main declared goal is to “continuously monitor and improve the quality of data links within the Linking Open Data cloud”. The existence of such large-scale incentives demonstrate the potential of LOD for the semantic enrichment of collections maintained in LAM. 5

A number of cultural institutions have experimented with NER and EL over the last decade. The Powerhouse Museum in Sydney has implemented OpenCalais within its collection management database, although no evaluation of the entities has been performed.[7] Lin et al. also explore NER in order to create a faceted browsing interface for users of large museum collections [Lin et al. 2010], while Segers et al. offer an interesting evaluation of the extraction of people, locations and events from unstructured text in the collection management database of the Rijksmuseum in Amsterdam [Segers et al. 2011]. Maturana et al. showed how LOD could be successfully integrated in a museum platform to enhance the experience of end users [Maturana et al. 2013]. Their innovative semantic platform MisMuseos, a meta-museum aggregating 17 000 works from seven Spanish cultural institutions, offers users a facet-based search module, semantic content creation and graph navigation. 6

In the specific domain of archives, Rodriquez et al. compared the results of several NER services on a corpus of mid-20th-century typewritten documents [Rodriquez et al. 2012]. A set of test data, consisting of raw and corrected OCR output, was manually annotated with people, locations, and organisations. This approach allows an evaluation of the different NER services against the manually annotated data. Their methodology was generalised for LAM by van Hooland et al. in the context of the *Free Your Metadata* project[8] [van Hooland et al. 2015], and extended to other languages such as French, by Hengchen et al. [Hengchen et al. 2015]. The BBC also set up a system to connect its vast archive with current material through Semantic Web technologies [Raimond et al. 2013]. 7

Bingel and Haider compared the performance of various entity classifiers on the DeReKo corpus of contemporary German [Bingel and Haider 2014] [Kupietz et al. 2010], which they say exhibits a “strong dispersion [with regard to] genre, register and time”. However, the authors concede that newspaper documents are largely prevailing and that “relatively few texts reach back to the mid-20th century”. This casts doubt over the actual strong temporal dispersion of this corpus. Moreover, although the study of NER in German is particularly challenging due to its use of capital letters for all common nouns, their evaluation remains monolingual and does not offer any insights as to how the classifiers would perform on a linguistically diverse corpus. Agirre et al. and Fernando and Stevenson considered how to adapt entity linking to cultural heritage content, but both focus exclusively on English data and did not take advantage of the multilingual structure of the Semantic Web [Agirre et al. 2012] [Fernando and Stevenson 2012]. Frontini et al. exploited the French DBpedia and combined it with the BnF Linked Data[9] in order to extract mentions of less known authors, but their graph-based approach also remained monolingual [Frontini et al. 2015]. 8

On the more focused task of extracting place names, Speriosu and Baldridge used a corpus of 20th century newswire articles and 19th century American Civil War texts to demonstrate that relying on information available in the text being processed is more effective than using external data [Speriosu and Baldridge 2013]. DeLozier et al. tackled the task of annotating a historical text corpus with geographic references [DeLozier et al. 2016], while Leidner proposed an evaluation method for different systems [Leidner 2007]. 9

Finally, the periodical Aggregation and Indexing Plan for Europeana periodicals, was launched in 2005. It produced metadata for 18 million pages of news and full-text from OCR for around 10 million pages, also including a NER component performed by the National Library of the Netherlands.[10] A new website[11] was introduced in 2014, allowing users to cross-search and reuse over 25 million digital items and over 165 million bibliographic records. However, this European Library does not use LOD resources to enrich documents, using instead its own ontology developed specifically for the project, a methodology that few institutions could afford to follow. 10

3. Case study

The *Historische Kranten*[12] project involved the digitization, OCR processing and online publication of over a million articles compiled from 41 Belgian newspapers published between 1818 and 1972. Such a large scope allows researchers to gather information on day-to-day activities in the Ypres region during WWI and WWII, thus offering a great potential in the context of CENDARI and other war-related Digital Humanities projects. The project has been launched under the impulse of Erfgoedcel CO7[13], a Flemish organisation aiming to shed light on the cultural features of the Ypres region. The target audience of such a project is thus broad: it includes scientists, WWI historians, and also simply individuals interested in the history of their region. Analysing the needs of one's target audience is central to digitisation projects, as illustrated by recent initiatives such as a Europeana user requirements group or, more recently, the Belgian Science Policy (BELSPO) funded project MADDLAIN[14]. 11

Articles in the *Historische Kranten* corpus are written in Dutch, French, and English, and focus mainly on the city of Ypres and its neighbourhood. Currently, the full texts of the *Historische Kranten* corpus have been indexed, which means that searches for particular mentions in the periodicals suffer from both noise and silence. For instance, a query on the string "Huygens" returns correct results about Christiaan Huygens: 12

Example 1. Links zien wij Christiaan Huygens die met zijn slingeruurwerk de oplossing bracht voor het meten van de tijd 13

But one also gets results that are not relevant in this context (noise): 14

Example 2. La reconnaissance du cadavre de la veuve Huygens, faite par les hommes de l'art, a fait constater l'existence de neuf blessures sur la tête 15

Moreover, interesting results are lost due to variations in spelling (silence): 16

