UNIVERSITÉ LIBRE DE BRUXELLES
FACULTÉ DE LETTRES, TRADUCTION ET COMMUNICATION

ULB

From Information Extraction to Knowledge Discovery: Semantic Enrichment of Multilingual Content with Linked Open Data

Thèse présentée en vue de l'obtention du grade académique de
Docteur en Information et communication

[Max De Wilde]

Promoteur
Prof. Isabelle Boydens

Co-promoteur
Prof. Pierrette Bouillon

Année académique
2015–2016

"*Prise entre deux langues, la pensée n'est pas réduite à un double servage ; au contraire, elle tire de cette confrontation permanente un regain de liberté.*"
Hélène Monsacré

"*We don't need more information. We need more meaning.*"
Paul Salopek

Contents

Acknowledgements

I would like to express my gratitude on the one hand to Pierrette Bouillon whose Master course of *Ingénierie linguistique* allowed me to get acquainted with natural language processing – and who was also the first one to suggest I take the job as a teaching assistant which financed this thesis – and on the other hand to Isabelle Boydens and Françoise D'Hautcourt for making a bet on me after a quick, last-minute interview one evening in late August, 2009. Isabelle and Pierrette proved to be patient, understanding, and complementary supervisors: I sincerely thank them both for their precious counselling and continuing support.

My interest in NLP subsequently led me to enrol, concurrently with my doctorate, for an advanced Master in computational linguistics at Universiteit Antwerpen, where Walter Daelemans successfully managed to rid me of all the administrative obstacles I encountered. Under his committed supervision and the mentoring of Roser Morante, I completed a highly rewarding intensive course in biomedical information extraction and attended my first international workshop in Cambridge in 2011. Later on, in 2013, I met Tobias Blanke during another workshop at the CEGES and was pleased to discover that we shared many research interests, from named-entity recognition to OCR and from the digital humanities to graph-based approaches. Walter and Tobias readily accepted to sit on my thesis committee, which I must admit makes me feel honoured.

My thanks go to Liesbeth Thiers, Hilde Cuyt, and Renee Mestdagh from the CO7 Erfgoedcel, and especially to Jochen Vermote from the Ypres City Archive, for welcoming me in their work environment and entrusting me with their dataset and analytics, even allowing me on first encounter to drive home with their 1TB hard drive of data, to be brought back the following week! They had to wait for a few years before getting any tangible results, I hope they are not disappointed with the outcome of the thesis. Thanks also to Robert Tiessen and Walter Tromp from Picturae for inviting me to their headquarters in Heiloo, near Amsterdam, for a presentation and in-depth talk.

Cheers to my colleagues Seth van Hooland and Ruben Verborgh for all our trips and talks as part of the Free Your Metadata world tour, from 8:00 a.m. pastrami with Amalia in a deserted diner to Marvin Gaye in cheap hotel rooms, and from Whit Monday library sequestration to ukulele recordings in Seth's living room. Special thanks to Ruben for asking me to be co-author of *Using OpenRefine*, which was a great collaborative experience and also a (morally, if not financially) rewarding achievement.

Thanks to my fellow PhD students Laurence, Raphaël and Simon (in alphabetical order) for partially relieving me of my teaching and administrative chores in the final months of my redacting, I doubt I could have made it in time without them. Having worked alone for the first four years, I can appreciate the added value of team work and constant brainstorming on matters ranging from the best definition of ontologies to the choice of the cutest lemur picture for a new research project.

I am definitely indebted to my family for providing constant support – to my dad in particular for technical assistance, from language detection to Python helpdesk, and to my mum for reading this work three times over without getting bored to death. To my brother Sam and his fiancée Vané then for delaying their wedding just after my submission (or is it the other way round?), and to my brother Tom for our not-so-productive-but-still-rewarding working nights while he wrote his Master thesis. Big thanks also to my other diligent proof-readers: Simon again, Lionel, and my uncle Marc.

Of course my gratefulness goes to my partner Hélène for encouraging, supporting and putting up with me during these six long years. She anticipated right from the start that I should have begun the redaction much sooner, so I'm extra grateful for her having granted me long undisturbed periods of writing time that proved crucial for the finalisation of this dissertation. I deeply admire her for the patience she could keep with the whole process.

I feel these acknowledgements would not be complete without a quick nod to the *7e tasse* teashop which provided me the fuel to go on. I obviously cannot count the number of cups gulped since 2009 but 13 000 is my rough guess, which averages to about 50 for every page written. Music-wise, Mr. Django remained as great a source of inspiration in late night hours, as ever.

And last but not least, this one goes out to my 2-year-old daughter Jeanne for regularly diverting me from my work, which could sometimes be annoying but in the end proved a blessing in disguise, allowing me to put things into perspective...

List of Figures

List of Tables

List of Abbreviations

ACE	Automatic Content Extraction
ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
AI	Artificial Intelligence
API	Application Programming Interface
CLEF	Cross-Language Evaluation Forum
CLTE	Cross-Lingual Textual Entailment
CoNLL	Conference on Computational Natural Language Learning
DBMS	Database Management System
DH	Digital Humanities
DIKW	Data, Information, Knowledge & Wisdom
EDL	Entity Discovery and Linking
GATE	General Architecture for Text Engineering
GSC	Gold-Standard Corpus
GUI	Graphical User Interface
HMM	Hidden Markov Model
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IE	Information Extraction
IR	Information Retrieval
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KB	Knowledge Base
KBP	Knowledge Base Population

LAM	Libraries, Archives & Museums
LCSH	Library of Congress Subject Headings
LOD	Linked Open Data
LREC	Language Resources and Evaluation Conference
MBL	Memory-Based Learning
MERCKX	Multilingual Entity/Resource Combiner & Knowledge eXtractor
MET	Multilingual Entity Task
MUC	Message Understanding Conference
NEE	Named-Entity Extraction
NER	Named-Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OCR	Optical Character Recognition
OWL	Web Ontology Language
POS	Part of Speech
RDF	Resource Description Framework
SAM	Strong Annotation Match
SER	Slot Error Rate
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol And RDF Query Language
SQuaRE	Systems and software Quality Requirements and Evaluation
TAC	Text Analysis Conference
TDQM	Total Data Quality Management
TEI	Text Encoding Initiative
TREC	Text REtrieval Conference
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VIAF	Virtual International Authority File
W3C	World Wide Web Consortium
WSD	Word Sense Disambiguation
XML	eXtensible Markup Language

Introduction

Outline

This first part reviews the key features of our doctoral work and provides the reader with the background material necessary to grasp the potential impact and significance of the thesis, but also its inevitable limitations. The stakes and challenges are broadly sketched out in order to give a good overview of what will be accomplished later on in this dissertation, with a special focus on incentives, goals, and structure.

Section 1 starts with the assertion of the main thesis defended within the dissertation, followed by a thorough presentation of our various motivations for undertaking this work. We also expose the interdisciplinary theoretical framework in which this research is set up and offer some justification for the current relevance of information extraction and semantic enrichment in a multilingual environment.

The objectives of our PhD are then outlined in Section 2. To sustain our premise, we formulate four research questions that we propose to investigate in the course of the thesis (Section 2.1) and define our methodology to answer them (Section 2.2). In order to illustrate our point effectively, we introduce a specific use case based on the semantic enrichment of a trilingual (Dutch/French/English) archive (Section 2.3).

Finally, the articulation of the five chapters is presented in Section 3: from information extraction to knowledge discovery through semantic enrichment of empirical content[1] with Linked Data, including constraints of data quality, language and time. Along the way, our governing principle will be the constant confrontation of theoretical considerations with the operational reality of our case study and the practical needs of end users.

[1] As opposed to deterministic content, see Chapter III.

Contents

1 Motivation

In this dissertation, we maintain that a high degree of generalisation is often more beneficial to the discovery of knowledge than an excess of specialisation. This affirmation is in clear contrast with decades of mainstream work into the development of more and more refined approaches to information extraction for a single language (Bender, 2011), domain (Chiticariu et al., 2010), or type of content. Likewise, we insist that the increased compartmentalisation of disciplines has led to an artificial partition of knowledge acquisition techniques, even inside a single field of study. This observation holds particularly true for empirical domains which will be introduced in detail in Chapter III.

To counter this restrictive tendency, we believe in the added value of interdisciplinary work which integrates advances from a broad range of disciplines in order to reason outside the narrow boxes of their respective communities. We explicitly propose to overcome the limitations of traditional information extraction by augmenting its output with Semantic Web technologies (Hitzler et al., 2009), along with insights from history, data quality, formal logic, and the philosophy of language.

Our ultimate aim is the automated harvesting and extraction of relevant new knowledge out of multilingual content and its free and open dissemination to end users. Finding information is no obvious task, even less so online. The loose structure of the Web, which has contributed to its worldwide success by letting billions of non-expert users participate in the global experience in a completely decentralised manner, also raises many questions in terms of coherence and completeness.

The Web, as a matter of fact, is far from being a consistent place, with lots of redundant or contradictory facts in addition to an endless variety of formats being in use simultaneously without any form of centralised control. On the other hand, this redundancy does not ever guarantee a web search to be fully complete, as the very same lack of a central organisation hampers comprehensiveness. The strengths of the Web, it appears, are also deemed to be its weaknesses.

Relying on search engines for information retrieval is essential due to the amount of online material to process in a very short time, but it also raises the question of the suitability of these engines for the task. While dealing with unstructured text is natural for human users, it remains a highly challenging task for machines which need to rely on formal patterns. Tried and tested retrieval algorithms and indexation practices that worked well on limited collections of data often fail to scale up to the dimension of the Web (Banko and Brill, 2001).

Over the last decades, we have therefore witnessed huge progress in the automated processing of text written in various human languages (Jurafsky and Martin, 2009), and important efforts have also been made to standardise the Web and make it more machine-friendly (Berners-Lee et al., 2001; Bizer et al., 2009a). At the same time, lots of institutions from the cultural sector and elsewhere have invested time and money into the massive digitisation of billions of documents (Coyle, 2006; Hahn, 2008; Tanner et al., 2009), prompting the information overload that led to the Big Data phenomenon.

However, the seamless integration of these different aspects is far from optimal: unstructured documents still coexist with Semantic Web resources and metadata from digitisation projects with primary sources, without necessarily interacting with one another efficiently. One of the contributions of this dissertation is therefore to offer a conclusive case study teaching us what can be gained for end users from an integrated approach to online collections of documents, focusing primarily on historical and multilingual material.

The coexistence of hundreds of languages on the Web is indeed both a challenge and an opportunity. A challenge because the risk is to build a new digital Babel where resources in multiple languages accumulate separately without ever communicating. But also a huge opportunity because it has the potential to unlock access to other cultures and to build bridges between them in a way that would have been unthinkable before the advent of the Internet.

The success of cross-lingual collaborative projects such as Wikipedia is heartening in this respect, but they raise the important issue of data quality and of the control of information. To what extent can the sum of knowledge be distributed and crowd-sourced, and to what extent should it be under the responsibility of a central authority? Whereas dictionaries and encyclopaedias used to be the strict prerogative of scholars – by no means a guarantee of objectivity but at least a form of peer-reviewed control –, anyone can now contribute to them online.

In a world ruled by data, the mapping of knowledge is less than ever a trifling task but carries a significant impact on our worldview, as emphasised by the exhibition currently ongoing at the Mundaneum, the visionary "paper Google" dreamed by pioneers Paul Otlet and Henri La Fontaine at the end of the 19th century. In this age of information overload, let us ask, how can we make *sense* of the growing mass of documentation available? How do we transcend Big Data to get at their *meaning*? How do we discover new *knowledge*?[2]

[2]Knowledge being an elusive concept in philosophy, we deliberately adopt the pragmatic definition of Davenport and Prusak (1998): "Knowledge is a fluid mix of framed experience, values, contextual information, expert insight and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information".

Like all interdisciplinary work, this doctoral research falls within the scope of not one single but several interconnected theoretical frameworks. Our starting point is information extraction (IE), an application of natural language processing related to artificial intelligence. IE emanates from computational linguistics and includes various tasks such as named-entity recognition (NER), relation detection and event extraction. From this point, we move on to explore potential improvements thanks to the contribution of several other fields.

Pioneer work in IE, mainly conducted in the United States, primarily focused on the English language, often gratuitously assuming a degree of generalisability to other languages without demonstrating it in practice. In the nineteen nineties, Palmer and Day (1997) argued that "there ha[d] been little discussion of the linguistic significance of performing N[amed] E[ntity] recognition". The 2000s saw a shift in this trend, with the appearance of many IE systems specialised for one or another language, incorporating basic language-specific resources – like diacritic characters and stopwords – but also more complex ones such as specific syntactic rules and semantic features.

Such systems were often adapted from previous ones that were specifically designed for English, requiring a huge effort to eliminate features that were hard-coded with only English in mind, thereby laying bare the Anglocentric myth of the intrinsic universality of this particular language. Other researchers chose to start from scratch in order to better accommodate the specificities of another language (say, French or Dutch), but their systems in turn became harder to further adapt because of this high level of linguistic specialisation.

Few approaches, however, have considered handling natural language as a whole (i.e. any human language, as opposed to formal languages), relying only on common, universal features. The idea of a universal grammar is not new, going as far back as the 13th century with the English philosopher Roger Bacon observing that all languages share a common structure. This theory was mainly developed by Noam Chomsky (1956) and taken over by Steven Pinker (1994) who both argued that the ability to acquire language is hard-coded in the brain, making Universal Grammar a biological evolutionary trait.[3] While not allowing to capture some subtleties, a language-independent conception of IE offers the advantage of being widely reusable and adaptable to new resources in a variety of languages.

[3] Although most linguists accept Universal Grammar, a vocal minority rejects the theory for various reasons. See Hinzen (2012) for a summary of these criticisms. A recent paper revived the debate by claiming evidence of a universal feature in 37 languages (Futrell et al., 2015).

Several recent initiatives show that IE-related technologies are attaining maturity and are therefore ready to be scaled up in a business context, as the awareness and interest of the industry in these technologies are increasing. Shortcomings, however, remain to be addressed, such as the volatility of evaluation relative to the application domain (Alexopoulos et al., 2015) and the quality of knowledge bases with regards to industrial needs.

In another respect, the importance of multilingualism in our globalised society hardly needs to be highlighted, even less so in the context of the European Union with its 24 official working languages.[4] Several European projects focus on linguistic diversity, and some of them would definitely benefit from general-purpose, language-independent IE tools accessible to the public. The scale of these projects, and the time and money invested in them, also testify of the utter relevance of these research topics today.

Combining these two thematics, the FREME project,[5] co-funded by the H2020 Framework Programme for Research and Innovation,[6] precisely aims to "build an open innovative commercial-grade framework of e-services for multilingual and semantic enrichment of digital content". The semantic enrichment of digital content in a multilingual context is thus a deeply contemporary and relevant concern.

According to Tang et al. (2015), "recognizing entity instances in documents according to a knowledge base is a fundamental problem in many data mining applications". In a recent and thorough survey of entity linking techniques, Shen et al. (2015) also emphasise the current stakes for information extraction:

> "The amount of Web data has increased exponentially and the Web has become one of the largest data repositories in the world in recent years. Plenty of data on the Web is in the form of natural language. However, natural language is highly ambiguous, especially with respect to the frequent occurrences of named entities. A named entity may have multiple names and a name could denote several different named entities."

This intrinsic ambiguity makes language technologies more relevant than ever if we are to take advantage of the huge potential of the Web in an automated manner. In what follows, we will therefore define concrete objectives and a clear methodology in order to bring the techniques of information extraction one step further towards the ideal of knowledge discovery.

[4] `http://ec.europa.eu/languages/policy/linguistic-diversity/official-languages-eu_en.htm`

[5] `http://www.freme-project.eu/`

[6] `http://ec.europa.eu/programmes/horizon2020/`

2 Objectives

The main objective of the present work is to investigate the limitations of specialised information extraction techniques and to offer generic alternatives in a multilingual context. In order to address this issue, we will consider several areas of enquiry, from traditional named-entity recognition to disambiguation of online resources and from the specificities of empirical content to problems related to quality, language, and time. Let us introduce these various aspects of our work.

Information extraction is commonly defined as the identification and classification of relevant information in unstructured text. In its classical sense, IE does not involve any disambiguation phase but rather stops right after categorisation, which is problematic because natural language is intrinsically ambiguous. To overcome this limitation, we propose to rest on external referents providing unique identifiers. This will allow us to bring IE a step further with techniques of entity linking and semantic relatedness.

The problem is made even more complex by a number of external constraints. First, empirical information is not deterministic and is subject to human interpretation, raising concerns about objectivity that have to be tackled upstream. Second, monolingual IE tools designed for English are not necessarily well suited to operate in our increasingly multilingual information society. Third, natural languages are not static objects that can be processed once and for all, but rather "dynamic and variable systems" (Diller, 1996) that evolve over time. All of these constraints have important consequences which are insufficiently addressed in the IE literature, so we offer to remedy this.

Following the work of Blanke and Kristel (2013) for the European Holocaust Research Infrastructure (EHRI), our final goal is to "use information extraction services to enrich the researchers' experience". Concretely, this amounts to provide additional knowledge to users in a Web search interface, through a mechanism of semantic annotation, disambiguation and enrichment. This can only be achieved by considering IE from a broader point of view than its traditional epistemological conception.

As highlighted in Section 1, there is currently a crying need for solution-oriented approaches to semantic enrichment. Focusing on the *fitness for use* principle, which will be introduced in detail in Chapter IV, will ensure we keep this objective in mind in order to deliver operational recommendations and tools to meet the actual needs of people looking for information, without being carried of by a theoretical approach to knowledge discovery disconnected from the reality of field work.

2.1 Research questions

To sustain our main thesis, we will articulate our argumentation around four central research questions which are closely linked to the first four chapters of this dissertation, as detailed below.

Question 1 Is the philosophical and linguistic distinction between named entities (proper nouns) and terms (common nouns) justified? Does it not maintain an arbitrary separation between objects of thought? For instance, is an encyclopaedia entry about Henri La Fontaine intrinsically different from en entry about the concept of peace, notwithstanding the ontological differences between the two? Practically speaking, is there a valid reason for using different tools for the extraction of these two types of semantic units?

Question 2 Does interdisciplinarity[7] improve our understanding of the world? Can Linked Data and Semantic Web technologies (and other disciplines that are not traditionally associated with computational linguistics) contribute to bring information extraction (IE) a step further through full disambiguation? Can the transposition of concepts from history or philosophy to natural language processing help overcome problems that have been left unaddressed by lack of interest or short-term vision?

Question 3 Are generic approaches to IE more productive than domain-specific ones? Does it hold true for any application domain? Does compartmentalisation of knowledge into disciplines not impede the quest for integrated, comprehensive search on any topic? Does the cost of manually specialising a tool for a given domain not outweigh the benefits obtained? Even with automated or semi-automated approaches, domain specialisation of IE tools remains a resource-consuming task, so to what extent is it profitable?

Question 4 Similarly, are language-independent approaches to IE more sensible than language-specific ones? If so, in which cases? Is the gain in accuracy obtained by the fine-tuning of systems for a given language (say, English) counterbalanced by the loss in portability to other languages? With over 7 000 languages spoken on the planet[8] (of which almost 300 have their own version of Wikipedia), is it not self-defeating to restrict IE systems to one or a few languages if we intend to access the sum of knowledge?

[7] We use the terms *interdisciplinary* and *interdisciplinarity* rather than "multidisciplinary" and "multidisciplinarity" to indicate a higher level of interaction between disciplines than mere juxtaposition, in much the same way that interculturalism goes further than multiculturalism.

[8] `http://www.ethnologue.com/` lists 7 102 of them as of May 18, 2015.

These research questions will guide our reflection on IE and related topics throughout the thesis, and will influence the way in which we address the existing literature. More specifically, Question 1 is about air-tight semantic categories and will be investigated in Chapter I, which presents the traditional approach to information extraction from a computational linguistics perspective. Question 2 is about the added value of mixing input from different disciplines and will mainly, but not exclusively, be discussed in Chapter II which shows what can be learned from the Semantic Web. Question 3 is about the specificities of application domains and will be addressed in Chapter III which deals with the typology of sciences and the particularities of empirical data. Finally, Question 4 is about multilingualism, one of the core issues tackled in Chapter IV. Investigating these research questions will then allow us to discuss practical implementation of a knowledge discovery system in Chapter V.

Our questions are not straightforward, i.e. they cannot be easily dismissed with a plain yes or no. On the contrary, they are complex interrogations underpinning whole visions of the world and therefore need to be evaluated in terms of costs and benefits, balancing each element against the others. Accounting for these different worldviews and their operational implementations for the representation, extraction and discovery of knowledge is at the heart of the approach adopted in this dissertation.

In order to answer these questions, we will strive to reconcile the traditional view of IE in computational linguistics with a more pragmatic, results-oriented approach that has gained momentum with the Semantic Web and the ability to disambiguate entities by linking them to a knowledge base. For cultural institutions and other players interested in the enrichment of unstructured content, these techniques, grouped under the common denomination of the "digital humanities", constitute an opportunity to gain value out of existing material at a low cost.

We are confident that our research questions are helpful to build a better approach of IE, while humbly recognising with Woody Allen that confidence is often what you have before you understand the problem.[9] Inevitably, some answers will be only half-answers and will maybe appear inconclusive. Yet as the American poet and novelist Don Williams, Jr. said: "our lessons come from the journey, not the destination".[10] To sum up, this thesis could be considered the utopian quest for the perfect cross-lingual, cross-domain IE system that will massively enhance the daily search experience of end users. If we do not find it, then let us hope that the search will be worth the trouble...

[9] In James Geary's Guide to the World's Great Aphorists (2007, p. 9).

[10] In Eric J. Keese's Timeless Words for Living Gigantic: 1001 Quotations (2015, p. 82).

2.2 Method

In order to achieve our objectives and to help answering our research questions, we develop an original methodology relying on several components which will be introduced progressively. Our method involves some basic linguistic processing steps such as tokenisation (lexical segmentation of a text into individual words or tokens) and relies on formal patterns to detect entity candidates, but it does not include any language-specific analysis.

To go beyond entity identification and classification, we propose to use Linked Data resources singled out by unique IDs. By building comprehensive gazetteers of labels (i.e. concepts lexicalised in natural languages), we aim to disambiguate a wide range of textual mentions and to filter them by type. This approach, commonly called entity linking, is evaluated on the basis of a case study (*Historische Kranten* project) introduced in the next section.

To overcome the difficulties raised by the empirical nature of historical content, we design a gold-standard corpus (GSC) of places (by far the most popular type of entities among users of the `http://historischekranten.be/` website) and evaluate popular NER and entity linking tools against it. This evaluation puts into light some recurring shortcomings of such tools: they work unequally well on different languages, suffer from cross-lingual ambiguity and often retrieve either too much or too little information.

We therefore propose a custom tool called MERCKX and compare its performance with existing tools, showing how simple language- and domain-independent insight can improve the extraction and disambiguation process. In order to handle noisy content resulting from optical character recognition (OCR), we suggest to rely on clustering algorithms such as the Levenshtein distance, although this functionality is still under development.

The usefulness of the results obtained is assessed with regards to user search statistics ranging over four years of existence of the project's website, showing that the extracted information matches the needs of the users. A qualitative analysis shows that this approach has the potential of improving the search experience of users at a reduced cost. In our research perspectives, we contemplate further validation by getting direct feedback from the users on the relevance of the entities extracted.

Finally, we report on the portability of our method and its generalisation to other languages, domains, and types of content. The lightweight approach we adopt allows to add support for components unrelated to our original case study in an agile way, as we demonstrate with two extra corpora. A proof-of-concept implementation of knowledge discovery applications is offered to demonstrate the usability of this workflow.

The idea of semantic enrichment is not new, so the originality of our work resides less in the topic itself than in the way we address it from an innovative angle, guided along the way by our four research questions. Our personal contribution can be articulated around three central themes which will be introduced in detail in Chapter IV.

Quality Quality control originates from the industry and has been applied to a variety of domains including modern-day databases, but data quality has seldom been considered seriously as an issue affecting either the output of information extraction or the Semantic Web. Using an original temporal model detailed below, Isabelle Boydens developed a framework for monitoring the quality of administrative databases and showed that it could be generalised to any empirical system, with concrete operational consequences measurable with a cost-benefit analysis. We propose to build upon this model in order to apply it to another kind of empirical information: digitised cultural heritage and the contents of knowledge bases.

Language The multilingual aspect of our work is equally important. In order to mine the World Wide Web in all its linguistic variety and to amass knowledge, semantic enrichment algorithms cannot be limited to one or a few languages. Although our corpus is quite modest in this respect, with only three Western European languages represented, it nevertheless allows us to experiment with cross-lingual approaches which can subsequently be validated on other, more diverse languages. Developing language-independent techniques is particularly challenging due to linguistic ambiguity, but also to the absence of one-to-one correspondence between concepts in different languages.

Time When dealing with historical material spreading over a century and a half, the temporal dimension cannot be ignored. Accounting for the drifting of concepts over time and tracking the emergence of new terms is one of the ideas explored later on in this dissertation. Building upon Fernand Braudel's framework of stratified timescales (*temporalités étagées*) but also on Norbert Elias' concept of evolving continuums (*Wandlungskontinua*) which allows to account for the bidirectionality of temporal fluxes,[11] we propose a generic model for language evolution and show that knowledge representation must be apprehended dynamically.

[11] Both these theories were originally transposed by Boydens (1999) to grasp the interactions between reality and the norm in the context of Belgian social security.

2.3 Use case

"T'as p'têt' fait le tour exotique : celui de Tournai à Doornik..."
André Bialek, *Visite guidée*

Throughout this dissertation, examples will be drawn whenever possible from the *Historische Kranten* corpus which will be presented in detail in Chapter III, allowing us to illustrate our point effectively while anchoring our research in a concrete case. This trilingual (Dutch/French/English) archive, provided by the city of Ypres, consists of over a million files from 41 Belgian periodicals published between 1818 and 1972 and focuses primarily on the Westhoek region.

The limits of specialisation are indeed better put into light with a specific use case. Let us imagine a researcher interested in the port of the Belgian coastal city of Ostend, especially in its evolution over time in all respects (biodiversity and ecology, tourism and economy, etc.).

Traditional full-text search is clearly not satisfactory since words such as "port" and "Ostend" are naturally ambiguous: *port* can refer to sweet Portuguese wine, while cities named *Ostend* also exist in England, Germany and New Zealand. Not taking these cases into account introduces noise in the search. The presence of another Ostend in a Belgian corpus may seem as unlikely as the occurrence of *Paris, Texas* in French literature, but the periodicals we used show that this possibility is not completely remote:

> « Nous apprendrons probablement à nos lecteurs qu'il existe dans le comté d'Essex un village que l'on appelle Ostend. Il n'existe apparemment aucune similitude ethnographique entre notre "Ostende" et "Ostend" en Angleterre et nous ne trouvons aucune raison linguistique susceptible d'avoir conduit les Anglais à donner ce nom à un village. (*Le Sud*, Sunday 4 August 1935) »

Moreover, concepts are often named differently from one language to another (*haven* in Dutch, *puerto* in Spanish) but also within a single language with synonymy (harbour). Even for proper nouns, local variants can arise: Ostend is called *Ostende* in French and *Oostende* in Dutch. The failure to account for this phenomenon generates silence.

Example 1. Special week-end trips to Ostend (without passports)

Example 2. Deux grandes fêtes à Ostende

Example 3. Een boot met wapens aangeslagen te Oostende

This issue is maybe more severe in our case, since a literal search for the term “Ostend” only returns 476 hits, such as Example 1, from a million-document archive. Using the French spelling returns 8 396 results (among which Example 2). With the Dutch one we get 39 104 more (e.g. Example 3). To counter the twin issues of noise and silence, we will demonstrate that Linked Data can bring us a step further towards full disambiguation and enrichment of user queries. An example of this process is shown in Figure 1 for the disambiguation of the city of Ostend.

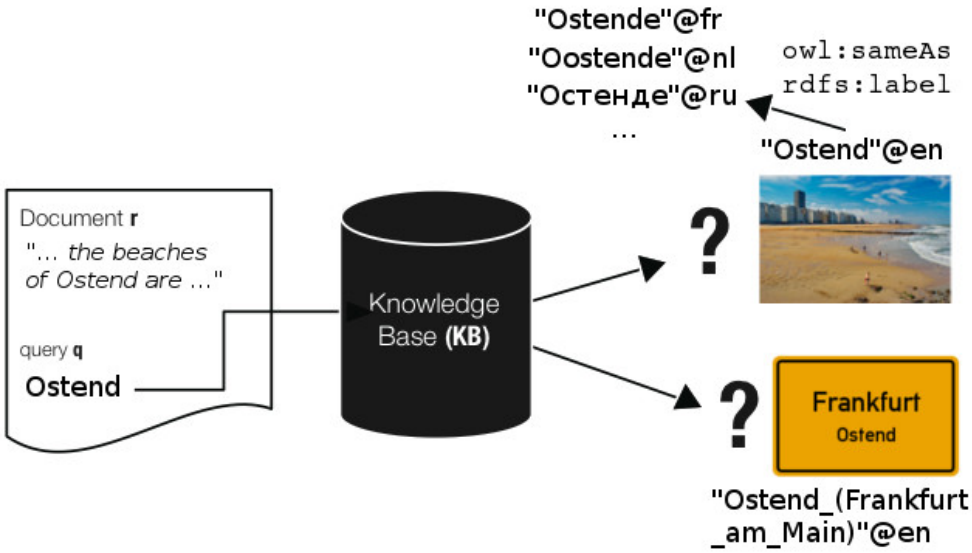


Figure 1: Disambiguation and enrichment with a knowledge base

We see that document *r* contains a reference to the beaches of Ostend. After performing named entity recognition, the extracted entity “Ostend” can be categorised as a location but remains ambiguous due to the existence of multiple places with the same name. Querying a knowledge base with *q* allows to select the correct uniform resource identifier (URI) based on frequency of use or the surrounding context (the mention of beaches for instance). Moreover, multilingual enrichment can be performed thanks to the `owl:sameAs`[12] and `rdfs:label` properties once the entity has been properly disambiguated.

[12]This property, allowing to express identity between resources, will be presented in Chapter II while problems arising from its use will be discussed in Chapter IV.

Despite what Figure 1 might suggest, knowledge bases do not operate as "black boxes",[13] i.e. they are not opaque systems taking a query as input and sending back a response without offering clues to how this response was reached. On the contrary, they are fully transparent repositories allowing to track the enrichment process at every step in the pipeline, from a simple string of characters to a fully disambiguated concept.

But technical approaches are nevertheless always in danger of becoming black boxes to neophyte users, as underlined by Ramsay and Rockwell (2012): "either the technique is encapsulated inside the black box of magical technology or it is unfolded in tedious detail obscuring the interpretation – tedious detail which ends up being a black box of tedium anyway". The devil, as ever, is in the details, and the clarity of arguments should remain a priority.

Most knowledge bases (KBs) are constructed in part by collaborative work of non-expert users and in part by the input of robots.[14] Depending of the application domain, bots can be more or less prominent compared to human beings (Steiner, 2014). In both cases, quality is a central issue: KBs are not curated vocabularies maintained by professionals and can therefore suffer from important downsides.

While many KBs contain knowledge in multiple languages, the structure used for the representation of this knowledge can vary from one KB to the other. Some KBs (like DBpedia) have different pages for each language that are linked to one another, while others (like Wikidata) prefer to keep a single central page that can be served in various languages. However, the absence of a systematic one-to-one correspondence between concepts in some languages makes both models vulnerable.

The French resource "port", for instance, links to "haven" in Dutch. "Haven", however, is also an English resource, a synonym for "harbour". A harbour can be either natural or artificial, but this resource in turn links to "havre" in French and "natuurlijke haven" in Dutch. This partial matching of close terms makes disambiguation and enrichment even more arduous. Nevertheless, KBs offer an unequalled coverage of human knowledge accessible at virtually no cost.

Although the *Historische Kranten* corpus will only be presented thoroughly in Chapter III in the broader context of the digital humanities and empirical content, every chapter will draw on the material of this case study to emphasise the usefulness of the various components introduced in the course of the thesis in order to reach our objectives.

[13] Contrary to most commercial NER tools, as will be seen in Chapter V.

[14] A notable exception is the expert knowledge base OpenCyc: `http://opencyc.org/`.

3 Structure

The dissertation is divided into five chapters covering both the state of the art and our personal contribution to the field, at a conceptual and technical level. We start with information extraction to reach knowledge discovery through the exploitation of Linked Open Data for semantic enrichment, and show how the union of these domains helps to overcome known limitations related to the handling of empirical content, focusing along the way on data quality, multilingual approaches, and the evolution of concepts over time.

