# Named-Entity Recognition: A Gateway Drug for Cultural Heritage Collections to the Linked Data Cloud?

Seth van Hooland[◇]*, Max De Wilde[◇], Ruben Verborgh[†],
Thomas Steiner[‡] and Rik Van de Walle[†]

[◇]Université libre de Bruxelles (ULB)
Information and Communication Science Department
Avenue F. D. Roosevelt, 50 – CP 123
B-1050 Brussels, Belgium
{svhoolan,madewild}@ulb.ac.be

[†]iMinds – Multimedia Lab – Ghent University
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium
{ruben.verborgh,rik.vandewalle}@ugent.be

[‡]Universitat Politècnica de Catalunya – Department LSI
Carrer Jordi Girona, 29
E-08034 Barcelona, Spain
tsteiner@lsi.upc.edu

**Abstract**

Unstructured metadata fields such as 'description' offer tremendous value for users to understand cultural heritage objects. However, this type of narrative information is of little direct use within a machine-readable context due to its unstructured nature. This paper explores the possibilities and limitations of Named-Entity Recognition (NER) to mine such unstructured metadata for meaningful concepts. These concepts can be used to leverage otherwise limited searching and browsing operations, but they can also play an important role to foster Digital Humanities research. In order to catalyze experimentation with NER, the paper proposes an evaluation of the performance of three third-party NER APIs through a comprehensive case study, based on the descriptive fields of the Smithsonian Cooper-Hewitt National Design Museum in New York. A manual analysis is performed of the precision, recall, and F-score of the concepts identified by the third party NER APIs. Based on the outcomes of the analysis, the conclusions present the added value of NER services, but also point out to the dangers of uncritically using NER, and by extension Linked Data principles, within the Digital Humanities. All metadata and tools used within the paper are freely available, making it possible for researchers and practitioners to repeat the methodology. By doing so, the paper offers a significant contribution towards understanding the value of NER for the Digital Humanities.

This is a pre-print version of an article submitted for publication. Please cite as:
*van Hooland, S., De Wilde, M., Verborgh, R., Steiner T., and Van de Walle, R.,*
*Named-Entity Recognition: A Gateway Drug for Cultural Heritage Collections to the Linked Data Cloud?*

*Corresponding author

# 1 Introduction

## *1.1 Linked Data and the Potential of NER for the Digital Humanities*

The combination of increasing budget cuts and growing electronic collections is currently forcing cultural heritage providers to rethink the ways in which they provide access to their resources. The traditional model of manual cataloging and indexing practices has already been under pressure for a number of years. The eContentplus[1] funding program of the European Commission, for example, explicitly did not fund the development of metadata schemas and the creation of metadata itself (van Hooland *et al.*, 2011). Funding bodies and grant providers expect short-term results and encourage cultural heritage institutions to gain more value out of their own existing metadata by linking them to external data sources.

It is precisely in this context that the concepts of Linked and Open Data (LOD) have gained momentum. Recent initiatives such as OpenGLAM[2] and LOD-LAM[3] illustrate how these evolutions are percolating into the cultural heritage domain. Both the US and the EU flagship digital library projects, respectively the Digital Public Library of America[4] and Europeana[5], are currently embracing Linked Data principles (Berners-Lee, 2006). The semantic enrichment and integration of heterogeneous collections can be facilitated by using subject vocabularies for cross-linking between collections, since major classifications and thesauri (*e.g.* LCSH, AAT, DDC, RAMEAU) have been made available following Linked Data principles. Reusing these established terms for indexing cultural heritage resources represents a big potential for the cultural heritage sector. Van Hooland *et al.* (2013) provide a state-of-the-art regarding the use of Linked Data within the cultural heritage sector and illustrate how collection managers can use non-expert tools to successfully reconcile their local vocabularies with the LCSH and the AAT. By doing so, collection holders can hook up their holdings within the Linked Data cloud. Hands-on tutorials, specifically geared towards non-IT experts from the cultural heritage domain, have been developed in the framework of the Free Your Metadata project[6] in order to demonstrate how interactive data transformation tools (IDTs) can be used to clean up and reconcile metadata.

The reconciliation of local vocabularies, or even uncontrolled keywords, can be a first logical step towards publishing metadata as Linked Data. This paper explores a complementary approach by mining the unstructured narrative offered in descriptive fields for meaningful concepts through the use of named-entity recognition (NER). For clarity's sake, we will refer to such fields throughout the paper by using the Dublin Core element 'description' defined as *'an account of the resource'*, which *'may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource'*[7].

## *1.2 Research Question and Outline of the Paper*

This paper aims to examine the possibilities and the limits of applying NER to derive more value out of existing unstructured metadata content from the description. More precisely, we will consider and answer the following two questions:

**Quantitative analysis:** How do the different services score in terms of precision and recall when compared to a manually annotated gold standard corpus?

**Qualitative analysis:** Are the results provided useful and relevant for cultural heritage practitioners?

The article starts out with an overview of how NER developed and what directions the field is currently taking in collaboration with the Semantic Web community, including previous work on NER within the cultural heritage sector (Section 2). We then describe the case study and the methodology used within the paper to evaluate the outcomes of NER (Section 3). In Section 4, we present the actual results of the study, and proceed with a discussion of opportunities and risks in costs/benefits terms (Section 5) before concluding and setting forth future challenges in Section 6.

# 2 Context and Related Work

## *2.1 Background and Early Developments Regarding NER*

Originally developed by computational linguists as an information extraction subtask, named-entity recognition and disambiguation has subsequently attracted the attention of researchers in various fields such as biology and biomedicine, information science, and the Semantic Web. The original concept of a 'named entity' (NE), proposed by Grishman and Sundheim (1996), covered names of people, organizations, and geographic locations as well as time, currency, and percentage expressions. Similarly, named entities were defined for the 2002 Conference on Computational Natural Language Learning shared task as *'phrases that contain the names of persons, organizations, locations, times, and quantities'* (Tjong Kim Sang, 2002).

As a result of the diversification of NER applications, this rather loose definition was further extended to include products, events, and diseases, to name but a few types recognized today as valid named entities. Nadeau and Sekine (2007) also note that the term 'named' in 'named entity' is effectively restricting the sense to entities refered to by rigid designators, as defined by Kripke (1982): *'a rigid designator designates the same object in all possible worlds in which that object exists and never designates anything else'*. According to this view, a distinction should be made between a named entity and a plain (unnamed) entity, but this nuance is ignored by most researchers who use 'entities' and 'named entities' interchangeably.