Example 3. [...] en op het uurwerk toegepast door den Hollander Huyghens (1629-1695). 17

In order to get a clearer picture of the interests of users, we tracked individual queries on http://www.historischekranten.be/ with Google Analytics over a 4-year period, yielding 124 510 results. About 4 200 unique keywords were used at least three times, of which the ten most popular are shown in Table 1. We can see that locations are especially favored by the users, which prompted us to focus preliminary work on this type of entities. Most are related to the First World War, since Ypres was the scene of four major battles during that conflict. 18

Interestingly, we notice that the list also contains the term *oorlog* ("war" in Dutch) which does not constitute a valid named entity in the sense of Kripke: "a rigid designator designates the same object in all possible worlds in which that object exists and never designates anything else" [Kripke 1982, 77]. This makes the case for a more integrated approach to information extraction able to tackle proper nouns and common nouns in a single workflow, which is precisely one of the advantages of working with Linked Open Data resources, as will be shown in Section 4. 19

#	Term	Hits	Category
1.	Zillebeke	398	Location
2.	Passendale	351	Location
3.	Westouter	259	Location
4.	Ieper	197	Location
5.	oorlog	178	Concept
6.	Reninghelst	163	Location
7.	Bikschote	149	Location
8.	Merkem	127	Location
9.	Geluveld	125	Location
10.	Wijtschate	121	Location

Table 1. Top 10 search terms on Historische Kranten

4. MERCKX: A Knowledge Extractor

20 The idea behind entity linking is that knowledge bases can be leveraged to perform a full disambiguation of entities through URIs. A correct disambiguation of a mention of Huygens in a text with DBpedia URI dbr:Christiaan_Huygens would encompass mentions of “Christian Huyghens” (French spelling) while excluding information about the Belgian painter Léon Huygens (which has his own unique URI: dbr:Léon_Huygens) or the crater on Mars named after the Dutch astronomer, dbr:Huygens_(crater). In this section, we present MERCKX (Multilingual Entity/Resource Combiner & Knowledge eXtractor), a tool that we designed in order to extract entity mentions from documents and to link them to DBpedia [Bizer et al. 2009]. As the basis of most semantic web projects and an extremely diverse, multilingual, and extensive ontology, DBpedia was the obvious choice for providing the URIs.

21 The workflow of MERCKX for the extraction and disambiguation of entities consists of three phases: downloading resources, building the dictionary, and annotating mentions with positions and URIs. The first two steps can be time-consuming depending on the chosen entity type and additional languages, but they need to be performed only once.

4.1. Downloading resources

22 In order to simplify the download and decompression of the DBpedia dump, we provide a shell script doing this automatically.[15] This script invokes another one written in Python which extracts all the URIs matching a given type in the DBpedia ontology. Instances (or resources) are linked to corresponding types in the form of RDF triples (subject – predicate – object):

```
<http://dbpedia.org/resource/Autism>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Disease>

<http://dbpedia.org/resource/Aristotle>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Philosopher>

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/resource/Alabama>
<http://dbpedia.org/ontology/AdministrativeRegion>

<http://dbpedia.org/resource/Alabama>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Place>
```

23 A concept can be categorised by several types, as illustrated by “Alabama” in the sample above, which is both an “Administrative Region” and a “Place”.

4.2. Mapping labels to URIs

24 In the second phase, MERCKX maps all the labels to their corresponding URIs in the selected languages. The relationship between URIs and labels is also expressed by triples:

```
<http://dbpedia.org/resource/South_Africa>
<http://www.w3.org/2000/01/rdf-schema#label>
"Afrique du Sud"@fr

<http://dbpedia.org/resource/Andorra>
<http://www.w3.org/2000/01/rdf-schema#label>
"Andorre"@fr

<http://dbpedia.org/resource/Angola>
<http://www.w3.org/2000/01/rdf-schema#label>
"Angola"@fr

<http://dbpedia.org/resource/Saudi_Arabia>
<http://www.w3.org/2000/01/rdf-schema#label>
"Arabie saoudite"@fr
```

25 To make this more legible and reduce the size of the file, the initialisation script converts this to a cleaner format, using the dbr: prefix instead of the full URI starting with http://dbpedia.org/resource/, removing the recurrent “label” predicate, and inverting the order of the original triples:

```
Afrique du Sud      dbr:South_Africa
Andorre             dbr:Andorra
Angola              dbr:Angola
Arabie saoudite     dbr:Saudi_Arabia
```

26 When using more than one language, the order in which they are loaded in this lookup table is important because a label can only point to a single URI. For instance, the French label "Liège"@fr predictably corresponds to the city of dbr:Liège, but the Dutch label "Liège"@nl redirects to the homonymy page dbr:Liège_(disambiguation) which is not a valid place: it contains references to the French municipality of Le Liège and to the Liège metro station in Paris for instance. If the languages are combined in that order, the conflict between URIs will result in a decrease in recall. To reduce problems due to conflicting labels, MERCKX applies the following strategy (text in parentheses provides a concrete example for every step):

1. Load the label files for each language, one by one (EN > NL > FR).
2. Check for each label if it corresponds to the chosen type (dbo:Place).
3. If the label already exists, check if the type remains the same ("Avant"@nl is already listed as a place, but is "Avant"@fr also a place?).
4. If the type is the same, update the URI (yes > URI FR replaces URI NL).
5. If the type is different – i.e. multilingually ambiguous – remove the label (no > suppress “Avant” from the file).[16]