The governing principle underlying the whole thesis is the designing of a functional knowledge extraction tool to improve the search experience of users of the `http://historischekranten.be/` website. The argument will therefore repeatedly refer to this unifying theme, anchoring our research in a concrete application case. But our aim in doing so is of course to show the broader potential of semantic enrichment technologies, not only for similar digital humanities projects but also for a range of domains dealing with empirical content on a daily basis. Generalisation is therefore an essential component of our work, which will constantly oscillate between high-level conceptual theorisation and down-to-earth practical implementation.

Chapter I starts with a historical account of the appearance of information extraction within the community of natural language processing and artificial intelligence. It goes on with a discussion of the achievements reached over the years in its three main subtasks: named-entity recognition (NER), relation extraction and event extraction. NER gets a particular treatment since it has been seen by many IE researchers as the crystallisation point of natural language ambiguity (Ehrmann, 2008; Hoffart et al., 2011), and has therefore become a classical task for competitions with its own conventions, although the ontological description of a named entity has remained quite elusive.

Chapter II introduces the Semantic Web – a structured layer designed to allow a better exploitation of the Web by machines – and the progressive formation of the Linked Open Data (LOD) cloud under the efforts of Tim Berners-Lee. Whereas traditional NER was necessarily limited to plain recognition of an entity's boundaries and its classification in broad semantic categories, the explosion of online information resulting in the Big Data phenomenon has enabled the full disambiguation of entities through the mechanism of URIs.[15]

[15] A special variety of unique identifiers of which traditional Web URLs are a subclass. The distinction between URIs and URLs, along with the specificities of URIs compared to traditional unique IDs, will be discussed in Chapter II.

Central to the LOD project are knowledge bases, which come with their own organisations (ontologies) and work as big repositories of structured data in the RDF format. From the fusion of IE and LOD, the new task of entity linking is born, paving the way for the semantic annotation and enrichment of textual data.

Chapter III discusses the intrinsic characteristics of empirical data, especially in the humanities but also elsewhere, and reflects on how these differ from deterministic data. After a critical presentation of the digital humanities (DH), including the taking into account of the concept of distant reading (Moretti, 2005), we reveal an original case study based on the digitisation and online publication of over one million historical documents from a local archive, already sketched in Section 2.3. This project is situated in the context of similar initiatives in the cultural heritage sector, which has taken a special interest in the use of Semantic Web technologies and LOD for the promotion of their collections.[16] The needs of the different stakeholders are carefully established in order to collect a number of specifications that will serve as guidelines for the implementation part.

Chapter IV gives special attention to internal and external constraints related to semantic enrichment projects at various levels. Data quality is an important issue affecting both input material (poor OCR, inconsistent XML, etc.) and the tools used in the enriching process (incomplete and/or incoherent Linked Data, irrelevant NER results, etc.). The challenges of multilingualism are also an important dimension of our work that will be dealt with extensively. The implications of a language-independent context for IE tools will be investigated, along with the specificities of cross-lingual corpora and the recent evolutions of the Semantic Web in an increasingly multilingual online reality. Finally, some issues related to the evolution of language and concepts over time will be taken into account, concept drift being a sizeable problem when dealing with historical material ranging over decades or even centuries.

[16]The financial constraints often undergone in this field makes it a particularly interesting object of study. Whereas better-funded application domains manipulating empirical data – such as social security, medicine, defence, law, finance, etc. – can afford to pay for more controlled knowledge sources, the cultural sector and related DH fields are compelled by lack of funds to reuse existing online material. We can also argue that the consequences of bad quality are less dire for cultural metadata than in domains where huge monetary losses, reputation or even human lives are at stake, although this is not an absolute hierarchy but rather depends on the values of our civilisation (Elias, 1996).

Chapter V finally puts into practice the lessons learned from the previous four chapters in order to construct a workable model for knowledge discovery. We first evaluate a variety of tools related to NER and semantic annotation. By retaining interesting components and adding a few insights of our own, we propose an original tool called MERCKX (Multilingual Entity/Resource Combiner & Knowledge eXtractor) for the disambiguation of relevant entities and the discovery of related information. In order to evaluate our approach, we compare our results against those of related tools and show where there is still room for improvement. We conclude this final chapter with a proof-of-concept implementation of our work for the *Historische Kranten* project and discuss how this model could profitably be generalised and transposed to other languages and application domains.

After these five chapters, we close up the dissertation with a summary of the main findings and outcomes – including tentative answers to our research questions – and a discussion of the limitations of our work. We conclude by providing concrete recommendations for transposing our methodology to other corpora and examining a few potential future applications for the model developed.

Chapter I

Information Extraction

Outline

In order to start our thesis on sound theoretical foundations and to provide building blocks for the semantic enrichment of content, this chapter explores the emergence of information extraction (IE) as a sub-field of natural language processing (NLP), along with its relationship to artificial intelligence (AI).

Section 1 provides some historical background about NLP and IE, showing how the latter was progressively driven to the forefront, as high expectations regarding the potential of NLP had to be lowered in order to meet more realistic goals. Specifically, we draw a distinction between IE and several related, partially overlapping fields and justify the choice for the terminology used.

We then focus in Section 2 on one of the main IE tasks – named-entity recognition (NER) – providing a thorough survey of state-of-the-art research into NER. A particular attention is given to the definition of the NER task, but also to the typology of named entities, different types of NER systems, and finally to the question of named-entity ambiguity and disambiguation.

In Section 3, we explore other relevant IE applications: relation detection, event extraction, and template filling. Relation detection allows to find links between entities extracted during the NER stage, event extraction deals with actions in a temporal context, while template filling focuses on the situations recurring in a fixed, predefined format.

All these IE subtasks are relevant to our main objective since they are not domain-specific but allow to extract realities from a broad range of subjects. However, IE has some known limits which have to be addressed, including the absence of unique identifiers, a poor handling of multilingualism, and some quality issues. We therefore ground our work in formal linguistics while already pointing to epistemological limitations related to current challenges.

Contents

1 Background

This first section provides some context about natural language processing in general (Section 1.1) and information extraction in particular (Section 1.2) in order to put the next sections into perspective. For historical material, we are greatly indebted to Jurafsky and Martin (2009), whose classic handbook remained a source of inspiration for us throughout this thesis. In Section 1.3, we then proceed to distinguish IE from related tasks such as data mining, text analytics and term extraction.

1.1 Natural language processing

The idea of NLP can be tracked back to the designing of the Turing (1950) test. Instead of answering the question "Can machines think?" – which he considered "too meaningless to deserve discussion"[1] – Turing designed a test he called the Imitation Game and proposed to replace the original question by another: "are there imaginable digital computers which would do well in the imitation game?"

In order to pass the test, a machine has to fool an examiner into believing it is a human being by answering their questions as a real person would. If the computer is able to trick the interrogator, Turing argues, then it can be considered "intelligent" (see Figure I.1):

> "I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning." ”

This famous test, which is also the starting point of artificial intelligence (Russell and Norvig, 2009), has been heavily criticised over the last half-century for a range of reasons,[2] but it remains the first attempt to demonstrate what a *thinking machine* would be able to do. Turing had anticipated most the objections to his "imitation game", which makes his visionary work

[1] Although Turing adds: "Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted". This prediction, among many others of Turing's, has been proved right: French speakers will typically say *l'ordinateur réfléchit* while the hourglass (wait) cursor is displayed on the screen.

[2] See, for instance, Searle (1980) and his "Chinese room" challenge to AI, or the repeated critiques of Dreyfus (1972); Dreyfus and Dreyfus (1986).

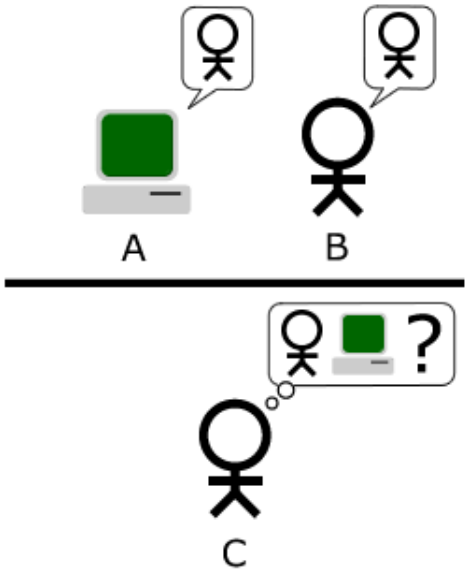


Figure I.1: Illustration of the Turing test, adapted from Saygin et al. (2000)

all the more remarkable, although the rather simplistic specifications of the test make it relatively easy to pass.[3]

Among these objections, we find the *arguments from various disabilities* which Turing says are of the form: "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one do X". X can be replaced by a variety of things, some of which do not seem so remote today: be beautiful (think of Apple's design efforts), make someone fall in love with it,[4] learn from experience (the purpose of machine learning), etc.

The list also contains the item "use words properly". What *properly* exactly means remains of course interpretation-prone, but the objection clearly stems from the fact that computers traditionally rely on formal languages rather than natural ones. Even Turing (1950) concedes: "Needless to say [instructions do] not occur in the machine expressed in English".[5]

[3]The question whether the Eugene Goostman computer program of the University of Reading actually passed the Turing test in 2014 is still open for debate: see `http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx` and `http://www.wired.com/2014/06/turing-test-not-so-fast/` for opposite views on that matter.

[4]A theme convincingly exploited in Spike Jonze's movie *Her*, among other works of fiction.

[5]And later: "It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak *English*" (emphasis added). For all his clear-sightedness, Turing did not envision that computers could have been interested in other languages.

From there on, the idea of computers being able to understand and produce human language was pursued in parallel in several fields, the main ones being computer science (which called it *natural language processing*) and linguistics (which called it *computational linguistics*). The early history of NLP is marked by the division between these two main approaches: a probabilistic, stochastic paradigm in computer science versus a more symbolic and formal/logical approach in linguistics.

The computer scientists used Bayesian models (Bledsoe and Browning, 1959) and developed the hidden Markov model (HMM), while Chomsky and fellow linguists applied finite-state automata to natural language in order to create the generative grammar (Chomsky, 1956) and subsequently developed parsing algorithms (Harris, 1962).

The stochastic approach further expanded in the 1970s and 1980s, notably with the work of researchers working at IBM's Thomas J. Watson Research Center and at AT& T's Bell Laboratories, while Chomsky's context-free grammar inspired the designing of unification grammars such as the Definite Clause Grammar (Pereira and Warren, 1980) and the Lexical Functional Grammar (Bresnan and Kaplan, 1982).

The 1990s saw the rise of hybrid systems, as the two historically disjointed branches of the field began to come together. Rule-based systems started to incorporate more and more probabilistic and data-driven models, as the development of the Web made available amounts of data previously unheard of. This tendency increased in the 2000s which saw the rise of machine learning algorithms, with systems becoming increasingly unsupervised[6] thanks to the sophistication of statistical techniques and the advent of Big Data (Banko and Brill, 2001).

In the 2010s, purely linguistic systems have become almost extinct, most approaches nowadays combining expert rules with some amount of statistical calculus (such as maximum entropy models and support vector machines) or being purely data-driven. Today, NLP remains a very prolific area of research boasting its own professional society (Association for Computational Linguistics), international conferences (International Joint Conference on Natural Language Processing, Conference on Empirical Methods in Natural Language Processing, Conference on Computational Natural Language Learning, etc.) and journals (Computational Linguistics, ACM Transactions on Speech and Language Processing, Linguistic Issues in Language Technology, etc.).

[6] Unsupervised learning does not require any manual input, whereas supervised techniques derive knowledge from labelled training data.

1.2 Information extraction

Extracting information from unstructured text is a critical task in order to make sense of the vast amount of data that we have been accumulating in the digital era. The aim of information extraction (IE) is therefore to extract useful, structured information from human-readable documents and to store it in data structures (such as relational and graph databases), thereby allowing their exploitation by machines.

In the field of medicine, for instance, practitioners would need to read around the clock to keep up with the literature of their speciality: in total, over a million medical articles are published every year. "Faced with this flood of information, clinicians routinely fall behind, despite specialization and subspecialization" (Gillam et al., 2009). IE allows scores of scientific papers to be processed and mined, reducing the workload imposed on physicians and other specialists.

The exact definition of information extraction is subject to some debate among scientists and its scope varies with the perspective in which it is considered. Building upon former attempts to define IE by Cowie and Lehnert (1996) and Riloff and Lorenzen (1998), Moens (2006, p. 4) offers the following definition (emphasis added):

> "Information extraction is the *identification*, and consequent or concurrent *classification* and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks."

The two main steps of IE are thus:

1. identification: recognising a given text string as useful information
2. classification: putting this information into a pre-established semantic category

Note that *disambiguation* is totally absent from the definition: this particular task is indeed not on the agenda of traditional IE, although this limitation can be overcome as will be shown in Chapter II. Moens also stresses the need for *suitability*: IE is not an end in itself but rather a way to achieve further processing of the information extracted. This is an important dimension to which we will come back in detail in chapters III and IV, when introducing the notion of *fitness for use*.

Information extraction is not to be confused with information retrieval, the task performed by search engines consisting in finding sets of relevant *documents* (Figure I.2), whereas IE focuses on the relevant *facts* in order to store them in a structured way (Figure I.3).[7] Transforming the information contained in a document into usable knowledge is also an essential part of the process of semantic enrichment which is at the heart of this dissertation.

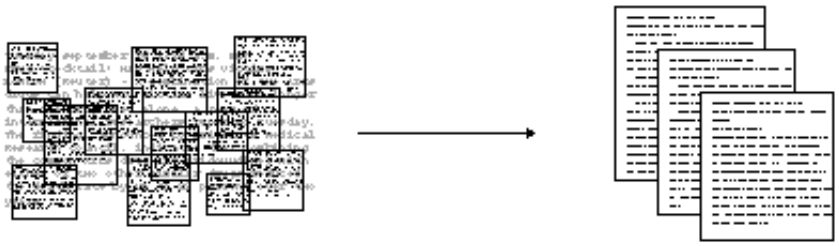


Figure I.2: Information retrieval

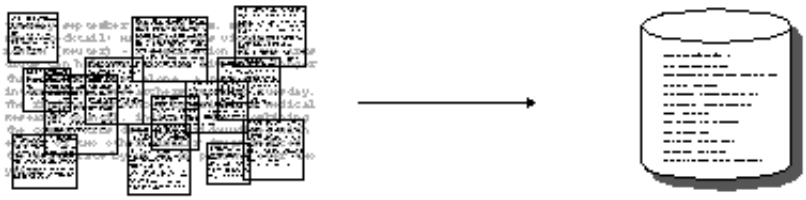


Figure I.3: Information extraction

Information retrieval mostly relies on the "bag-of-words" model which uses a simplified view of language as a loose sequence of unrelated words, while IE aims to take the syntactic and semantic dimensions of language into account. This makes IE systems more challenging to design in a multilingual context, but also more apt to capture knowledge in all its linguistic diversity.

Itself a sub-field of natural language processing, information extraction can be further divided into more concrete applications. Jurafsky and Martin (2009, pp. 761–791) list named-entity recognition, relation detection, event extraction, temporal analysis, and template filling as examples of such tasks, all of which will be covered to some extent in the remainder of this chapter.

[7] Both figures are reproduced from `https://gate.ac.uk/ie/` (CC BY-NC-SA).

1.3 Related fields

The choice of the term *information extraction* over other valid designations to describe the starting point of this dissertation is significant and justified by our integration into the NLP tradition. IE is also the oldest term to refer to the process of transforming unstructured text into processable information. This field of study should be distinguished from related tasks, which overlap in parts with IE but sometimes originate from another period, in other domains or from other communities of researchers. These different fields are briefly presented below and summarised in Table I.1.

Data mining covers knowledge discovery from structured sources such as databases. Interest in this topic started in the late 1980s with the Association for Computing Machinery's (ACM) Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) and was studied in depth during the 1990s, in the work of Fayyad et al. (1996) notably.

Terminology extraction (also called glossary extraction, term extraction, (automatic) term recognition or terminology mining) is interested in terms (common nouns) rather than entities (proper nouns). It can be traced back to the work of Bourigault et al. (1996) for machine translation. Multilingual and language-independent approaches were successfully applied to terms thanks to sub-sentential linguistic alignment, allowing the extraction of complex multi-word terms (Lefever et al., 2009).

Text mining (first described by Hearst (1999) as text data mining) is a broader field making use of IE techniques "at the intersection of natural-language processing, machine learning, data mining and information retrieval" (Mooney and Nahm, 2003). Data mining can be applied on structured information acquired from text by IE techniques, and in turn used to improve IE, as shown in Figure I.4.

Content extraction is the wording used in Automatic Content Extraction campaigns[8] from 1999 onwards. It is closely related to IE since the ACE campaigns focus on entities, relations and events. The term is seldom used outside the Automatic Content Extraction programme, except for a few authors.[9]

[8] `http://www.itl.nist.gov/iad/mig/tests/ace/`

[9] Gupta et al. (2005) for instance.

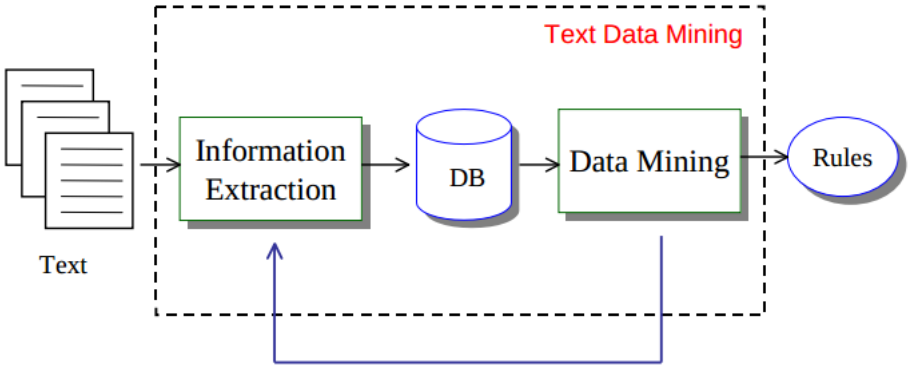


Figure I.4: Text mining, reproduced from Mooney and Nahm (2003)

Ontology learning (also ontology acquisition, ontology extraction or ontology generation) appeared in the early 2000s (Maedche and Staab, 2001) and is similar to terminology extraction, but focused on the semi-automatic creation of an ontology[10] for a given domain.

Text analytics is treated as an equivalent to information extraction by Jurafsky and Martin (2009, p. 759), while some other authors consider it to be closer to text mining (Feldman and Sanger, 2007). It is the favoured term in corpus linguistics.

Table I.1 gives a summary of all the fields presented above along with an approximate date of appearance of the term and broad domain of origin.

Field	Date	Domain
Information extraction	1982	Natural language processing
Data mining	1989	Computer science
Terminology extraction	1996	Machine translation
Text mining	1999	Computational linguistics
Content extraction	1999	ACE campaign
Ontology learning	2001	Information science
Text analytics	2004	Corpus linguistics

Table I.1: Summary of related fields

[10] In the information science sense, see Section 2.2 in Chapter II for more details.

In this section, we have introduced natural language processing and information extraction which offer a general framework for the later achievements to be carried out in this thesis. We have shown how IE differentiates itself from the simpler task of information retrieval, but also from a range of closely related, partially overlapping fields. The next section will now focus on named-entity recognition, which is a central task in IE.

2 Named-Entity Recognition

Named-entity recognition (NER) is a core subtask of information extraction consisting in the automatic identification and classification of named entities (proper nouns) from unstructured documents. It is an essential component of any semantic enrichment or knowledge discovery system and therefore requires to be discussed in detail before moving on to these more complex tasks. Consider Example 4 which shows a typical sentence from a newspaper:

Example 4. Captain Stuart Oswald, an old member of the Ypres League, died yesterday in Amiens.

This example contains entities of various type such as a person (Captain Stuart Oswald), an organisation (the Ypres League) and a geographical location (Amiens). The aim of NER is to recognise all of them. NER has several concrete applications beyond IE, for instance in machine translation (Babych and Hartley, 2003) and question answering (Mollá et al., 2006). After giving some initial context on the scope of NER (Section 2.1), we will survey common classifications of both named entities and NER systems (Section 2.2), before tackling the core subjects of named-entity ambiguity and disambiguation (Section 2.3).

2.1 Task definition

The concept of a "named entity", proposed by Grishman and Sundheim (1996), originally covered names of people, organisations, and locations as well as expressions of time, amounts of money, and percentages. Similarly, named entities were defined for the CoNLL 2002 and 2003 shared tasks as "phrases that contain the names of persons, organizations, locations, times, and quantities" (Tjong Kim Sang, 2002).

Over the years, NER has attracted the attention of researchers in many fields such as financial journalism (Farmakiotou et al., 2000), biology and biomedicine (Ananiadou and McNaught, 2006), and business intelligence (Saggion et al., 2007). As a result of this diversification of NER, the original range of named entities was progressively extended to include brands, genes, and even college courses (McCallum, 2005).

Nadeau and Sekine (2007), however, denounce this misuse of language and suggest that the term "named" in "named entity" should effectively be restricting its sense to entities referred to by rigid designators, as defined by Kripke (1980, p. 77): "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 today by most NER applications, which use "entities" and "named entities" interchangeably.

There is, as a result, no real consensus on the exact definition of a (named) entity, which remains largely domain-dependent. For Jurafsky and Martin (2009, pp. 725–726), a named entity is anything that can be referred to with a proper name, but they note that "what constitutes a proper name and the particular scheme used to classify them is application specific".

This absence of scientific agreement is by no way an oddity: the relative nature of empirical data makes them open to interpretation, as will be seen in detail in Chapter III. The *fitness for use* principle (introduced in Chapter IV), originating from data quality, is a good practical guide to entity validity: if an entity is useful for our needs, then it can be argued that it is valid, notwithstanding epistemological considerations. This is better understood with the help of an example:

Example 5. Excessive consumption of Côte d'Or chocolate has been shown to cause type-2 diabetes.

A NER application interested in companies or brands will count "Côte d'Or" as a valid entity, but will probably tend to disregard "diabetes" altogether. In contrast, a biomedical application will extract "type-2 diabetes" as an entity of type DISEASE, for instance, while ignoring the brand: the term "chocolate" could possibly be recognised as a CAUSE or disease-inducing entity, but the company producing it is probably irrelevant in this context.

Ehrmann (2008) provides a comprehensive overview of the evolution of the concept of a named entity from linguistics to NLP. She offers the following justification for the distinction between NER and terminology (p. 174):

> « Il importe ici de souligner la différence entre la reconnaissance d'entités nommées et la terminologie : celle-ci s'intéresse aux termes d'une langue en tant qu'ils représentent les traces de concepts d'un domaine donné, celle-là s'intéresse à des expressions linguistiques en tant qu'elles représentent les traces de référents d'un modèle donné. S'il n'y a pas de modèle, il n'y a pas d'entités nommées. »

However, the argument seems flawed since the model used could as validly choose the concepts as referents. Returning to Example 5, the term *diabetes* could be seen as the mention of a concept of a given domain, or alternatively be considered a referent from a given model. The distinction between entities and terms is therefore not as clear-cut as Ehrmann would like us to think.

Another interesting approach was proposed by Chiticariu et al. (2010) who drew up a list of criteria for the domain customisation of NER, including entity boundaries, scope and granularity among others. The definition of a named entity, according to them, is never fixed but depends both on the data to process and on the application processing it.

However, the building of a reference corpus for evaluation purposes (see Chapter V) requires entity categories to be well defined in order to give precise instructions to human annotators. In fact, named entities being empirical by nature, the objective evaluation of their correctness is practically impossible, except if some external referent is substituted for the real world.

Using such a gold-standard corpus (GSC) allows to evaluate entities in a pseudo-deterministic manner. By disregarding some subtleties inherent to the richness of reality, this artificial construction becomes the ground truth under the closed-world assumption: all entities in the GSC are considered to be necessarily correct, while those not in the GSC are necessarily incorrect. We will now examine some named-entity classifications that have been used for similar purposes in the past.

2.2 Typologies

In this section, we offer common classifications of both named entities and NER systems. These typologies are useful for the evaluation of applications, as explained above, but they also reduce the flexibility that the users would be entitled to expect from an extraction tool in order to meet their needs. The aim is therefore to understand how and why these classifications were proposed in order to build on their strengths, while discarding some of their rigidity. In doing so, we will partially answer our first research question, and thereby strengthen our thesis according to which generalisation and decompartmentalisation can lead to better NER.

2.2.1 Named-entity typology

Although several typologies of named entities have been proposed over the years, “there is no fixed and final set of classes for named entities” (Bingel and Haider, 2014). Stern (2013, p. 16) even argues that every NER task defines its own typology. Their levels of granularity vary widely: for instance, some tools distinguish between Countries, Cities, Mountains etc. (Sekine et al., 2002), whereas other authors group all these entities into a single Location category. Some of the most widespread typologies, developed in the framework of vast evaluation programmes, are detailed below.

Message Understanding Conferences (MUC) started in the late 1980s and lasted until 1997, aiming to encourage innovative approaches to IE. A special focus on NER and co-reference resolution was added for MUC–6 and MUC–7. In their retrospective paper on MUC, Grishman and Sundheim (1996) reflect on the division of expressions into TIMEX (temporal expressions: date and time mentions), NUMEX (numeric expressions: amounts and percentages), and ENAMEX (entity names), the latter comprising Locations, Organisations and Persons.

Conferences on Computational Natural Language Learning (CoNLL) were launched in 1997 at the initiative of the Association for Computational Linguistics's special interest group on natural language learning, and were held up to 2008. From 1999 onwards, they included a shared task for participants to compete on a given IE topic. The CoNLL 2002 and 2003 shared tasks were centred on the question of language-independent NER, for the purpose of which named entities were divided into person names (PER), organisations (ORG), locations (LOC) and miscellaneous names (MISC) (Tjong Kim Sang, 2002). The MISC category is a handy catch-all for rare types of entities but obviously raises epistemological questions regarding its validity and usability.

Automatic Content Extraction (ACE) was a programme run from 1999 to 2008 for developing IE technologies and centred on the extraction of entities, relations and events. ACE entities were categorised into seven types: Geo-political entity, Facility, Location, Organisation, Person, Vehicle, and Weapon (Doddington et al., 2004). These categories allowed for more fine-grained distinction between entities (with Geo-political entity accounting for cases where it would be difficult to choose between Location and Organisation) but also included entities that were very specific and not strictly proper nouns (like vehicles and weapons), pushing the case for a more inclusive approach.

National and regional programmes also introduced their own specificities. In France, for instance, the ESTER[11] campaign added Function (political, military...) and Human Production (award, work of art...) to PER, ORG and LOC. In the Netherlands, the SoNaR[12] project (Oostdijk et al., 2008) divided named entities into six broad categories: PER, ORG, LOC, PRO (products), EVE (events) and MISC, which were then further divided into 19 subcategories. Figure I.5 shows the SoNaR categories in order to illustrate this diversity.

[11] `http://www.afcp-parole.org/camp_eval_systemes_transcription/`

[12] Loose acronym for STEVIN (Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands) Nederlandstalig Referentiecorpus: `http://lands.let.ru.nl/projects/SoNaR/`.

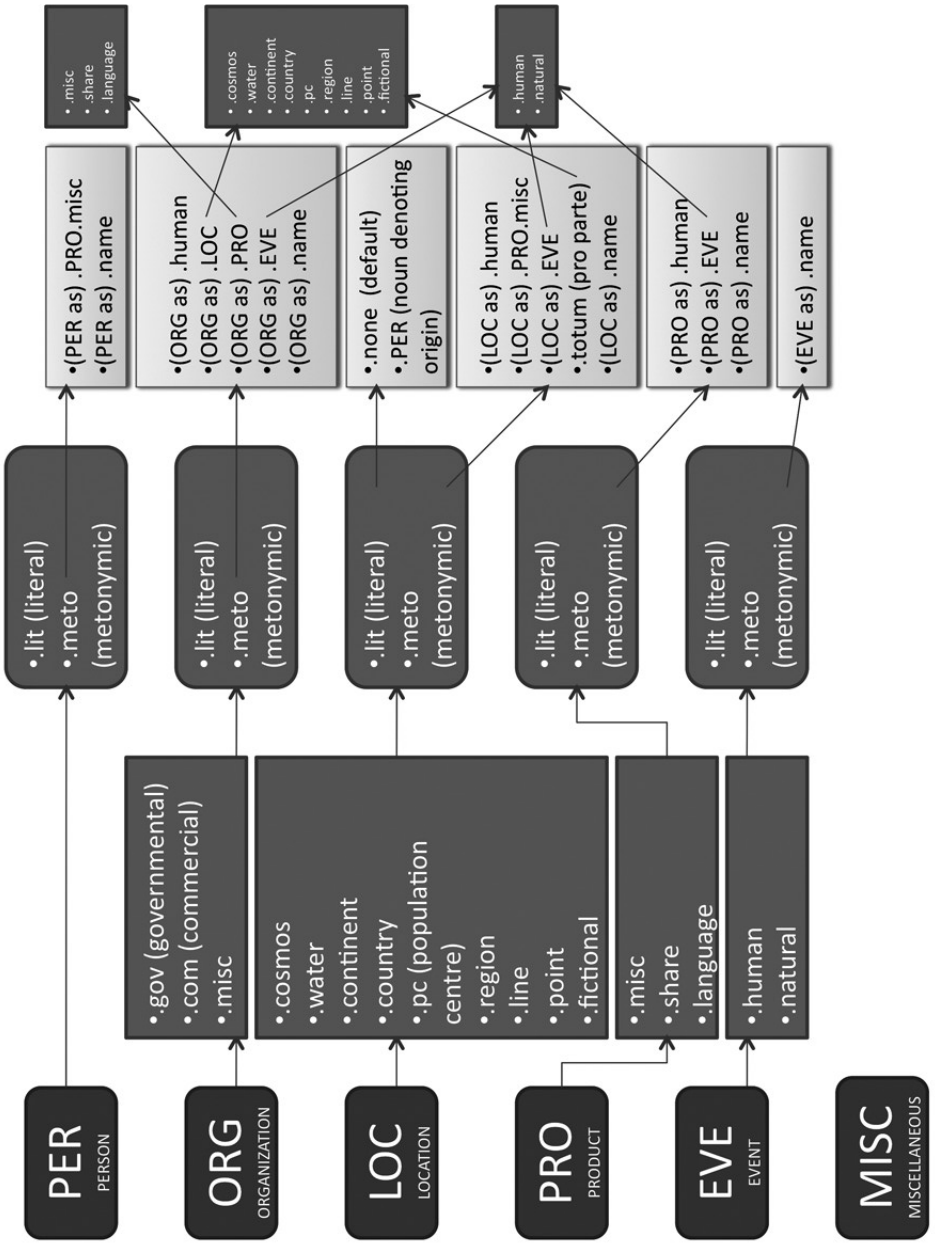


Figure I.5: SoNaR named entity typology, reproduced from Hoste (2009)

2.2.2 Types of NER systems

Broadly speaking, all NER systems can be divided into linguistic, probabilistic and hybrid approaches. This tripartite division originates from the development of NLP in general, as detailed in Section 1.1. In order to comprehend the present challenges faced by NER, we quickly outline this evolution and show how more recent approaches make the best of two worlds in order to gain in robustness and generalisability, showing a clear shift in the direction of the main thesis defended within this dissertation.

Linguistic or rule-based approaches The first NER systems designed in the early nineties were largely symbolic, relying on refined linguistic knowledge and rules. As these rules had to be encoded manually by linguists, these approaches were extremely time-consuming and costly to maintain. Some linguistic systems also made use of gazetteers, i.e. large lists of common first names or locations for instance, which helped to increase the coverage of the rules. For other types of entities, however, such as organisations or products, gazetteers are hardly capable of providing an exhaustive list of existing possibilities,[13] and patterns aiming at recognising them are always incomplete to some extent.

Probabilistic or data-driven approaches The availability of huge corpora prompted a shift towards more statistical models, taking advantage of machine learning techniques (such as support vector machines and conditional random fields) on large corpora. Rau (1991), for instance, automatically extracted thousands of company names from over one million words of financial news, reporting a 25% increase in recall compared to human annotators.[14] However, totally unsupervised approaches often suffer from a lack of precision, leading researchers to experiment with semi-supervision (human intervention in the automated process). Ji and Grishman (2006) make use of two semi-supervised learning algorithms to improve name-tagging, i.e. the recognition of the names of persons (and therefore a subset of NER). Contrary to Banko and Brill (2001), they conclude that more data does not always mean better results, some amount of human intervention being necessary for the feature selection process.