There is, nonetheless, no real consensus on the exact definition of a (named) entity, which remains largely domain-dependent. An useful approach was adopted recently by Chiticariu *et al.* (2010) who proposed a list of criteria for the domain customization of NER, including entity boundaries, scope and granularity. They observe, for instance, that some NER tools choose to include generational markers (*e.g.* 'IV' in 'Henry IV'), whereas other do not. The definition of a named entity, according to them, is never clear-cut, but depends both on the data to process and on the application. In Section 3.2.2, we will therefore make explicit our own acceptation of a NE in the framework of this article.

## *2.2 NER and the Semantic Web*

The NER task is strongly dependent on the knowledge bases used to train the NE extraction algorithm. Leveraging on the use of DBpedia, Freebase, and YAGO, recent methods have been introduced to map entities to relational facts exploiting these fine-grained ontologies.

In addition to the detection of a NE and its type, efforts have been made to develop methods for disambiguating information units with a Uniform Resource Identifier (URI). Disambiguation is one of the key challenges in natural language processing, giving birth to the field of word-sense disambiguation (WSD), since natural languages (as opposed to formal or programming languages) are fundamentally ambiguous. For instance, a text containing the term Washington may refer to the George Washington or to Washington DC, depending on the surrounding context. Similarly, people, organizations, and companies can have multiple names and nicknames. These methods generally try to find clues in the surrounding text for contextualizing the ambiguous term and refine its intended meaning. Therefore, a NE extraction workflow consists of analyzing input content for detecting named entities, assigning them a type weighted by a confidence score and by providing a list of URIs for disambiguation.

However, as will be demonstrated in Section 5.5, a URI can not be taken at face value. We will therefore refer to the four rules Tim Berners-Lee defined in a W3C Design Issue to assess the quality of Linked Data (Berners-Lee, 2006):

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

The services described in Section 3.1.3 were selected on the basis of confirming at least partially to rule number three. For example, the well-known service OpenCalais mostly provides custom URIs that do not deliver useful information, and was therefore not included in our analysis.

Initially, the Web mining community has harnessed Wikipedia as the linking hub where entities were mapped (Hoffart *et al.*, 2011; Kulkarni *et al.*, 2009). A natural evolution of this approach, mainly driven by the Semantic Web community, consists in disambiguating named entities with data from the Linking Open Data (LOD) cloud. Several Web APIs such as AlchemyAPI, DBpedia Spotlight, Evri, Extractiv, Yahoo! Term Extraction, and Zemanta, provide services for named-entity extraction and disambiguation within the LOD cloud. These APIs take a text fragment as input, perform named-entity extraction on it, and then link the extracted entities back to the LOD cloud. In order to facilitate the evaluation of different NER services, Rizzo and Troncy (2011) have developed a tool that facilitates the examination of the outcomes of multiple services in parallel.

### *2.3 Previous Use of NER within the Digital Humanities*

A number of research projects and cultural institutions have experimented with NER over the last years. The Powerhouse museum in Sydney has implemented OpenCalais within its collection management database (Chan, 2008). The feature has been appreciated both by the professional museum world and end-users, but no concrete evaluation of the NE has been performed. Lin *et al.* (2010) explore NE in order to offer a faceted browsing interface to users of large museum collections. On the basis of interviews with a limited test group, the relevance of the extracted NE is assessed, but this evaluation is not based on a statistically significant sample. Segers *et al.* (2011) offer an interesting evaluation of the extraction of event types, actors, locations, and dates from non-structured text from the collection management database of the Rijksmuseum in Amsterdam. However, the test corpus consists of 3,724 historical Wikipedia articles, whose form and content may be inherently more suited for NER than descriptive metadata fields from a museum collection. Also, the NER process is highly customized and requires a substantial amount of programming effort.

Rodriquez *et al.* (2012) discuss the application of several third party NER services on a corpus of mid-20th-century typewritten documents. A set of test data, consisting of raw and corrected OCR output, is manually annotated with people, locations, and organizations. This approach allows a comparison of the precision, recall, and F1 score of the different NER services against the manually annotated data. The methodology applied by Rodriquez *et al.* (2012) is very much in line with the approach of this paper. This allows to position the outcomes of our analysis with the results obtained there. The corpus and the NER services used within this paper are sufficiently different in character in order to offer a significant added value to the discussion regarding the value of NER for cultural heritage collections.

## 3 Methodology

The main goal of the paper is to catalyze more experimentation and research regarding the use of NER within the Digital Humanities context. Linked Data has become an important topic for digital humanists, but the use of NER has been limited to large-scale projects. Ramsay and Rockwell (2012) recently underlined the importance of hands-on experimentation in order to come to grips with technology and to work towards an epistemology of building tools and research infrastructures. If the Digital Humanities truly want to foster such an epistemology, tools need to be made more accessible for humanities scholars, but also the methodologies to asses the outcomes of those tools.

Previous research provides an introduction on the topic of vocabulary reconciliation (van Hooland *et al.*, 2013), making it possible for scholars and metadata practitioners to interconnect cultural heritage collections across the Web with the help of a browser-based graphical interface. The current paper builds on top of this previous work, as NER allows to detect concepts which can, at a later stage, be used for vocabulary reconciliation, using the methodology presented by van Hooland *et al.* (2013). With the help of a comprehensive case study based on a freely available corpus and tools, the current paper delivers all necessary components for digital humanities scholars to repeat the analyses performed. The following sections will describe in detail the building blocks of the case study: the framework for NER services, the corpus, and the sample.

### *3.1 Open-source Framework for NER services*

#### 3.1.1 Context of Interactive Data Transformation Tools and the Use of OpenRefine

IDTs are similar in appearance to common spreadsheet interfaces. While spreadsheets are designed to work on individual rows and cells, IDTs operate on large amounts of data at once. These tools offer an integrated and non-expert interface through which domain experts can perform both the cleaning and reconciliation operations. Several general-purpose tools for interactive data transformation have been developed over the last years, such as Potter's Wheel ABC[8] and Wrangler[9]. In this paper, we will focus on OpenRefine[10] (formerly Freebase Gridworks and Google Refine), as it has recently gained a lot of


Fig. 1 Illustration of the NER OpenRefine extension

popularity and is rapidly becoming the tool of choice to efficiently process and clean large amounts of data in a browser based interface. OpenRefine further allows to reconcile data with existing knowledge bases, creating the connection with the Linked Data vision.