27 Table 2 shows a summary of the number of places extracted (URIs, labels by language, and combined labels).

URIs	EN	NL	FR	ALL
735,062	709,357	194,208	186,483	857,911

Table 2. Summary of the extracted places

28 In total, 735,062 unique locations were found in the DBpedia dump of August 2014.[17] Only 709,357 of them have a corresponding English label, leaving over 25,000 without a proper lexicalised form in this language. This can be explained by the fact that English speakers do not always find it useful to mention explicitly in a Wikipedia infobox (from which DBpedia extracts structured information) that the term to refer to the city of Ypres is “Ypres” for instance. In other words, a mapping from the label "Ypres"@en to the URI dbr:Ypres may seem redundant but makes sense in a multilingual perspective, taking non-native speakers into account.

29 The numbers of labels for Dutch and French are dramatically lower, 194,208 and 186,483 respectively. The explanation is similar: users of the English Wikipedia/DBpedia seldom take time to encode labels in alternative languages, while speakers from these other languages are often more keen to fill information on their “own” language chapters (http://nl.dbpedia.org or http://fr.dbpedia.org for instance) rather than perform this tedious work for the benefit of the English central version. This state of affairs constitutes one of the major downside of the current structure of DBpedia, which is simply replicated from Wikipedia rather than organised in a language-independent manner. The overall number of labels (857,911) is not equal to the sum of the individual languages but a much lower number, since several labels were either replaced or suppressed during the steps 4 and 5 described above.

30 At initialisation time, all the labels and URIs are loaded into a Python dict (json-like dictionary) data structure, allowing instant lookup during the spotting phase. After this last transformation, the data in memory look like this:

```
{
    "Afrique de Sud"   :  "dbr:South_Africa",
    "Andorre"          :  "dbr:Andorra",
    "Angola"           :  "dbr:Angola",
    "Arabie saoudite"  : "dbr:Saudi_Arabia",
}
```

31 At this stage, everything is in place to process textual content with MERCKX.

4.3. Tokenizing, spotting, and annotating

32 The next step is to tokenize the documents we want to enrich with the NLTK WordPunctTokenizer[18] and to perform a simple greedy lookup[19] of entities up to three tokens in length. Tokens shorter than three characters are ignored in order to reduce the noise they are likely to induce, although this comes at the price of losing locations like the municipality of Y in the Somme department.

33 For the entities present in the dictionary, the longest match is chosen and annotated with its first and last characters, in addition to the corresponding URI, thereby disambiguating these entities completely. For instance, the expression “East Yorkshire” has a match in DBpedia, and is therefore preferred to the shorter “Yorkshire”. It appears in the sample from character 626 to 640, and links to the URI http://dbpedia.org/resource/East_Riding_of_Yorkshire. This corresponds to the format of the Entity Discovery and Linking track[20] at the Text Analysis Conference.[21] Once the URI is known, contextual knowledge about the entities (such as the date of birth of people and the geographic coordinates of a place, for instance) can be retrieved seamlessly from the Linked Open Data cloud, enriching the original content.

5. Evaluation

34 In this section, we describe the methodology used in order to design the reference corpus used for evaluation purposes

(Section 5.1), before performing a benchmarking of some related tools (Section 5.2) and providing the results obtained along with a quantitative and qualitative analysis of errors (Section 5.3).

5.1. Gold-standard corpus

In order to compute the precision, recall and F-score of MERCKX, we needed a manually constructed gold-standard corpus (GSC). In information science, precision, recall and F-score are metrics intended to measure the performance of a retrieval system. Precision is how *correct* a retrieval system is: it determines whether results retrieved are relevant. Recall measures the thoroughness of the system – by comparing how many relevant items are in the corpus and how many are retrieved by the system, it becomes possible to assess the performance of the system. The F-score is the harmonic mean of precision and recall: it balances out the two above-mentioned metrics into a single one – it measures a system's reliability. Although some GSC are available online for the evaluation of entity linking, none of them is centred on digitised newspapers or the cultural heritage sector. Making the same observation, Rodriquez et al. built their own GSC for the evaluation of NER on raw OCR text, but using very different data: testimonies and newsletters, which do not compare to newspapers archives [Rodriquez et al. 2012]. We therefore used a sample from our own archival corpus and asked trained annotators to indicate all valid places. 35

5.1.1. Sample selection

Since the *Historische Kranten* corpus contains 1,028,555 articles, we calculated with the help of an online tool[22] that a sample of at least 96 articles was needed to reach a 95% confidence level with a 10% confidence interval. This means that with 96 articles, we are 95% certain that our sample is representative of the overall corpus with a deviance of maximum 10%. The confidence interval is actually much smaller (about 5%), since the probability of a word being a location is not 50% but rather 2–3%. We therefore generated a random sample of 100 documents, divided over the three languages proportionally to the overall distribution: 49 French documents, 49 Dutch ones and 2 English ones.[23] The documents range from 1831 to 1970, every decade being covered by at least two documents. We then annotated all mentions of places manually with their positions in the text (first and last character) and manually disambiguated them with their corresponding DBpedia URIs, yielding a total of 662 locations in the following format: 36

```
187   198   Bouvancourt        37
199   205   Fismes             38
561   565   Pévy               39
626   640   East Yorkshire     40
1076  1082  Trigny             41
1145  1151  Muizon             42
1200  1205  Vesle              43
```