[13] Although some resources can become available in a closed-world perspective. In Belgium, companies could be disambiguated by their number in the Crossroads Bank for Enterprises.

[14] While linguistic approaches typically achieve good precision thanks to the manual work of language experts, they lack in coverage due to the difficulty to predict all future situations. Probabilistic approaches, with their ability to process much more information, can make up for this shortcoming and achieve better recall (see Chapter V for a discussion of these metrics).

Hybrid approaches Today, most IE systems combine some linguistic insight with the power of data-driven algorithms (Watrin, 2006). An example of such a technique is maximum entropy (MaxEnt) in which potentially relevant features are listed by a human being but the combining and weighting tasks are left for the machine to compute. For instance, one could hypothesise that two capitalised words in a row is a potential indicator of a person name (Forename Surname) without having to precisely define rules and exceptions that can rapidly become burdensome. Examples of systems making use of the MaxEnt algorithm can be found in Chieu and Ng (2002) and Bender et al. (2003) among others. Curran and Clark (2003) show that this approach is also promising for language-independent NER (cf. Chapter IV). For a more complex approach combining different types of classifiers, see Florian (2002); Florian et al. (2003).

Kozareva et al. (2007) also combine three approaches – Hidden Markov Models (HMM), Maximum Entropy (MaxEnt) and Memory-based learning (MBL) – using a voting strategy, thereby achieving a 98.5% accuracy for named-entity identification and almost 85% for classification on Spanish data. It is unclear, however, how their system would fare on other languages as the algorithms were trained on language-specific data. Brun et al. (2007, 2009) present a hybrid method for the resolution of named-entity metonymy (see Section 2.3.3), in particular names of places and organisations. In the framework of the SemEval 2007 challenge, their XRCE-M system developed at the Xerox Research Center combines a syntactic parser with distributional analysis. Hybrid approaches are also favoured by entity linking tools, as will be seen in Chapter V.

2.3 Entity ambiguity and disambiguation

Natural language is intrinsically ambiguous.[15] As a result, several cases of ambiguity can arise while performing NER. This section will briefly discuss the main forms of ambiguity to be found at named-entity level, including synonymy, homonymy, polysemy and metonymy. While some of these may seem to be one and the same, subtle variations of meaning can lead to differences in their handling later on. The accurate disambiguation of entities is a critical challenge for several NLP-related applications, and a key dimension of the general approach defended within this dissertation. In fact, using uniform resource identifiers (URIs) to disambiguate entities is not domain-specific nor language-dependent, as we will show in Chapter II.

[15] This ambiguity is reinforced by the fact that "natural language is its own metalanguage", as noted by Boydens (2011, p. 125): clarifying the use of a word or construction can only be done with more words that are in turn ambiguous, *ad nauseam*.

In the context of semantic enrichment, particularly, entity ambiguity can be a major problem, because accurate NER is a prerequisite for further processing such as entity linking and related search recommendations. Errors occurring early in the pipeline can result in the attribution of a wrong URI to an entity, and in turn lead to irrelevant (or even, in the worst-case scenario, offensive) suggestions for the users.

For instance, a user interested in finding information about the right-wing mayor of French town Châtenay-Malabry, Georges Siffredi, may enter the search keywords `siffredi UMP`. A wrong identification of the French snowboarder Marco Siffredi may cause unwanted suggestions about fellow Everest-climbers which is relatively harmless, but a disambiguation as the Italian pornographic actor Rocco Siffredi could have more damaging consequences. Similarly, a mistaken understanding of the abbreviation UMP as Universal Music Publishing or Universale Maschinenpistole (instead of "Union pour un Mouvement Populaire") could trigger a negative user experience.

By investigating different types of language ambiguity, we reach a level of understanding which will prove crucial later on for the correct disambiguation and reuse of entities in real-world applications. Several typologies of lexical relations have been proposed (see Allan (1986) for instance). Our own order of presentation is based on the effects on IR: silence on the one hand (caused by synonymy) and noise on the other hand (caused by homonymy, polysemy, or metonymy).

After this short survey of the various types of ambiguity to be found in natural language, Section 2.3.4 will focus on the disambiguation process (which is not explicitly included in IE in the traditional sense) and will quickly investigate ways to overcome this limitation, a point that will be further developed in Chapter II. As already emphasised, disambiguation is central for the semantic enrichment task applied to our case study of multilingual periodicals, and will also bring us a step closer to generalisation, in accordance with our main thesis, by granting an equal status to all information units without any distinction of language, domain, or type of content.

2.3.1 Synonymy

Just like common nouns can be synonymous to some extent and share part or whole of their meanings, a single entity can be referred to with different, overlapping names.

Example 6. The city of Ypres (Dutch Ieper) was nicknamed "Wipers" by British troops during World War I.

As shown by Example 6, synonymy occurs within a single language (Ypres versus Wipers) but even more pervasively across languages (Ypres versus Ieper). Moreover, different references to the same entity can (and often will) coexist inside a given document. Co-reference resolution, including the subtask of anaphora resolution, is therefore an essential part of NER:

Example 7. JOZEF DE VEUSTER was born in 1840 in Tremelo, Belgium. *He* is better known as FATHER DAMIEN, *his* chosen religious name.

In Example 7, the two named entities in small caps refer to the same person. Furthermore, two pronouns in italics also refer to Damien. The task of linking JOZEF DE VEUSTER to FATHER DAMIEN is the task of co-reference resolution, while understanding to whom the pronouns *he* and *his* refer is the task of anaphora resolution.

2.3.2 Homonymy and polysemy

Since words often hold several meanings, homonymy can be seen as the main cause of ambiguity in text, generating noise in search results. Homonymy comes in different forms:

Proper vs. common noun : The same word can be used to refer to a company (*Apple*) or a fruit (*apple*), the latter being a common noun and thus not a valid named entity in the stricter sense. Such cases are the easiest to resolve since the two words differ by their capitalisation (consider, however, the use of a common noun at the beginning of a sentence, such as in "Apple is my favourite fruit.").

Cross-type homonymy : A named-entity mention can refer to types of entities completely different. For instance, *Bern* that refers either to the capital of Switzerland or to French journalist Stéphane Bern. In those cases, named-entity detection is not sufficient: a classification into categories is needed to differentiate between the two uses.

Same-type homonymy : Finally, a name can also refer to different entities of the same type, such as *Paris* (capital of France) and *Paris*, Texas or *George Bush* (father and son). These cases are the more difficult to handle, as even a classification into type LOC or PER is not sufficient to unambiguously identify the entity actually mentioned in the text. Full disambiguation is necessary here, as we will discuss in Chapter II.

Polysemy is closely related to homonymy but differs from it by the fact that polysemes share a common origin, while homonyms are coincidental. Consider the following examples:

Example 8. The Benevolent Bank is located on the bank of the Yser.

Example 9. South American people still believe in the American dream.

In Example 8, the two senses of *bank* are homonyms because their sharing the same spelling is coincidental. Compare this with Example 9 where the term *American* is polysemous since it can refer either to a person living on the continent America, or to a citizen (or, here, value) of the United States, the two senses being etymologically related. Most of the time, however, there is no practical difference in the handling of these two types of ambiguity.

2.3.3 Metonymy

Cross-type ambiguity frequently arises from the use of metonymy, a figure of speech consisting in replacing a name by another. Typical cases include geopolitical entities such as *Ypres* that can be used either as locations or as organisations depending on the context, as can be seen in the following examples:

Example 10. No one was allowed into [*LOC* Ypres] on any account.

Example 11. Josse Destrée got financial support from [*ORG* Ypres].

Capital cities are also commonly used to refer to their country:

Example 12. [*ORG* Brussels] agreed to fund our research project.

The case of *Brussels* is even more ambiguous since its metonymic use can refer either to the Belgian government or to the European institutions.

Following the work of Markert and Nissim (2006), Brun et al. (2007, 2009) have studied in detail the various cases of metonymy and its implications for NER in the context of the SemEval 2007 campaign.[16] This type of ambiguity can arise in more subtle ways and is generally more difficult to detect, especially due to the diversity of linguistic variations.

2.3.4 Disambiguation

According to pioneers Palmer and Day (1997), "the goal of the N[amed] E[ntity] task is to automatically *identify* the boundaries of a variety of phrases in a raw text, and then to *categorize* the phrases identified" (emphasis added).

[16] `http://nlp.cs.swarthmore.edu/semeval/`

This echoes the steps of identification and classification we presented in Section 1.2, as defined by Moens (2006) for information extraction in general.

Similarly, competitors in the CoNLL 2002 shared task such as Carreras et al. (2002) divide the more general task that they call Named Entity Extraction into two main subtasks: recognition and classification. They choose to address these tasks separately, but note that "an open line of research to be addressed is the simultaneous approach of named-entity recognition and classification tasks, so each decision may take advantage of the synergy between both knowledge levels".

These definitions, however, restrict the NER task to a mere classification: to categorise Washington as a PER allows to rule out the LOC meaning of Washington, DC (cross-type homonymy), but it fails to resolve to more insidious same-type homonymy: are we talking about president George Washington, the singer Dinah Washington or some other person still? Ehrmann (2008, p. 162) sees lexical disambiguation as the finer-grained level of classification:

> « Le second usage du sens en [Traitement Automatique des Langues], la reformulation, est une opération davantage interne à la langue : il s'agit de donner, pour des mots, des paragraphes ou des textes, des mots ou paraphrases équivalents sémantiquement. Au niveau lexical, la reformulation correspond à de la désambiguïsation et peut être mise en œuvre pour les noms d'une part, et pour les autres catégories, verbes, adjectifs, etc. d'autre part. La désambiguïsation nominale rejoint en quelque sorte l'opération de catégorisation dans la mesure où on s'intéresse ici au niveau le plus fin de la classification : pour le mot barrage par exemple, il s'agit d'identifier l'objet du monde dont il est question, un barrage de police ou un barrage hydraulique. »

Likewise, Stern (2013, p. 17) notes that named-entity categorisation is insufficient for the needs of semantic annotation, in which entities have to be treated as unequivocally identifiable, extra-linguistic individuals. In order to know *what* a document is really *about* (and not only what kind of entity it is about), a referential link between a mention and its conceptual representation (in the form of a knowledge base entry, for instance) is needed. Although a way to achieve fully automated disambiguation was unthinkable when NER appeared in the nineties, it can now be envisioned with the help of knowledge bases such as DBpedia, as will be detailed in Chapter II. This method, however, also raises issues about the quality of reference material, a problem that will be dealt with in Chapter IV.

3 Relations and Events

In addition to NER, two IE applications have retained our attention as potential components of a semantic enrichment system: the extraction of relations between entities and the processing of events (including in its simpler form involving templates). Both rely on NER for their success, since a relationship is generally established between two entities, while events involve temporal and geographical components expressed by named entities. Example 13 gives a better notion of what is involved in the two tasks:

Example 13. LORD FRENCH TO THE NATION
This month, the seventh anniversary of the birth of the Ypres Salient,[17] has been signalized by the following letter to the Press by F.-M. Earl French, the President of the Ypres League.

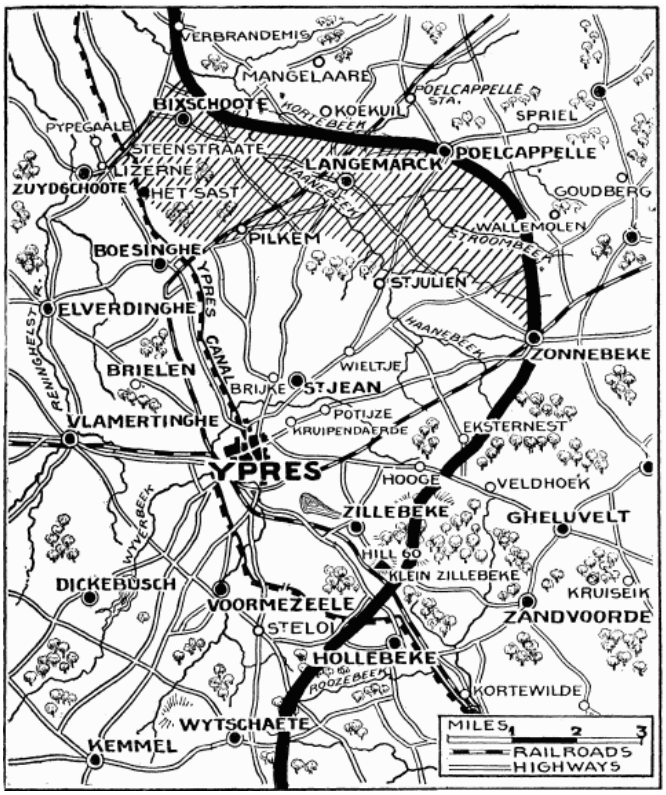


Figure I.6: Illustration of the Ypres Salient

[17] In military terms, a salient is a battlefield feature that projects into enemy territory (see `https://en.wikipedia.org/wiki/Ypres_Salient`). The Ypres Salient is shown in Figure I.6.

The fact that Lord French was the president of the League is a relation between two entities that can be represented in the following manner:

[PER Lord French] IS-PRESIDENT-OF [ORG Ypres League]

Note that extracting this relationship requires, as a preliminary step, to resolve a coreference between the two entities "Lord French" and "F.-M. Earl French" which refer to the same person. In addition to this relation, two events can be deduced from the text by using the date of the article (1 October 1923):

$$\text{EVENT-1:} \begin{bmatrix} \text{EVENT-TYPE:} & \text{birth} \\ \text{SUBJECT:} & \text{Ypres Salient} \\ \text{DATE:} & \text{October 1916} \end{bmatrix}$$

$$\text{EVENT-2:} \begin{bmatrix} \text{EVENT-TYPE:} & \text{announcement} \\ \text{ANNOUNCER:} & \text{F.-M. Earl French} \\ \text{RECIPIENT:} & \text{the Press} \\ \text{DATE:} & \text{October 1923} \end{bmatrix}$$

We hereafter briefly present these tasks as they will be of interest to us in the next chapters. More specifically, relation detection will be associated with semantic relatedness between concepts in Chapter II, and the extraction of events and temporal analysis will be of decisive use in the historical context detailed in Chapters III. Of course, relations and events are also affected by multilingualism: attention will be duly paid to this important question in Chapter IV (Section 2.1.2). Eventually, understanding how entities relate to one another – and how they are progressively constructed over time – will help us to handle them efficiently in order to build useful applications in Chapter V.

Starting with relation detection in Section 3.1, we will discuss typologies and systems before doing the same for event extraction and temporal analysis in Section 3.2, finishing with template filling in Section 3.3.

3.1 Relation detection

Relation detection and classification, also called relation extraction, focuses on the relationships between entities discovered in the NER phase. These relations can be of very diverse types and are sometimes unpredictable, but most systems rely on the recurrence of pre-established patterns. We will now focus on typologies of relations and on a number of relation detection systems in order to learn what can be gained for our semantic enrichment purpose, and how these can be improved with Linked Open Data.

3.1.1 Typology of relations

Several models have been put forward to categorise relations between entities. The ACE programme (Doddington et al., 2004) included 24 subtypes of relations grouped into five general types: `located`, `near`, `part`, `role` and `social`. Aone and Ramos-Santacruz (2000) went even further by proposing an ontology of 37 relation types extended from MUC, shown in Table I.2.

Relations	
Place Relations	**Artifact Relations**
Place–Name&Aliases	Artifact–Name&Aliases
Place–Type	Artifact–Type
Place–Subtype	Artifact–Subtype
Place–Descriptor	Artifact–Descriptor
Place–Country	Artifact–Maker
	Artifact–Owner
Organization Relations	**Person Relations**
Org–Name&Aliases	Person–Name&Aliases
Org–Descriptor	Person–Type
Org–FoundationDate	Person–Subtype
Org–Nationality	Person–Descriptor
Org–TickerSymbol	Person–Honorific
Org–Location	Person–Age
Org–ParentOrg	Person–PhoneNumber
Org–Owner	Person–Nationality
Org–Founder	Person–Affiliation
Org–StockMarket	Person–Sibling
	Person–Spouse
	Person–Parent
	Person–Grandparent
	Person–OtherRelative
	Person–BirthPlace
	Person–BirthDate

Table I.2: Relation ontology, adapted from Aone and Ramos-Santacruz (2000)

Jurafsky and Martin (2009) list generic relations such as employment (PER works for ORG; Example 14), family (PER is the son of PER; Example 15), part-whole (ORG is a division of ORG; Example 16) and geo-spatial relations (PER/ORG is based in LOC; Example 17).

Example 14. De 30-jarige Serge Zaidman, die werkt voor diamanthandel Bativier uit Antwerpen, werd te Hongkong door 2 mannen overvallen.

Example 15. Major Tubb was the son of Harry and Emma E. Tubb, of St. Helena, Longwood East, Victoria, Australia.

Example 16. La Verrerie de Zeebrugge, filiale de l'Union des Verreries mécaniques de Dampremy a été fermée jeudi pour un temps indéterminé.

Example 17. [...] via Nieuport and Dunkerque (where we saw the first bomb fall on that peaceful town) to Poperinghe, where Corps H.Q. were established in the Convent School [...]

3.1.2 Relation detection systems

We present hereafter a few notable systems in chronological order. Seminal research in relation identification and extraction was conducted in the late 1990s. At the IBM Watson Research Center, Byrd and Ravin (1999) outlined the key features of the task: the use of patterns, frequency filters, selectional restrictions, and the organisation of the relations in a lexical network.

Bouillon et al. (2001) used part-of-speech and semantic tagging in order to "automatically acquire N[oun]-V[erb] pairs whose components are linked by one of the qualia structure roles". Incidentally, their approach applies to terms (common nouns) rather than entities (proper nouns), but we can notice that the underlying principle is very similar.

In the biomedical domain, Ramakrishnan et al. (2006) focused on explicit and implicit relationships between entities and developed a rule-based method for the extraction of these relations from unstructured text and their conversion into RDF triples. They used Medical Subject Headings identifiers as URIs for entities and aimed to uncover hidden, valuable relationships from what they call "Undiscovered Public Knowledge" (in English only, however).

Auer and Lehmann (2007) exploited Wikipedia templates in order to discover relations between concepts. These templates include People, Organisations and Geographic entities (broadly equivalent to our named-entity classes), but also Plants and Education, demonstrating once again the artificiality of the distinction between entities and terms.

Kramdi et al. (2009) offer a generic approach to relation extraction based on ontologies and the Semantic Web. They propose the extraction of relations between concepts or instances without domain-specific knowledge, making their approach adaptive and robust, by modifying the $(LP)^2$ algorithm (Ciravegna, 2001) in order to account for relevant context. The system is largely entity-agnostic, the aim being to build ontologies and annotate texts for the general domain, whereas most systems are domain-specific.

Akbik and Broß (2009) explore the idea of building large knowledge bases, extracting semantic relations with the help of dependency grammar patterns. In discussing the general applicability of their approach, the authors list a number of advantages for building a system on the Wikipedia model, including "the *coverage and diversity*, the *high quality of content*, the *actuality*, the *factual language* and the *internal link structure*" (italics theirs). However, they note that "the hypothesis underlying [their] approach is a purely linguistic one and therefore is only bound to English language".

Wong et al. (2009, p. 267) offer an original view of relation detection relying on ontologies rather than on entities and patterns. They note that "relation acquisition techniques which require named entities have restricted applicability since many domain terms with important relations cannot be easily categorised. In addition, the common practice of extracting triples using only patterns and grammatical structures tends to disregard relations between syntactically unrelated terms". To overcome these shortcomings, they propose a hybrid approach divided into two phases: term mapping and term resolution.

More recently, Nebhi (2013) investigated rule-based methods with syntactic parsing for the extraction of relations from DBpedia, while Augenstein et al. (2014) discarded knowledge bases as too incomplete and preferred to extract the relations directly from unstructured Web text. These alternative methods will be investigated in Chapter II, while a critical assessment of their quality will be performed in Chapter IV.

3.2 Event extraction and temporal analysis

Event extraction consists in finding and analysing *what happened*. According to Hogenboom et al. (2011), an event can be represented as "a complex combination of relations linked to a set of empirical observations from texts". Events are key ways to apprehend knowledge: an event can become "révélateur de réalités autrement inaccessibles" (Pomian, 1984, p. 35). Similarly to named entities and relations, several types of events exist and can be divided into thematic categories. The REES system of Aone and Ramos-Santacruz (2000), for instance, covers 60 types of events, as shown in Table I.3.

Events	
Vehicle	**Transaction**
Vehicle departs	Buy artifact
Vehicle arrives	Sell artifact
Spacecraft launch	Import artifact
Vehicle crash	Export artifact
	Give money
Personnel Change	**Business**
Hire	Start business
Terminate contract	Close business
Promote	Make artifact
Succeed	Acquire company
Start office	Sell company
	Sue organization
	Merge company
Crime	**Financial**
Sexual assault	Currency moves up
Steal money	Currency moves down
Seize drug	Stock moves up
Indict	Stock moves down
Arrest	Stock market moves up
Try	Stock market moves down
Convict	Stock index moves up
Sentence	Stock index moves down
Jail	
Political	**Conflict**
Nominate	Kill
Appoint	Injure
Elect	Hijack vehicle
Expel person	Hold hostages
Reach agreement	Attack target
Hold meeting	Fire weapon
Impose embargo	Weapon hit
Topple	Invade land
Family	Move forces
Die	Retreat
Marry	Surrender
	Evacuate

Table I.3: Event ontology, adapted from Aone and Ramos-Santacruz (2000)

3.2.1 Event annotation

For the annotation of events, Pustejovsky et al. (2003) introduced TimeML,[18] a robust specification language for temporal expressions in natural language. In 2009, ISO recognised TimeML as a standard for time and event markup and annotation.[19] However, the XML-like syntax used by TimeML, along with the high number of possible attributes, makes it a complex markup language that can be daunting to master. For instance, a sentence as simple as:

Example 18. John taught 20 minutes every Monday.[20]

will be formalised in TimeML as:

```
John
<EVENT eid="e1" class="OCCURRENCE">
taught
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB" tense="PAST"
 aspect="NONE" polarity="POS"/>
<TIMEX3 tid="t1" type="DURATION" value="P20TM">
20 minutes
</TIMEX3>
<TIMEX3 tid="t2" type="SET" value="xxxx-wxx-1" quant="EVERY">
every Monday
</TIMEX3>
<TLINK timeID="t1" relatedToTime="t2" relType="IS_INCLUDED"/>
<TLINK eventInstanceID="ei1" relatedToTime="t1" relType="DURING"/>
```

This complexity explains the limited success of TimeML outside academia, although the annotation of events remains very much an issue.

3.2.2 Event extraction systems

For the authors of the SoNaR corpus (Oostdijk et al., 2008), events are just another type of named entities (see Section 2.2). However, as events are often more complex in form than people or places, we chose to handle them separately. Most NER tools are not well equipped to locate events: Buitinck and Marx (2012) for instance report a relatively low F-score of 71% on the EVE category against 90% for locations.

[18] `http://www.timeml.org/`

[19] ISO 24617-1:2009, later integrated into the SemAF framework (ISO 24617-1:2012).

[20] Example borrowed from `http://www.timeml.org/`.

A complete overview of event extraction techniques can be found in Hogenboom et al. (2011). Defining the task as an application of text mining consisting in "deducing specific knowledge concerning incidents referred to in texts", the authors offer a survey of various event extraction methods divided into data-driven, knowledge-driven and hybrid methods. Data-driven approaches rely on probabilistic models and Big Data, but fail to establish valid meaning for events. In the words of Hogenboom et al. (2011), "statistical relations do not necessarily imply semantically valid relations, nor relations that have proper semantic meaning".

In contrast, knowledge-driven approaches require a certain level of expert domain knowledge in order to build linguistic patterns, either lexico-syntactic (using regular expressions) or lexico-semantic (using gazetteers or ontologies). Hybrid approaches combine the use of large amounts of data with expert knowledge, thereby overcoming the limitations of both methods (lack of semantics and pattern maintainability). The authors conclude that knowledge-based event extraction techniques are better suited and easier to grasp for casual users, whereas data-driven and hybrid methods are more powerful for advanced users.

The extraction of temporal events has gained particular attention in the field of biomedicine, since events are crucial in determining how proteins and genes interact with one another. Like NER, biomedical event extraction has evolved from fully linguistic systems (Yakushiji et al., 2001) to more robust, semi-supervised (or even unsupervised) approaches (Zhou and He, 2011). In 2009, the BioNLP conference shared task focused on event extraction (Kim et al., 2009), with more than twenty participants competing to build the best system able to detect complex biomolecular events such as binding and regulation. Miwa et al. (2010) used machine learning with rich features in order to design an event detector that outperforms the best system from the BioNLP09 shared task challenge.

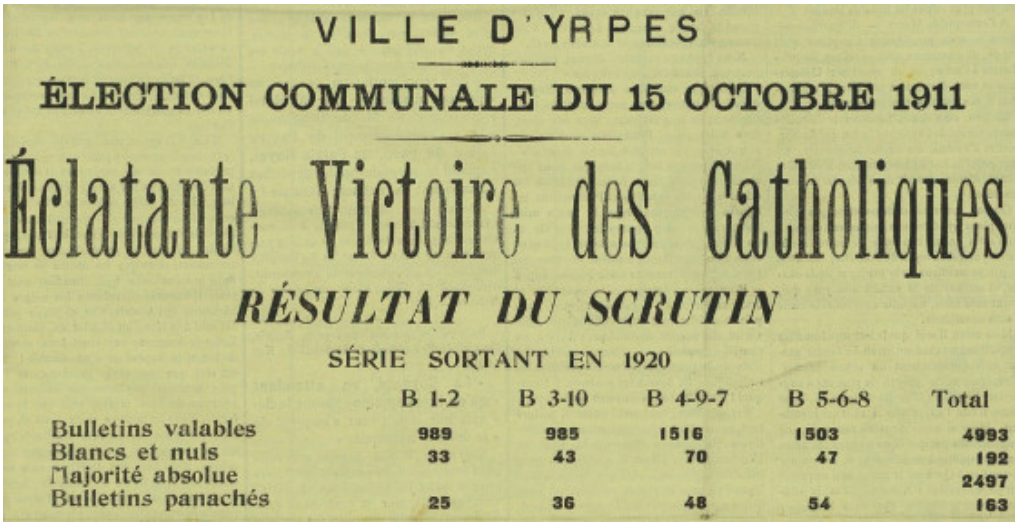
However, events are also important in other domains. In their T-REX system, Albanese and Subrahmanian (2007) use RDF[21] to extract cultural information from various domains. By extracting instances associated to user-specific RDF schemas, T-REX is able to find relevant information about Pakistani-Afghan tribes, but also occurrences of totally unrelated violent events, at the rate of nearly 50 000 web pages per day. Despite favouring this domain-independent approach, the authors insist on using language-specific extraction rules, as they argue that "the current state of the art of Machine Translation does not guarantee sufficiently high quality translations".

[21] RDF will be presented in detail in Chapter II.

While translating all news items to English before analysing them would clearly be questionable, the introduction of very specific rules also has its drawbacks, as it makes the system more complex and harder to maintain. An alternative would be to have a system that is both domain-independent and language-independent, bearing in mind that there always exists a trade-off between generalisation and extraction quality. Using such a multilingual approach, Busemann and Krieger (2004) develop a system based on the shallow information extraction platform SProUT (Becker et al., 2002) in order to assess travelling risks from press releases published on British, French, and German governmental websites. Research perspectives along these lines will be contemplated in Chapter IV, and further elaborated upon in our conclusions.

3.3 Template filling

Fortunately, the information contained in text is not completely random and often follows known patterns representing stereotypical situations in the real world. IE can take advantage of this recurring structure of information to anticipate what is to be found by providing templates containing a fixed number of slots that need to be filled. Slot filling, or template filling, is thus one of the simpler subtasks of IE, merely consisting in finding values in text for a given set of attributes. For instance, elections at various levels happen fairly often, especially in a country with such complex political structures as Belgium. An example from the *Journal d'Ypres* is provided in Figure I.7.



VILLE D'YPRES

ÉLECTION COMMUNALE DU 15 OCTOBRE 1911

Éclatante Victoire des Catholiques

RÉSULTAT DU SCRUTIN

SÉRIE SORTANT EN 1920

	B 1-2	B 3-10	B 4-9-7	B 5-6-8	Total
Bulletins valables	989	985	1516	1503	4993
Blancs et nuls	33	43	70	47	192
Majorité absolue					2497
Bulletins panachés	25	36	48	54	163

Figure I.7: Candidate event for template filling

Elections typically involve an administrative level, a geographical entity, a period of time, and a winner. Knowing this in advance enables looking for these particular elements in text, and to fill the template accordingly:

$$\text{ELECTION:} \begin{bmatrix} \text{LEVEL:} & \text{municipal} \\ \text{ENTITY:} & \text{Ypres} \\ \text{DATE:} & \text{1911-10-15} \\ \text{WINNER:} & \text{Catholic Party} \end{bmatrix}$$

In this way, template filling allows for better comparison of similar events. In most cases, however, it remains very difficult to predict in advance what will have to be extracted in an unknown document. Generalisable, open extraction methods will therefore achieve better results than narrow templates, except for very specific domains.

Summary

In this first chapter, we provided an overview of the field of information extraction (IE), mainly concentrating on the core task of named-entity recognition (NER) but also discussing relation detection, event extraction, and template filling. While the focus was more on theoretical and epistemological considerations than on practical matters, these components will later serve as building blocks for the operational implementation of our knowledge discovery tool.

Going back to the origins of artificial intelligence and natural language processing, we started from the Turing test in order to understand the deep motivations behind the drive to build thinking machines. Addressing the counterargument anticipated by Turing (1950) that computers could never "use words properly", decades of linguists and computer scientists strived to design programs able to extract the vast amount of information contained in unstructured documents in an automated way.

At the forefront of this large-scale initiative is NER, a task concerned with the identification and subsequent (or concurrent) categorisation of proper names present in a text. Natural language being intrinsically ambiguous, getting formal machines to learn what constitutes a valid entity is not trivial, and much work has been invested to reach this goal. While the classification into semantic categories (such as persons, organisations, and locations) allows to solve some cases of ambiguity, "un typage, aussi fin et précis soit-il, ne constitue pas l'établissement explicite d'un lien de référence entre une mention et une entité" (Stern, 2013). The only realistic path toward reconciling the signifier and the signified therefore seems to involve the full disambiguation of entities and concepts.

In the next chapter, we will introduce the Semantic Web and Linked Data as alternative approaches to information extraction, exploiting the vast amount of knowledge present on the Web in a variety of languages to annotate and disambiguate entities. This approach, however, will not come without its lot of difficulties, from the elusive nature of empirical content to the quality of external references and from language-independence to the evolution of concepts over time. While these issues will now temporarily be set aside, it is only to come back to them in chapters III and IV, where they will be discussed extensively in order to assess their operational impact on the claim to generalisability defended in our main thesis.

Chapter II

Semantic Enrichment with Linked Data

Outline

Chapter I introduced information extraction with its strengths and limits. Named-entity recognition, in particular, is a crucial task to identify and categorise proper nouns in unstructured text. Full disambiguation, however, calls for new computational techniques to be applied on digitised or born-digital corpora, and traditional information extraction systems are currently not well-equipped to provide them. The building-up of the Semantic Web, as dreamt by Tim Berners-Lee (2000, pp. 157–158), offers a framework to go beyond entity recognition and classification.

Section 1 exposes the reasons behind the structuring of the Web, from the original vision to new standards for data interoperability and the transition towards the Web of Data, laying the foundations for semantic enrichment. Linked Open Data are considered as a gateway toward better accessibility and visibility for cultural heritage players that will be introduced in Chapter III.

In Section 2, we focus on various resources that will be useful for our purposes, especially knowledge bases such as DBpedia but also ontologies behind them. The crucial role of URIs, which are used as unique identifiers for disambiguation, is discussed there, along with some known limitations related to their use that will be further developed in Chapter IV.

Finally, Section 3 introduces semantic enrichment of content in itself, which mainly relies on the task of entity linking. Since the ambiguity of what constitutes a valid (named) entity has already been highlighted in the previous chapter, we also extensively discuss how entities relate to terms, and terms to concepts. By critically assessing entity linking and its technical foundations, we go a step further toward the designing of an original methodology for knowledge discovery that will be implemented in Chapter V.