### 3.1.2 Development of an OpenRefine NER Extension

While OpenRefine supports reconciliation, *i.e.* mapping single- or multi-word terms to a unique identifier, it does not offer native NER capabilities on full-text fields. In contrast, several third-party companies provide Web services that offer NER functionality. Unfortunately, those services can be difficult to access without a technical background, and it is unpractical to invoke them repeatedly on multiple text fragments. Furthermore, each service has a different, proprietary interaction model. An ideal solution would be to integrate them into an existing workflow, hiding the low-level details from users.

To this end, we have developed an open source extension for OpenRefine, which is freely available for download.[11] This extension provides an integrated front-end, illustrated in Fig. 1, that gives access to multiple NER services from within OpenRefine, thereby providing two levels of automation: 1) only a single user interaction is required to perform NER on multiple records; 2) each record can be analyzed by multiple NER services at the same time. The implementation of the extension abstracts every NER service into a uniform interface, minimizing the amount of code necessary to support additional services. It also allows users to manage their service preferences, ensuring consistency between NER operations on different datasets. The extension makes NER part of a common toolkit of data operations, offering the full potential of NER in a single, accessible operation.

### 3.1.3 Currently Supported Services

The initial version of the extension supports three services out-of-the-box: AlchemyAPI, DBpedia Spotlight, and Zemanta. Despite the excellent results delivered by Stanford NER in (Rodriquez *et al.*, 2012), we decided not to include this service as Stanford NER limits itself to standard recognition and does not provide disambiguation with URIs. For similar reasons, it was decided not to include OpenCalais, as the URIs it provides are unfortunately proprietary ones and only a fraction of the returned entities link to other sources from the LOD cloud.

- AlchemyAPI[12]: capable of identifying people, companies, organizations, cities, geographic features, and other typed entities within textual documents. The service uses statistical algorithms and NLP to extract semantic richness embedded within text. AlchemyAPI differentiates between entity extraction and concept tagging. AlchemyAPI's concept-tagging API is capable of abstraction, *i.e.* understanding how concepts relate and tag them accordingly ('Hillary Clinton', 'Michelle Obama' and 'Laura Bush' are all tagged as 'First Ladies of the United States'). In practice, the difference between named-entity extraction and concept tagging is subtle. As a consequence, we treat entities and concepts in the same way. Overall, AlchemyAPI results are often interlinked to well-known members of the LOD cloud, among others with DBpedia (Auer *et al.*, 2007), OpenCyc (Lenat, 1995), and Freebase (Markoff, 2007). AlchemyAPI offers free use of their services for research and non-profit purposes. On registration, users receive an API key allowing a default amount of 1,000 extraction operations per day. Upon request, non-profit users receive 30,000 operations per day.

- DBpedia Spotlight[13]: a tool for annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linking Open Data cloud through DBpedia. DBpedia Spotlight performs named-entity extraction, including entity detection and disambiguation with adjustable precision and recall. DBpedia Spotlight allows users to configure the annotations to their specific needs through the DBpedia Ontology[14] and quality measures such as prominence, topical pertinence, contextual ambiguity, and disambiguation confidence. DBpedia Spotlight can be used for free as a Web service.

- Zemanta[15]: allows developers to query the service for contextual metadata about a given text. The returned components currently span four categories: articles, keywords, photos, and in-text links, plus optional component categories. The service provides high-quality entities that are linked to well-known datasets of the LOD cloud such as DBpedia or Freebase. Zemanta also offers free use of their services for research and non-profit purposes. Upon registration, users receive an API key allowing a default amount of 1,000 operations per day. Upon request, non-for-profit users receive 10,000 operations a day.

## *3.2 Case study: Smithsonian Cooper-Hewitt National Design Museum*

### 3.2.1 Description of the Corpus and the Sample

The Smithsonian Cooper-Hewitt National Design Museum is the world's largest design museum and holds over 200,000 objects, 60% of which are documented within the online database. The collection management team has been very active to get the most value out of the existing metadata and to enrich them with outside sources in an automated manner. Fig. 2 illustrates the front-end of the collection database, which was published as an alpha release in the fall of 2012 and is available on `http://collection.cooperhewitt.org/`. In parallel, the museum offers a complete dump of its metadata on GitHub, publicly available for download on `https://github.com/cooperhewitt/collection/`.

We believe the descriptive fields from the Cooper-Hewitt museum are representative for the type of metadata created by professional catalogers within the context of a large cultural heritage institution. Out of the 123,756 records available from the GitHub download, only 33,640 records contain a description. Some of them being identical, this leaves us with 25,007 unique descriptions. On the basis of a confidence level of 95% and a confidence interval of 5, a representative sample of 378 records was selected through a simple random sampling method.

### 3.2.2 Methodology for the Elaboration of the Manually Annotated Gold Standard Corpus

There is, to our best knowledge, no freely available corpus that can be used as a gold standard corpus (GSC) for the evaluation of NER in the cultural heritage sector. Making the same observation, Rodriquez *et al.* (2012) 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 object descriptions. Even if museum-oriented GSC would exist, it would still be useful to develop multiple manually annotated corpora for different application domains, the task of NER being largely domain-dependent, as already noted in Section 2.1.

For these reasons we decided to annotate the sample ourselves. Obviously, a concrete set of NE types was required in order to perform this annotation. An analysis of the data showed that the most relevant categories in our metadata were persons (PER), locations (LOC) and historical events (EVE)[16]. All capitalized names were considered valid NE candidates, and categorized according to this typology. Organizations, although a common NE type for journalistic corpora, are less frequent in cultural heritage data, so they were bundled together with other miscellaneous entities (MISC).