The median number of locations by document is 4.5, ranging from 1 to 62. Most places comprise only one word, but 38 of them contain two and 9 have three words or more. The annotation is partly subjective: one could judge that the correct place is “Yorkshire” instead of “East Yorkshire” for instance, every location having five matching candidates in the dictionary on average. We thus had to validate the list with extra annotators before using it as a GSC. 44

5.1.2. Cohen's kappa

The Cohen's kappa coefficient measures inter-rater agreement on a scale between 0 and 1, 0 being zero agreement and 1 total agreement [Cohen 1960]. A value of K greater than .8 is generally considered sufficiently reliable to draw sound conclusions based on the annotation [Carletta 1996]. The kappa is computed as follows: 45

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Figure 1. Pr(*a*) stands for the relative agreement between two raters and Pr(*e*) for the probability of random agreement.

46 Our sample of 100 documents contained 30 186 tokens in total, spread over the three languages. For each language, in addition to our own annotation (A), an external annotator (B) was asked for every token to decide whether it was part of a place name or not. Locations containing OCR errors were accepted as long as the annotator could be reasonably sure that it was a place name. Table 2 presents the raw annotation counts, along with the kappa by language.

Lang.	Both	A	B	None	Tot	Pr(a)	Pr(e)	*K*
EN	20	2	2	678	702	.994	.939	.906
FR	197	46	8	13422	13673	.996	.968	.877
NL	384	13	27	15387	15811	.997	.950	.949

Table 3. Cohen's kappa for our GSC

47 The average kappa of .91 shows a high agreement that is largely sufficient to consider the GSC reliable. This good score can be explained in part by the relative straightforwardness of the annotation task (LOC versus NON-LOC) compared to more complex ones involving several types of entities, and in part by the detailed instructions provided to the external annotators prior to the task. After some corrections, insertions and deletions, we were left with 654 locations that we mapped to their corresponding DBpedia resources, producing our GSC in the TAC KBP/EDL format, which is slightly different from the one used to annotate the sample:

```
gsc3.txt   187   198   Bouvancourt
gsc3.txt   199   205   Fismes
gsc3.txt   561   565   Pévy
gsc3.txt   626   640   East Yorkshire
gsc3.txt   1076  1082  Trigny
gsc3.txt   1145  1151  Muizon
gsc3.txt   1200  1205  Vesle
```

48 The impact of OCR quality on entity linking also needed to be evaluated. To do so, we manually corrected the 120 places (out of 654) containing OCR errors and produced a second reference. The original sample and both GSC are also available on the GitHub page of the project.

5.2. Benchmarking

49 To compare MERCKX to related systems presented in this section, we used the *neleval* tool[24] which is a collection of Python evaluation scripts for the TAC[25] entity linking task and related Wikification, named-entity disambiguation, and cross-document coreference tasks. This utility allowed us to specify different GSC (raw OCR versus corrected), systems (DBpedia Spotlight, Zemanta, Babelfy & MERCKX), and measures (simple entity match versus strong annotation match) to compare.

According to Ruiz and Poibeau “the E[ntity] L[inking] literature has stressed the importance of evaluating systems on more than one measure” [Ruiz and Poibeau 2015]. Following the evaluation framework made available by the authors,[26] we used the distinction between simple entity match (ENT), i.e. without alignment, and strong annotation match (SAM) which is stricter on entity boundaries : 50

> SAM requires an annotation’s position to exactly match the reference, besides requiring the entity annotated to match the reference entity. ENT ignores positions and only evaluates whether the entity proposed by the system matches the reference. [Cornolti et al. 2013]

Both measures will be used to evaluate our results in Section 5.3 in order to get a more nuanced picture of what can be achieved by entity linking tools. 51

5.2.1. DBpedia Spotlight

DBpedia Spotlight[27] allows to find entities in text and link them to DBpedia URIs. Interestingly, the authors pay special attention to quality issues and fitness for use: “DBpedia Spotlight allows users to configure the annotations to their specific needs through the DBpedia Ontology and quality measures such as prominence, topical pertinence, contextual ambiguity and disambiguation confidence”. The four stages of its workflow are spotting, candidate selection, disambiguation, and configuration (emphasis preseverd): 52

> The *spotting* stage recognizes in a sentence the phrases that may indicate a mention of a DBpedia resource. *Candidate selection* is subsequently employed to map the spotted phrase to resources that are candidate disambiguations for that phrase. The *disambiguation* stage, in turn, uses the context around the spotted phrase to decide for the best choice amongst the candidates. The annotation can be customized by users to their specific needs through *configuration* parameters [...]. [Mendes et al. 2011]

However, the original Spotlight was designed for English only, as is the case for many tools. To counterbalance this limitation, Daiber et al. developed a new multilingual version of DBpedia Spotlight, which they claim is faster, more accurate, and easier to configure [Daiber et al. 2013]. This statistical version has been adopted for the online demo.[28] In addition to English, their language-independent model was tested on seven other languages: Danish, French, German, Hungarian, Italian, Russian, and Spanish. The authors reported accuracy scores for the disambiguation task ranging from 68% to 83%. 53