Contents

1 Making Sense of the Web

The democratisation of the World Wide Web in the 1990s revolutionised our relationship to documents and propelled mankind into the Information Age, fulfilling the original vision of Paul Otlet (1934, p. 428):[1]

> “ De là une [...] hypothèse, réaliste et concrète celle-là, qui pourrait, avec le temps, devenir fort réalisable. Ici la Table de Travail n'est plus chargée d'aucun livre. À leur place se dresse un écran et à portée un téléphone. Là-bas au loin, dans un édifice immense, sont tous les livres et tous les renseignements, avec tout l'espace que requiert leur enregistrement et leur manutention, avec tout l'appareil de ses catalogues, bibliographies et index, avec toute la redistribution des données sur fiches, feuilles et en dossiers [...] Le lieu d'emmagasinement et de classement devient aussi un lieu de distribution, à distance avec ou sans fil [...] De là on fait apparaître sur l'écran la page à lire pour connaître la réponse aux questions posées par téléphone [...] ”

The exponential explosion of information, however, has made it increasingly difficult for human agents to perform exhaustive search on the material available online. Tim Berners-Lee therefore pushed the utopia one step further and imagined a Web that would be meaningful not only to humans but also to computers, an evolution that Otlet (1934, p. 428) had also predicted (emphasis added):

> “ Une telle hypothèse, un Wells certes l'aimerait. Utopie aujourd'hui parce qu'elle n'existe encore nulle part, mais elle pourrait bien devenir la réalité de demain pourvu que se perfectionnent encore nos méthodes et notre instrumentation. Et ce perfectionnement pourrait aller peut-être jusqu'à rendre automatique l'appel des documents sur l'écran (simples numéros de classification, de livres, de pages); automatique aussi la projection consécutive, *pourvu que toutes les données aient été réduites en leurs éléments analytiques et disposées pour être mises en œuvre par les machines à sélection.* ”

The next three sections look into the emergence of the Semantic Web in the early 2000s and its later developments until today, focusing on the original vision of its co-founders, the underlying mechanisms for data structure and interoperability, and the progressive transition to the Web of Data.

[1] For a detailed analysis of the parallel between Otlet's vision and the workings of Hypertext in the context of the Web, see Rayward (1994).

1.1 The original vision

In their seminal article, Berners-Lee et al. (2001) imagined a new form of the Web that would make sense to machines and could be automatically exploited by them in order to meet new challenges posed by information overload :

> “ Most of the Web’s content today is designed for humans to read, not for computer programs to manipulate meaningfully. [...] The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.
>
> The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and “understand” the data that they merely display at present. ”

The vision of Tim Berners-Lee is often described as a shift in focus from raw data and information (i.e. unstructured documents loosely aggregated) to superior levels of understanding called knowledge or even wisdom. Figure II.1 shows the original DIKW (Data, Information, Knowledge, & Wisdom) Pyramid inspired by Ackoff (1989). Leal et al. (2012) sum up the project of the Semantic Web as follows:

> “ Currently, the Web is a set of unstructured documents designed to be read by people, not machines. The semantic web – sponsored by W3C – aims to enrich the existing Web with a layer of machine-interpretable metadata on Web resources so that computer programs can predictably exchange and infer new information. This metadata is usually represented by a general purpose language called Resource Description Framework (RDF). ”

American journalist Clay Shirky (2003), however, is pessimistic about the fulfilment of this ideal: “the Semantic Web, with its neat ontologies and its syllogistic logic, is a nice vision. However, like many visions that project future benefits but ignore present costs, it requires too much coordination and too much energy to effect in the real world, where deductive logic is less effective and shared worldview is harder to create than we often want to admit”.

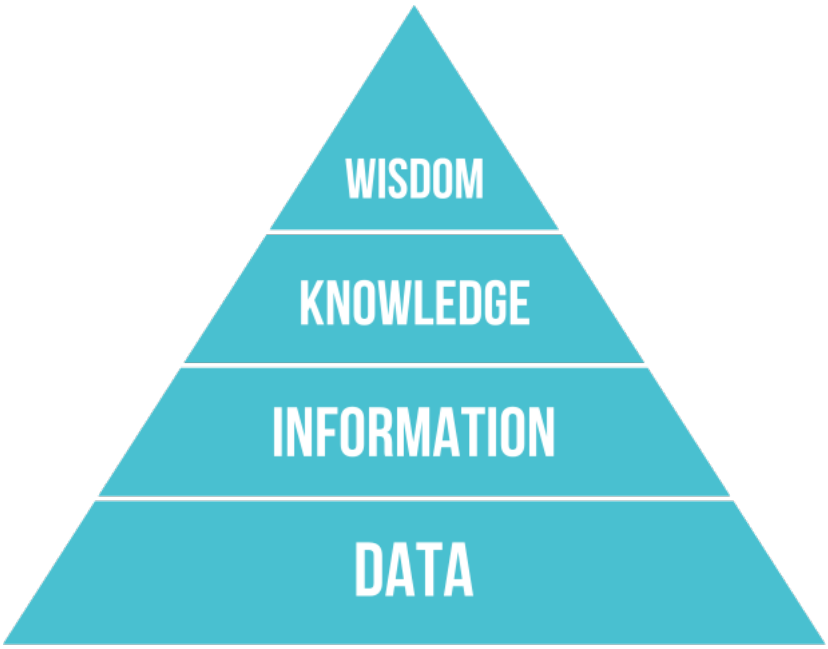


Figure II.1: DIKW Pyramid, reproduced from Chris Alvarez (CC BY-SA)

Do these criticisms entail that the Semantic Web is doomed to fail or useless? We do not think so and answer with Greenberg and Méndez (2007, p. 3):

> “ Criticism is useful for addressing current shortcomings and planning the next step in developing a Semantic Web. The downside of criticisms is that they often fail to note where important progress has been made.
>
> What is important and stands as evidence of major progress is the wide range of communities with a growing interest in information standards, data interoperability, and open information. Never in our time has there been a more universal interest in producing structured, standardized information. The idea of the Semantic Web initiative will, at the very least, help many more initiatives to benefit from standardized organization and access to information. ”

While realistically acknowledging that the Semantic Web is no panacea, we propose to build upon its technical foundations and to gain advantage of what semantic resources, such as knowledge bases and ontologies, have to offer in order to improve information extraction techniques.

1.2 Data structure and interoperability

In order to make the Web of documents exploitable by machines, data need to be organised in a structured form but also published according to standards making them interoperable. In this section, we will review the fundamental languages behind the Web of Data: XML and RDF, SPARQL, and SKOS.[2]

1.2.1 XML and RDF

Standards are an integral part of our case study: XML is used to encode the news articles in the corpus while RDF is the format of the facts we will use to enrich them. Berners-Lee et al. (2001) herald the importance of these recommendations of the World Wide Web Consortium (W3C) for the Semantic Web:

> “Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). [...] XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean. Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence.”

In place of the verb, the term *predicate* is used to encompass a broader set of relations: `S:Ypres P:postalCode O:8900`, for instance, is a valid triple despite the fact that “postal code” is not a verb. These *(s,p,o)* triples are the basis of knowledge bases such as DBpedia. They are also known as *facts*, and the action of identifying them in text as *fact spotting*, a task which can “help derive valuable training data for entity linking and relation extraction tasks” (Tylenda et al., 2014). In order to identify the different parts of a triple unequivocally, the Semantic Web makes use of unique identifiers called uniform resource identifiers (URIs), as explained by Berners-Lee et al. (2001):

> “In RDF, a document makes assertions that particular things (people, Web pages or whatever) have properties (such as “is a sister of”, “is the author of”) with certain values (another person, another Web page). This structure turns out to be a natural way to describe the vast majority of the data processed by machines. Subject and object are each identified by a Universal Resource Identifier (URI), just as used in a link on a Web page.”

[2] The Web Ontology Language (OWL) is also an important component of the Semantic Web but its discussion will be delayed until Section 2.2 which deals with ontologies.

What makes RDF more useful and more dynamic, however, is that not only entities are represented by URIs but also the properties and relationships linking them together. Berners-Lee et al. (2001) continue:

> “The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just by defining a URI for it somewhere on the Web. [...] The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web.”

Creating concepts out of nowhere “just by defining a URI” may sound idealistic, but the important part is that RDF does not rely on predefined properties.

The DIKW pyramid finds its technical counterpart in the Semantic Web stack (or layer cake), which describes the different technology layers needed to achieve our goal (Figure II.2). The building blocks are URIs (see Section 2.3) and Unicode (crucial to deal with diacritics), while higher abstraction levels include XML and RDF, but also SPARQL (introduced in Section 1.2.2) and OWL (Section 2.2.1). The other layers will not be discussed here for lack of space.

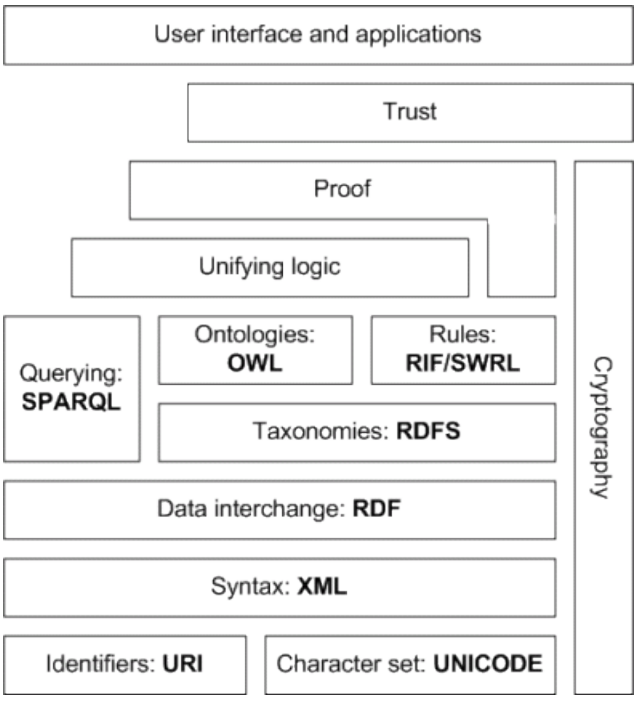


Figure II.2: The Semantic Web layer cake

XML relies on the traditional alternation between an intensional model and an extensional document. The intension is realised by an XML schema (`.xsd` file), while the actual document appears in a `.xml` file. This model is pervasive on the Semantic Web, being also used to articulate ontologies and knowledge bases (see Section 2.2). In contrast, the simplicity of the RDF model nearly amounts to schemalessness, or rather schema neutrality: "By simplifying to a maximum the data model, all of the semantics are made explicit by the triple itself. By doing so, there is no longer a need for a schema to interpret the data." (van Hooland and Verborgh, 2014, p. 44).[3]

As the authors note, however, "schema-neutral does not mean that no schema-related issues remain". We still depend on schemas, but the methods and tools to use them are becoming increasingly open and standardised. Moreover, although XML is widely used in the cultural heritage domain in order to embed metadata in text, it has also been criticised for a number of inadequacies. Schmidt (2010) lists the following potential problems with markup:

- the impossibility of overlapping tags
- the inclusion of potentially obsolescent technical and subjective information into texts that are supposed to be archivable for the long term
- the manual encoding of information that could be better computed automatically
- the obscuring of the text by highly complex technical data

Since we will be using millions of XML files in our case study presented in Chapter III, we need to be aware of these limitations, which can potentially arise in the context of any digitisation project.

Spaeth (2004) warns against the temptation of using XML for its own sake in cases where keeping data in their original format would have been sufficient, although he argues it can still be useful as a data exchange format. Respecting Text Encoding Initiative (TEI) standards can also make projects so complex that they defy understanding, as illustrated in Figure II.3: twenty distinct steps are required in order to publish diplomatic documents online.

[3]The idea of schemalessness is pervasive since the rise of popularity of alternatives to relational databases. While evaluating the added value of graph databases for the humanities, Blanke and Kristel (2013), for instance, write that "NoSQL databases [...] often give up on the schema-centricism of traditional databases". See blog entry `http://blog.jooq.org/2014/10/20/stop-claiming-that-youre-using-a-schemaless-database/` for a critique of the term *schemaless* as used by MongoDB NoSQL DBMS. The capabilities of graph databases for implementation will be discussed in Chapter V and again in our conclusions.

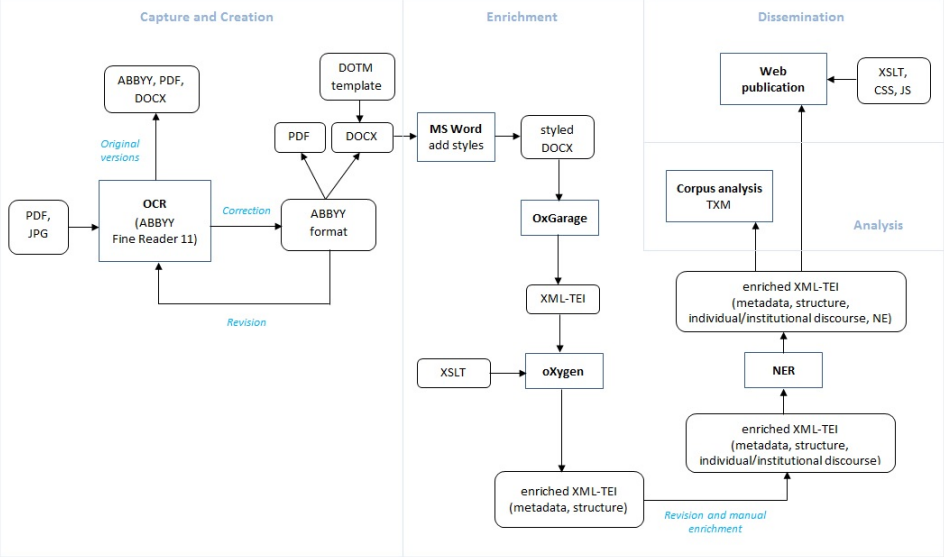


Figure II.3: Overly complex TEI workflow, reproduced from the research blog `http://cvcedhlab.hypotheses.org/98`

1.2.2 SPARQL

The SPARQL Protocol And RDF Query Language[4] (recursive acronym) is a W3C recommendation allowing to query Linked Open Data (LOD) resources published in RDF (such as those contained in knowledge bases) in a simple, straightforward way. It requires a SPARQL endpoint (RDF triple store) to be set up. SPARQL comes in various implementations, e.g. OpenLink Virtuoso.[5] For instance, querying DBpedia's endpoint for finding all persons sharing the same occupation as Jacques Cartier and having the term "Canada" mentioned in their biography (i.e. all Canadian navigators or explorers of Canada) can be achieved as shown below:

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?person WHERE {
  dbr:Jacques_Cartier dbo:occupation ?job .
  ?person dbo:occupation ?job .
  ?person dbo:abstract ?biography .
  FILTER regex(?biography, "Canad") }
```

[4] `http://www.w3.org/TR/sparql11-overview/`
[5] `http://virtuoso.openlinksw.com/`

The first two lines of code declare that the prefixes `dbr` and `dbo` will be used for `<http://dbpedia.org/resource/>` and `<http://dbpedia.org/ontology/>` respectively in order not to have to repeat the whole URIs every time.[6] The remainder of the query selects all distinct persons with the following conditions:

- Jacques Cartier has an occupation we will name `?job`
- the person we seek must also have this occupation
- furthermore, we look into the person's summary and call it `?biography`
- we then filter this biography with a simple regular expression to see if it contains the string "Canad" (covering the terms *Canada* and *Canadian*)

Note that the names of the variables (`?person`, `?job`, `?biography`) have no special meaning for SPARQL and are left to the personal choice of the user. Selected results for this query are shown in Table II.1.

person
http://dbpedia.org/resource/Samuel_de_Champlain
http://dbpedia.org/resource/David_Thompson_(explorer)
http://dbpedia.org/resource/Joseph_William_McKay
http://dbpedia.org/resource/John_Davis_(English_explorer)
http://dbpedia.org/resource/John_Rae_(explorer)
http://dbpedia.org/resource/Leif_Erikson
http://dbpedia.org/resource/Antoine_de_la_Mothe_Cadillac
http://dbpedia.org/resource/Mathieu_de_Costa
http://dbpedia.org/resource/Peter_Fidler_(explorer)
http://dbpedia.org/resource/Thomas_Button

Table II.1: Results of a SPARQL query for Canada-related explorers

SPARQL queries are essential to access knowledge bases programmatically and restrict results to types of interest, as will be shown again in Section 2.2.1. This mechanism is also an integral part of MERCKX, the knowledge extractor that we will be introducing in Chapter V.

[6] Actually these are very common, predefined prefixes so it is not even necessary to declare them explicitly. See `http://dbpedia.org/sparql?nsdecl` for a full list of DBpedia predefined namespace prefixes.

1.2.3 SKOS

The Simple Knowledge Organization System[7] is a W3C recommendation for the representation of controlled vocabularies such as thesauri, taxonomies and subject headings (e.g. the Library of Congress Subject Headings) within the framework of the Semantic Web (Miles and Bechhofer, 2009). Conceived as a help for vocabulary maintainers to publish their taxonomies as Linked Data but also as a tool to manage them, SKOS aims to foster interoperability.

Meunier (2014) offers an insightful critique of SKOS, underlining some of its contradictions raised by the tension between these two goals: publication on the Web and closed-world representation. Despite these shortcomings, SKOS remains an opportunity for libraries, archives, and museums wishing to maintain their collections in a sustainable manner. In particular, the `skos:Concept` property allows to query linked datasets for relevant concepts.

1.3 From the Semantic Web to Linked Data

Although some of the predictions of Tim Berners-Lee have come true thanks to the efforts of the W3C, the Semantic Web envisioned in 2000 remains largely underdeveloped fifteen years on. The main cause for this partial failure can be found, ironically, in the very foundational paper of the Semantic Web: the Web has notoriously always been resistant to any form of regulation, so while browsers can more or less be wooed into adopting standards, coercing users into adopting semantic markup is doomed to fail.

While individual web pages remain largely un-semantic, more and more collections of structured data have been published online and interconnected, a phenomenon now known as Linked Data. Linked Data can be seen as the next best thing to the Semantic Web: it does not yet fulfil the old dream in AI of intelligent agents[8] managing our schedules and making appointments on our behalf, but it nonetheless provides a *Web of Data* that can be browsed by machines in order to discover new knowledge. In 2006, Tim Berners-Lee acknowledged the non-advent of the Semantic Web and restated its goals in a less ambitious manner, formulating the four principles[9] of Linked Data:

[7] `http://www.w3.org/2004/02/skos/`

[8] In AI, an intelligent (or rational) agent is an agent able to "select an action that is expected to maximize its performance measure, given the evidence provided by [its sensors] and whatever built-in knowledge the agent has" (Russell and Norvig, 2009, p. 37). According to Heylighen (2014), "the task of the intelligent agent is [...] to transform or process the input information (problem, initial state, 'question') via a number of intermediate stages into the output information (solution, goal state, 'answer')".

[9] `http://www.w3.org/DesignIssues/LinkedData.html`

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.

Bizer et al. (2009a) explore the concept and technical principles of Linked Data and establish a list of related initiatives. The authors use *Semantic Web* and *Web of Data*[10] more or less interchangeably, but concede that the latter in its current state is more modest than the original idea of the Semantic Web:

> “By publishing Linked Data, numerous individuals and groups have contributed to the building of a Web of Data, which can lower the barrier to reuse, integration and application of data from multiple, distributed and heterogeneous sources. Over time, with Linked Data as a foundation, some of the more sophisticated proposals associated with the Semantic Web vision, such as intelligent agents, may become a reality.”

As of August 2014, just over 1000 datasets were officially listed in the “State of the LOD cloud” report.[11] Table II.2 shows the distribution of these datasets across domains. Note that the quality of these datasets is uneven, an issue that will be examined with a special attention in Chapter IV.

Topic	**Datasets**	**%**
Social web	520	51.28%
Government	183	18.05%
Publications	96	9.47%
Life sciences	83	8.19%
User-generated content	48	4.73%
Cross-domain	41	4.04%
Media	22	2.17%
Geographic	21	2.07%
Total	**1 014**	**100.00%**

Table II.2: Linked datasets by topical domain

[10] `http://www.w3.org/2013/data/`
[11] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`

As a result of the multiplication of these datasets (from 295 in 2011 to 1014 in 2014), the representation of the LOD cloud itself has grown increasingly difficult to read as a whole. Indeed, Valsecchi et al. (2015) deplore the fact that "researchers that are not experts in Semantic Web technologies often lose themselves in the intricacies of the Web of Data". Figure II.4 shows a small subset of it, centred on DBpedia.[12]

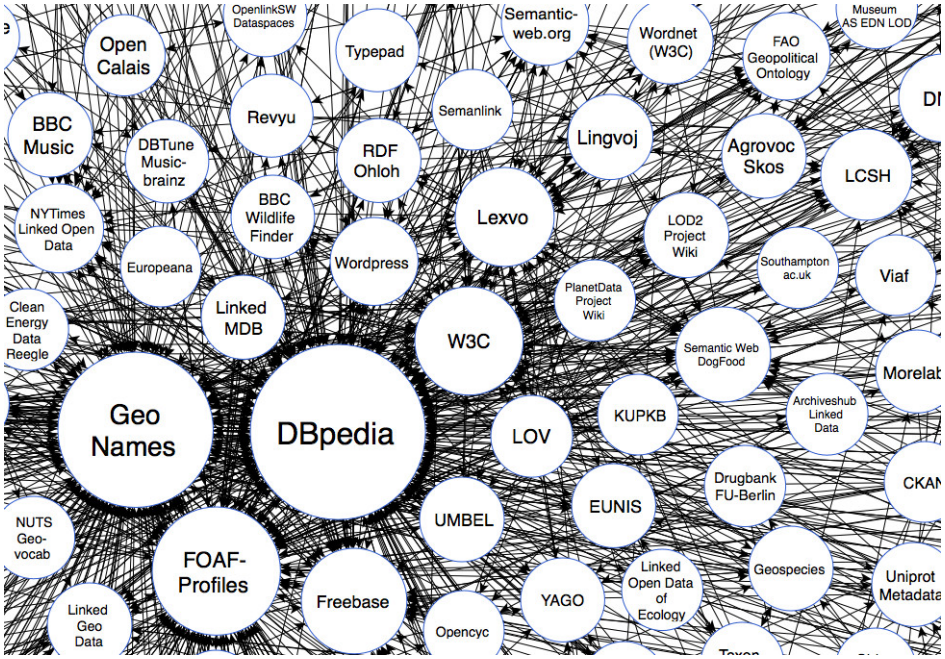


Figure II.4: Snapshot of the Linked Open Data cloud

What strikes is the tight connectedness of the graph around the central resources, along with the prominence of GeoNames[13] which grew from a marginal resource in 2011 to a major one in 2014. Other useful datasets for IE and NER include OpenCalais (top left), Freebase and YAGO (bottom) and the Virtual International Authority File (VIAF, right). Notice that these other datasets are more loosely connected than DBpedia, with the exception of Freebase. To understand how these datasets work and the stakes attached to their use in the context of semantic enrichment, the next section will focus in depth on semantic resources such as knowledge bases, ontologies, and URIs, and show how they can be leveraged to achieve entity linking.

[12] Adapted from the Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. `http://lod-cloud.net/` (CC BY-SA)

[13] `http://www.geonames.org/`

2 Semantic Resources

In order to make the Semantic Web functional, information needs to be more tightly structured than on the Web of documents. This requirement is not new: knowledge representation has been striving for a long time to represent data in a processable way. The novelty of the Semantic Web, however, is to achieve this without a single central authority, but rather by linking several authorities or knowledge bases together (Berners-Lee et al., 2001):

> “ Information varies along many axes. One of these is the difference between information produced primarily for human consumption and that produced mainly for machines. [...] To date, the Web has developed most rapidly as a medium of documents for people rather than for data and information that can be processed automatically. The Semantic Web aims to make up for this. [...]
>
> For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. [...]
>
> Knowledge representation [...] contains the seeds of important applications, but to realize its full potential it must be linked into a single global system. ”

Resources available in order to achieve this goal include knowledge bases (“structured collections of information”) and ontologies (“sets of inference rules” + taxonomies) that will be presented in sections 2.1 and 2.2 respectively. Both rely on the mechanism of URIs, which will be discussed in Section 2.3.

2.1 Knowledge bases

Knowledge bases (KBs) are central to the Linked Data approach: they contain the information, or rather the *knowledge* (in the practical sense defined in our introduction) necessary to semantically enrich content. KBs differ from traditional relational databases in part by their structure which is often graph-based,[14] but mainly by their function: their aim is not to store data but rather to derive new information from established facts.

In what follows, we present a few of the most popular KBs, starting with DBpedia on which we will be relying for knowledge extraction in Chapter V.

[14] The idea behind graph databases is to provide a more flexible model than relational databases by allowing each record to have its own schema, using Euler’s graph model with nodes (resources) and edges (relationships).

2.1.1 DBpedia

DBpedia[15] is "a community effort to extract structured information from Wikipedia and make this information available on the Web" (Lehmann et al., 2015). Presented as a crystallisation point for the Web of Data (Bizer et al., 2009b) and often considered the *de facto* centre of the Linked Open Data Cloud, DBpedia is maintained in 128 languages and covers over 38 million topics (3 billion facts). Figure II.5 shows an example of a DBpedia resource.

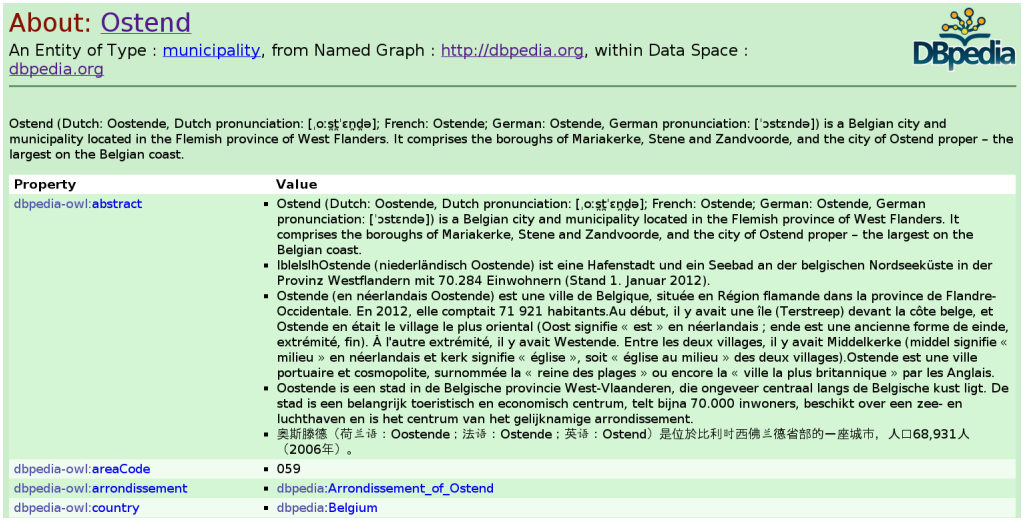


Figure II.5: Preview of the "Ostend" resource in DBpedia

DBpedia is designed to be browsed by either humans or machines. The documentation about the resource representing the city of Brussels, for instance, can be read in HTML format at `http://dbpedia.org/page/Brussels` (where the corresponding resource redirects, see Section 2.3) but also harvested programmatically in XML/RDF or JSON[16] among other formats. As such, DBpedia is an effective implementation of the Write Once, Publish Many strategy. Additionally, it provides a SPARQL endpoint[17] to perform more complex queries on the knowledge base as a whole.

However, every resource does not have an equivalent in every language, which raises some questions about the multilingual structure of DBpedia. We will come back to this important issue in Chapter IV, where we will also tackle the problem of the evolution of concepts.

[15] `http://dbpedia.org/`

[16] `http://dbpedia.org/data/Brussels.rdf` / `http://dbpedia.org/data/Brussels.json`

[17] `http://dbpedia.org/sparql`

2.1.2 YAGO

YAGO[18] (for Yet Another Great Ontology) is a semantic knowledge base built upon the aggregation of Wikipedia, GeoNames, and the English lexical database WordNet.[19] As of April 2015, it contains over ten million topics and more than 120 million facts (Mahdisoltani et al., 2015). The latest version (YAGO3) includes a special focus on multilingual data, harvesting information from Wikidata (see below) and from Wikipedia categories and infoboxes of ten languages[20] with a reported precision over 95%.

Maintained at the Max Planck Institute for Informatics, YAGO is less ubiquitous than DBpedia in the LOD cloud but offers the interesting characteristic of discarding the distinction between concepts and terms, demonstrating that the approach hinted at in our first research question is workable. The multilingual version of YAGO is indeed merged with the English WordNet (Miller, 1995; Fellbaum, 1998) to create a coherent and comprehensive KB containing both proper and common nouns.

2.1.3 Freebase

Created by Metaweb and later acquired by Google, Freebase[21] presented itself as a "community-curated database of well-known people, places, and things". With over 47 million topics and nearly 3 billion facts, Freebase was one of the biggest KB available and a serious challenger to DBpedia, raising the often overlooked question of the economy behind knowledge representation (i.e. public versus business-owned).

In January 2015, however, Google announced the "retirement" of Freebase as of June 30, 2015[22] and the transfer of its content to Wikidata (see below) in an effort to avoid the multiplication of LOD sources. The move is not atypical from Google – which has been known to speed-withdraw other products in the past[23] – but questions the permanence of URIs which is key to Linked Data in general and entity disambiguation in particular.

[18] `http://www.yago-knowledge.org/`

[19] `http://wordnet.princeton.edu/`

[20] In their experiment, Mahdisoltani et al. (2015) ran the YAGO extraction system on English, German, French, Dutch, Italian, Spanish, Romanian, Polish, Arabic, and Farsi. The deliberate inclusion of non-European, non-Latin script languages is particularly interesting in a language-independent context.

[21] `https://www.freebase.com/`

[22] `https://goo.gl/rxyCvn`

[23] Compare the radical U-turn on Google Wave which saw the massive creation of 100 000 accounts in 2009 followed by the sudden discontinuation of the project as early as 2010.

2.1.4 Wikidata

Wikidata[24] is an initiative by the Wikimedia Foundation to gain value from the content of Wikipedia into a structured form by providing a knowledge base that can be read and edited by both humans and machines (Vrandečić, 2012; Vrandečić and Krötzsch, 2014). As of July 2015, it contains over 14 million topics (110 million facts). In contrast to its competitor DBpedia which uses Wikipedia article titles for its URIs, Wikidata relies on numeric unique identifiers for each concept documented in the knowledge base. While this is more language-neutral and ensures a better robustness in case of article renaming, it is done at the cost of a loss of transparency for the user: `https://www.wikidata.org/wiki/Q76` for instance is clearly less meaningful than `dbr:Barack_Obama`, although both refer to the same real-world entity and can be linked to each other with the `owl:sameAs` property (see Section 2.2).

Table II.3 shows the corresponding items for the IDs ranging from Q1 to Q80, starting with *universe* all the way to *Tim Berners-Lee*. Notice that some place-holders were left blank (Q6, Q7, Q9–12, etc.), perhaps for future usage. From the point of view of ID integrity, this practice gives an illusion of objectivity but is in fact as bad as – or even worse than – semantically rich DBpedia URIs: sequentially ordered, humanly-controlled numbers do not offer the stability of randomised computer-generated IDs.[25]

2.1.5 ConceptNet

ConceptNet[26] presents itself as a "semantic network containing lots of things computers should know about the world", but is essentially a KB (Havasi et al., 2007; Speer and Havasi, 2012). It contains common sense information such as:

Example 19. saxophone – UsedFor → jazz

ConceptNet covers 3.6 million topics (15 million facts) in over 1000 languages. As noted on the project's homepage, "it would not adequately represent human knowledge if it didn't contain other languages besides English, as well":

Example 20. книга – MadeOf → бумага[27]

ConceptNet also ignores the distinction between terms and entities, which confirms the hypothesis formulated in our first research question that there is no necessary division between the two.

[24] `https://www.wikidata.org/`
[25] See Section 2.3 for a comparison of URIs with other types of unique IDs.
[26] `http://conceptnet5.media.mit.edu/`
[27] A book is made of paper.

ID	Item	ID	Item
Q1	universe	Q41	Greece
Q2	Earth	Q42	Douglas Adams
Q3	life	Q43	Turkey
Q4	death	Q44	beer
Q5	human	Q45	Portugal
Q6		Q46	Europe
Q7		Q47	
Q8	happiness	Q48	Asia
Q9		Q49	North America
Q10		Q50	
Q11		Q51	Antarctica
Q12		Q52	Wikipedia
Q13	triskaidekaphobia	Q53	Club-Mate
Q14		Q54	all your base are belong to us
Q15	Africa	Q55	Netherlands
Q16	Canada	Q56	lolcat
Q17	Japan	Q57	Never Gonna Give You Up
Q18	South America	Q58	penis
Q19	cheating	Q59	PHP
Q20	Norway	Q60	New York City
Q21	England	Q61	Washington, D.C.
Q22	Scotland	Q62	San Francisco
Q23	George Washington	Q63	
Q24	Jack Bauer	Q64	Berlin
Q25	Wales	Q65	Los Angeles
Q26	Northern Ireland	Q66	Boeing
Q27	Ireland	Q67	Airbus
Q28	Hungary	Q68	computer
Q29	Spain	Q69	Courrendlin
Q30	United States of America	Q70	Bern
Q31	Belgium	Q71	Geneva
Q32	Luxembourg	Q72	Zürich
Q33	Finland	Q73	IRC
Q34	Sweden	Q74	Breighton
Q35	Denmark	Q75	Internet
Q36	Poland	Q76	Barack Obama
Q37	Lithuania	Q77	Uruguay
Q38	Italy	Q78	Basel
Q39	Switzerland	Q79	Egypt
Q40	Austria	Q80	Tim Berners-Lee

Table II.3: First 80 items of Wikidata

Table II.4 recapitulates the numbers of languages, topics/concepts, and facts/assertions covered by the five knowledge bases presented above.

KB	# lang.	# topics	# facts
DBpedia	128	38.3M	3040M
YAGO	10	10M	120M
Freebase	18	48.3M	2980M
Wikidata	279	14.6M	110M
ConceptNet	1000+	3.6M	15.2M

Table II.4: Comparison of popular knowledge bases

KBs are essentially build by hand by scores of contributors (crowdsourcing), but automatic computer programs (bots) also play a significant role in the editing and maintaining of semantic resources (Steiner, 2014), although they still struggle in terms of coverage (Russell and Norvig, 2009, p. 1047):

> “ There is great promise in using the Web as a source of natural language text [...] to serve as a comprehensive knowledge base, but so far machine learning algorithms are limited in the amount of organized knowledge they can extract from these sources. ”

The evolution of knowledge over time is indeed not trivial and calls for constant monitoring, as will be discussed in depth in Chapter IV and in the research perspectives mentioned in our conclusions.

2.2 Ontologies

In order to establish a one-to-one correspondence between the various knowledge bases presented in the previous section, the Semantic Web uses *ontologies*[28] which formally define the relationships between concepts. An ontology is "an explicit specification of a conceptualization" (Gruber, 1995); that is, a way to translate a conceptual model into a meaningful set of terms that can be shared by a community of users.[29]

Ontologies and knowledge bases are closely related, the former formalising "the intensional aspects of a domain, whereas the extensional part is provided by a knowledge base that contains assertions about instances of concepts and relations as defined by the ontology" (Buitelaar et al., 2005). In other words, ontologies provide the building bricks to establish meaningful links between the facts stored in knowledge bases.[30]

Ontologies consist of a taxonomy and a set of inference rules, and provide equivalence relations with one another, allowing interoperability. They also offer a potential solution to the problem of language ambiguity, since each distinct concept is given a different URI in a KB, and consistency between URIs about the same concept in various KB is assured by the equivalence relation `owl:sameAs`.[31]

Buitelaar et al. (2005) give an overview of several methodologies that can be used in order to derive ontologies from unstructured text. The authors build upon Gruber's definition provided above, adding three restrictions:

1. the ontology should be *formal*, i.e. machine-readable
2. the conceptualisation should be *shared*, i.e. accepted by a group or community
3. it should be restricted to a given *domain of interest*, i.e. ontologies are useful only for a particular application domain

[28] Although distantly related to the philosophical study of the nature of existence, ontologies in their modern sense only appeared in the second half of the 20th century, while the use of the term itself is generally credited to Gruber (1995).

[29] Note how this definition also goes against Ehrmann's argument, discussed in Section 2.1 of Chapter I, that the model is what distinguishes entities from terms.

[30] Although the OWL guide (`http://www.w3.org/TR/owl-guide/`) affirms that an "ontology may include `[...]` instances", we claim with Buitelaar that this is a corruption of the original meaning of an ontology, which should by nature only be concerned by the conceptual level.

[31] See `http://www.w3.org/TR/owl-ref/#sameAs-def` for a full definition of this property. According to Moura and Davis (2014), however, few LOD sources correctly use this property. Similarly, Halpin et al. (2010) note that the boundary between *sameness* and *similarity* is sometimes porous, to the extent that the misuses of `owl:sameAs` outnumber its "correct" uses on the Web of Data. See Chapter IV for a discussion of the consequences of the blind use of this property.

While the first two are easily agreed upon, we contest the third restriction which is in opposition with the main thesis defended within this dissertation, i.e. the usefulness of general-domain resources for information extraction. The DBpedia ontology,[32] for instance, is clearly convenient for any domain covered, although some properties are necessarily more specific than others. We therefore maintain that general ontologies are workable, even if they are arguably not well-adapted for very specialised domains.

In what follows, we present a brief overview of the Web Ontology Language used widely on the Semantic Web, with concrete examples of how it can be implemented for our needs. We then sum up some of the main critiques that have been formulated against ontologies.

2.2.1 Web Ontology Language

The Web Ontology Language (OWL)[33] is a W3C recommendation ensuring the coherence between intension and extension. For instance, `dbr:Place` describes what a place is, whereas `dbo:Place` puts the class of places in a hierarchy, as a subclass of things and a superclass of populated places for instance. Using the prefixes introduced in Section 1.2.2, the fact that Ypres is a place can be represented by:

Example 21. `dbr:Ypres rdf:type dbo:Place`

Extracting all places contained within DBpedia can then easily be achieved with a basic SPARQL query,[34] as explained is Section 1.2.2:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?place WHERE {
 ?place a dbo:Place
}
```