We first converted the sample into a 14,000-line text file with one word per line[17]. A cross-annotation was then carried out, every word being categorized by two persons in order to reduce errors. We used

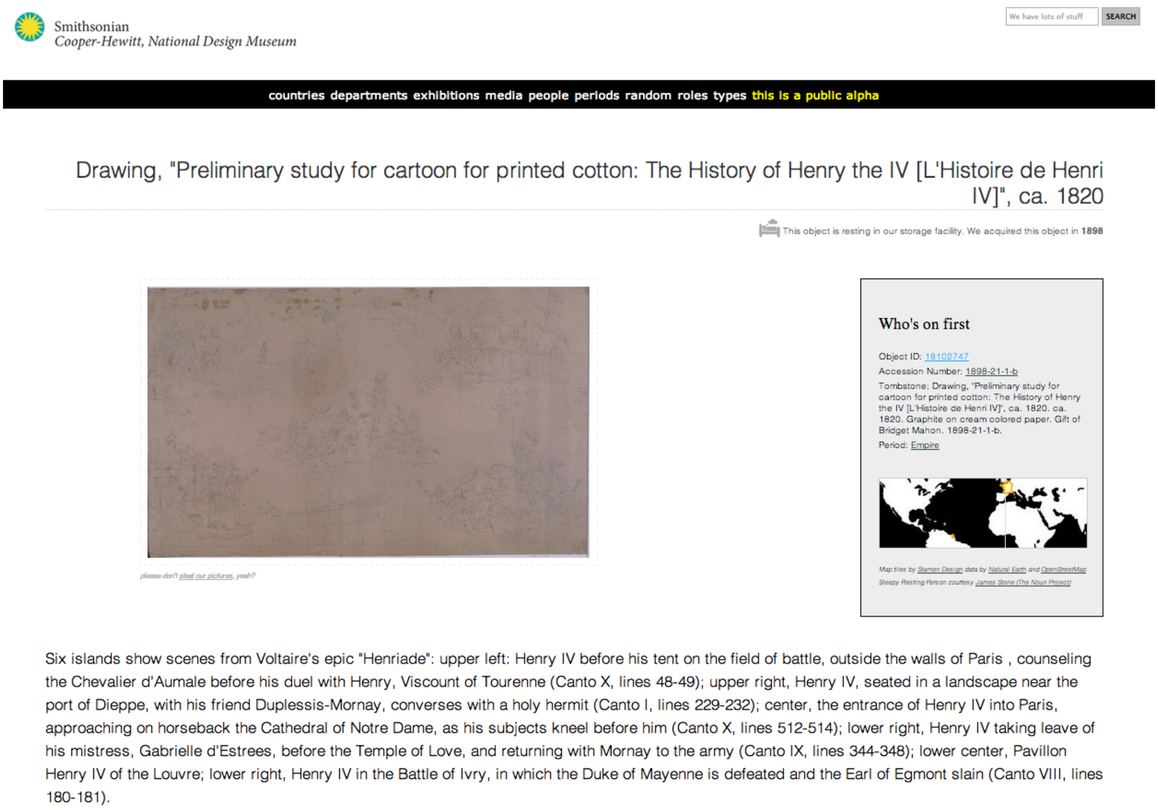


Fig. 2 Front-end display of the descriptive field

a variant of the widely-used IOB format (Ramshaw and Marcus, 1995), producing content such as the following:

```
Lincoln B-PER
delivered O
an O
effective O
political O
speech O
at O
Cooper-Union B-LOC
, O
Feb. B-EVE
27 I-EVE
, I-EVE
1860 I-EVE
. O
```

This annotated sample was then used as a GSC, allowing us to compute the precision, recall, and F-score by service and category. These results are presented in the following section.

## 4 Analysis of Precision and Recall

Using the annotated sample described in Section 3.2, we performed a quantitative analysis of the services in terms of precision and recall. It should be noted that, for this purpose, our annotation was considered a gold standard, *i.e.* an absolute reference as to what is a valid NE and what is not. As a consequence, terms that could be considered useful by collection holders (such as the material in which an object is made) were explicitly excluded and treated as errors when retrieved by a NER service. These shortcomings, unavoidable for the computation of recall, are accounted for in Section 5 where a more qualitative analysis of results is offered.

Out of the 186 entities we identified in the sample (detailed by NE type in Table 1), AlchemyAPI retrieved 60, DBpedia only 14, and Zemanta 82. Alchemy also incorrectly tagged 38 extra entities, DBpedia 44, and Zemanta 20. Using these data, we computed the precision, recall, and F1-score for each service. The results are summarized in Table 2.

The results show that, on our 378-object sample, Zemanta performed best (almost 60% F-score), followed by AlchemyAPI (about 40%), while DBpedia is lagging behind (only just above 10%). Persons and locations are generally better recognized than other NE types, although Zemanta scores over 50% on the heterogeneous MISC category. Although events and dates are an important dimension of object descriptions in historical collections, they are generally more difficult for these services to spot, a few of them being correctly identified (yielding 100% precision scores) but most being ignored, as shown by the low recall figures.

Overall, precision is better than recall, which could be surprising since many common terms found by the services were tagged as incorrect since they did not fit in our closed categories. In this respect, DBpedia was more affected than the two others. Recall does not hit the 50% mark for any service, which means that they failed to identify more than half of the NE we judged relevant. To sum up, while these results show that silence overbears noise, AlchemyAPI and Zemanta provide a meaningful input for cultural heritage collections.

| **Type** | **#** | **%** |
|---|---|---|
| PER | 50 | 26.9 |
| LOC | 37 | 19.9 |
| EVE | 24 | 12.9 |
| MISC | 75 | 40.4 |
| *Total* | 186 | 100 |

Table 1 Distribution of entities across NE types in our sample

| **Service** | **Type** | **P** | **R** | **F1** |
|---|---|---|---|---|
| AlchemyAPI | PER | .80 | .56 | .66 |
| | LOC | .69 | .54 | .61 |
| | EVE | 1 | .08 | .15 |
| | MISC | .31 | .13 | .18 |
| | *Total* | .61 | .32 | **.42** |
| DBpedia | PER | .86 | .12 | .21 |
| | LOC | .50 | .05 | .09 |
| | EVE | 1 | .04 | .08 |
| | MISC | .11 | .07 | .09 |
| | *Total* | .24 | .08 | **.11** |
| Zemanta | PER | .97 | .56 | .71 |
| | LOC | .73 | .51 | .60 |
| | EVE | .80 | .17 | .28 |
| | MISC | .74 | .41 | .53 |
| | *Total* | .80 | .44 | **.57** |

Table 2 Results of the services by category

# 5 Discussion

Section 4 presented a clearly delineated and standardized approach on the precision and recall of NE, which can be compared to results of other publications using the same methodology. However, this approach excludes from the analyses a large number of generated entities which do not belong to one of the categories defined in Section 3.2.2 and used to annotate the gold standard corpus. Nouns or adjectives identified by the NER services, such as *epigraphy* or *gold* for example, obviously hold a potential value. This issue opens the door to a number of important questions, which all directly or indirectly refer to the question of how we can assess the overall *quality* of the outcomes of the NER services.