For the spotting phase, the authors experiment with two methods: a language-independent (data-driven) one based on gazetteers and a language-dependent (rule-based) one relying on more heavy linguistic processing using Apache OpenNLP models.[29] Surprisingly, the language-dependent implementation does not improve the results significantly: it only outperforms the language-independent implementation by less than a percentage point. 54

The subsequent steps are also fully language-independent: candidate selection is done by computing a score for each spot candidate as a linear combination of features with an automated estimation of the optimal cut-off threshold; disambiguation is performed by using the probabilistic model proposed by Han and Sun [Han and Sun 2011]; finally, configuration allows users to refine the results obtained by setting their own confidence and relevance thresholds, these scores being computed independently of the language. 55

5.2.2. Zemanta

Developed as a Web content enrichment platform, Zemanta[30] offers a NER API among other services for bloggers. It gained worldwide attention in 2010 when an evaluation campaign showed that it outperformed other state-of-the-art systems for entity disambiguation,[31] an assessment confirmed by later studies [van Hooland et al. 2015] [Hengchen et al. 2015]. Zemanta was subsequently integrated into the NERD framework [Rizzo and Troncy 2012] and into the OpenRefine NER extension.[32] Zemanta requires an API key in order to use its services,[33] but the webpage to apply for one seems to have been down for a long time, preventing new users from registering (although older keys still work). 56

Despite the fact that it officially only supports English text, Zemanta has proved in our own experience to work reasonably well on French and Dutch.

5.2.3. Babelfy

57 Moro et al. introduce Babelfy,[34] a system bridging entity linking and word sense disambiguation and based on the BabelNet[35] multilingual encyclopaedic dictionary and semantic network which is constructed as a mash-up of Wikipedia and WordNet [Moro et al. 2014]. Aiming to bring together “the best of two worlds”, Babelfy also uses a graph-based approach but relies on semantic signatures to select and disambiguate candidates. The use of these dense subgraphs is very effective to collectively disambiguate entities that would have proven almost impossible to identify separately. Relying on a large-scale multilingual network, Babelfy officially supports 267 languages, in addition to a language-agnostic option.

5.3. Results

58 Tables 3 and 4 present the results for simple entity match (ENT) and strong annotation match (SAM) respectively, with the best figures indicated in bold. MERCKX outperforms the three other systems evaluated, except for precision where Zemanta scores best.[36] The columns marked “Raw” show the results obtained on the original GSC, while those marked “Corr” indicate scores obtained on corrected OCR. Preliminary results of this experiment are presented in [DeWilde 2015].

System	Precision		Recall		F-score	
	Raw	Corr	Raw	Corr	Raw	Corr
Spotlight	.466	.468	.192	.207	.272	.287
Zemanta	**.887**	**.898**	.333	.371	.485	.525
Babelfy	.656	.688	.376	.446	.478	.541
MERCKX	.712	.744	**.488**	**.559**	**.579**	**.638**

Table 4. Simple entity match (ENT)

System	Precision		Recall		F-score	
	Raw	Corr	Raw	Corr	Raw	Corr
Spotlight	.235	.287	.190	.251	.210	.268
Zemanta	**.867**	**.888**	.278	.362	.421	.515
Babelfy	.662	.711	.321	.399	.433	.511
MERCKX	.782	.805	**.443**	**.517**	**.566**	**.629**

Table 5. Strong annotation match (SAM)

5.3.1 Quantitative analysis

59 Precision is consistently ahead of recall, with Zemanta reaching scores between 85% and 90%. The harder task of strong annotation match (taking into account the exact position of each entity in the text) does not affect precision: Babelfy and MERCKX actually improve on their scores, although Spotlight’s precision is cut by a factor of 2. This can be explained by recurrent entities that are correctly identified in all cases. In contrast, all recall scores decrease when considered from the SAM perspective. MERCKX outperforms other systems on recall, but it peaks at 49% (ENT) and 44% (SAM) only.

60 Low recall scores under 50% can be explained by the multilingual context and by the lack of coverage of DBpedia for some types of locations. Whereas these would be unacceptable in a medical context where failing to retrieve a

document can have dramatic consequences, a better precision is generally preferred in less critical applications. MERCKX reaches a F-score just under 60%, a ten-point improvement on both Zemanta and Babelfy which have F-scores under 50%. Spotlight fares disappointingly, with F-scores around the 25% mark. Results on corrected OCR (columns marked “Corr”) will be discussed separately in Section 5.3.3.