This simple mechanism allows us to build a comprehensive gazetteer of locations that will be used in Chapter V to extract place mentions from our corpus. Of course, this methodology can easily be adapted to different needs by simply selecting another relevant property from the ontology, making it quickly portable to other application domains.

[32] `http://dbpedia.org/ontology/`

[33] `http://www.w3.org/TR/owl-ref/`

[34] With a limit of 10 000 results at a time, which makes it necessary to run the query several times with different offsets (a task that can be automated with a loop).

2.2.2 Limits of ontologies

When they exist for a given domain, ontologies are a powerful way to classify documents. In many cases, however, "such [a] classification does not exist and the cost of creating and maintaining an ontology would be unbearable" (Leal et al., 2012).

Moreover, Enache and Angelov (2010) warn that "developing large scale ontologies is always an error-prone process", especially due to the lack of integrity constraints: "most [ontology description] languages are based on some kind of untyped logic which allows to assert axioms which are not well-formed. In contrast, even the simplest database systems are equipped with some database schemas which rule out incorrect records". We will see in Chapter IV that database integrity constraints are not sufficient to prevent incorrect values from arising, but their complete absence indeed makes ontologies all the more vulnerable to formal errors of various kinds.

Shirky (2003, italics his) is similarly disparaging about the possibility of a general, worldwide ontology:

> "Any attempt at a global ontology is doomed to fail, because metadata describes a worldview. The designers of the Soviet library's cataloguing system were making an assertion about the world when they made the first category of books "Works of the classical authors of Marxism-Leninism." Melvyl Dewey was making an assertion about the world when he lumped all books about non-Christian religions into a single category, listed last among books about religion. It is not possible to neatly map these two systems onto one another, or onto other classification schemes – they describe different *kinds* of worlds."

On the other hand, ontologies that are too specialised are also self-defeating and cannot easily be reused for purposes not foreseen by their creators. As noted by Maturana et al. (2013), "ontologies are deeply focused on solving problems and satisfying interests of [...] professional groups" rather than user-oriented. The question of maintenance also tips the scale in favour of lightweight ontologies as opposed to very specific ones: robust ontologies with general properties about the world are less likely to break over time than those conceived for specialised technical domains which are subject to significant concept drift, a topic that will be covered in Chapter IV (Section 3.3).

2.3 Identifiers

As the name suggests, identifiers are used to establish the identity of either a unique object (instance) or a unique class of objects. The concept of a "unique identifier" may thus seem redundant at first sight, but it is justified by this double role: model number `VPCZ13M9E` is an identifier of a class of computers, while serial number `27536461 5000245` is a *unique* identifier of a single machine. This distinction sends us back to the opposition between named entities and plain entities discussed in Chapter I, to which we will come back in Section 3.1.

The classic literature about unique IDs warns about the pitfalls of using numbers that have any semantic attached to them. Instead, most authors recommend the use of "unintelligent numbers", i.e. "purely random number[s] which can only be interpreted by reference to a central database; examining the number itself tells you nothing about the object which it identifies" (Green and Bide, 1996). The difficulty to maintain "intelligent" (semantic) IDs in the long run has been put into light repeatedly (Paskin, 1999).

In Belgium, for instance, social security numbers contain a 3-digit sequence which is even for women and odd for men. The distinction seemed stable enough when the practice was extended to the whole population in the 1980s. However, transgender persons now challenge this dichotomy,[35] which raises problems that were unforeseen at the time of the creation of the system but would have been prevented by the adoption of semantically neutral IDs.

The depletion of IPv4 addresses, and the costly transition to IPv6, also emphasise the need to adopt a very cautious stance when designing new identifiers: the pool of 4 billion IPv4 addresses created in the 1980s was in fact exhausted in thirty years. When designing identifiers destined to be used on a large scale, the rule of thumb remains extreme prudence.

2.3.1 Uniform resource identifiers

For Web resources, a persistent issue is the difficulty to draw a line between a URI representing something on the one hand, and the actual documentation about this thing on the other hand. For instance, should we consider that the URI `http://www.wikidata.org/wiki/Q80` is disambiguating Tim Berners-Lee the person, or the webpage about TBL? Since a URI must be unique by definition, this seemingly double reference (known as the httpRange-14 issue) is problematic.

[35] The German constitutional court officially recognised a third gender in 2013 (source: `https://www.smalsresearch.be/data-simplification-and-abstraction-part-i/`).

Bunescu and Pasca (2006) offer an exemplification of this commonplace confusion when they say that "because each article [of Wikipedia] describes a specific entity or concept,[36] the remainder of the paper sometimes uses the term 'entity' interchangeably to refer to both the article and the corresponding entity." While that may appear innocuous, this state of affairs challenges the uniqueness of URIs since they are used to represent two resources at the same time.

DBpedia settled (or rather circumvented) the issue by duplicating each URI: while `http://dbpedia.org/resource/Tim_Berners-Lee` identifies the human being, `http://dbpedia.org/page/Tim_Berners-Lee` identifies the documentation about him, the former automatically redirecting to the latter. This naive solution gives the illusion of solving a century-old question with a technical fix, while leaving the user none the wiser about which properties refer to the subject, to the document, or to both.

These epistemological reflections on the nature of the difference between an object and its representation send us back to the classic distinction between the signifier and the signified (French *signifiant et signifié*) formalised by de Saussure (1916, p. 99):

> « Nous proposons de conserver le mot signe pour désigner le total, et de remplacer concept et image acoustique respectivement par signifié et signifiant [...]
>
> Le lien unifiant le signifiant et le signifié est arbitraire, ou encore, puisque nous entendons par signe le total résultant de l'association d'un signifiant à un signifié, nous pouvons dire plus simplement : le signe linguistique est arbitraire. »

Wismann (2012, pp. 232–233) goes even further by attributing the origin of the argument to Heraclitus: "dans le langage, il y a le signifié et le signifiant, la chose dont on parle et cette chose qui parle de la chose dont on parle. Et la thèse d'Héraclite, c'est qu'il y a une différence insurmontable entre ce que le langage dit et le dire même du langage". If the actual thing and the way to talk about it cannot be reconciled, communication becomes impossible. But the link between the two remains arbitrary, and a simple Web redirection from one URI to another will not solve this old philosophical problem. To complicate matters even further, some identifiers are also used for other uses, as detailed in the next section.

[36] Note that entities and concepts are not formally distinguished but coexist at the same level. More about that in Section 3.1.

2.3.2 Identifiers and locators

Uniform resource locators (URLs) should not be confused with URIs: URLs are a subtype of URIs fulfilling a different role. Shirky (2003), notes that "the fact that a URL itself doesn't have to mean anything is essential – the Web succeeded in part because it does not try to make any assertions about the meaning of the documents it contained, only about their location".

In the real world, locators are seldom used as unique identifiers. A library call number like `2SIC 025.04 BOYD` will allow to locate a book within a library, but not to identify it universally since the same number could refer to another book in another library. Conversely, an ISBN number like `978-2-80271268-8` will completely disambiguate a book, but will offer no clue as to where it can be found in practice. The same analogy is true for human beings: a postal address allows to locate someone, but not to identify them uniquely since several people can live at the same house, whereas a social security number is a unique ID but does not include information relative to the person's whereabouts.

On the Web, however, the separation between the ID and the locator is sometimes blurred by the double function of the URI/URL. DBpedia, for instance, uses URIs like `http://dbpedia.org/resource/Holy_Grail`, which also happens to be a URL.[37] Although this is in accordance with the second Linked Data Principle ("Use HTTP URIs so that people can look up [the] names [of things]"), the fact remains that this state of affairs can be confusing to the uninformed user.

Despite these shortcomings that will be duly addressed in Chapter IV, the added value of URIs as unique identifiers to disambiguate entities and concepts is not to be denied. In Chapter V, we will see that most semantic enrichment tools rely on such resources for their linking components, as we will also do with our own knowledge extractor.

[37] A locator for the documentation about the resource, not for the resource itself, alas!

3 Enriching Content

Looking for information on the Web is not as straightforward as sometimes assumed. According to Leal et al. (2012), "searching effectively on a comprehensive information source [...] usually boils down to using the right search terms". But does it? Finding the answer to a simple question such as "How old was Dutch astronomer Christiaan Huygens when he died?", for instance, involves at least four non intuitive steps:

1. converting the question formulated in natural language to a suitable request made of keywords, or *searchese* ("`christiaan huygens date birth death`" for instance)

2. browsing the hundreds of thousands of results for a relevant page

3. locating the dates of birth and death in the text

4. mentally subtracting the numbers to get the age of the person

This makes the case for the integration of more linguistic knowledge into search engines, in order to improve the performance of information retrieval systems (Bouillon et al., 2000; Moreau et al., 2007), something that has been made easier by the advent of Linked Data and the availability of general purpose linguistic resources such as WordNet (see Section 2.1.2).

Today, Google's Knowledge Graph[38] is capable, to a limited extent, to mine structured knowledge bases such as DBpedia and Freebase in order to serve the answer to the user directly. As a result, a simple search for "`age of Huygens?`" or even "`age huygens`" will yield the infobox displayed in Figure II.6. The fact that Huygens died at 66 is represented by an RDF triple of the form (Christiaan Huygens, Age at death, 66). Google correctly identifies the string "huygens" in the user's query as the person Christiaan Huygens (subject) and the string "age" as the predicate (or property) Age at death, which prompts it to return the object (or value) 66 along with the dates used to infer this number. Along the way, additional information is provided about related persons whom the user might be interested in.

While Google is able to afford this kind of implementation of question answering on such a massive scale, this is clearly not the case for smaller companies, even less for cultural institutions and other players from the humanities.

[38] `http://www.google.be/insidesearch/features/search/knowledge.html`

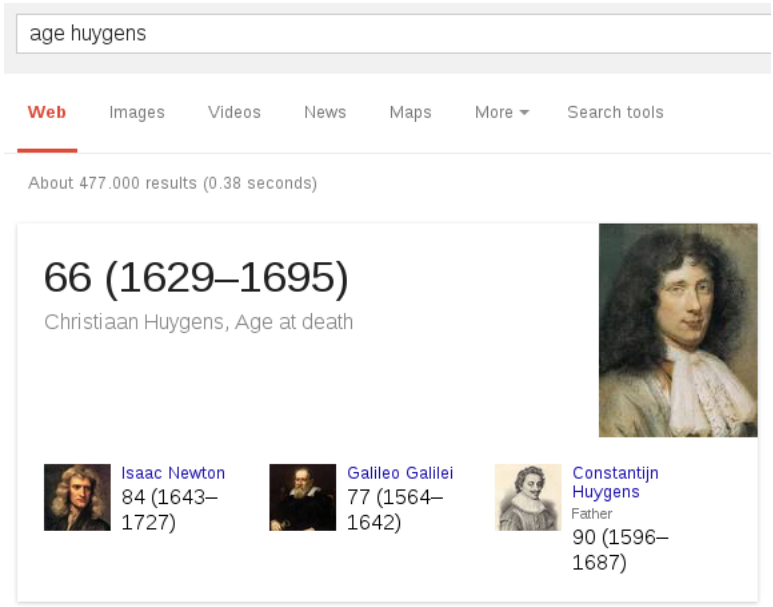


Figure II.6: Google's Knowledge Graph

However, libraries, archives, and museums can take advantage of several initiatives heralded under the common name of digital humanities[39] in order to make the best of available semantic techniques. Chapter III will investigate what the digital humanities have to offer for the semantic enrichment of a multilingual archive, to what extent, at what cost, and to whose benefit.

3.1 Terminology

Building a gold-standard corpus (GSC) for evaluation purposes requires to manually annotate potentially relevant content, using clear-cut categories. As we have seen in Chapter I, named entities (proper nouns) are traditionally separated from terms (common nouns) for this purpose. Failing to do so makes it extremely difficult to reach sufficient agreement between annotators to what constitutes a valid entity (van Hooland et al., 2015).

In the absence of an unequivocal frame of reference, a GSC is the next best thing to compare the output of an automated system to "reality". But it should be kept in mind that there is no necessary isomorphism between the GSC and reality itself: the artificial reference is always subject to interpretation by the

[39]This quite recent term covers a reality that is far from new but goes back to older quantitative approaches in the humanities. Rather than a genuine field of study, we see the digital humanities as a community of scholars with a keenness for computational practices.

human annotators, even when there is a high degree of agreement between them.

The distinction between terms, concepts, entities, mentions, labels, etc. is indeed not always entirely clear. For instance, Leal et al. (2012) define terms as "labels of concepts in an ontology", which somewhat blurs the whole picture. In order to define the task of entity linking in Section 3.2, we first investigate the relation between terms and concepts on the one hand (Section 3.1.1), and between terms and entities on the other hand (Section 3.1.2).

3.1.1 Terms and concepts

Much as the distinction between resources and pages, the opposition between concepts and terms can be traced back to de Saussure's division of the linguistic sign into signifier and signified (see Section 2.3), illustrated in Figure II.7.

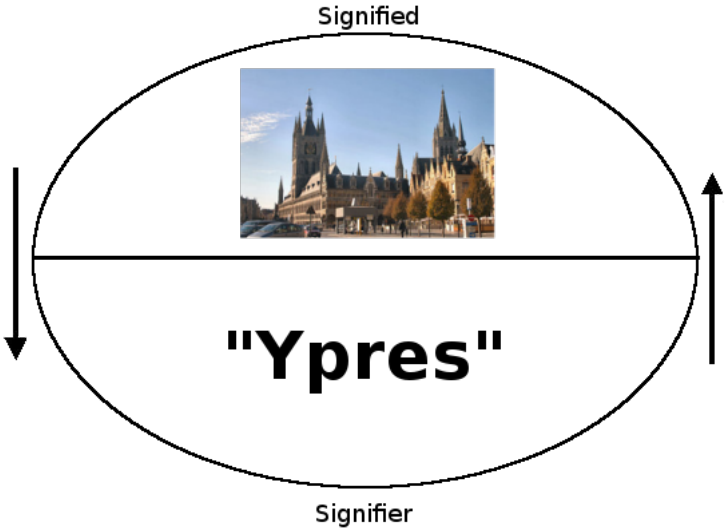


Figure II.7: Arbitrariness of the linguistic sign

The linguistic sign being intrinsically arbitrary (as shown by the fact that the same object has different names in different languages, or even in the same language), Frege (1960, p. 60) drew a distinction between the idea, the sense, and the reference of an object:

> "The reference of a proper name is the object itself which we designate by its means; the idea, which we have in that case, is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself."

To illustrate his point, Frege (1960, p. 57) argues that "The reference of 'evening star' would be the same as that of 'morning star', but not the sense". To clarify this further, he provides the analogy of someone observing the Moon through a telescope. The Moon itself is the reference, the physical object, whereas the optical image projected through the glass in the interior of the telescope can be compared to the sense. This image is clearly not the Moon, but it is also distinct from the retinal image of the observer (which Frege calls the idea or experience).

Must a concept necessarily be abstract or can it also cover a more concrete reality? For Prost (1996, pp. 126–127), "on hésite à parler à propos [des désignations d'époque] de concepts, car ces termes ont un contenu concret indiscutable. [...] pour qu'un mot devienne un concept, il faut qu'une pluralité de significations et d'expériences entre dans *ce seul* mot" (italics his). We immediately notice the subjectivity of the distinction between a "plain" word and a full concept, which can vary with the context and/or the interpretation supplied.

The ISO 25964 norm (ISO, 2011a, p. 3) also distinguishes terms from concepts and defines the latter as follows: "Concepts can often be expressed in a variety of different ways. They exist in the mind as abstract entities independent of terms used to express them." This definition is criticised by Meunier (2014) for a number of reasons, the least not being that *abstract entities in the mind* are not very useful when dealing with very concrete thesaural relations in an operational context.

The community of computational linguists and NLP scholars is by no means immune to this concept/term dissociation fallacy, as it can be noticed from the point of view adopted by the organisers of the Automatic Content Extraction (ACE) evaluation in their guidelines for competitors, as reported by Ahn (2006):

> "Within the ACE program, a distinction is made between entities and entity mentions (similarly between event and event mentions, and so on). An entity mention is a referring expression in text (a name, pronoun, or other noun phrase) that refers to something of an appropriate type. An entity, then, is either the actual referent, in the world, of an entity mention or the cluster of entity mentions in a text that refer to the same actual entity. The ACE Entity Detection and Recognition task requires both the identification of expressions in text that refer to entities (i.e., entity mentions) and coreference resolution to determine which entity mentions refer to the same entities."

Whereas the pervasive synonymy phenomenon resulting in an entity being referred to by various literal forms (as discussed in Chapter I) is not disputed here, what constitutes "the actual referent, in the world" and how this relates to the "cluster of entity mentions" remains entirely unclear. The nature of the link between the signifier (term) and the signified (concept) is as elusive as ever.

3.1.2 Terms and entities

Although terminology extraction and named-entity recognition have been conducted in the past as distinct research fields (see Section 1.3 of Chapter I), there is a strong case for considering them together since the two tasks share a number of similarities. From a strictly practical point of view, there is no intrinsic difference between the DBpedia resources for Ypres (`dbr:Ypres`) and for ruins (`dbr:Ruins`): both are identified by a URI and share common properties. Although they are listed as different types of resources (`dbo:Place` and `yago:Decay` respectively), both are part of classes that can ultimately be traced back to the `owl:Thing` superclass, which covers all the resources of DBpedia.

The blending of entities and terms can seem confusing at first because of the traditional distinction between common and proper nouns adopted by most lexicographers, but the arbitrariness of the separation is laid bare when looking at it from an information retrieval perspective: when searching for information, common and proper nouns are used indistinctly, and the previously insurmountable distinction between capitalised and uncapitalised words promptly disappears before search engines.

Let it suffice to look at the query proposed at the beginning of Section 3 to persuade ourselves of this fact: "`christiaan huygens date birth death`" indeed mixes terms and concepts without any regard for semantic distinctions. Even more convincingly, the field survey presented in Chapter III will show that top search terms include both terms and entities, and that Ypres and the Titanic stand alongside war and ruins in the preoccupations of users.

Moreover, some mentions are simply too ambiguous to be successfully categorised as either a term or an entity: they lie in-between. Prost (1996, p. 131) proposes the example of "Crise économique d'Ancien Régime" and calls it a half-proper noun or imperfect common noun. In fact, the expression is too general to be considered a named entity in the strict sense defined by Kripke (see Chapter I), but still more precise than the plain word "crise". Considering such an empirical concept as one or the other is not a deterministic choice, but rather a matter of interpretation relative to a given usage, as will be argued in Chapter III.

Some NLP researchers explicitly acknowledged this issue. Alfonseca and Manandhar (2002), for instance, consider that plain NER is too restrictive by only recognising persons, organisations and locations. Instead, their methodology extends to all kinds of concepts contained in WordNet, blurring the distinction between entities and terms/concepts. Moreover, it being unsupervised makes their approach applicable to various languages and domains. The authors nevertheless maintain the distinction and argue that "it would be desirable that each synset [of WordNet] had a flag indicating whether it represents an instance or a concept" (a synset being a cluster of synonymous terms). Kulkarni et al. (2009) also criticise former work on entity annotation for being "biased toward specific entity types like persons and places".

In order to account for this reality, we need tools that are not confined to the extraction of specific categories of words, as strict NER systems and terminology extractors are, but are more focused on what users actually search for (*user-oriented*) and are interested to discover about (*result-oriented*). Some of the tools presented in Chapter V clearly go in this direction, and we will push the case for following this path.

While accepting that there may be some circumstances where the distinction between terms and entities is still productive, we reckon that this discussion allows us to answer our first research question conclusively in the context of general-purpose semantic enrichment: all concepts are equal when considered from the perspective of casual end users.[40] Determining which mentions are relevant in a document should therefore remain the prerogative of users in relation to their needs, rather than be imposed by restrictive technology.

3.2 Entity linking

Entity linking is a new task that has emerged over the last few years and has also been called entity resolution (Alexopoulos et al., 2015), named-entity extraction (NEE) (Fafalios et al., 2015) or record linkage (Tylenda et al., 2014). Interestingly, Fafalios et al. (2015) note that "Entity Linking is also considered a way of Named Entity Disambiguation (NED), since a resource (e.g. a URI or a Wikipedia page) can determine the identity of an entity".

Whereas classic NER limits the disambiguation of an entity to a categorisation, entity linking goes further by trying to resolve the meaning of the entity with a unique identifier in a knowledge base. According to Fafalios et al. (2015), "a major challenge for the Semantic Web is the extraction of structured data through the development of automated NEE tools".

[40] Although Orwell would probably consider that some concepts are more equal than others...

In this section, we will investigate three tasks that have appeared separately but that can now be reconciled under the common heading of entity linking: disambiguation to Wikipedia or *Wikification*, semantic annotation of documents, and Knowledge Base Population. All these early attempts go in the direction of the generalisation and decompartmentalisation of IE defended in this thesis, and we will build on them in Chapter V. For a complete survey of state-of-the-art entity linking techniques, see Shen et al. (2015).

3.2.1 Wikification

As stated in Chapter I, natural language is intrinsically ambiguous. Despite improvements in the field of word sense disambiguation (Moro et al., 2014), this ambiguity remains a major challenge for all kinds of NLP applications. Efforts in this direction using Wikipedia as a component are detailed below.

In the context of *geoparsing*, for instance, which involves the identification of place names in unstructured text, Moura and Davis (2014) note that "place names are often ambiguous with other place names and with nouns used to designate people and objects". Accordingly, the authors make a distinction between *geo/geo ambiguity* (i.e. a place mention ambiguous with other place names) and *geo/non-geo ambiguity* (a place ambiguous with other entities). This situation is by no means restricted to places and reflects the various types of ambiguity detailed in Chapter I (Section 2.3). To solve this problem, Moura and Davis (2014) argue that "Wikipedia is a good external source of evidence, both for recognition and for disambiguation".

For Kulkarni et al. (2009), "the challenge of Web mining systems is to harness the chaotic 'wisdom' of the crowds into relatively clean knowledge" (reminding us of the DIKW pyramid of Ackoff (1989), see Section 1.1). Their aim is to "identify textual references (called 'spots') to named entities and annotate the spots with unambiguous entity IDs (called 'labels') from a catalog". To do so, the authors propose a general collective disambiguation approach for the aggressive open-domain annotation of Web pages, by treating entity mentions globally at document-level rather than individually. To illustrate their method, they give the following example: while *Michael Jordan* and *Stuart Russell* are fairly ambiguous names than can refer to a lot of different persons, "a page where both *Michael Jordan* and *Stuart Russell* are mentioned is almost certainly about computer science, disambiguating them completely".

Charton and Torres-Moreno (2009) operate what looks like a complete reversal of the problem of named-entity disambiguation: instead of using the full potential of Wikipedia to disambiguate entities, they try to classify them back in categories in order to fit in the traditional definition of the NER task.

In this approach, the requirements of the ESTER evaluation campaign (see Chapter I, Section 2.2.1) clearly supplant the potential added value for users.

Finally, Singh et al. (2011, 2012) use an impressive "labeled corpus of 1.5 million disambiguated mentions in Web pages by selecting link anchors referring to Wikipedia entities" in order to train their entity linking system. This approach is biased, however, since "Wikipedia mentions are disproportionately likely to have corresponding Wikipedia pages" compared to mentions from general text according to Ratinov et al. (2011), who also identify the "key remaining challenge: determining when mentions refer to concepts *not* captured in Wikipedia", a task similar to NIL clustering (see Section 3.2.3).

All these approaches are limited by the fact that Wikipedia is not a real knowledge base but an encyclopaedia: most of the information it contains is in unstructured form, and links to external sources of knowledge are sparse. In order to improve the coverage and reliability of the information extracted to enrich content, a larger panel of resources needs to be considered, which can only be achieved with Linked Data.

3.2.2 Semantic Annotation

According to Stern (2013), semantic annotation is the task consisting in linking annotated elements in text to resources from the Semantic Web, such as knowledge bases. In agreement with the vision of Tim Berners-Lee, its aim is to take advantage of the potential of human knowledge hidden in unstructured data, and to connect it in a way that is workable for intelligent agents. Semantic annotation can be seen as the operational counterpart of the Semantic Web, linking information diluted in unstructured text to models of formalised knowledge. Tamilin et al. (2010), for instance, used this technique to semantically enrich an Italian news archive.

IE can therefore be seen as a necessary component of semantic content enrichment, since it provides the technical framework for the extraction of relevant information and its automated exploitation. Wilks and Brewster (2009) in fact opened the way to this interdisciplinary approach by considering natural language processing as an essential foundation of the Semantic Web.

Annotating existing content with semantic metadata allows to improve their future interpretation. To do so, semantic annotation needs to cross the border between linguistic data and formal representation. It remains that annotation is not an end in itself but should serve a tangible goal. While Stern (2013) deplores the fact that the Semantic Web suffers from a lack of annotated objects, Wilks (2008) conversely warns against the "apotheosis of annotation" which risks to obfuscate the real meaning of documents.

3.2.3 Knowledge Base Population

Although efforts to map natural language text onto databases have existed for a long time (Mazlack and Feinauer, 1980), Knowledge Base Population (KBP) is a relatively new area of research consisting either in enriching an existing KB with new facts and relations extracted from unstructured text, or even in populating an entirely new KB from scratch (cold-start KBP). It differentiates between non-collective approaches which process each entity mention separately, and collective approaches which gain advantage of the global coherence of a document to disambiguate related entities simultaneously.

KBP emerged as a separate track in the 2009 edition of the Text Analysis Conference (TAC),[41] with the aim to "promote research in and to evaluate the ability of automated systems to discover information about named entities and to incorporate this information in a knowledge source".[42] Since 2014, TAC-KBP includes an entity discovery and linking (EDL) task, which "requires a system to take raw texts as input, automatically extract entity mentions, link them to a knowledge base, and cluster NIL mentions"[43] (Ji et al., 2014). This task was initially focused on English but has been extended to Chinese and Spanish for the 2015 campaign, opening the way to multilingual entity linking (although mapping remains unidirectionally linked to an English KB).

Whereas a simpler entity linking task (aiming at linking a given named-entity mention to a KB) had been part of the TAC-KBP track since 2009, EDL introduces the idea of an end-to-end pipeline, thereby recognising the need to merge NER with entity linking. In this sense, EDL is related to Wikification but remains restricted to traditional named-entity categories: persons, organisations, and geo-political entities.

EDL also differs by the importance given to NIL clustering: whereas simple Wikification can afford to ignore entities without a Wikipedia entry, EDL (and KBP in general) is more demanding in that it precisely aims to enrich a KB with new concepts. This can prove particularly tricky when a similar entry is present within the KB, which requires the EDL algorithm to assign a stronger weight to the absence of link (NIL) than to the best candidate. Let us consider the following example:

Example 22. Guido Calogero (1904–1983) enseigna l'histoire de la philosophie à l'Université de Rome. Calogero diriga également la revue *Panorama*.

[41] `http://www.nist.gov/tac/2009/`

[42] Quoted from the task description of TAC-KBP.