How can quality be defined in the context of information systems? We can refer to the ISO 9000 definition, which describes quality as the *'totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs'* (ISO, 2005). Therefore, the quality of an information system denotes its adequacy with respect to the purposes assigned to it, which can be referred to as the 'fitness for use' principle. 'Total quality' does not exist, since the concept is relative: on the basis of a cost-benefit analysis, the most pertinent quality criteria – which can include the timeliness of information and the speed of data transmission or of user access – must be adopted in a given context (Boydens and van Hooland, 2011). To tackle the issue of quality at a more fundamental level, one needs to clearly distinguish deterministic data from empirical data. As Boydens clearly points out, deterministic data are *'characterized by the fact that there is, at any moment, a theory which makes it possible to decide whether a value (v) is correct. This is the case with algebraic data: in as much as the rules of algebra do not change over time, we can know at any time whether the result of a sum is correct. But for empirical data, which are subject to human experience, theory changes over time along with the interpretation of the values that it has made possible to determine'* (Boydens, 2011, p. 113).

Cultural heritage metadata, such as those of the Cooper-Hewitt case study, are empirical by nature and equally lack a direct frame of reference for testing their correctness. Their appropriateness to the needs of the field can be determined only indirectly, by considering the relative relevance of the information with respect to the objectives pursued (Boydens and van Hooland, 2011). Drucker also refers to this tension between deterministic and empirical realities, which often brings us back to the clash between the humanities and the hard sciences: *'probability is not the same as ambiguity or multivalent possibility within the field of humanistic inquiry. The task of calculating norms, medians, means, and averages will never be the same as the task of engaging with anomalies and taking their details as the basis of an argument'* (Drucker, 2012, p. 90).

In the following subsections, we will pose a number of interrelated questions which will help us to evaluate in a more qualitative way, when compared to Section 4, the quality of the outcomes of the NER services. By doing so, a more global perspective on the added value of NER for the Digital Humanities can be developed.

### *5.1 Are Identified Entities Relevant?*

The first general question to be asked on the totality of the retrieved entities of the sample, is whether they are *relevant* with regards to the description. A manual inspection of all retrieved entities within the sample allowed to assess whether an entity is closely connected or appropriate to the description. This resulted in the following observations for the three different services:

- AlchemyAPI: 124 NE in total, out of which one is irrelevant ('della mura')
- DBPedia: 372 NE in total, all of which are relevant
- Zemanta: 452 NE in total, out of which 29 are irrelevant (*e.g.* 'Table tennis' and 'Far right politics')

On the whole, the relevance of the entities is very high. Zemanta scores lower than the two other services, as its attempts at detection of hyperonyms sometimes fail. A representative example is the entity *White ground technique* which is rendered on the basis of the description 'Floral sprays on white ground'. Other errors are more difficult to explain, such as the entity *Table tennis* associated with the description *'Oval base decorated with band of overlapping acanthus leaves, applied leaf design above, holds ink pot with open lid, the front showing a mask with protruding tongue. Pen holders, in shape of a horn, flank the pot'.*

### *5.2 Do Entities Refer to Specific or General Concepts?*

Knowing that the large majority of entities is relevant in regards to the description, the next step is to analyze whether the entities represent a discriminatory value. Variance of the application domain, but also of the type of use, makes it impossible to differentiate in an absolute manner low- from high-level semantics. For example, words considered as stop words in one context can be considered to be useful in others, as 'the' and 'who' could be discriminatory in the music domain when querying for 'The Who'.

However, certain objective indications can provide indirect insights. An analysis of the syntactic structure of the entities, for instance, delivers useful information about their complexity. In order to assess the internal structure of the entities retrieved, a part-of-speech (POS) analysis was performed with the help of the Natural Language Toolkit[18], a collection of modules for advanced text analytics, providing among other tools a probabilistic (maximum-entropy) POS tagger. The used tags originate from the Penn Treebank project[19], which is the most widely established reference in the field of Natural Language Processing.

Table 3 shows the five most common structures, with figures and percentages for each service (NNP stands for proper noun; NN for singular or mass noun; NNS for plural noun and JJ for adjective). Terms consisting of a single proper noun (*Japan*) account for about a third of Alchemy entities, a quarter of Zemanta's but less than 5% of entities from DBpedia, which recognizes much more common nouns, both singular (*silver*) and plural (*cartoons*), explaining its lower score on our sample. Entities composed of two proper nouns (*Abraham Lincoln*) are also frequent, especially in Alchemy, and so are singles adjectives (*rectangular*) to a lesser extent.

In total, Alchemy and DBpedia identified roughly the same number of patterns, 20 and 23 respectively (with a large overlap), whereas Zemanta recognized thrice as much (64 patterns), demonstrating an ability to cover more diverse entities. These include very rare structures such as NNP NNP JJ NN (*New York Public Library*) and NNP CD IN NNP (*Louis XVI of France*, CD standing for cardinal number and IN for preposition), but also common ones such as JJ NN (*classical ballet*) that Alchemy and DBpedia generally fail to detect.

| | | **Alchemy** | | **DBpedia** | | **Zemanta** | |
|---|---|---|---|---|---|---|---|
| **POS tags** | **Example** | # | % | # | % | # | % |
| NNP | *Japan* | 40 | 32.3 | 17 | 4.6 | 118 | 26.1 |
| NN | *silver* | 16 | 12.9 | 108 | 29.0 | 12 | 2.7 |
| NNP NNP | *Abraham Lincoln* | 28 | 22.6 | 3 | 0.8 | 26 | 5.8 |
| NNS | *cartoons* | 8 | 6.5 | 38 | 10.2 | 8 | 1.8 |
| JJ | *rectangular* | 2 | 1.6 | 12 | 3.2 | 8 | 1.8 |

Table 3 Parts of speech used in the entities

It should be mentioned that only a minority of the reconciled single-word concepts relate to very broad and general types of objects (*e.g.* 'Brown' or 'windows'), whereas the majority of them deliver sufficient discriminatory value to perform interesting queries over large, heterogeneous metadata sets (*e.g.* 'Brooch', 'anemones' or 'gilt', which identify highly specific object types).

### *5.3 Are the Entities Correctly Disambiguated?*

One of the main selection criteria for the inclusion of the three specific NER services within our framework is their ability to disambiguate through the provision of URIs. A manual inspection of the concepts retrieved within the sample allowed to asses how well the different NER services disambiguate, and more in particular what the impact of polysemy is:

- AlchemyAPI: 124 NE in total, no issue of polysemy was found
- DBPedia: 372 NE in total, two issues of polysemy were found ('doubles' and 'swatch')
- Zemanta: 452 NE in total, nine issues of polysemy were found (*e.g.* 'Blue flower' and 'Pink Ribbon')

We can conclude that only a few cases of polysemy were detected. In most cases, the literal sense of an entity ('Blue flower', *i.e.* a flower which has the color blue) is mistaken for the figurative sense ('Blue flower' as the symbol of the joining of human with nature, rendered popular by German romanticism). Such cases are seldom problematic, but could yield embarrassing annotations (*e.g.* for 'groin vault').