5.3.2. Qualitative analysis

MERCKX is heavily dependent on the quality of DBpedia, on which it relies for the disambiguation of entities. The errors of our system can be grouped into three categories, following the typology of Makhoul et al.: insertions, deletions, and substitutions [Makhoul et al. 1999].[37] 61

Insertions (spurious entities or false acceptances) are entities in the system output that do not align with any entity in the reference. A common factor causing this is multilingual ambiguity. The French adjective “tous”, for instance, when written with a capital “T”, can be incorrectly mapped to the town of dbr:Tous,_Valencia. The type check performed during the construction of the dictionary normally avoids such cases, but some problems can remain when a disambiguation page is missing: in this case, the French resource dbpedia-fr:Tous also points to the Spanish city, with no reference to the adjective. Another frequent mistake occurs when places are mentioned in the name of streets. For instance, the “rue de Lille” in Ypres does not really refer to the French city of Lille, and should therefore not be disambiguated with dbr:Lille. A more elaborate algorithm could try to detect such cases in order to exclude them, but it would be difficult to implement it in a language-independent manner without explicitly blacklisting words such as “rue”, “straat”, “street”, etc. 62

Deletions (missing entities or false rejections) are entities in the reference that do not align with any entity in the system output. One of the main causes for this is the absence of the dbo:Place RDF type in the resource of a location. For instance, dbr:East_Riding_of_Yorkshire is described as a owl:Thing which is very general and therefore not helpful. However, it is also tagged as a yago:YagoGeoEntity which is more precise. Using multiple types instead of just dbo:Place could improve the recall. Another cause is the absence of a particular label (e.g. when an old spelling is used). The resource dbr:Reims, for instance, does not include a label “Rheims” in any of the three languages used. However, the resource dbr:Rheims does exist and redirects to dbr:Reims. Including redirections in addition to labels could also help to limit the number of missing entities.[38] 63

Substitutions (incorrect entities) are entities in the system output that do align with entities in the reference but are scored as incorrect. These cases are far more rare than insertions and deletions. Substitutions can be due to the wrong detection of entity boundaries: “Jette” instead of “Jette-Saint-Pierre”, “Flanders” instead of “West-Flanders”. The greedy lookup mechanism of MERCKX normally prevents that, but extra spaces (“West- Flanders”) or long entities (“Jette-Saint-Pierre” contains five tokens because hyphens are tokenized separately) can prove tricky. Another possibility is the attribution of a wrong URI when two places have the same name. No case was detected in our system, but the output of DBpedia Spotlight contains an occurrence of this type of mistake: mapping “Vitry” to “Vitry-le-François” instead of “Vitry-sur-Seine”. 64

5.3.3. Impact of OCR

In similar work on Holocaust testimonies, Rodriquez et al. found that “manual correction of OCR output does not significantly improve the performance of named-entity extraction” [Rodriquez et al. 2012]. In other words, even poorly digitized material with OCR mistakes could be successfully enriched to meet the needs of users. The confirmation of these findings would mean a lot to institutions that lack the funding to perform first-rate OCR on their collections or the manpower to curate them manually. 65

However, contrary to this study, we see that OCR correction improves the results of all systems. Precision goes up by 1 to 3% on ENT and 5% on SAM in the case of Babelfy. Recall improvement reaches 7% on ENT and over 8% on SAM for Zemanta. Accordingly, F-scores get improved by up to 6% on the corrected version, with MERCKX crossing the 60% mark on both ENT and SAM. This state of affairs can be explained by a number of factors. First, the quality of the OCR seems to be much worse in the case of the *Historische Kranten* corpus than in the testimonies used for their study: the authors report a word accuracy of 88.6% and a character accuracy of 93.0%, whereas in the case of our sample these 66 scores were somewhat lower: 81.7% (word accuracy on places only) and 85.2% (character accuracy). The overall word accuracy, tested on a subset of the sample, was much lower still: a mere 68.3%. Secondly, the entity linking task is harder than simple named-entity recognition: full disambiguation with an URI is more prone to suffer from OCR mistakes. Using a fuzzy matching algorithm such as the Levenshtein distance could help increase the results without needing manual correction of the OCR. Preliminary experiments with this algorithm indicate that it could lead to an improvement of about 5% F-score, bringing MERCKX close enough to the performance achieved on the corrected version of the sample, although this would come at the expense of efficiency since the Levenshtein distance has an exponential time complexity.

6. Conclusion and Future Work

67 In this paper, we have shown how named-entity recognition and entity linking could easily be adopted by LAM in order to semantically enrich their collections. Compared to existing NER tools, the approach of MERCKX is to take advantage of available Linked Data resources to perform a full disambiguation of entities. This also allows to handle multiple languages seamlessly, although it comes at the price of losing some precisions in case of multilingual ambiguity.

68 A fully language-independent system would obviously need to incorporate labels from multiple knowledge bases, relying on a fallback mechanism when an entity does not exist in a specific language. To address the issue of low recall, one could experiment will the combination of several knowledge bases instead of DBpedia only. For place names, the aggregation of GeoNames[39] and GeoVocab[40] looks promising. By linking historical locations to their corresponding URIs (and, by extent, coordinates), we allow the semi-automatic creation of maps and other visualisations of the dataset – other entry points to the data than the traditional close reading approach. The system should also be tested with other corpora, and steps have been taken forward in collaborating with other cultural heritage institutions, namely the AMSAB-ISG[41] – which holds a vast collection of Flemish socialist newspapers.

69 We are also experimenting with topic modelling [Blei et al. 2003] – a task which has gained momentum in the last years, [Newman et al. 2007]– to further optimise the end users' search experience. By discovering latent topics in the dataset, disambiguating the topics with DBpedia concepts and grouping related news articles together – thus allowing faceted search capabilities –, we intend to suggest users with results related to their original queries. Using topic modelling to extract keywords and then harvesting the multilingual characteristics of Linked Data [Hengchen et al. 2016] we hope to further improve cross-lingual search capabilities. We will then apply this methodology to other collections in order to further demonstrate the added value of natural language processing techniques in the context of Digital Humanities projects undertaken by cultural heritage institutions.