[43] A NIL mention is a found named entity that is considered to be without a corresponding entity in a given KB. NIL clustering is central to KBP, since it allows to group variants of a yet undocumented concept before deciding to create a new entry in the KB.

While the Italian philosopher Guido Calogero has its own page on the Italian version of Wikipedia,[44] it lacks a counterpart in either English or French (i.e. `https://en.wikipedia.org/wiki/Guido_Calogero` and `https://fr.wikipedia.org/wiki/Guido_Calogero` do not exist). The challenge for an EDL system would be to recognise both "Guido Calogero" and "Calogero" as valid entity mentions, to cluster them together but to link them to NIL with respect to the French Wikipedia, rather than establishing an erroneous link to `https://fr.wikipedia.org/wiki/Calogero` which refers to the French singer–songwriter.

A related task is the extension of ontologies, sometimes called knowledge acquisition. Alfonseca and Manandhar (2002) propose an approach to do so that is completely unsupervised, based on what they call *General Named Entity Recognition*, "a task that covers, and is harder than both Named Entity Recognition and Word Sense Disambiguation".

Finally, Exner and Nugues (2012) propose a method to automatically transform information contained in unstructured text into RDF triples, mapping them to the DBpedia namespace.[45] However, several questions are left unanswered in their approach, such as the quality of the reference, the adaptation of this method to other languages or the evolution of the triples over time, all of which will be tackled in Chapter IV.

3.3 Semantic relatedness

According to Leal et al. (2012), "extracting the semantic relatedness of terms is an important topic in several areas, including data mining, information retrieval and web recommendation". The relationships existing between terms send us back to the relations between entities discussed in Chapter I, but computing semantic relatedness will also allow us to design an alternative approach to search suggestions in Chapter V.

Pioneer work in this field was conducted by Gabrilovich and Markovitch (2007) who proposed an original method based on Wikipedia and named it "explicit semantic analysis". One of the interesting features of their approach is that they do not make a difference between the lexical level and the document level: their method "treats both words and texts in essentially the same way" (compare with Blanke and Kristel (2013)). Considering words in context, the authors then compare their occurrence patterns across a large collection of natural language documents in order to obtain an accurate representation of the meaning of a text as a vector of Wikipedia-based concepts.

[44] `https://it.wikipedia.org/wiki/Guido_Calogero` (also available in Basque and Polish).
[45] The authors made their extracted corpus available at `http://semantica.cs.lth.se`.

Wubben and van den Bosch (2009) investigate the difference between semantic similarity (synonymy) and semantic relatedness. They argue that while WordNet shortest paths are well suited for the former, they do not export well to other types of networks such as Wikipedia and ConceptNet. The authors propose a new estimate metric based on a bidirectional breadth-first search algorithm run on an open graph structure extracted from the KBs. While the results for ConceptNet are lower due to its lack of coverage, the score achieved on a Wikipedia dump significantly outperforms former approaches, showing that "free link structure in conceptual networks is better suited for finding semantic relatedness than hierarchical structures organized along taxonomic relations as WordNet" (Wubben and van den Bosch, 2009).

Building on these insights, Leal et al. (2012) propose a novel approach based on a graph extracted from DBpedia. They introduce the notion of *proximity*, including a connectedness component, to replace the traditional measure of *distance* between two nodes in a graph: "rather than focusing solely on minimum path length, proximity also balances the number of existing paths between nodes". To achieve their goal, the authors rely on Apache Jena,[46] a free and open source Java framework for building Semantic Web and Linked Data applications. Unfortunately, their approach is self-defeated by the incompleteness of DBpedia for the specific domain they planned to cover, Portuguese alternative music.[47]

In a more ambitious study, Mikolov et al. (2013) from Google Research show that high quality vector representations can be derived using simple model architectures on a huge corpus (1.6 billion words) at a low computational cost. Their approach outperforms more complex ones such as neural networks and significantly improved the state-of-the-art performance for measuring semantic similarity.

Although these different experiments allow to improve on traditional relation extraction by leveraging the links already established between semantic resources in knowledge bases, semantic relatedness measures can also suffer from the shortcomings of Linked Data already put into light, as shown by the failed project of Leal et al. (2012) for instance.

Before introducing our case study in the context of the digital humanities and empirical sciences in general, let us take stock of what has been achieved in this second chapter.

[46] `http://jena.apache.org/`

[47] See Chapter IV for a detailed analysis of the common quality problems related to DBpedia.

Summary

Chapter II introduced the idea of a Web that can be consumed by machines, a Web that makes its semantics explicit (or at least less obscure than in the Web of documents). From the original vision of Berners-Lee (2000), the project evolved into an aggregation of individual collections of structured facts, known as Linked Open Data or as the Web of Data.

We have seen, however, that many innovations of the Semantic Web are in fact regressions compared to well-established database practices: the use of URIs as unique identifiers in knowledge bases raises unforeseen issues of coherence, while the lack of integrity constraints in ontologies makes them more error-prone that traditional relational databases.

Despite these known limitations, KBs offers the potential to improve on information extraction techniques by allowing the disambiguation of concepts through the mechanism of entity linking. While named-entity recognition stopped at the classification of entities into broad semantic categories, entity linking goes one step further and attempts to establish a one-to-one correspondence between a mention in text and a semantic resource identified by a URI. This operation is far from trivial, however, and can fail in numerous ways, from lack of context to data sparsity, as will be amply illustrated in Chapter IV.

It remains that when entity linking succeeds, a whole world of possibilities is opened in terms of semantic enrichment and knowledge discovery: new properties can be learned about the entities extracted, and interesting relationships to other entities can be uncovered in an automated way, as we will see in Chapter V. While the quality of this material can remain a blocking factor for some critical applications, empirical domains in general – and the humanities in particular – can benefit from this gold mine to enrich existing content at a reduced cost, as we will now explore in the next chapter.

Chapter III

The Humanities and Empirical Content

Outline

In this chapter, we show how humanities players can take advantage of Linked Open Data in order to gain more value out of existing content. In particular, cultural heritage institutions such as libraries, archives, and museums can benefit from extraction tools to increase the visibility of their collections. Without targeting a specific sub-domain, we make the case for the exploitation of semantic technologies in any empirical context, showing that a high level of generalisation is not incompatible with useful operational results.

Section 1 starts by establishing the distinction between deterministic and empirical information: while the former can always rely on a stable model to check its validity at any time, the latter is subject to human interpretation and therefore needs to be considered from a different perspective allowing for a subjective arbitration between conflicting needs.

As shown in Section 2, the humanities epitomise empirical science, and as such constitute a playground for the computational techniques highlighted in the first two chapter of this dissertation. Under the common denomination of the digital humanities, we group all kinds of practices designed to exploit documents in an automated manner, providing complementary insights to the traditional intellectual exploitation by scholars. Since these practices do not enjoy unanimous support from humanities scholars, we address critiques that have been formulated against them to get things into perspective.

In order to demonstrate the general applicability of information extraction and semantic enrichment techniques for cultural heritage, we then introduce in Section 3 a case study based on the *Historische Kranten* project at the Ypres City Archive, an effort to publish online a collection of over one million Belgian periodical articles written in three languages over the course of a century.

Contents

1 Empirical Information

> "One reason why mathematics enjoys special esteem, above all other sciences, is that its laws are absolutely certain and indisputable, while those of other sciences are to some extent debatable and in constant danger of being overthrown by newly discovered facts."
> Albert Einstein (1922)

The border between human sciences and exact sciences is not airtight. As Boydens (1999) reminds us, the same object can be approached simultaneously from several disciplines depending on the scientific method adopted: "la différence entre sciences exactes de la nature et sciences inexactes de l'homme et du vivant n'est pas une différence de fond mais une différence de choix" (Moles, 1995, p. 39). Similarly, Rickert (1986, p. 54) tells us that "empirical reality becomes nature when we conceive it with reference to the general. It becomes history when we conceive it with reference to the distinctive and the individual." Ladrière (1984) divides science into three categories:

1. formal sciences, (mathematics, logic, etc.)
2. social sciences and the humanities (law, history, literature, etc.)
3. natural sciences (chemistry, physics, medicine, etc.)

The particular status of formal sciences, i.e. the perfect isomorphism between the model and reality, allows them to draw a bijective function (one-to-one correspondence) between the object of study and its representation. New discoveries occur, but they never contradict the pre-existing model. However, as Boydens (1999, p. 141) notes, a database (or knowledge base) containing exclusively mathematical or logical formulas has little meaning whatsoever, which rules out formal sciences as an application domain. This leaves us with categories 2 and 3 – that we group under the common tag of "empirical sciences" – for which there is always a necessary gap between the model and the observations. In other words, theory in empirical sciences always constitutes a "reconstruction conjecturale de la réalité" (Ladrière, 1984, p. 39).

The present discussion about the nature of empirical sciences will allow us to anticipate specific issues related to our case studies, but also to detect common traits with related domains in order to generalise our approach. Section 1.1 investigates the difference between deterministic and empirical information. Section 1.2 will then envision application domains at the crossover of empirical and formal sciences, while Section 1.3 focuses on the specificities of the humanities.

1.1 Deterministic and empirical data

Formal knowledge, such as the laws of arithmetic, exhibits a relative stability that is not constantly questioned. The validity of the equation 1+1=2, for instance, is considered universal and not subject to change any time soon. In contrast, empirical data, whether in structured form (databases, XML files) or unstructured form (text), is always open to interpretation because of the absence of a reference: "il n'existe aucun référentiel absolu en vue de valider la correction de l'information représentée dans une base de données relative à un domaine d'application empirique et humain" (Boydens, 1999, p. 143).

We have seen in Chapter II that knowledge units (facts) consist of a subject, a predicate and a object formalised as a *(s,p,o)* RDF triple. Similarly, a data element can be considered a triplet *(i,d,v)* consisting of a unique identifier *i*, a domain of definition *d* and a value *v*. The differences between deterministic and empirical data, directly reminiscent of the distinction between formal sciences on the one hand and social and natural sciences on the other hand, are emphasised in the work of Isabelle Boydens (2011, p. 118, italics hers):

> "It is important to distinguish *deterministic data* from *empirical data*. The first 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. [...] 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."

Empirical concepts do not follow any hard-coded rules. On the contrary, they are "construits par une série de généralisations successives et définis par l'énumération d'un certain nombre de traits pertinents, qui relèvent de la généralité empirique et non de la nécessité logique" (Prost, 1996, p. 129). Crucially, the absence of a referential makes empirical information impossible to assess independently of the reality it represents (Boydens, 1999, p. 469):

> "La question de l'exactitude de l'information empirique est en elle-même dépourvue de sens. [...] Afin de valider l'information répertoriée dans une base de données, il faudrait idéalement connaître *a priori* une réalité qu'elle seule nous permet de connaître."

The same holds true for most of the facts contained in knowledge bases in the form of RDF triples: their empirical nature make them immune to simple validation procedures without access to the underlying reality. This is not to say that empirical data escape control completely, but their quality is always relative to usage, as will be detailed in chapter IV.

Drucker (2012, p. 90) also notices the palpable tension between deterministic and empirical realities: "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". Some domains, however, lie in-between and combine both types of tasks.

1.2 Crossover application domains

It is commonly admitted that data from the social sciences are empirical and therefore subject to human interpretation. This property makes them particularly interesting to study with new techniques of analysis and visualisation, as we will do in Chapter V. The reverse does not hold true, however: empiricism is not intrinsically related to social sciences but is also inherent to disciplines from the natural sciences, ranging from medicine to physics and aeronautics.

These "empirico-formal" sciences rely on formal models but deal on a daily basis with empirical data that they need to confront with the model. In contrast to the laws of arithmetic that are immutable, models from empirico-formal sciences evolve over time with the "boomerang of reality". In particle physics, for instance, the Standard Model, on which the whole theory of nuclear interactions relies, could have been abandoned if the existence of the Higgs boson had been disproved by experiments at the Large Hadron Collider (Guasch and Sola, 1998).[1]

Although they rely on complex formalisms, natural sciences do not escape human interpretation, as shown by plenty of examples where observations of real-world phenomena led to theory change. In the case of stratospheric databases, for instance, measurements of low ozone levels recorded by the NASA were initially discarded as anomalies because the corresponding theory at the time could not account for them: the integrity constraints of the database ensured they were rejected. Only after ozone depletion was discovered a decade later were the data reassessed and reinterpreted in light of the new theory (Wiener, 1994, p. 37), as will be further discussed in Chapter IV.

Life sciences, and particularly biomedicine, have shown great interest in information extraction techniques (Ananiadou and McNaught, 2006), as we already mentioned in Chapter I. The evolutions presented in Chapter II to disambiguate concepts and enrich content with Linked Data would definitely benefit the field, since most biomedical concepts (such as diseases and genes) are indeed empirical and subject to changes in their interpretation.

[1] As it happened, the existence of the boson was confirmed and the Standard Model left reinforced by the discovery, earning François Englert and Peter Higgs a joint Nobel prize.

As Zheng et al. (2014) remark, however, "simply applying a news-trained entity linker [on biomedical information] produces inadequate results". This inadequacy can have far more damaging consequences in biomedicine, where public health is at risk, than when handling empirical content from the humanities. In order to extract knowledge efficiently while ensuring that its quality measures up to the needs of users, other sources can be investigated but the methodology remains the same:

> "Wikipedia is a popular knowledge base that is often used for entity linking because it contains structured information such as titles, hyperlinks, infoboxes as well as unstructured texts. However, in order to take advantage of richer structures and domain knowledge which are not offered by Wikipedia, we constructed a knowledge base from 300 biology-related ontologies from BioPortal.[2] Based on the rich structure contained in these ontologies, we created a web of data (WOD)."

In this way, Linked Data can improve the experience of users from a broad range of domains, although some remain understandably less keen to engage with crowdsourced, uncurated resources when financial losses are at risk or human lives at stake.

The incorporation of domain knowledge might seem to contradict the ideal of generalisation formulated in our third research question, but switching knowledge bases does not amount to specialisation since the underlying technology (OWL and RDF) remains the same, and so does the linking principle. In fact, the mechanism of identity implemented by the `owl:sameAs` property will even allow expert systems and popular knowledge bases to coexist and to complement one another.[3]

In Chapter V we will see examples of tools allowing to use various ontologies selectively without any need to redesign the whole application for each application domain. X-Link, for instance, offers a fully configurable model which can be used in a wide range of contexts, and is actually implemented in two very different projects related to marine activity[4] and patent search[5] (Fafalios et al., 2014).

[2] `http://bioportal.bioontology.org/`

[3] Although this raises the important question of provenance, which is often underestimated in the context of Linked Data (Hartig and Zhao, 2010).

[4] `http://www.i-marine.eu/`

[5] `http://www.perfedpat.eu/`

1.3 Specificities of the humanities

The subjective dimension of the humanities is generally accepted as a matter of course, with historical views on a past event regularly changing and interpretations of literary works varying at the whim of new theories. However, the formal modelling of empirical knowledge often indulges in postulating a direct bijection between the real object and its formalised representation. As Boydens (1999, pp. 142–143) explains, the fallacy of the isomorphism between the real and the represented can be brought into focus by the very nature of social sciences, for which reflexivity is a key component:

> « L'hypothétique rapport biunivoque entre le réel étudié et sa représentation [en sciences humaines] est d'autant plus illusoire qu'à la différence des sciences dites empirico-formelles, le sujet observant est de même nature humaine que l'objet observé auquel il est immanent. »

Similarly, Rickert (1986, p. 114) argues that "the logical distinctiveness of an empirical science is to be understood in terms of the relationship the content of its concepts bears to empirical reality in its unique and distinctive form". The positivist tradition in history long attempted to pretend that its methods were as objective as those of exact sciences: "j'étais devant mon sujet comme devant la métamorphose d'un insecte" wrote Hippolyte Taine (1875, p. 5). But this myth was deconstructed by later historians conscious of their own necessary subjectivity.

Ginzburg (1989, p. 179) remarks that this puts the humanities before an unpleasant dilemma: "ou assumer un statut scientifique faible pour arriver à des résultats marquants, ou assumer un statut scientifique fort pour arriver à des résultats négligeables". The humanities have regularly oscillated between those two extremes. In the next section, we try to reconcile both approaches by embracing what the digital humanities have to offer to a certain extent, while keeping a critical eye on the unnecessary hype that often accompany technological advances presented as new although they bear a striking likeness to well-worn considerations.

While focusing on the humanities, we will keep in mind that the border between social and natural science is porous and that this compartmentalisation of empirical sciences is somewhat artificial: "Nous avons pris l'habitude de diviser conceptuellement l'univers selon les lignes de partage des différents domaines universitaires de spécialisation" (Elias, 1996, p. 97). A concrete consequence of this observation is that the results obtained can easily be generalised to other empirical application domains.

2 Digital Humanities

This section introduces the field of the digital humanities (DH), an interdisciplinary area of research at the crossroads of computer science and social sciences. DH strives to put modern technology at the service of humanities students and scholars, thereby opening up new possibilities for the analysis of empirical objects. The purpose of the section is to show how expert tools and methods can be popularised in order to benefit this larger audience. More specifically, it will consider the application of computing techniques on corpora stemming from traditional humanities subfields, paving the way for the exploitation of a historical archive to be introduced in Section 3.

We first show how the DH have gained momentum with the advent of the Web of Data, redefining disciplines by making huge corpora available online (Section 2.1), before taking a closer look at the notion of "distant reading" coined by literary scholar Franco Moretti in opposition to the traditional practices of *close reading* (Section 2.2). The origins of the latter, in the context of the New Criticism literary movement, are investigated before addressing the more recent concept of distant reading, along with the debate on the end of literary theory. Both are then reconciled to overcome the artificial division between these two complementary approaches, opening the way for an integrated exploitation of our archive.

The polemic claims of some DH evangelists have nonetheless been met with scepticism, and a number of counterarguments have been formulated in order to refute the most excessive stances. The heated debate around DH illustrates the fact that the enthusiasm for computational methods must be handled with care (Section 2.3), although distant reading and related techniques can nevertheless prove a useful complement to human analysis when comprehensiveness is rendered impractical by the volume of data to process.

2.1 Context

Faced with increased budget cuts, libraries, archives, and museums are forced to become more pragmatic with their metadata creation and management. Funding bodies and grant providers expect short-term results and push cultural heritage institutions to gain more value out of their own existing metadata by linking them to external data sources (van Hooland et al., 2013).

It is precisely in this context that Linked Open Data (LOD) have gained momentum in the cultural heritage sector (van Hooland and Verborgh, 2014). For small institutions, the perspective of reusing existing knowledge to bridge their collections to the Web has important implications in terms of visibility.

2.1.1 From humanities computing to digital humanities

Although some literary scholars and historians have been using computing techniques for a long time, the field of humanities computing – as imagined by pioneer Roberto Busa (1980) – remained relatively niche until the nineteen-nineties when the Internet became available to the general public.

In their comprehensive introduction to the field, Schreibman et al. (2008) chose to use the term *digital humanities* rather than *humanities computing* in order to put the stress on the revolution that has taken place over the last decades: instead of a fringe branch of computing dedicated to humanists, DH mark the entering of humanities scholarship into the digital era. The distinction may appear insignificant, but it symbolises the re-appropriation of a whole research area by domain experts, in contrast to a technocratic approach to the humanities.

Rather than a brand new domain, DH can therefore be conceived as a re-definition of the traditional field of the humanities with a view to encompassing new methods made available by recent advances in computer science, thereby empowering it and enabling it to deal with new challenges.

2.1.2 The era of digitisation

For libraries, archives and museums (LAM), the advent of mass-digitisation has proven both a tremendous opportunity and a major challenge (Coyle, 2006; Hahn, 2008). While it admittedly enabled these institutions to achieve unprecedented visibility on the Web through the publication of their collections, it also raised the issue of data quality and accessibility for users, since traditional search tools are not necessarily optimised for the retrieval of large chunks of unstructured text (Tanner et al., 2009).

With limited funding, LAM are often unable to invest in developing and maintaining costly classification schemes such as thesauri, and are put under growing pressure to gain more value from their existing data (van Hooland et al., 2015). In this context, DH constitute a real opportunity to exploit rich material that has accumulated over the years but proves too time-consuming to process manually.

The quality of optical character recognition (OCR) involved in digitisation projects can also be subject to a considerable amount of variations, with small LAM unable to afford state-of-the-art OCR for large collections or to control the quality of the output produced by unscrupulous third-parties, often outsourcing the job to subcontractors.

Google Books[6] is arguably the best-known, largest-scale digitisation project, but several institutions launched their own initiatives, such as the Dutch Royal Library that digitised over 8 million newspaper pages ranging over three centuries and made them available online. In Belgium, unfortunately, the case for open access to digitised material is less pervasive and plagued by copyright issues, as lamented by Thomas Crombez (2015, p. 197):

> “Less recommendable is the newspaper digitization initiative of the Belgian Koninklijke Bibliotheek “Albert I” (Royal Library). Although the institution digitized c. 3.2 million pages from seventy periodicals of the nineteenth and twentieth century, it is only possible to consult this invaluable resource through “five special PCs in the reading room.” ”

Fortunately, smaller institutions sometimes have a more open-minded approach to open dissemination of their material, which allowed us to work on a concrete case study – introduced in Section 3 – without having to sacrifice the local anchorage nor the multilingual dimension.

2.1.3 Information extraction for cultural heritage

Named-entity recognition (NER) and more advanced IE techniques such as entity linking have gained attention from DH enthusiasts, since they allow small institutions to enrich their collections with semantic information at a relatively low cost. According to Blanke and Kristel (2013), “semantically enriched library and archive federations have recently become an important part of research in digital libraries and archives”. The growing of the Linked Open Data cloud and the availability of free online tools have facilitated the access to IE for librarians, archivists and collections managers that are not IT experts but are eager to experiment with new technologies.

The LOD Around The Clock (LATC) project of the European Commission, for instance, was started to “help institutions and individuals in publishing and consuming quality Linked Data on the Web”.[7] Its main declared goal is to “continuously monitor and improve the quality of data links within the Linking Open Data cloud” (see Chapter IV for a detailed account of the impact of poor quality on LOD). The mere existence of such large-scale incentives demonstrates the potential of LOD for the semantic enrichment of collections maintained in libraries, archives and museums.

[6] https://books.google.com/

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

A number of cultural institutions have therefore experimented with NER 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. Lin et al. (2010) also explore NER in order to create a faceted browsing interface for users of large museum collections, while Segers et al. (2011) 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.

Maturana et al. (2013) showed how LOD could be successfully integrated in a museum platform to enhance the experience of end users. Their innovative semantic platform MisMuseos, a meta-museum aggregating 17 000 works from seven Spanish museums, offers users a facet-based search module, semantic content creation and graph navigation among other functionalities.

In the specific domain of archives, Rodriquez et al. (2012) compared the results of several NER services on a corpus of mid-20th-century typewritten documents. 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, as explained in Chapter I. The BBC also set up a system to connect its vast archive with current material through Semantic Web technologies (Raimond et al., 2013).

Bingel and Haider (2014) compared the performance of various entity classifiers on the DeReKo corpus of contemporary German (Kupietz et al., 2010), which they say exhibits a "strong dispersion [with regard to] genre, register and time". However, the authors later concede that newspaper texts are largely prevailing and that "relatively few texts reach back to the mid-20th century", which 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. (2012) and Fernando and Stevenson (2012) 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. Frontini et al. (2015) exploited the French DBpedia and combined it with the BnF Linked Data[8] in order to extract mentions of less known authors, but their graph-based approach[9] also remained monolingual.

[8] `http://data.bnf.fr/semanticweb`

[9] See Chapter II (Section 2.1) for a short presentation of graph databases.

Finally, the periodical Aggregation and Indexing Plan for Europeana periodicals, which produced metadata for 18 million pages of news and full-text from OCR for around 10 million pages, also includes a NER component performed by the National Library of the Netherlands.[10] A new website[11] was launched in December, 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 can afford to follow. For registered users, a personal key is also provided to interact with the Application Programming Interface (API). Although some basic documentation[12] is provided, it does not reflect the actual state of the API. For instance, the PDF mentioned above says that a query on "Romanov" with a JSON output can be performed with `http://data.theeuropeanlibrary.org/opensearch/catalogue?q=Romanov&format=json&apikey=yourkey` whereas the correct URL at the time of writing is `http://data.theeuropeanlibrary.org/opensearch/json?query=Romanov&apikey=yourkey`. This state of affairs is sadly not uncommon with APIs, which rely too heavily on underlying technology and are therefore exposed to architecture changes. For a critique of the API craze, especially by Europeana and the Digital Public Library of America, see Verborgh et al. (2015).

2.2 Close and distant reading

This section offers a brief historical account of the appearance of the concept of *close reading*, and of its converse, i.e. *distant reading*. The two approaches are then reconciled to transcend the false opposition between them and move toward a more enlightened understanding of the humanities.

2.2.1 Close reading and New Criticism

Close reading consists in carefully interpreting individual texts by paying close attention to words and syntax, postulating literary self-sufficiency. Although the practice of close reading can be traced back to the Victorian era and beyond, the formal concept is generally attributed to I. A. Richards who developed it in the context of the New Criticism literary movement (Richards, 1929).

[10] `http://blog.kbresearch.nl/2014/03/03/ner-newspapers/` reported on a preliminary experiment on Dutch, French and German, accounting for about half the corpus.

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

[12] `http://www.theeuropeanlibrary.org/confluence/download/attachments/8880494/TheEuropeanLibrary_API_V2+0.pdf` (accessed on January 22, 2015).

The New Critics opposed the alternative current of Historical Criticism which took biographical and historical elements into account in order to interpret a text. In his collection of essays making the eulogy of close reading, Cleanth Brooks (1947) – another major figure of the New Criticism – famously described a literary work as a *well wrought urn*: i.e. an autonomous artefact able to speak for itself without any recourse to biographic or exegetic material.

2.2.2 Distant reading or the end of theory

The notion of distant reading, as defined by literary scholar Moretti (2005), has been gaining considerable attention. Instead of traditional methods of close reading, consisting of manually reading and interpreting a very limited corpus, cultural heritage institutions are increasingly experimenting with NLP to allow distant reading practices by end users, using analysis and visualisation techniques such as graphs, maps, and trees.

When confronted with huge volumes of documents, Moretti argues, one should distance oneself from the text and stop to consider it as the only object worth of attention, as in the tradition of close reading. To emphasise his point, he coined the opposite approach of *distant reading* which consists of making sense of texts without actually reading them but by making use of a variety of computational techniques (Moretti, 2005).

Initiatives such as the one led by librarian Eric Lease Morgan and other digital humanists at the Center for Research Computing at the University of Notre-Dame[13] allow readers to explore books by either close or distant reading (see Figure III.1). Selecting the "Do distant reading" option enables browsing frequent entities such as Names and Organisations, as seen on Figure III.2.

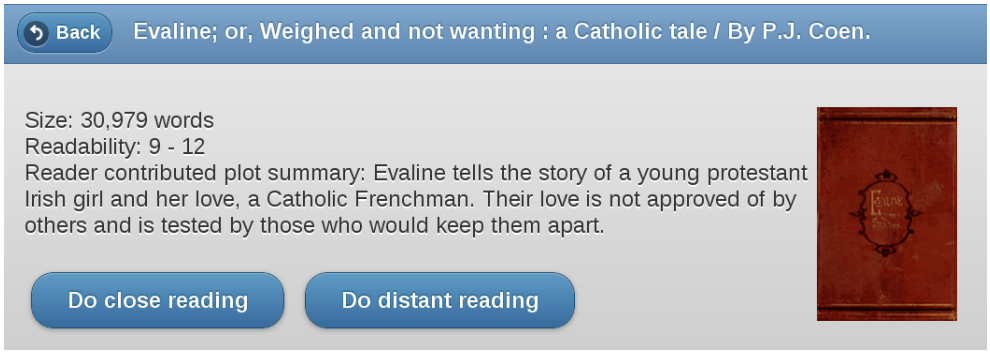


Figure III.1: Close and distant reading

[13] http://dh.crc.nd.edu/

Figure III.2: Browsing named entities

Anderson (2008) goes further by stating that data will supplant science: "We can stop looking for models. We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot". This extreme view has contributed to make distant reading unpopular among traditional literary scholars, giving the impression that reading books was no longer desirable nor necessary.

2.2.3 Reconciling the two approaches

The tension between close and distant reading, exacerbated by some DH scholars, is not as dichotomous as they pretend: as shown in Figure III.1, the two approaches can complement each other fruitfully. Nevertheless, it raises the question of the right distance from which to look at a historical object, already raised by Ginzburg (2002) in the context of micro-history.

In the domain of archives, distant reading can take another dimension and become a prerequisite of close reading: getting a mental overview of the composition of archive collections – for instance in terms of the most common named entities or geographical dispersion – is a useful preparation before actually visiting the institution physically. This is especially true when high-quality digital copies of the archives are not available online.[14]

[14] A good example of this is the Perelman archive at the Université libre de Bruxelles. Its website (`http://perelman.ulb.be/`) makes clear that "la base de données 'Archives Perelman' n'a pas vocation à remplacer entièrement la consultation des archives 'papier'. Elle se veut un outil de recherche qui puisse en rendre compte de manière significative, et permettre un accès direct vers une sélection de pièces". See Chapter V for a generalisation of our method to this corpus.

While the idea of distant reading is an original concept that can bring a new depth to existing collections, other developments associated directly or indirectly with the digital humanities have not been met with the same enthusiasm, but have sparked off indifference at best and bare contempt at worst. In the next section, we review some of the objections DH has been subject to.

2.3 Critiques

Predictably, polarised arguments about the superiority of distant reading have faced sharp criticism, both from within and from without the humanities. We will look at an example of each, starting with the dangers of over-zealousness when applying computational techniques, and moving on to the more general issue of the hype surrounding the adopting of any new technology. Balancing views, we will then adopt a more moderate stance towards the usefulness of the digital humanities and take stock of what can be gained from it.

2.3.1 Over-interpretation

While the digital humanities clearly offer new possibilities to exploit larger corpora, there is always a significant risk linked to the automatic, unsupervised analysis of literary works.

In an insightful critique of the excesses of DH, literary theorist Stanley Fish issues a warning about the semi-automatic detection of pattern and their over-interpretation in the context of literary analysis.[15] To illustrate his view, Fish takes an example from Areopagitica, John Milton's pamphlet on freedom of speech and expression.

Milton writes about Presbyterians resenting the censorship of bishops but becoming censors themselves at the same time. As a result, Fish argues, "Bishops and Presbyters are the same to us both name and thing". More than a likeness in their actions, their *names* actually look alike.

Indeed, the words *Bishops* and *Presbyters* contain the consonants "b" and "p", forming a chiasmus. Both are labial plosives in phonological terms, which reinforces the similarity. What is more, the abstract contains a myriad of related words containing the same consonants: prelaty, pastor, parish, Archbishop, books, pluralists, bachelor, parishioner, private, protestations, chop, Episcopacy, palace, metropolitan, penance, pusillanimous, breast, politic, presses, open, birthright, privilege, Parliament, abrogated, bud, liberty, printing, Prelatical and people.

[15] `http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/`

Such patterns could easily be detected by specific tools and offer the ground for a new understanding of Milton's work, but the point is, as Fish concludes after a long argument, that there are only twenty-six letters in our alphabet and therefore that the co-occurrence of those "b's" and "p's" is purely coincidental. In other words, DH provides ways to find facts that were previously undetectable, but it does not necessarily follow that these new discoveries are relevant or worthy of attention.

2.3.2 The Hype cycle

> "The computer industry is the only industry
> that is more fashion-driven than women's fashion."
> Larry Ellison, CEO of Oracle[16]

The Hype cycle is a term coined by IT consulting firm Gartner[17] for representing the degree of maturity of technologies (Fenn and Raskino, 2008). As shown in Figure III.3, a "new" technology almost always raises exaggerated expectations when first launched, provoking subsequent disillusionment among users when they realise that the solution they have been sold does not live up to their initial, unrealistic hopes.