### *5.4 What is the Overlap and Complementarity in between NER Services?*

An obvious question is to what extent an overlap and a complementarity exists between the three different NER services. Fig. 3 gives a synthetic overview of the statistics. 56.5% of the NE of our manually annotated gold standard corpus were identified by either AlchemyAPI, DBpedia Spotlight or Zemanta. A surprisingly low 2.2% of the entities were found by all three services, illustrating a very small global overlap. When we have a closer look at the figures, we clearly see that DBpedia Spotlight delivers a very limited value, as only 1.1% of the NE are only identified by this service, all the others being also retrieved by Zemanta. The figures regarding AlchemyAPI and Zemanta do make a case for a parallel use.

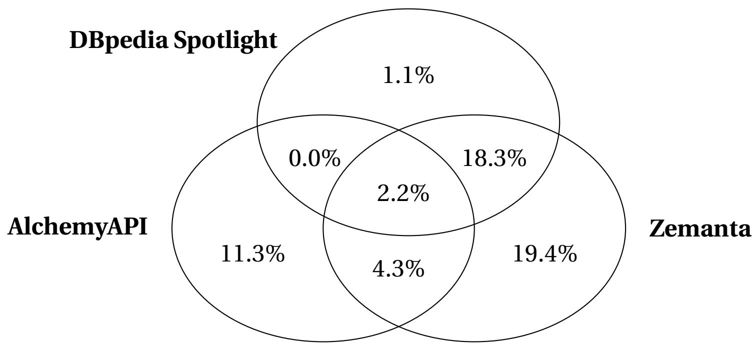


Fig. 3 The overlap between NER results of different services

### *5.5 Do URIs Refer to Resources or their Descriptions?*

Understanding what a URI is actually referring to is conceptually probably the most challenging question. Before referring to examples of the case study, the topic needs to be positioned within the broad debate in the Web community on whether a URI should be understood as a reference to a document or a resource. For example, does the URI `http://en.wikipedia.org/wiki/Richard_Nixon` identify the former US president, or does it identify a document *about* this person? Clearly, they are distinct entities: they can have separate values for the same property (*e.g.* the age of a person is different from the age of a document about that person) and one entity can evolve independently of the other. Since one URI can only identify a single resource (Berners-Lee *et al.*, 1994, 2005), a concept and its describing document(s) should necessarily have different identifiers. The question of what is identified by a URI has been an long-standing issue for the W3C's Technical Architecture Group (TAG), and has been known as 'HTTP-range 14' (Berners-Lee, 2002*c*). The conceptual difficulty arises because HTTP URIs serve a double purpose: on the one hand, they identify a resource, and on the other hand, they can provide the address to obtain a representation of that resource. The Linked Data principles (Section 1.1, Berners-Lee, 2006) demand that both functions are effectuated to ensure all URI-identified resources have a representation at their own address.

Berners-Lee (2002*a*,*b*) initially suggested to distinguish between URIs without and with fragment identifier. The former (*e.g.* `http://en.wikipedia.org/wiki/Richard_Nixon`) would identify documents, and the latter (*e.g.* `http://en.wikipedia.org/wiki/Richard_Nixon#richard`) would identify a concept (within that document). This distinction is also referred to as the difference between *information resources* and *non-information resources*. The compromise ultimately chosen by the TAG was to make this distinction by inspecting the return code when the URI is dereferenced (Fielding, 2005). While this is an acceptable solution for some, the debate still goes on (Rees, 2012).

This issue and the discussion surrounding it is very relevant for the digital humanities community, because it determines how identifiers for documents and concepts should be used. In particular with NER, we should be careful not to consider a link to a document *about* a resource as an identifier for that resource. Unfortunately, not all APIs makes this distinction. While AlchemyAPI and Zemanta differentiate between various link types and sources (attaching labels such as 'dbpedia', 'yago', and 'website'), there is no explicit indication whether the link points to an information or a non-information resource, although any given link type should consistently produce one or the other. DBpedia Spotlight returns DBpedia URIs, which always point to the concept. Still, it is important that distinct extracted entities have a unique URI to determine whether two pieces of content refer to the same entities. Continuing the earlier example, a text about Richard Nixon and a text about a document that describes president Nixon handle a different topic. However, if a NER service assigns the document's URI as an identifier of the person, that URI cannot be used to identify the document itself, leading to a paradoxical situation.

Let us bring back the discussion to our case study. The issues mentioned above are clearly illustrated by the various URIs referring to the fashion designer Isaac Mizrahi. AlchemyAPI provides `http://www.freebase.com/view/en/isaac_mizrahi`, a link to the biography of Mizrahi available in Freebase and therefore a document *about* the subject. On the other hand, Zemanta provides a URI to `http://www.lyst.com/isaac-mizrahi/`, bringing us to an online catalog of objects made by Mizrahi. Another example of a URI to an information resource is `http://www.lastfm.fr/music/Lulu`, providing access to the music of the artist. In general, we see many non-information URIs and few to none information URIs.

# 6 Conclusions and Future Work

Within this article, we focused on the evaluation of three services (AlchemyAPI, DBPedia Spotlight, and Zemanta) in order to assess the added value of NER within the Digital Humanities field. In order to calculate the precision, recall, and F1-score of the different services, a manually annotated gold standard corpus was created, based upon a sample from the Smithsonian Cooper-Hewitt National Design Museum. The results clearly identified Zemanta as the best-performing service (almost 60% F-score), followed by Alchemy (about 40%), with DBpedia largely lagging behind (only just above 10%). Persons and locations were generally well-recognized. Unfortunately, events and dates remained largely unidentified. Generally speaking, recall did not hit the 50% mark for any service, which means that they failed to identify more than half of the NE judged relevant. Resuming, these results show that silence overbears noise, although Alchemy and Zemanta clearly provide a meaningful input.

A large part of the entities identified by the NER services (such as the material out of which an object is made) do not belong to one of the categories (PER, LOC, EVE, and MISC) explicitly defined to allow the computation of recall. However, as the terms excluded from the strictly defined categories potentially hold value for search and retrieval purposes, we focused within the discussion in Section 5 on a more qualitative analysis of all entities identified by the services, irrespective of the formal categories used to annotate the gold standard corpus.