Acknowledgements

70 Simon Hengchen was supported by CENDARI[42] during the course of this research (funded by the European Union's Seventh Framework Programme [FP72007-2013] under grant agreement n° 284432). The material used was provided courtesy of the Ypres City Archive and the CO7 Heritage Cell.

Notes

[1] Semantic enrichment is the process of adding an extra layer of metadata to existing collections.

[2] http://www.ehri-project.eu/

[3] http://www.cendari.eu/

[4] http://www.europeana-newspapers.eu/

[5] https://github.com/europeananewspapers/ner-corpora

[6] http://cordis.europa.eu/project/rcn/95552_en.html

[7] http://www.freshandnew.org/2008/03/opac20-opencalais-meets-our-museum-collection-auto-tagging-and-semantic-parsing-of-collection-data/

[8] http://freeyourmetadata.org/

[9] http://data.bnf.fr/semanticweb

[10] http://blog.kbresearch.nl/2014/03/03/ner-newspapers/ reported on a preliminary experiment on Dutch, French and German.

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

[12] http://www.historischekranten.be/

[13] http://erfgoedcelco7.be/

[14] https://www.maddlain.iminds.be

[15] The source code of MERCKX and related material used in this paper is available at https://github.com/madewild/MERCKX.

[16] Another option would have been to keep multiple meanings in parallel, but this is currently incompatible with the design of MERCKX.

[17] Note that this number is constantly fluctuating: as of August 2015, the figure has decreased to 725,546, which means that almost 10,000 places have been suppressed from DBpedia over the course of a year.

[18] See http://www.nltk.org/api/nltk.tokenize.html for details about how it works.

[19] A greedy algorithm always takes the best immediate solution available at each stage.

[20] http://nlp.cs.rpi.edu/kbp/2015/

[21] http://www.nist.gov/tac/

[22] http://www.surveysystem.com/sscalc.htm

[23] Documents in the sample contain 1430 characters on average, which is comparable to the corpus used by [Milne and Witten 2008].

[24] https://github.com/wikilinks/neleval

[25] http://www.nist.gov/tac/

[26] https://sites.google.com/site/entitylinking1

[27] http://spotlight.dbpedia.org/

[28] http://dbpedia-spotlight.github.io/demo/

[29] https://opennlp.apache.org/

[30] http://www.zemanta.com/

[31] See this blog post for a detailed report on the Entity Extraction & Content API Evaluation: http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/.

[32] http://freeyourmetadata.org/named-entity-extraction/

[33] http://papi.zemanta.com/services/rest/0.0/

[34] http://babelfy.org/

[35] http://babelnet.org/

[36] Consistently with results reported by Rizzo and Troncy [Rizzo and Troncy 2011]. Since Zemanta operates as a black box, it is difficult to

learn from it in order to improve precision. Our guess is that it simply uses a higher confidence threshold, at the expense of recall.

[37] To give a rough idea, Zemanta suffers from 9 insertions and 134 deletions on the corrected GSC with ENT measure; in comparison, MERCKX has 41 insertions but only 94 deletions (substitutions being close to nil).

[38] Although the risk is then to introduce more noise: dbr:Cette, for instance, redirects to dbr:Sète because the spelling of the French town changed in 1927. The danger of confusion with the French determiner *cette* is obvious.

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

[40] http://geovocab.org/

[41] http://www.amsab-isg.be. The AMSAB-ISG is a cultural heritage centre that focuses on social, humanitarian and ecological movements.

[42] http://www.cendari.eu/

Works Cited

Agirre et al. 2012 Agirre, E., Barrena, A., De Lacalle, O. L., Soroa, A., Fernando, S., and Stevenson, M. (2012). “Matching Cultural Heritage Items to Wikipedia.” In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 1729–1735.

Bingel and Haider 2014 Bingel, J. and Haider, T. (2014). “Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers.” In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Bizer et al. 2009 Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). “DBpedia – A Crystallization Point for the Web of Data.” *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

Blanke and Kristel 2013 Blanke, T. and Kristel, C. (2013). “Integrating Holocaust Research.” *International Journal of Humanities and Arts Computing*, 7(1–2):41–57.

Blei et al. 2003 Blei, D.M., Ng, A. Y., and Jordan, M. I. (2003). “Latent dirichlet allocation.” *The Journal of Machine Learning Research*, 3:993–1022.

Carletta 1996 Carletta, J. (1996). “Assessing Agreement on Classification Tasks: The Kappa Statistic.” *Computational Linguistics*, 22(2):249–254.

Cohen 1960 Cohen, J. (1960). “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement*, 20(1):37–46.

Cornolti et al. 2013 Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). “A Framework for Benchmarking Entity-Annotation Systems.” In *Proceedings of the 22nd International Conference on theWorldWideWeb*, pages 249–260.

Daiber et al. 2013 Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). “Improving Efficiency and Accuracy in Multilingual Entity Extraction.” In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM.