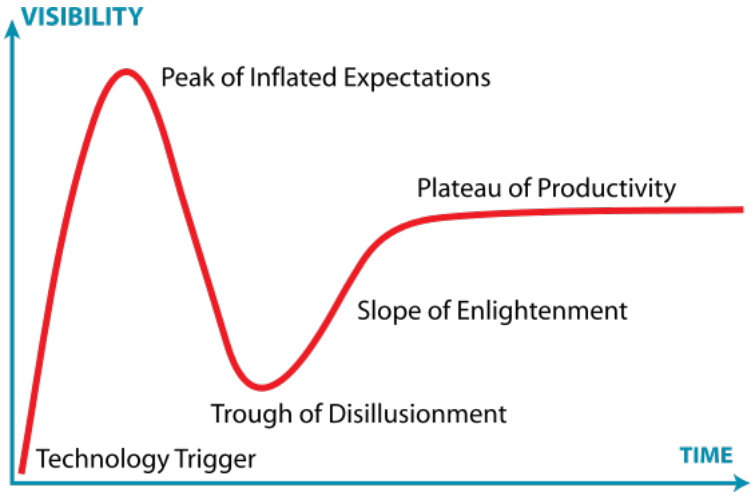


Figure III.3: The Hype cycle, reproduced from Jeremy Kemp (CC BY-SA)

[16] Quoted by Jim Finkle on `http://blogs.reuters.com/mediafile/2008/09/25/what-on-earth-is-cloud-computing/`

[17] `http://www.gartner.com/`

The digital humanities do not escape this pattern, with enthusiasts claiming it will revolutionise the humanities and sweep out former methods: "l'historien de demain sera programmeur ou il ne sera plus" dramatically proclaimed Emmanuel Le Roy Ladurie (1973, p. 14).[18] This peak of inflated expectations was inevitably followed by a trough of disillusionment, which is well illustrated by the vitriolic remarks of Stone (1979): "whenever possible, sampling by hand is preferable and quicker than, and just as reliable as, running the whole universe through a machine". In this light, Anderson's suggestion to "throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns" sound strangely hollow.

Stone insists and drives the point home:

> "It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the most disappointing."

This echoes the "unpleasant dilemma" faced by the humanities according to Ginzburg (1989), evoked in Section 1.3. And Stone again: "the sophistication of the methodology has tended to exceed the reliability of the data, while the usefulness of the results seems – up to a point – to be in inverse correlation to the mathematical complexity of the methodology and the grandiose scale of data-collection". Unfortunately, the plateau of productivity sometimes seem to have been substituted for fresh excessive expectations by a new generation of scholars. Havelange (1993) regrets this state of affairs in these terms:

> "c'est là sans doute l'un des principaux écueils d'une histoire quantitative qui, née paradoxalement du rejet explicite de la tradition positiviste, n'a pas toujours su éviter le piège d'un nouveau positivisme, d'un nouveau formalisme, d'un nouveau scientisme, aussi désséché que le premier."

If History repeats itself, how can one guard against the negative effects of computerisation for its own sake while still gaining benefits from state-of-the-art data processing and visualisation techniques? The text section will attempt to answer this important question.

[18] Ironically, one of the leading websites promoting the digital humanities today is called The Programming Historian: `http://programminghistorian.org/`.

2.3.3 Picking the low-hanging fruit

Despite its several shortcomings highlighted in the previous section, we argue that cultural institutions should try to make the best of what the DH have to offer. In the same mindset, Svensson (2010) provides an overview of current research in the field of the digital humanities at large, while distancing himself from the hype inevitably linked to the birth of a (seemingly) new discipline.

The Free Your Metadata project[19] was launched to demonstrate that players from libraries, archives, and museums can effectively use computational methods and tools in order to enhance the management of their collections (van Hooland et al., 2013). Without lapsing into DH evangelism, this initiative encourages cultural institutions to take a pragmatic stance towards metadata management and to reuse material freely available online in order to energise and promote their collections at a relatively low cost.

While metadata management may seem a long way from our main objective and only distantly related to information extraction, the two tasks are not without similarities. Some authors indeed consider named entities and other resources derived from semantic enrichment to be *metadata* in the broader sense, i.e. content about the content (Stern, 2013, p. 11, italics hers):

> “ À la différence des métadonnées de documents entendues au sens usuel, telles que les informations de date, d'auteur ou de propriété associées au document mais distinctes du contenu informatif, ces métadonnées sont *ancrées* dans le contenu textuel et relient les éléments marqués à des ressources extérieures au document, par le mécanisme des URI [...]. ”

Obtaining this meta-information to enrich documents on the largest scale necessarily requires the use of computational techniques. This is not always a bad thing: if we carefully mind our Ps and Bs, the digital humanities can still prove an inspiring way to deal with a sizeable collection of documents, as will be demonstrated in the next section introducing our case study. All things considered, the idea of distant reading makes a lot of sense when facing the monumental task of extracting knowledge from millions of digitised articles containing empirical information in a multilingual context.

[19] `http://freeyourmetadata.org/`

Figure III.4: *Historische Kranten* homepage

3 Historische Kranten

The specificities of empirical information, and the usefulness of the digital humanities as a community of practices, better reveal themselves when considered from a practical perspective. This section presents an original case study based on a trilingual (Dutch/French/English) archival corpus provided by the city of Ypres, Belgium (Figure III.4).[20] Like Scholz (2010), "notre corpus est construit sur le présupposé selon lequel le champ discursif ne s'arrête pas aux frontières des langues": we consciously refrain from adopting a comparative methodology in favour of a more unifying approach.

This *Historische Kranten* corpus[21] consists of 1 144 516 XML files from 41 Belgian periodicals published between 1818[22] and 1972, totalling 3.2 GB of text. The composition of the corpus is detailed in Table III.1. From now on, we will refer to periodical titles by their 3-letter code. Section 3.1 presents the general structure and Section 3.2 covers the linguistic distribution of the corpus, while Section 3.3 discusses the people involved and their expectations.

[20]We gratefully acknowledge the invaluable assistance of the *Stadsarchief Ieper* and of the *Erfgoedcel CO7* which granted us unconditional access to this corpus for research purposes.

[21]`http://www.historischekranten.be/`

[22]Strictly speaking, periodicals ranging from 1818 to 1829 cannot be called "Belgian" since the Independence of Belgium only occurred in 1830, but this nuance is ignored here.

periodical title (years covered)	**Lang.**	**Code**	**# files**
Journal d'Ypres (1866–1913)	FR	JDY	173 451
Het Ypersche – La région d'Ypres (1920–1944)	NL/FR	HYP	155 335
Het Ypersch nieuws (1930–1971)	NL	HYN	135 976
Le Progrès (1841–1912)	FR	PRG	134 899
Le Propagateur (1819–1871)	FR	PRP	102 102
Het Wekelijks Nieuws (1946–1953)	NL	HWN	96 780
De Halle (1925–1940)	NL	DHA	71 262
Nieuwsblad van Yperen (1872–1912)	NL	NVY	54 816
De Toekomst (1862–1894)	NL	DTO	42 597
Le Sud (1934–1940)	FR	LSU	40 903
Het Weekblad van IJperen (1886–1906)	NL	HWY	37 882
Het Ypersche Volk (1910–1932)	NL	HYV	22 967
De Weergalm (1904–1914)	NL	DWE	19 287
L'Opinion (1863–1873)	FR	LOP	13 976
De Strijd – La Lutte (1894–1899)	NL/FR	LUT	10 914
De Ypersche bode (1927–1928)	NL	DYB	8 040
The Ypres Times (1921–1936)	EN	TYT	5 167
De Kunstbode (1880–1883)	NL	DKU	4 350
De Poperinghenaar (1940)	NL	DPO	3 147
De Dorpsbode van Rousbrugge (1856–1862)	NL	DVR	3 104
Gazette van Yperen en Poperinghe (1957–1961)	NL	GVY	2 096
Tuinklokke (1928–1940)	NL	TUI	1 811
Gazette van Ypre (1857–1858)	NL	GYV	1 732
Liberté (1947)	FR	LIB	580
Het Poperinghenaartje (1915–1918)	NL	HPO	554
De Raadselbode (1901–1909)	NL	RAA	199
De Grensgalm (1895–1904)	NL	GRE	96
Le Courrier d'Ypres (1858–1911)	FR	COU	93
Het Veld (1914)	NL	VEL	93
Le Messager d'Ypres (1890)	FR	MES	49
L'Annonce d'Ypres (1854–1859)	FR	ANN	41
De Handboog (1889)	NL	HAN	40
Burgersbelang (1910)	NL	BUR	28
L'Indicateur (1861)	FR	IND	25
La Commune d'Ypres (1849)	FR	COM	22
La Publicité (1840)	FR	PUB	20
De Herbergier (1901)	NL	HER	19
De Volksvriend (1859)	NL	VOL	17
Den Klappenden Ekster (1850)	NL	KLP	16
De Yperling (1853)	NL	YPE	16
La Vérité (1857)	FR	VER	14
Total			**1 144 516**

Table III.1: Periodical titles by number of XML files

3.1 Structure

The XML files from the *Historische Kranten* corpus can be divided into two categories: page files containing no raw text but only general properties (metadata) on the one hand, and article files with actual news content (data) on the other hand. Since a page from a periodical typically contains several short articles, the second category is necessarily much more populated than the first.

Page files are organised by periodical code, year, month, date and 3-digit page number. For instance, the second page of the 3 October 1859 edition of "L'Annonce d'Ypres" will be found under `ANN-18591003-002.xml`. Most article files are built upon a page filename with the 3-digit article number appended. The fourth article of the page above would therefore be `ANN-18591003-002004.xml`.

However, not all files follow this convention, although no explanation is provided for the discrepancy. A large number of articles have an additional zero in their page number, making it `ANN-18591003-0002004.xml` instead of `ANN-18591003-002004.xml` for instance (making it more difficult to link it to the parent page `ANN-18591003-002.xml`). A few articles (only 24 from two periodicals) have a shorter format with no initial zeros but an extra zero in the article number: `HYP-19381224-20001.xml` for instance.

Moreover, a significant number of page files have a completely different format: `DKU-01_0012-001-18800530-0109.xml` for instance. These unexplained variations in format makes it more complex to process the whole corpus programmatically since several cases have to be taken into account. Table III.2 synthesises the various file formats, whereas Table III.3 shows the breakdown of the number of XML files by periodical into the number of pages and articles (sorted on the latter). As we immediately notice, there are gross disparities between the number of articles by periodical, ranging from 10 to well over 100 000.

# chars	# files	example	type
20	63 531	ANN-18540826-001.xml	page
22	24	HYP-19381224-20001.xml	article
23	482 466	DKU-18800530-001001.xml	article
24	546 065	ANN-18540826-0001003.xml	article
33	52 430	DKU-01_0012-001-18800530-0109.xml	page

Table III.2: Different file formats used

Code	**# files**	**# pages**	**# articles**
JDY	173 451	15 716	157 735
HYP	155 335	17 363	137 972
HYN	135 976	11 006	124 970
PRG	134 899	17 991	116 908
HWN	96 780	4 070	92 710
PRP	102 102	17 425	84 677
DHA	71 262	3 531	67 731
NVY	54 816	3 733	51 083
DTO	42 597	5 442	37 155
LSU	40 903	4 228	36 675
HWY	37 882	3 365	34 517
HYV	22 967	1 581	21 386
DWE	19 287	1 759	17 528
LOP	13 976	2 108	11 868
LUT	10 914	1 086	9 828
DYB	8 040	713	7 327
DKU	4 350	482	3 868
TYT	5 167	2 053	3 114
DPO	3 147	182	2 965
DVR	3 104	658	2 446
GVY	2 096	296	1 800
GYV	1 732	282	1 450
TUI	1 811	443	1 368
LIB	580	64	516
HPO	554	244	310
RAA	199	36	163
GRE	96	16	80
COU	93	16	77
VEL	93	16	77
MES	49	4	45
ANN	41	6	35
HAN	40	8	32
BUR	28	4	24
IND	25	4	21
COM	22	4	18
PUB	20	4	16
HER	19	6	13
VOL	17	4	13
KLP	16	4	12
YPE	16	4	12
VER	14	4	10
Total	**1 144 516**	**115 961**	**1 028 555**

Table III.3: Breakdown of files into pages and articles

The XML structure of a typical article file is as follows:

```
<?xml version="1.0" encoding="iso-8859-1"?>

<clip id= type=>

  <source>
    <pub id=></pub>
    <date year= month= day=/>
    <page number=/>
    <coords>
      <page width= height= swidth="" sheight=""/>
      <coord pageid= btype= order= xl= xr= yb= yt=/>
    </coords>
  </source>

  <content>
    <headers>
      <header/>
    </headers>
    <subheaders/>
    <bylines/>
    <body> </body>
    <pictures number=/>
    <captions/>
  </content>

  <meta>
    <author/>
    <subject/>
    <language> </language>
    <surface height= size= unit=/>
    <operator> </operator>
    <pages> </pages>
    <book> </book>
    <title> </title>
  </meta>

</clip>
```

After the declaration of the XML version and character encoding, each article is encapsulated in a <clip> tag (root element). A clip consists of three sections:

1. source: information on the provenance of the article (paper, issue date, page and geometric position on the page)
2. content: actual article, including headers, pictures and caption (main text being in the <body> and sometimes divided into paragraphs)
3. meta: metadata about the article such as the title (sometimes distinct from the printed header), author (when known) and language

3.2 Linguistic distribution

Given the linguistic context of Belgium on the one hand and the history of city of Ypres/Ieper on the other hand, it should come as no surprise that the corpus is multilingual, with French and Dutch (including Flemish dialects) more or less equivalently represented, plus the appearance of an English newspaper after World War I (during which Ypres was the theatre of three major battles). In the section, we experiment with different methods in order to get a clearer picture of the language distribution across the corpus. Since the multilingual dimension is key to our work, we devoted a fair amount of resources to this task, as it conditions the soundness of our claim to language independence.

3.2.1 Hard-coded language tag

Our first intuition was to look for the `<language></language>` tag in the XML meta section, but it proved not to be reliable at all: with a few exceptions showed in Table III.4, almost all articles (82.6% of the corpus) were labelled "nl" regardless of the actual language used in the article (including those from the English paper TYT), making this piece of information meaningless.

Only three periodicals (accounting for a mere 17.4% of the corpus) are annotated with the "fr" tag denoting French content: JDY (the larger in the collection), LOP, and LUT (despite the latter being overtly bilingual). Other oddities include the isolated appearance of a supposedly English article in HYN,[23] and of 15 of them in HWN. There is even one article titled "Football" (with no body) which is labelled "it", which raises the question of the integrity constraints used for the language tag encoding: if any language could appear, why just in this single pseudo-Italian article? Finally, two articles in NVY have no language declared at all, having just a single empty `<language/>` tag.

[23] Which is in fact in Dutch.

Code	# articles	nl	fr	en	it	n/a
JDY	157 735		157 735			
HYP	137 972	137 971			1	
HYN	124 970	124 969		1		
HWN	92 710	92 695		15		
NVY	51 083	51 081				2
LOP	11 868		11 868			
LUT	9 828		9 828			

Table III.4: Periodicals with declared non-Dutch articles

Openly French titles such as PRG and PRP were considered Dutch-only, while bilingual titles HYP and LUT were deemed monolingual (Dutch- and French-only respectively), and the English title TYT was not taken into account despite the (irrelevant) appearance of the "en" language tag elsewhere in the corpus. All these inconsistencies clearly show that the language metadata cannot be relied on. We therefore looked for alternative ways of assessing the distribution of languages in the corpus.

3.2.2 Periodical titles

In the light of the evidence provided above, the titles of the newspapers seemed a promising track to follow, since the title often reflects the language a paper is written in. Table III.5 shows that Dutch (including its local varieties) accounts for almost two thirds of the titles (25 out of 41). However, it only corresponds to just over 45% of the articles (far from the 82% computed from the XML language tags), while French covers almost 40%. Two papers (HYP and LUT) are overtly bilingual, accounting for just under 15% of the corpus, and another one (TYT) is written in English but covers only 0.3%.

Language	# titles	# articles	Global %
Dutch	25	469 040	45.6%
French	13	408 601	39.7%
Dutch+French (bilingual)	2	147 800	14.4%
English	1	3 114	0.3%

Table III.5: Distribution of languages across periodicals

This summary is, unfortunately, an oversimplification of reality. Most periodicals, despite having either a French or a Dutch title, do not stick to one particular language but rather mix articles in both languages. In practice, articles written in French can be found in Dutch-named periodicals and conversely: ANN contains news pieces in Dutch, while BUR has articles in French.

Additionally, the actual proportion of each language in the bilingual titles could not be determined. An estimation postulating that Dutch and French each account for a half of articles in these publications would lead to the 14.4% being split evenly amongst them, increasing the overall percentages to 52.8% and 46.9% for Dutch and French respectively, with English staying at 0.3%.

3.2.3 Language detection

Since it is impossible to read over one million articles individually in order to detect the language used in each of them, we decided to rely on a language detection algorithm. The best-known algorithm of this kind is the one included in the Google Translate web tool[24], but unfortunately its access through an API has been subjected to a fee since 2011. The task has been the focus of recent research (Zampieri et al., 2014), with state-of-the-art systems achieving over 95% accuracy even when the languages to distinguish between are closely related (Goutte et al., 2014).

We selected the tool `langid.py`[25] (Lui and Baldwin, 2012) which recognises 97 languages, offers good robustness on a range of datasets (from 89% accuracy on very noisy corpora to almost 99% accuracy on the more standardised EuroGOV corpus) and is implemented in Python, which makes it easy to call from within our other Python scripts used to load the corpus and parse the XML files.[26] We needed to evaluate, however, to what extent the low OCR quality of our dataset would not negatively affect the language detection output. For this purpose, we created a random sample of 800 articles representative of the corpus. We then manually annotated each text with its corresponding language, removing undecidable cases (i.e. OCR too poor to even read the text), blanks and mixed languages (title in one language and body in other for instance), leaving us with a gold-standard corpus (GSC) of 779 language-disambiguated texts (466 Dutch, 311 French and 2 English).

[24] `https://translate.google.com/`

[25] `https://github.com/saffsd/langid.py`

[26] Another choice could have been TextCat (Cavnar and Trenkle, 1994), but Derczynski et al. (2015) indicate that `langid.py` outperforms it on French and English, while being only marginally less accurate on Dutch. Moreover, it is implemented in Perl which makes its integration in Python less straightforward.

The `langid.py` utility allows us to restrict the number of possible languages from 97 to a more practical subset, thereby decreasing the risk of error. We first set this parameter to Dutch, French and English with `langid.set_languages(["nl","fr","en"])`, obtaining satisfactory accuracy. However, several errors arose from a confusion between Dutch and English in very short texts. Since English is used in only a small fraction of the corpus, we tried to reduce the allowed languages to just "nl" and "fr", further reducing the error rate.

Tables III.6 and III.7 show the confusion matrices for both attempts, while Table III.8 summarises the results, showing a state-of-the-art accuracy of 97% on our GSC with the binary classifier.

		Gold		
	Lang.	NL	FR	EN
Guess	NL	452	8	0
	FR	5	293	1
	EN	9	10	1

Table III.6: Confusion matrix with three languages: Gold represents the human annotation and Guess the machine's guess (for instance, 8 articles were written in French but wrongly identified as Dutch)

		Gold		
	Lang.	NL	FR	EN
Guess	NL	453	9	1
	FR	13	302	1

Table III.7: Truncated confusion matrix with only two languages: the two English articles in the sample are *de facto* ignored, leading to one more error but dramatically reducing the error rate by not allowing French or Dutch articles to be marked "en"

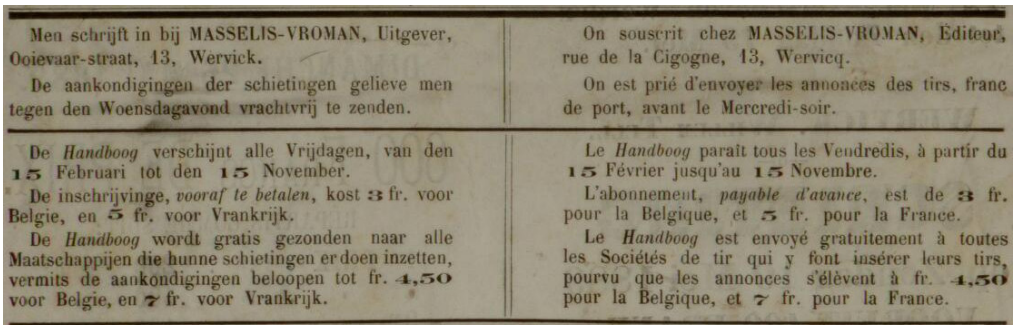
Languages	**# Errors**	**Accuracy**
NL/FR/EN	33	95.76%
NL/FR	24	96.92%

Table III.8: Accuracy scores for language detection

In order not to penalise the small proportion of English articles by limiting the range of choices to only two languages when launching the full analysis of the 1 028 555 articles, a compromise was found: `langid.py` was run with all three languages on TYT[27] and with Dutch and French only on the rest of the corpus. This allowed us to limit the number of English false positives while ensuring maximum accuracy.

Table III.9 shows the estimated distribution of languages across the corpus after the language detection process, corroborating the estimation by periodical titles with a margin of error of 0.4% only: 52.4% instead of 52.8% for Dutch and 47.3% instead of 46.9% for French. The results also confirm that the gap between the number of articles in Dutch and French is only 5%, much less than would have appeared from the 2:1 ratio for periodical titles in Table III.5.

The identification of the language in which each article is written with a degree of certitude of about 97% opens the door for a language-specific approach to NER. However, difficulties remain because some articles are simply inconsistent or deliberately bilingual, as illustrated in Figure III.5. Although uncommon, such odd cases make the case for a language-independent approach to IE, already outlined in chapters I and II. How this additional constraint of language can be dealt with in practice will be detailed in Chapter IV.



Men schrijft in bij MASSELIS-VROMAN, Uitgever, Ooievaar-straat, 13, Wervick.

De aankondigingen der schietingen gelieve men tegen den Woensdagavond vrachtvrij te zenden.

On souscrit chez MASSELIS-VROMAN, Éditeur, rue de la Cigogne, 13, Wervicq.

On est prié d'envoyer les annonces des tirs, franc de port, avant le Mercredi-soir.

De *Handboog* verschijnt alle Vrijdagen, van den 15 Februari tot den 15 November.

De inschrijvinge, *vooraf te betalen*, kost 3 fr. voor Belgie, en 5 fr. voor Vrankrijk.

De *Handboog* wordt gratis gezonden naar alle Maatschappijen die hunne schietingen er doen inzetten, vermits de aankondigingen beloopen tot fr. 4,50 voor Belgie, en 7 fr. voor Vrankrijk.

Le *Handboog* paraît tous les Vendredis, à partir du 15 Février jusqu'au 15 Novembre.

L'abonnement, *payable d'avance*, est de 3 fr. pour la Belgique, et 5 fr. pour la France.

Le *Handboog* est envoyé gratuitement à toutes les Sociétés de tir qui y font insérer leurs tirs, pourvu que les annonces s'élèvent à fr. 4,50 pour la Belgique, et 7 fr. pour la France.

Figure III.5: Bilingual article from De Handboog

After this extensive exploration of the contents, structure, and linguistic distribution of the *Historische Kranten* corpus, we will now focus our attention on the players involved in the digitisation project on the one hand, and on the expectations of the end users taking advantage of the online publication of this material on `http://www.historischekranten.be/` on the other hand.

[27] This was done in order to allow for the possibility of non-English articles in TYT, although in the event no such articles were detected, as shown by the third column of Table III.9.

Code	nl	fr	en
JDY	18 443	139 292	
HYP	88 747	49 225	
HYN	97 734	27 236	
PRG	10 654	106 254	
HWN	89 711	2 999	
PRP	19 582	65 095	
DHA	36 180	31 551	
NVY	44 961	6 122	
DTO	32 324	4 831	
LSU	916	35 759	
HWY	28 579	5 938	
HYV	20 858	528	
DWE	14 048	3 480	
LOP	11 399	469	
LUT	5 992	3 836	
DYB	7 060	267	
DKU	2 396	1 472	
TYT	20	74	3 020
DPO	2 805	160	
DVR	2 023	423	
GVY	1 582	218	
GYV	1 068	382	
TUI	1 355	13	
LIB	4	512	
HPO	61	249	
RAA	128	35	
GRE	56	24	
COU	55	22	
VEL	77		
MES	6	39	
ANN	15	20	
HAN	17	15	
BUR	19	5	
IND	15	6	
COM		18	
PUB		16	
HER	13		
VOL	9	4	
KLP	11	1	
YPE	12		
VER		10	
TOTAL	**538 935**	**486 600**	**3 020**
%	**52.4**	**47.3**	**0.3**

Table III.9: Estimated distribution after language detection

3.3 People and needs

After focusing on numbers in the previous two sections, we should remind ourselves that a semantic enrichment project, like any technological project, is first and foremost about human beings and their expectations rather than technical achievements for their own sake. In this section, we present the different people involved and the outcomes they expected from the project, and continue with an analysis of the search habits of users in order to derive specifications serving as guidelines for the designing of a specific tool in Chapter V.

3.3.1 Stakeholders

In order to satisfy a demand, we need to identify the actual wishes and needs of both the managers of the archive and its users, and ideally to reconcile both. Not to take user views into account presents the risk of being disconnected from reality and designing a system that does not answer a real need in the first place, what Ariès (1986, p. 216) calls "l'insupportable vanité du technicien qui demeure à l'intérieur de sa technique, sans jamais tenter de regarder au dehors". The point of view of the client is not monolithic but rather polyvocal since we have had several interlocutors in this project:

Ypres City Archive: The Ieper Stadsarchief[28] owns the collection and is in charge of its *preservation*. Our contact there was Jochen Vermote, who made us visit the physical archive and drew attention to potential quality problems due to poor conservation or chemical alteration of old newspapers.

Heritage Cell CO7: Erfgoedcel CO7[29] is the service in charge of the collection's *promotion*. Our contact was Liesbeth Thiers until she quit in 2013, then Hilde Cuyt. Access to and support for the Google Analytics account of `http://www.historischekranten.be/` was provided by Renee Mestdagh.

Picturae: The project's website was reassigned in 2014 to *Picturae*[30], now in charge of the online *propagation* and technical maintenance. Picturae used scans originally performed by X-CAGO[31] but did the OCR and indexing over again. Our contact there was Robert Tiessen, who invited us to their headquarters in Heiloo (The Netherlands) in order to discuss implementation.

[28] `http://www.ieper.be/`
[29] `http://erfgoedcelco7.be/`
[30] `https://picturae.com/`
[31] `http://www.x-cago.com/`

3.3.2 Field survey

In the context of Holocaust research, Blanke and Kristel (2013) argue that "research users often have more demands on semantics than is generally provided by archival metadata. For instance, in archival finding aids place names are often only mentioned in free-form narrative text and not especially indicated in controlled access points for places. Researchers would like to search for these locations". But does this observation apply to our particular context?

In order to understand the profiles of users and their search patterns, we collected statistics from the project website over a four-year period from March 2, 2010 to March 2, 2014. Google Analytics[32] (through its `Behaviour > Site Search` submenu) showed that, over this interval, 35 996 users performed a total of 123 984 queries on the website. All in all, a visitor out of two (49.5%) used the site search functionality and users spent four minutes on average exploring the results, visiting 3.86 pages. As shown in Figure III.6, the raw search terms are quite noisy and needed some streamlining before being suitable for analysis. The cleaning was performed with OpenRefine,[33] an open source tool for working with messy data (Verborgh and De Wilde, 2013).

5.	westouter	**171**	(0.14%)
6.	site_searchs	**154**	(0.12%)
7.	reninghelst	**112**	(0.09%)
8.	(Passchendaele)	**99**	(0.08%)
9.	ieper	**91**	(0.07%)
10.	(de)	**81**	(0.07%)
11.	(westouter)	**77**	(0.06%)
12.	(passchendaele)	**74**	(0.06%)
13.	(oorlog)	**68**	(0.05%)
14.	(	**67**	(0.05%)
15.	oorlog	**66**	(0.05%)

Figure III.6: Noisy Google Analytics data

[32] `http://www.google.com/analytics/`
[33] `http://openrefine.org/`

The top twenty-five search terms are displayed in Table III.10, along with the number of hits and categories (either entity type or generic "Concept" for common names). The data show that places are indeed preponderant among the preoccupations of the users: 30 of the 50 most popular search terms (60%) are locations, with only 3 persons (6%) and 2 organisations (4%).

These statistics also show that users do not operate a formal distinction between concrete named entities (proper nouns) such as Ieper and Hitler on the one hand, and more abstract concepts such as "war" (*oorlog*) and "murder" (*moord*) on the other hand: 12 out of 50 top terms (24%) are such concepts (the remaining 6% representing rarer cases such as Titanic).

#	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
11.	Zonnebeke	112	Location
12.	Hollebeke	108	Location
13.	moord	102	Concept
14.	Poperinge	98	Location
15.	Watou	93	Location
16.	Langemark	91	Location
17.	Proven	75	Location
18.	Titanic	75	Vehicle
19.	Abeele	73	Location
20.	Zandvoorde	71	Location
21.	Diksmuide	57	Location
22.	Becelaere	49	Location
23.	Hitler	49	Person
24.	Noordschote	49	Location
25.	1914	47	Time

Table III.10: Top 25 search terms on Historische Kranten

3.3.3 Specifications

Based on discussions with the stakeholders and the results of the survey, we drafted a bill of specifications for the conception of our semantic enrichment and knowledge discovery tool MERCKX, which will be presented in Chapter V. These specifications are summarised below. As emphasised in Section 3.3.1, the various players involved in the *Historische Kranten* project have different concerns, needs, and expectations. These preoccupations can be summed up with three P-words: preservation, promotion, and propagation.

The Ypres City Archive is concerned with preservation, and is therefore keen to have quality digital copies of the original periodicals that encounter the risk of deteriorating over time. While our project does not directly address this need, duplicating the material and reusing it in several forms participates in the effort of long-term conservation, especially in an open data perspective.

The CO7 Heritage Cell is responsible for the promotion of the work and for gaining more symbolic value out of it. Our initiative is therefore clearly in line with their objectives, as confirmed by personal communication with the team stating that they "want to encourage scientific research on the materials in their possession and [...] are very pleased with the cooperation as it exists today". In particular, data visualisation tools and the dissemination of research results to the wider scientific community appeals to the heritage cell.

Picturae takes care of the propagation part, and therefore expects to preserve its image of expertise without needing to invest too much development work in the maintenance of existing tools. A turnkey semantic enrichment solution would thus constitute a real opportunity for them, allowing to enhance the website's functionalities at a low cost with search suggestions and links to external resources, while ensuring maximum benefits for the users.

Building on the lessons from the field survey, the clear preference for places also encourages us to concentrate our efforts more specifically on the extraction of locations: in Chapter V, the system and evaluation will therefore focus on this type of entities, but without sacrificing the generalisability to other types. The observation that common and proper nouns freely intermix corroborates our assertion, formulated in our first research question and already confirmed in Chapter II (Section 3.1.1), that the separation between entities and terms in language processing is somewhat artificial and does not necessarily reflect the interests of the end users of the applications developed.

Summary

In Chapter III, we introduced the difference between formal sciences and the humanities, and the corollary distinction between deterministic and empirical information. We have seen that while the truth value of a logical or mathematical expression can always be determined unambiguously, the same does not hold true for domains that are subject to human experience, in which interpretation plays a decisive rule. In the absence of a formal referent (of which a gold-standard corpus is only an avatar), correctness and quality become void concepts per se and can only be apprehended meaningfully relatively to usage.

The humanities are particularly affected by this relativism, which has prompted some scholars to adopt more rigorous and statistical techniques in an attempt to objectify their practice. The digital humanities, a catch-all branding for all kinds of computational approaches to literary and historical material, is no exception to this tendency. But transforming social sciences into a deterministic object of study is a fallacy and is doomed to fail because of the intrinsic empirical nature of these disciplines. Instead, their inherent subjectivity must be acknowledged and the interpretative dimension of human analysis accepted as a necessary component.

To illustrate this view, we presented an original case study based on the *Historische Kranten* project that involved the digitisation and online publication of over one million historical documents from a multilingual historical archive. This corpus is analysed in depth and many examples are drawn from it in order to illustrate the relevance of our work. In Chapter IV, we will focus on three important dimensions of this collection that can be generalised to similar projects: data quality, multilingualism and the evolution of language and concepts over time.

Chapter IV

Quality, Language, and Time

Outline

In this chapter, we handle three important aspects that can alternatively be considered constraints, threats, or opportunities in the context of semantic enrichment of empirical content: data quality, the variety of language, and its evolution over time.

Section 1 introduces the notion of data quality, which is essential when dealing with uncurated data such as the material present on the Web. Through the prism of Isabelle Boydens' work, we critically assess the quality of the OCR output from cultural heritage projects (and *Historische Kranten* in particular), before evaluating the quality of Linked Open Data resources, mainly focusing on DBpedia which will constitute the backbone of our future system MERCKX.

In Section 2, we investigate the notion of multilingualism and what the handling of multiple languages entails for information extraction techniques. Language-independent methods are reviewed and compared to language-specific ones, with a special emphasis on multilingual corpora and the need to increase portability from one language to another.