First of all, a manual analysis of all the entities showed that their relevance is very high. Almost no entities were found that lacked relevance in regards to the descriptive field from which they were derived. An illustration of such an exceptional error is for example Zemanta, which proposes the entity 'Far right politics' based on the following part of a description *'To the very far right and closer to the foreground is a belltower with domed cupola'*. The identification of irrelevant entities necessarily has to be done manually, but one could crowd-source this process by inviting users to react when confronted with an irrelevant entity.

An analysis of the syntactic structure of the entities demonstrated that a large majority of the entities represent complex concepts but also allowed to differentiate the effectiveness of the different services to identify complex entities. Alchemy and DBpedia identified roughly the same number of syntactic structures, whereas Zemanta recognized thrice as much, demonstrating an ability to cover more diverse entities. These include very rare structures represented by terms such as 'New York Public Library' or 'Louis XVI of France'. The manual analysis also allowed to evaluate the capacity of the NER services to correctly disambiguate the entities. Only a few cases of polysemy were detected within the entities identified by Zemanta, caused by confusion between the literal and figurative sense of entities.

An obvious question is whether it makes sense to use three NER services in parallel. The Venn diagram depicted in Fig. 3 represents the overlap and complementarity between the services. Almost 60% of the NE of our manually annotated gold standard corpus were identified by either AlchemyAPI, DBpedia Spotlight or Zemanta, but only 2.2% were found by all three services, illustrating a very small global overlap. On the whole, DBpedia Spotlight delivers a very limited added value, but a parallel use of AlchemyAPI and Zemanta definitively allows to identify more NE.

The discussion finishes with the challenging issue of what exactly is identified by a URI: a resource or a document about this resource? This has been an long-standing issue for the W3C's Technical Architecture Group (TAG), known as "HTTP-range 14". The clarification of this issue will only become more urgent as Linked Data principles are being applied within the Digital Humanities field. There is a fundamental difference between how services refer to, for example, the fashion designer Isaac Mizrahi: AlchemyAPI provides a link to Mizrahi's biography in Freebase, whereas Zemanta provides a link to an online catalog of products designed by him. This issue also confronts us with a fundamental problem of metadata: they are ever-extendible, in the sense that every representation can be documented by another representation, becoming a resource in itself (Boydens, 1999). Distinguishing between information and non-information resources is therefore context-dependent.

Based on the results of the paper, we can affirm that NER provides relevant entities at a low cost, based on non-structured metadata from the description field. However, the analyses allow to raise awareness regarding potential difficulties or even outright dangers regarding the use of NER within the Digital Humanities. For example, if we take the NE 'Henry IV', Zemanta delivers `http://rdf.freebase.com/ns/en/henry_iv_of_france`, whereas AlchemyAPI `http://dbpedia.org/resource/Henry_IV_of_France`, `http://umbel.org/umbel/ne/wikipedia/Henry_IV_of_France` and `http://mpii.de/yago/resource/Henry_IV_of_France`. Confronted with the heterogeneity of information given by these four different knowledge bases, the famous Julian Barnes quote spontaneously comes to mind: *'History isn't what happened. History is just what historians tell us'* (Barnes, 1989, p. 86). Linked data evangelists will instantly point out that different descriptions of the same reality can be reconciled by cross-referencing URIs from competing knowledge bases and metadata schemes with OWL:sameAs. However, in reality and especially in a humanistic one, two things are hardly ever exactly the same. Schemes such as Dublin Core helped us over the last decade to aggregate for example sculptures and paintings by Picasso, by mapping the fields 'Sculptor' and 'Painter' from individual databases to an aggregator such as Europeana using the Dublin Core field 'Creator'. This approach is very useful, but has also opened the door for numerous metadata quality issues (Foulonneau and Riley, 2008). If the Digital Humanities community wants to apply Linked Data principles on a large scale, we need to be at least aware of these issues and learn lessons from the existing literature in the information science domain.

To conclude, the Digital Humanities need to launch a broader debate on how we can incorporate within our work the probabilistic character of tools such as NER services. Drucker eloquently states that *'we use tools from disciplines whose epistemological foundations are at odds with, or even hostile to, the humanities. Positivistic, quantitative and reductive, these techniques preclude humanistic methods because of the very assumptions on which they are designed: that objects of knowledge can be understood as ahistorical and autonomous.'* (Drucker, 2012, p. 86). The purely probabilistic nature of NER not only makes abstraction of the empirical nature of humanistic data but is also tremendously influenced by economical factors, which remain by and large opaque to the general public but also to researchers. Within the next years, the competition between knowledge bases (DBpedia, representing an open-source approach, versus Freebase, which has been acquired by Google) and metadata schemes (Schema.org, an initiative of Google, Bing, and Yahoo! versus the Open Graph Protocol, a Facebook initiative) will rise as Linked Data principles are applied. Whether we like it or not, a small number of competing players such as Google and Facebook are currently imposing their way of how to render semantics explicit within the Linked Data cloud. As a community, the Digital Humanities remain for the most part ignorant of these issues, as we are busy writing up grant proposals to hook up our research data into the Linked Data cloud. Instead of this hype-driven and opportunistic behavior, the Digital Humanities community should use its unique potential to stand up and launch a scientific and public debate on these matters.

## Notes

[1]http://ec.europa.eu/information_society/activities/econtentplus/closedcalls/econtentplus/, accessed January 20, 2013
[2]http://openglam.org, accessed January 20, 2013
[3]http://lodlam.net, accessed January 20, 2013
[4]http://dp.la, accessed January 20, 2013
[5]http://europeana.eu, accessed January 20, 2013
[6]http://freeyourmetadata.org, accessed January 20, 2013
[7]http://purl.org/dc/elements/1.1/description, accessed January 20, 2013
[8]http://control.cs.berkeley.edu/abc/, accessed January 20, 2013
[9]http://vis.stanford.edu/papers/wrangler/, accessed January 20, 2013
[10]https://openrefine.org, accessed January 20, 2013
[11]https://github.com/RubenVerborgh/Refine-NER-Extension, accessed January 20, 2013
[12]http://www.alchemyapi.com/api/entity/, accessed January 20, 2013
[13]https://github.com/dbpedia-spotlight/, accessed January 20, 2013
[14]http://wiki.dbpedia.org/Ontology, accessed January 20, 2013
[15]http://developer.zemanta.com/docs/, accessed January 20, 2013
[16]Although events were previously considered on their own, there is now a tendency to include them into NE. The Dutch SoNaR corpus (Oostdijk *et al.*, 2008), for instance, divides named entities into six categories: PER, LOC, ORG, EVE, PRO (products), and MISC (Buitinck and Maarten, 2012).
[17]The tokenization was performed with the Natural Language Toolkit's WordPunct Tokenizer.
[18]http://www.nltk.org/, accessed January 20, 2013
[19]http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html, accessed January 20, 2013

## Funding

This work was supported by Ghent University; the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT); the Fund for Scientific Research Flanders (FWO Flanders); and the European Union.