DeLozier et al. 2016 DeLozier, G., Wing, B., Baldridge, J., and Nesbit, S. (2016). “Creating a novel geolocation corpus from historical texts.” In *Proceedings of LAW X-The 10th Linguistic Annotation Workshop*, pages 188–198.

DeWilde 2015 De Wilde, M. (2015). “Improving Retrieval of Historical Content with Entity Linking.” In *New Trends in Databases and Information Systems*, Volume 539 of *Communications in Computer and Information Science*, pages 498–504. Springer.

Fernando and Stevenson 2012 Fernando, S. and Stevenson, M. (2012). “Adapting wikification to cultural heritage.” In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101–106. ACL.

Frontini et al. 2015 Frontini, F., Brando, C., andGanascia, J.-G. (2015). “SemanticWeb BasedNamed Entity Linking for Digital Humanities and Heritage Texts.” In *Proceedings of the 1st International Workshop on Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, pages 77–88, Portorož, Slovenia.

Han and Sun 2011 Han, X. and Sun, L. (2011). “A Generative Entity-Mention Model for Linking Entities with Knowledge Base.” In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, volume 1, pages 945–

954, Portland, OR, USA.

Hengchen et al. 2015 Hengchen, S., van Hooland, S., Verborgh, R., and De Wilde, M. (2015). *“L’extraction d’entités nommées: une opportunité pour le secteur culturel?” Information, données & documents*, 52(2):70–79.

Hengchen et al. 2016 Hengchen, S., Coeckelbergs, M., van Hooland, S., Verborgh, R., and Steiner, T. (2016). “Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the European Commission.” In *First Workshop “Computational Archival Science: digital records in the age of big data”,Washington DC*, volume 1.

Kripke 1982 Kripke, S. (1982). *Naming and Necessity*. Harvard University Press, Cambridge, MA, USA.

Kupietz et al. 2010 Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). “The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research.” In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Leidner 2007 Leidner, J. L. (2007). “Toponym resolution in text: annotation, evaluation and applications of spatial grounding.” In *ACM SIGIR Forum*, volume 41, pages 124–126. ACM.

Lin et al. 2010 Lin, Y., Ahn, J.-W., Brusilovsky, P., He, D., and Real, W. (2010). “ImageSieve: Exploratory Search of Museum Archives with Named Entity-Based Faceted Browsing.” *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10.

Makhoul et al. 1999 Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). “Performance Measures for Information Extraction.” In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 249–252.

Maturana et al. 2013 Maturana, R. A., Ortega, M., Alvarado, M. E., López-Sola, S., and Ibáñez, M. J. (2013). “Mismuseos.net: Art After Technology. Putting Cultural Data toWork in a Linked Data Platform.” LinkedUp Veni Challenge.

Mendes et al. 2011 Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). “DBpedia Spotlight: Shedding Light on theWeb of Documents.” In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, Graz, Austria.

Milne and Witten 2008 Milne, D. and Witten, I. H. (2008). “Learning to Link with Wikipedia.” In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518, Napa Valley, CA, USA.

Moro et al. 2014 Moro, A., Raganato, A., and Navigli, R. (2014). “Entity Linking Meets Word Sense Disambiguation: A Unified Approach.” *Transactions of the ACL*, 2.

Newman et al. 2007 Newman, D., Hagedorn, K., Chemudugunta, C., and Smyth, P. (2007). “Subject metadata enrichment using statistical topic models.” In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’07, pages 366–375, New York, NY, USA. ACM.

Raimond et al. 2013 Raimond, Y., Smethurst, M.,McParland, A., and Lowis, C. (2013). “Using the Past to Explain the Present: Interlinking Current Affairs with Archives via the Semantic Web.” In *The Semantic Web – ISWC 2013*, pages 146–161. Springer.

Rizzo and Troncy 2011 Rizzo, G. and Troncy, R. (2011). “NERD: Evaluating Named Entity Recognition Tools in the Web of Data.” In *Proceedings of the 1st Workshop on Web Scale Knowledge Extraction (WEKEX)*, Bonn, Germany.

Rizzo and Troncy 2012 Rizzo, G. and Troncy, R. (2012). “NERD: a Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools.” In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the ACL*, pages 73–76. ACL.

Rodriquez et al. 2012 Rodriquez, K. J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). “Comparison of Named Entity Recognition Tools for Raw OCR Text.” In *Proceedings of KONVENS 2012*, pages 410–414. Vienna.

Ruiz and Poibeau 2015 Ruiz, P. and Poibeau, T. (2015). “Combining Open Source Annotators for Entity Linking through Weighted Voting.” In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM)*, Denver, CO, USA.

Segers et al. 2011 Segers, R., van Erp, M., van der Meij, L., Aroyo, L., Schreiber, G., Wielinga, B., van Ossenbruggen, J., Oomen, J., and Jacobs, G. (2011). “Hacking History: Automatic Historical Event Extraction for Enriching Cultural Heritage Multimedia Collections.” In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP)*, Banff, Alberta, Canada.

Speriosu and Baldridge 2013 Speriosu, M. and Baldridge, J. (2013). “Text-driven toponym resolution using indirect supervision.” In *ACL (1)*, pages 1466–1476.

van Hooland et al. 2015 [van Hooland et al., 2015] van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., and Van de Walle, R. (2015). “Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections.” *Digital Scholarship in the Humanities*, 30(2):262–279.