## References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007), DBpedia: A Nucleus for a Web of Open Data, in *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, ISWC 2007 + ASWC 2007, Springer, pp. 722–735.

Barnes, J. (1989), *A History of the World in Ten and a Half chapters*, Picador.

Berners-Lee, T. (2002*a*), "The range of the HTTP dereference function", Maling list of the W3C Technical Architecture Group, available at http://lists.w3.org/Archives/Public/www-tag/2002Mar/0092.html (accessed January 20, 2013).

Berners-Lee, T. (2002*b*), "What do HTTP URIs identify?", available at http://www.w3.org/DesignIssues/HTTP-URI.html (accessed January 20, 2013).

Berners-Lee, T. (2002*c*), "What is the range of the HTTP dereference function?", Issue of the W3C Technical Architecture Group, available at http://www.w3.org/2001/tag/group/track/issues/14 (accessed January 20, 2013).

Berners-Lee, T. (2006), "Linked Data", available at http://www.w3.org/DesignIssues/LinkedData.html (accessed January 20, 2013).

Berners-Lee, T., Fielding, R. T. and Masinter, L. (1994), "Uniform Resource Identifier (URI): Generic syntax", IETF Request for Comments, available at `http://tools.ietf.org/html/rfc3986` (accessed January 20, 2013).

Berners-Lee, T., Masinter, L. and McCahill, M. (2005), "Uniform Resource Locators (URL)", IETF Request for Comments, available at `http://tools.ietf.org/html/rfc1738` (accessed January 20, 2013).

Boydens, I. (1999), *Informatique, normes et temps*, Bruylant.

Boydens, I. (2011), *Practical Studies in E-Government : Best Practices from Around the World*, Springer, chapter Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium, pp. 113–130.

Boydens, I. and van Hooland, S. (2011), "Hermeneutics applied to the quality of empirical databases", *Journal of Documentation*, Vol. 67, pp. 279–289.

Buitinck, L. and Maarten, L. (2012), Two-stage named-entity recognition using averaged perceptrons., in Bouma, G., Ittoo, A., Métais, E. and Wortmann, H. (Eds.), *NLDB*, Vol. 7337 of *Lecture Notes in Computer Science*, Springer, pp. 171–176.

Chan, S. (2008), "OpenCalais meets our museum collection: auto-tagging and semantic parsing of collection data", available at `http://www.freshandnew.org/2008/03/opac20-opencalais-meets-our-museum-collection-auto-tagging-and-semantic-parsing-of-collection-data/` (accessed January 20, 2013).

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F. and Vaithyanathan, S. (2010), Domain adaptation of rule-based annotators for named-entity recognition tasks, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts, USA, pp. 1002–1012.

Drucker, J. (2012), *Debates in the Digital Humanities*, Minesota Press, chapter Humanistic Theory and Digital Scholarship, pp. 85–95.

Fielding, R. T. (2005), "The range of the HTTP dereference function", Maling list of the W3C Technical Architecture Group, available at `http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html` (accessed January 20, 2013).

Foulonneau, M. and Riley, J. (2008), *Metadata for digital resources*, Chandos.

Grishman, R. and Sundheim, B. (1996), Message Understanding Conference-6: a brief history, in *16th International Conference on Computational Lingusitics*, pp. 466–471.

Hoffart, J., Yosef, A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S. and Weikum, G. (2011), Robust disambiguisation of named entities in text, in *Conference on Empirical Methods in Natural Language Processing*, pp. 782–792.

ISO (2005), Quality management systems – fundamentals and vocabulary (ISO 9000:2005), Technical report.

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

Kulkarni, S., Singh, A., Ramakrishnan, G. and Chakrabarti, S. (2009), Collective annotation of wikipedia entities in web text, in *15th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 457–466.

Lenat, D. B. (1995), "CYC: A large-scale investment in knowledge infrastructure", *Communications of the ACM*, Vol. 38, ACM, New York, NY, USA, pp. 33–38.

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", *Journal of the American Society for Information Science and Technology*, Vol. 47, pp. 1–10.

Markoff, J. (2007), "Start-Up Aims for Database to Automate Web Searching", available at `http://www.nytimes.com/2007/03/09/technology/09data.html` (accessed November 19, 2012).

Nadeau, D. and Sekine, S. (2007), "A survey of named entity recognition and classification", *Linguisticae Investigationes*, Vol. 30, pp. 3–26.

Oostdijk, N., Reynaert, M., Monachesi, P., Noord, G. V., Ordelman, R., Schuurman, I. and Vandeghinste, V. (2008), From D-Coi to SoNaR: a reference corpus for Dutch, in Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. and Tapias, D. (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco.

Ramsay, S. and Rockwell, G. (2012), *Debates in the Digital Humanities*, Minesota Press, chapter Developing things: notes towards an epistemology of building in the digital humanities, pp. 75–84.

Ramshaw, L. A. and Marcus, M. P. (1995), Text chunking using transformation-based learning, in *ACL Third Workshop on Very Large Corpora*, ACL, pp. 82–94.

Rees, J. (2012), "HTTP-range 14 webography", W3C Wiki pages, available at `http://www.w3.org/wiki/HttpRange14Webography` (accessed January 20, 2013).

Rizzo, G. and Troncy, R. (2011), NERD: evaluating named entity recognition tools in the Web of data, in *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11)*, Bonn, Germany.

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*, Vienna, pp. 410–414.

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'11)*.

Tjong Kim Sang, E. F. (2002), Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, in *Proceedings of CoNLL-2002*, Taipei, Taiwan, pp. 155–158.

van Hooland, S., Vandooren, F. and Mendéz, E. (2011), "Opportunities and risks for libraries in applying for European funding", *The Electronic Library*, Vol. 29, pp. 90–104.

van Hooland, S., Verborgh, R., Wilde, M. D., Hercher, J., Mannens, E. and Van de Walle, R. (2013), "Evaluating the success of vocabulary reconciliation for cultural heritage collections", *Journal of the American Society for Information Science and Technology*.