Finally, Section 3 raises the issue of language evolution and of its practical implications for a semantic enrichment system. Accounting for changes in the meaning of terms and concepts over time requires a temporal framework, which we borrow from Boydens once again in order to transpose it to our particular case.

The understanding of the three central thematics of quality, language, and time – along with their operational repercussions – will allow us to identify the missing links in our theoretical approach to knowledge discovery before implementing it on a practical level in Chapter V.

Contents

1 Data Quality

The notion of quality is central to any kind of evaluation, but the reality hidden behind this concept can vary significantly from one context to another. While everybody would agree that *good* quality – or at least a good price-quality ratio – is a must for business-grade applications or products, what exactly constitutes this quality is left to the judgement of the customer.

Following Juran (1951), we consider that quality is not an absolute ideal but rather always relative to usage. Juran coined the term *fitness for use* to stress the fact that the meaning of quality depends on the expectations and actual needs of users. The ISO 9000[1] family of quality management systems standards acknowledged this reality by adopting the analogous designation of *fitness for purpose*.

In the 1990s, Redman (1997) applied this conception of quality to information management, laying the foundations of modern data quality. With the growth of Big Data, quality became a pervasive issue and a major stake for companies. In 2001 already, the Data Warehousing Institute[2] estimated that poor data quality was costing US businesses in excess of 600 billion dollars annually. Several works testify of the severity of data quality problems in the industry (Olson, 2003; Batini and Scannapieco, 2006).

Despite the designation of the Total Data Quality Management (TDQM)[3] programme of Richard Wang (1998), the original assessment of Juran remains valid today: total quality does not exist, it is always relative to a number of criteria. This reality was recently reaffirmed in a collection of case studies from the computer science community (McCallum, 2012).

Depending on the context, the needs of the users and the purpose for which the data will be used, one or several components can gain more importance than the others. Quality is therefore not a unipolar measure but rather a complex property exhibiting many conflicting dimensions.

Boydens (1999) went beyond the limitations of the field by drawing the useful distinction between deterministic and empirical data, presented in Chapter III. This allowed her to reformulate the problem of "correct" data, insoluble in an empirical context in the absence of a stable referent, in terms of the permanent construction of data over time and their constant (re)interpretation. The *fitness for use* principle is introduced in Section 1.1 and is then applied to OCR (Section 1.2) and Linked Open Data (Section 1.3).

[1] `http://www.iso.org/iso/iso_9000/`
[2] `https://tdwi.org/`
[3] `http://web.mit.edu/tdqm/`

1.1 Fitness for use

We have seen that the underlying reality covered by the notion of quality is subject to important variations of interpretation depending on the context and the needs of the users. Per se, it is not such a useful concept since it cannot be measured efficiently and unequivocally. Quality therefore needs to be redefined for each purpose with regards to external constraints. In order to turn quality into an operational property, objective quality indicators have to be defined upstream of any evaluation process. These (mostly) quantifiable indicators will then allow to measure concrete quality deficiencies relative to a certain purpose and to implement improvement strategies. Specifying these indicators requires an understanding of the nature of a data element, and more importantly of what is considered a "correct" data element (Boydens, 2011).

As seen in Chapter III, any piece of data can be represented in the form of an *(i,d,v)* triple where i is its unique identifier, d its domain of definition and v its value. Whereas establishing the correctness of a given value does not raise any problem in the case of deterministic data for which a formal model exists, the analogous operation for empirical data is made tricky, in the inexorable absence of a stable referent, because of changes in theory following human interpretation. It entails that the closed-world assumption does not hold for empirical data whose underlying reality is constantly evolving. If there is no absolute reference against which to compare the correctness of a data element, we must deduce that data are not given once and for all but rather progressively constructed over time, as will be discussed in Section 3.

As a consequence, the appropriateness of empirical data to the needs of the users "can be determined only indirectly, via a series of lateral indicators" (Boydens, 2011, p. 121). Most of these indicators of quality are quantifiable (e.g. in terms of precision and recall of documents for a user, see Chapter V for more detailed evaluation metrics) but some useful indicators, such as relevance, are non-quantifiable and therefore subject to interpretation.[4]

Concretely, what constitutes data quality is always an arbitration, a trade-off between several conflicting components only measurable by balancing costs and benefits for the client rather than in absolute terms. This is pleasantly illustrated by the following business mantra: "our services are good, fast and cheap but you can only pick two: if it's good and cheap, it won't be fast; if it's fast and good, it won't be cheap; if it's cheap and fast, it won't be good". A representation of this quality trade-off conception is displayed in Figure IV.1.

[4] Although efforts have been made to reduce the subjectivity involved in the evaluation of NLP systems, see for instance Daelemans and Hoste (2002); van Zaanen and Freeman (2004), and more recently Snow et al. (2008) under this apt title: "Cheap and Fast—But is it Good?".

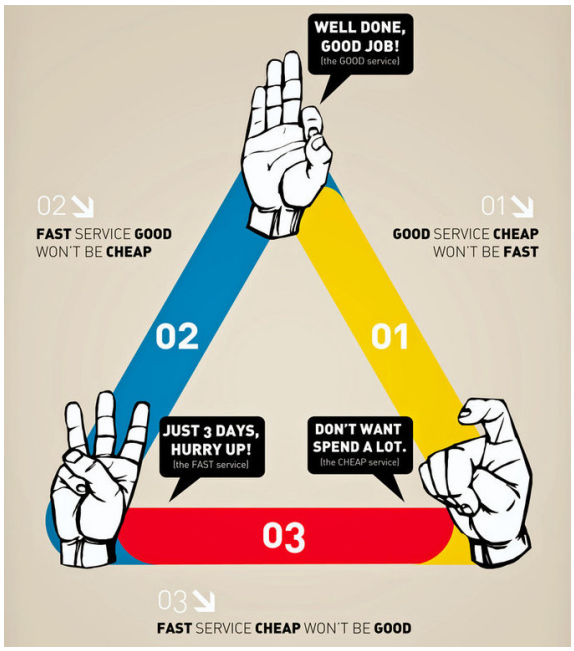


Figure IV.1: Quality trade-off, reproduced from R. Sabatini (CC BY-NC-ND)

Moreover, the notion of quality applied to empirical information is always subject to the "boomerang of reality": a well-tested model can suddenly prove to be outdated due to unforeseen observations that challenge its validity and force the whole representation of reality to be put into question. This dynamic between observed facts and their representation is essential in order to avoid being stuck with a model that is completely disconnected from reality and impervious to new discoveries. This issue is well illustrated by Wiener (1994, p. 37), already mentioned in Chapter III, who explains that before scientific theory accounted for the ozone hole, all NASA measures of extremely low ozone levels were systematically considered *bad data* because to validate them would have been unthinkable at the time: "quand on ne dispose pas d'une source indépendante d'information, il peut s'avérer très dangereux de rejeter un relevé parce qu'il ne correspond pas aux idées préconçues".

In what follows, we will apply this conception of quality to the concrete applications of OCR and Linked Open Data, focusing on the needs of the users of the *Historische Kranten* corpus presented in Chapter III. A particular attention will be paid to the problem of identity within Semantic Web resources, and to the impact of the quality of knowledge bases for operational purposes. In Section 3, the *fitness for use* principle will acquire a new dimension by being extended to language evolution.

1.2 Optical character recognition

The relevance of optical character recognition (OCR) quality for our work is obvious, since the whole *Historische Kranten* corpus has been OCRised prior to its online publication. As emphasised in Chapter III, the past decade has seen a tremendous transformation of the humanities due to the explosion of data, both born-digital and digitised. Scholarly practice has evolved accordingly, with ambitious digitisation projects – such as Google Books and Europeana Newspapers – transforming the relationship researchers entertain with their documents (Jankowski, 2009).

In this context, OCR, the technique allowing to convert scanned images into indexable text, has played a major role for the exploitation of archival collections and other historical material. Commercial solutions like ABBYY FineReader[5] have been shown to achieve over 98% accuracy on clean input text (Holley, 2009), but they are relatively expensive and small cultural institutions cannot always afford to hire the consultants to perform top-grade OCR.

Moreover, these commercial products invariably operate as black boxes that do not allow field workers to understand how the OCR pipeline works, nor to interfere with it before evaluation and correction at post-processing time. This state of affairs creates frustration among collection managers who know their material very well but are unable to offer any useful input during the digitisation process, in between early specifications and the final product. Blanke et al. (2012, p. 660) reflect on this problematic situation:

> “ This use of “black box” tools, and even more so the outsourcing of OCR processing, leads to a skills and knowledge gap among researchers and archives staff involved in digitisation, which results in a failure to appreciate the problems and opportunities that OCR approaches offer the scholarly community. ”

The alternative is in-house digitisation, which allows for greater control on the output, but often at the expense of quality (Alex et al., 2012). In the context of the *Historische Kranten*, the quality of the OCR performed by X-CAGO is very uneven and not quite state-of-the-art, as shown by Figure IV.2 and its transcript below.[6]

[5] `http://www.abbyy.com/finereader/`

[6] A second OCR was performed recently by the company Picturae with ABBYY FineReader 10, yielding better results, but we could not access the new files in time to establish a thorough comparison in this dissertation. Ideally, the implementation phase described in Chapter V will have to take the later OCR into account, although the extraction scores will not necessarily be dramatically affected, as explained below.


Dimanche, 19 Octobre 1890. Quatrième année. N° 42.

DISTRIBUTION GRATUITE.

TIRAGE A 1000 EXEMPLAIRES.

LE MESSAGER D'YPRES,

Journal d'Annonces des Notaires, Négociants, Maisons de Commerce, etc.

Ce JOURNAL paraît le Samedi soir, il est distribué et affiché dans la ville d'Ypres, envoyé dans les Maisons Communales et principaux estaminets de la province de la Flandre Occidentale. Il est, en outre, expédié à tous les Notaires, Huissiers et Agents d'affaires de la dite province, et partout où dans l'intérêt de la publicité des Annonces, sa distribution pourrait offrir une utilité particulière.

Le prix d'insertion est de DIX centimes la ligne. — Les affiches, circulaires et annonces imprimées au Bureau de ce Journal reçoivent une insertion GRATIS.

On traite à forfait pour les Annonces Commerciales pour trois mois, six mois et toute une année à des conditions très-avantageuses. — Tout ce qui concerne le MESSAGER doit y être adressé FRANCO.

Le Bureau est établi chez *LAMBIN-MATHÉE*, Imprimeur-Editeur, rue au Beurre, 20, Ypres.

Figure IV.2: OCRised article from the Messager d'Ypres

Dimanche,**M**9 Oc**l**obre 1890.
Quatrième année.
D'YPRES, Journal d'Annonces des Notaires, Négociants, Maisons de Commerce, etc.
Ce JOURNAL paraît le Samedi soir, il est distribué et affich**e** dans la ville d'Ypres, envoyé dans les Maisons Communales et principaux estaminets de la province de la Flandre Occidentale. **1**1 est, en outre, expédié a tous les Notaires, Huissiers et Agents d'affaires de la dite pr**b**vince, et partout o**ü** dans l'**intérêtdela** publicité des Annonces, sa distribution pourrait offrir une utilité particuli**e**re. Le prix d'insertion est de **IMX** centimes la ligne. **—** Les affiches, circulaires et annonces impri**ra**ées au Bureau de ce Journal re**g**oivent une insertion gratis. On traite **a** forfait pour les Annonces Commerciales pour trois mois, six mois et toute une année **a** des conditions très-avantageuses. **—** Tout ce qui concerne le Messager doit y **è**tre adressé franco. Le Bureau est établi chez **AAifiJ/iV-J/A r/f/sT?**, i**ni**primeur-Editeur, rue au Beurre, 20, Ypres.

While a quantitative analysis of a larger sample will be performed later on in Chapter V in order to evaluate the impact of OCR quality on entity linking, a qualitative assessment of the types of errors is equally relevant to understand the issues affecting OCR output.

Different quality problems can indeed be identified in this article:

- undetected chunks of text: vertical stamps but also mentions such as "N° 42." and the main heading "LE MESSAGER" (probably in too big a font to be handled)
- confusion involving digits ("M9" instead of 19, "11" instead of Il)
- problems with special character (long dash – converted into HTML entity "—")
- issues with alternative typographies ("IMX" instead of DIX in bold, "AAifiJ/iV-J/A r/f/sT?" instead of LAMBIN-MATHÉE in italics)
- incorrect diacritics ("oü" instead of où, "ètre" instead of être)
- various misrepresented characters ("prbvince" instead of province, "iniprimeur" instead of imprimeur)

In the late nineteen nineties already, Palmer and Day (1997) wondered "how the existing high-scoring [NER] systems would perform on less well-behaved texts, such as single-case texts, non-newswire texts, or texts obtained via optical character recognition". In this section, we explore this question in order to formulate hypotheses that will then be tested by the evaluation process in Chapter V.

By testing four NER tools on raw OCR data, Rodriquez et al. (2012) made the counterintuitive discovery that "manual correction of OCR output does not significantly improve the performance of named-entity recognition". This could be due to very good OCR quality from the start, but the authors also note that they used an open source OCR system, leading to a transcript quality that "would be considered quite low for human readers, but even when uncorrected can offer value for search indexing and other machine processing purposes".

Similarly, Tanner et al. (2009) note that the need for high-quality OCR is context-dependent: most search applications are able to retrieve text strings containing errors with the help of fuzzy matching algorithms, while human users seldom have access to the actual OCR output but will rather read from the original images. Thus, a suboptimal OCR performance score is not necessarily related to poor extraction results. Moreover, commercial OCR is not necessarily best-suited for cultural heritage applications and collections of historical documents in general, principally due to its lack of easy customisation. Blanke et al. (2012, p. 660) show that open source OCR can achieve a higher degree of flexibility on historical data (emphasis added):

> “ There are key differences between large-scale digitisation efforts (as attempted by Google or in the context of the Europeana initiative) and those that have concentrated more on historical material with a specific focus on humanities research. Firstly, there is the question of resources. Most research funding is project-based and notoriously limited, making it less likely that additional resources are available for buying in OCR expertise. More importantly, however, the material produced by humanities digitisation projects is known to be the *result of interpretation*. This means that there may be a need to revisit aspects of the digitisation process from time to time when new discourses about the source material emerge. ”

This interpretable nature of historical/empirical material brings us back to the practice of data quality set forth by Boydens. The evolving nature of digitisation is also emphasised, an issue that will be dealt with in depth in Section 3.

1.3 Linked Open Data

Linked Open Data (LOD) are key in the tasks of semantic enrichment and entity linking, on which we will build our methodology in Chapter V. Whereas the LOD trend has made huge datasets available online, the question of the quality of these data largely remains unaddressed. As we have exposed before, this quality can only be understood in terms of *fitness for use*: do the data live up to the standards of people actually using them? This open question is summed up humorously by the Pedantic Web Group (Hogan et al., 2010) in these terms:

> “ You publish RDF data on the Web, and thereby contribute to our shared passion: the emerging global information space that we call the Web of data. Thank you for that! Thank you for sharing your data!
>
> But your data is broken. Syntax errors, unescaped characters, encoding problems, broken links, ambiguous identifiers, undefined vocabulary terms, mismatched semantics, unintended inferences: if you publish anything on the Web of data, chances are that there is some problem.[7] ”

These shortcomings are inherent to the nature of Open Data: most of the information present on the Web is not curated and therefore seldom suitable for use in strategic contexts involving high human or monetary stakes, such as

[7] From `https://www.smalsresearch.be/linked-open-data-quality-around-the-clock/`

medicine, defence, law, or finance. In these domains, decision-making processes need to rely on high-confidence data, necessarily requiring the ad hoc creation of domain-specific resources like ontologies. LOD can nevertheless be useful in other domains where optimal reliability is less crucial, such as those described in Chapter III. As noted by Tylenda et al. (2014), "facts extracted through IE are not completely error-free; errors may result either from incorrect statements at the source or be induced by the extraction process".

The first two instances of the Workshop on Linked Data Quality[8] showed an increased interest from Linked Data researchers for issues of quality. In the remainder of this section, we will focus on the classical problem of identity between resources, and then on the specific case of DBpedia.

1.3.1 owl:sameAs and identity

According to the official definition,[9] "an `owl:sameAs` statement indicates that two URI references actually refer to the same thing: the individuals have the same 'identity'". For simple cases, the meaning of "identity" is quite obvious. We can, for instance, state that Charles Buls and Karel Buls are one and the same person:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:dbr="http://dbpedia.org/resource/"
         xmlns:dbrnl="http://nl.dbpedia.org/resource/">
  <rdf:Description rdf:about="dbr:Charles_Buls">
    <owl:sameAs rdf:resource="dbrnl:Karel_Buls"/>
  </rdf:Description>
</rdf:RDF>
```

In practice, however, very few cases are as straightforward. The resource for Brussels, for instance, is well represented in LOD datasets. Its DBpedia page (`http://dbpedia.org/page/Brussels`) lists 128 equivalent resources in various KBs, from Freebase to LinkedGeoData and to Global Administrative Areas (GADM). Among those, we can find two URIs from GeoNames: `http://sws.geonames.org/2800866/` and `http://sws.geonames.org/2800867/`, the former corresponding to Brussels as the capital of a political entity and the latter to Brussels Capital as a first-order administrative division. The first has a population of 1 019 022, the second of 1 830 000. So how can these two URIs *actually refer to the same thing*?

[8] `http://ldq.semanticmultimedia.org/`

[9] `http://www.w3.org/TR/owl-ref/#sameAs-def`

As seen in Chapter II, Halpin et al. (2010) noted that the boundary between *sameness* and *similarity* is sometimes porous, to the extent that the misuses of `owl:sameAs` outnumber its correct uses on the Web of Data. Although the meaning of this property seems intuitive enough for most Linked Data users, their intuitions "almost always violate the rather strict logical semantics of identity demanded by *owl:sameAs* as officially defined". The authors trace back the concept of identity to Leibniz's Law, also referred to as the identity of indiscernibles, which states that if two objects have all the same properties, then they are identical. This Law can be formalised logically as IV.1:

$$\forall x \forall y [\forall P(Px \leftrightarrow Py) \rightarrow x = y] \tag{IV.1}$$

The reverse (indiscernibility of identicals) is also true: if two objects are identical, then they share the same properties (IV.2).

$$\forall x \forall y [x = y \rightarrow \forall P(Px \leftrightarrow Py)] \tag{IV.2}$$

However, this definition is clearly too restrictive to be applied to empirical sciences as defined in Chapter III. Whereas first-order logic deals with stable, ethereal concepts, their transposition to mundane reality is conditioned by temporal context: is Karel Buls the adult the `sameAs` Karel Buls the child? Was *Brosella* in the Middle Ages the `sameAs` *Brussels* today? For Elias (1996, p. 53), "l'identité n'est pas tant celle d'une substance que celle de la continuité des transformations conduisant d'un stade au suivant".

Since we cannot enumerate all the properties of an object, we must abstract a subset of them. Identity based on property matching is thus necessarily under-specified. Even so, we cannot assume that different references across the LOD cloud will retain the same properties for a given object. As stated by Halpin et al. (2010), "the problems with *sameAs* start when we apply the principle of substitution to it, by inferring from a *sameAs* assertion that its subject and object share all the same properties".

In order to avoid transforming the Web of Data into "the semantic equivalent of mushy peas", the authors recommend to provide additional documentation and finer-grained properties to distinguish between a variety of nuances that are currently loosely covered by `owl:sameAs` alone: different senses with the same reference (Frege, 1960), matching,[10] similarity, and relatedness.

[10] As in `skos:exactMatch` which "indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications" (Miles and Bechhofer, 2009). Note how this definition is far more oriented toward operationality.

1.3.2 Quality of DBpedia

Using crowdsourcing, Zaveri et al. (2013) evaluated that about 12% of DBpedia triples have some quality problems, ranging from broken links to irrelevant information and objects incorrectly extracted from Wikipedia (see Table IV.1 for their full typology of 16 quality issues sub-categories). A common issue is related to the use of parentheses in Wikipedia, and hence DBpedia, to disambiguate concepts. Leal et al. (2012) illustrate this with an example: "`"Queen (band)"@en` is a different concept from `"Queen"@en`, but in a music setting the term in brackets is not only irrelevant but would disable the identification with the term `"Queen"` when referring to the actual band". These parentheses thus need to be removed in some contexts in order to improve the matching with mentions in text, but at the cost of an unavoidable loss in precision.

Wrong DBpedia properties can also create havoc in LOD-based semantic extraction systems. Let us consider the following example. When specifically restricting retrieved entities to places (`dbo:Place` property from the DBpedia ontology), one can notice that "January" still pops up as a valid entity. A quick lookup of `https://en.wikipedia.org/wiki/January_(disambiguation)` shows that January can be a first name, a surname, a song or a book as well as the first month of the year, but not a single toponym is mentioned, which makes it even more puzzling.

A closer look at `http://dbpedia.org/page/January` shows that January is indeed tagged with type `dbo:Place` (as well as with equivalent properties such as `http://schema.org/Place`, `d0:Location` and `dbo:Wikidata:Q532`), but the origin of the error remains unclear. Further investigation reveals that January is also considered a `dbo:part` of Dubravica, Zagreb County, Croatia. How this came to happen is open to questioning since `https://en.wikipedia.org/wiki/Dubravica,_Zagreb_County` does list ten settlements of this municipality, but nothing remotely close to "January".

Such property errors are not infrequent in DBpedia and other open KBs, and they introduce noise in the set of retrieved entities. But it does not entail that LOD are unusable in all cases: quality, as we have seen, is always defined relatively to a given task, according to the *fitness for use* principle. Zaveri et al. (2013) give a concrete illustration of this fact:

> "In the case of DBpedia, for example, the data quality is perfectly sufficient for enriching Web search with facts or suggestions about common sense information [...] For developing a medical application, on the other hand, the quality of DBpedia is probably completely insufficient."

Dimension	Category	Sub-category
Accuracy	Triple incorrectly extracted	Object value is incompletely extracted
		Special template not properly recognised
	Datatype problems	Datatype incorrectly extracted
	Implicit relationship between attributes	One fact encoded in several attributes
		Several facts encoded in one attribute
		Attribute value computed from another attribute value
Relevancy	Irrelevant information extracted	Extraction of attributes containing layout information
		Redundant attribute values
		Image related information
		Other irrelevant information
Consistency	Representation of number values	Inconsistency in representation of number values
Interlinking	External links	External websites
	Interlinks with other datasets	Links to Wikimedia
		Links to Freebase
		Links to Geospecies
		Links generated via Flickr wrapper

Table IV.1: DBpedia quality issues, adapted from Zaveri et al. (2013)

Since our task in this thesis (i.e. the semantic enrichment of a multilingual archive) is closer to the former task than to the latter, we consider DBpedia to be reasonably accurate for our means, while keeping in mind that quality issues can always occur. Moreover, it should be noted that a manually curated knowledge base (such as the late Freebase) would not necessarily yield better results, as shown by the experiments of Mooney and Nahm (2003):

> “ [...] an automatically extracted database will inevitably contain significant numbers of errors. An important question is whether the knowledge discovered from this “noisy” database is significantly less reliable than knowledge discovered from a cleaner database. This paper presents experiments showing that rules discovered from an automatically extracted database are close in accuracy to that discovered from a manually constructed database. ”

Another problem, which is not explicitly taken into account by Zaveri et al. (2013), is the *incompleteness* of KBs in general, and DBpedia in particular. When trying to compute semantic relatedness between musical concepts, Leal et al. (2012) are hindered by the fact that the ontology they extracted from DBpedia does not satisfactorily cover their domain of application. Augenstein et al. (2014) also observe that even the largest openly-accessible KBs, such as Freebase and Wikidata, are far from complete.

The Agile Knowledge Engineering and Semantic Web (AKSW) research group of the University of Leipzig pursue the work of Zaveri et al. (2013) with their DBpediaDQ project[11] which aims to monitor and improve the quality of DBpedia resources in a user-driven perspective, involving an evaluation campaign of quality issues.[12] Knuth (2014) also evaluates the quality of DBpedia, with a focus on change management.

In a recent talk,[13] Laura Hollink emphasised two related implications for the use of Linked Data: credibility (who created it and how?)[14] and frequency of update. She also underlined the fact that compared to other data sources, “the need for dataset evaluation is exacerbated when using linked data”. Maintaining a level of quality sufficient to meet user needs remains a key challenge for DBpedia and other LOD sources, while keeping a balance with coverage and automation is also necessary. As stressed by Fafalios et al. (2015), “both the number of the URIs that match an entity name and their quality (in terms of relevance) highly depend on the KBs that we exploit”.

[11] `http://aksw.org/Projects/DBpediaDQ.html`

[12] `http://nl.dbpedia.org:8080/TripleCheckMate/`

[13] `http://www.slideshare.net/LauraHollink/to-e-dhbenelux2015`

[14] Echoing the question on provenance evoked in Chapter III (Hartig and Zhao, 2010).

2 Multilingualism

"Without a bridging of the language gap,
Archimedes is just another naked Greek man shouting in his bathroom."
John Gallagher[15]

The quotation of Hélène Monsacré opening this dissertation, written as a foreword to Wismann (2012), argues that the thought trapped between languages is not condemned to double slavery but rather finds a space for extra freedom in the constant confrontation of these languages. Later in his book, Wismann (2012, p. 103) writes in the same vein that "l'espace que l'on peut voir naître entre les langues n'est pas prioritairement une ligne de transmission où communiqueraient les traditions, mais parfois un lieu étrange, où la confrontation de deux langues en engendre une troisième, irréductible : un espace de recréation". The purpose of this section is to investigate this "strange place" at the crossroads of languages in order to offer working solutions towards a bridging of the language gap.

As we have seen in Chapter III, the *Historische Kranten* corpus is more or less evenly distributed between French and Dutch titles, with English accounting for a small part. A common approach to such a corpus could be to treat the three languages separately, with ad hoc linguistic resources and parameters for each part. Indeed, several semantic enrichment tools that will be presented in Chapter V adopt this point of view.

We would like, however, to reflect on the relative necessity and opportunity of such a treatment. If, like we are keen to believe after Chomsky, all languages share a common structure that transcend individual particulars,[16] then to what extent can natural language processing exploit this common structure to create completely language-independent tools that are effortlessly portable from one language to the other?

The question we ask is not trivial and has important operational implications, in terms of costs and benefits for the development of new tools and semantic resources, but also more generally for the very perception of the discipline. Indeed, taking the argument a step further, we could foster a debate about the scope of NLP: does it aim to process natural language in general (French *langage*) or individual natural languages (French *langues*)? The nuance is not mere wordplay, as we would like to demonstrate here.

[15] The Guardian Weekly, 17–23 April 2015, p. 36.

[16] The theory of Universal Grammar, referenced in our introduction, recently gained broad empirical support from large-scale evidence of dependency length minimization in 37 languages (Futrell et al., 2015).

This section is structured to enlighten the content of the first three chapters with multilingual input. Section 2.1 looks at information extraction from the angle of language-independence, while Section 2.2 investigates the multilingual Semantic Web and Section 2.3 considers the specificities of a multilingual corpus, comparing the *Historische Kranten* with other such corpora.

2.1 Language-independent information extraction

In contrast to mainstream language specialisation, several initiatives have explicitly favoured the language-independent approach defended in this dissertation. Bender (2011) offers a thorough survey of the language-independence trend in NLP, with a special focus on the evaluation of such systems, and recommends to enrich machine learning algorithms with input from linguistic typology. According to her, the advantages of language independence are manifold (Bender, 2011):

> “ Truly language-independent NLP technology would be very valuable from both practical and scientific perspectives. From a practical perspective, it would enable more cost-efficient creation of NLP technology across many different language markets as well as more time-efficient creation of applications [...]. In addition, language independence means that technology is more likely to be deployed for languages that have less economic clout. NLP technology for so-called low-density languages also has scientific interest [...]. Finally, in the ideal scenario, language-independent NLP systems can teach us something about the nature of human language, and what human languages share in common. ”

But then, Bender asks, which languages? She lists “All currently spoken human languages” (about 7 000) and “All languages with established writing systems” (about half of that)[17] as reasonable answers, while noting that in practice language independence is often understood on a subset thereof, without this being explicitly specified.

Indeed, many studies implicitly assume that their methods tried on a given language could as well apply to other ones. While this may be true, Bender argues, “it is not the case that ‘apply’ entails ‘work’. [...] And if it doesn’t work reasonably well, then it is not truly language independent”. To make a strong claim of independence, systems then have to perform equally (or almost equally) well on every language.

[17] `http://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten` (accessed on May 18, 2015).

Even when no language-specific information (such as syntactic rules or stopwords) is intentionally included in a system, some features may be making unconscious assumptions about the structure of language because of the linguistic background of its makers. In fact, "languages show variation beyond that which one might imagine looking only at a few familiar (and possibly closely related) languages" (Bender, 2011). As a matter of fact, the range of languages covered by NLP research is much narrower than could be expected and not at all representative of world language diversity. In a survey of over 200 papers from the ACL and EACL (European chapter of the ACL), Bender found that a mere 8% of language families were effectively represented, with articles focusing on English only accounting for over half of the scientific production.

Even more disturbingly, a large proportion of these papers neglected to declare the studied language altogether, thereby assuming universality. All of them but one were about English. "The lack of specification of the language of study seems to follow from a sense that English is a suitably representative language, combined with the idea that any methodology not explicitly coded to include language-specific knowledge must therefore be language-independent" (Bender, 2011).

Leaving the language unspecified may seem an innocuous omission at first glance, but Bender argues convincingly that it gives the incriminated studies an (undeserved) aura of language-independence. Linguistic variation is not infinite, but it is still much greater that can be guessed from the knowledge of a few European languages.[18] Building truly language-independent tools therefore requires to make informed decisions based on an awareness of the full variety of human languages.

Finally, we should distinguish between three degrees of language independence. *Multilingual* tools simply work on two or more languages, possibly with distinct instructions for each of them. *Cross-lingual* applications go a step further by postulating some common features between the language covered, and a crossing of their linguistic borders. In machine translation, this can include the resort to an *interlanguage*. Real *language-independent* systems make the stronger claim by assuming an universal grammar which allows them (in theory at least) to perform equally well on any human language.

In the remainder of this section, we first focus on the task of named-entity recognition, for which the idea of multilingualism has already been tested for some time due to the relatively language-neutral morphology of proper nouns, before tackling cross-lingual relation detection and event extraction.

[18] Browsing the World Atlas of Language Structures (`http://wals.info/`) can be a humbling experience in this respect. For an application to directive speech acts, see De Wilde (2009).

2.1.1 Multilingual NER

Whereas very competitive NER results have been achieved on English data since the nineties, the topic of language-independent NER has only been attracting the attention of researchers recently. Noteworthy exceptions include the Multilingual Entity Task (MET)[19] (Merchant et al., 1996) and the double (2002 and 2003) Computational Natural Language Learning (CoNLL)[20] shared task which was dedicated to that very subject[21] (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

The adaptation of English systems to other languages, though, is not a straightforward task and can be complicated by occurrences of previously unambiguous entities suddenly displaying ambiguity in a multilingual context, as shown by the following examples:

Example 23. Avant, je travaillais à l'UCL.

Example 24. Dieser Mann hat sich viele Namen gegeben.

In Example 23, an English-only NER tool could mistakenly recognise the first word (French for "before") as the town of Avant, Oklahoma (or the American rapper Avant, see Figure IV.3), while the acronym would presumably be expanded to University College London rather than Université catholique de Louvain out of context. Similarly, capitalised common nouns in Example 24 could wrongly be interpreted as named entities by a system unaware of this German peculiarity: *Mann* (German for "man") could easily be mistaken as writer Thomas Mann or director Michael Mann, while *Namen* (German for "names") could be identified as the Dutch name for the Belgian city of Namur.

Efforts are being made on the part of government bodies to encourage the development of multilingual IE systems in general, and NER systems in particular. Lonsdale et al. (2010) list NIST TREC[22] in the U.S., CLEF[23] in Europe and NTCIR[24] in Japan as initiatives fostering research and evaluation of new, language-independent information extraction systems.

[19] `http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/met.htm`

[20] `http://ifarm.nl/signll/conll/`

[21] The 2002 shared task focused on Spanish and Dutch, while the 2003 shared task was concerned with English and German. See `http://www.clips.uantwerpen.be/conll2002/ner/` and `http://www.clips.uantwerpen.be/conll2003/ner/` for more details about the task involved and the results obtained. It should be noted that very little research has been conducted on French NER in a cross-lingual context.

[22] `http://trec.nist.gov/`

[23] `http://www.clef-initiative.eu/`

[24] `http://ntcir.nii.ac.jp/`

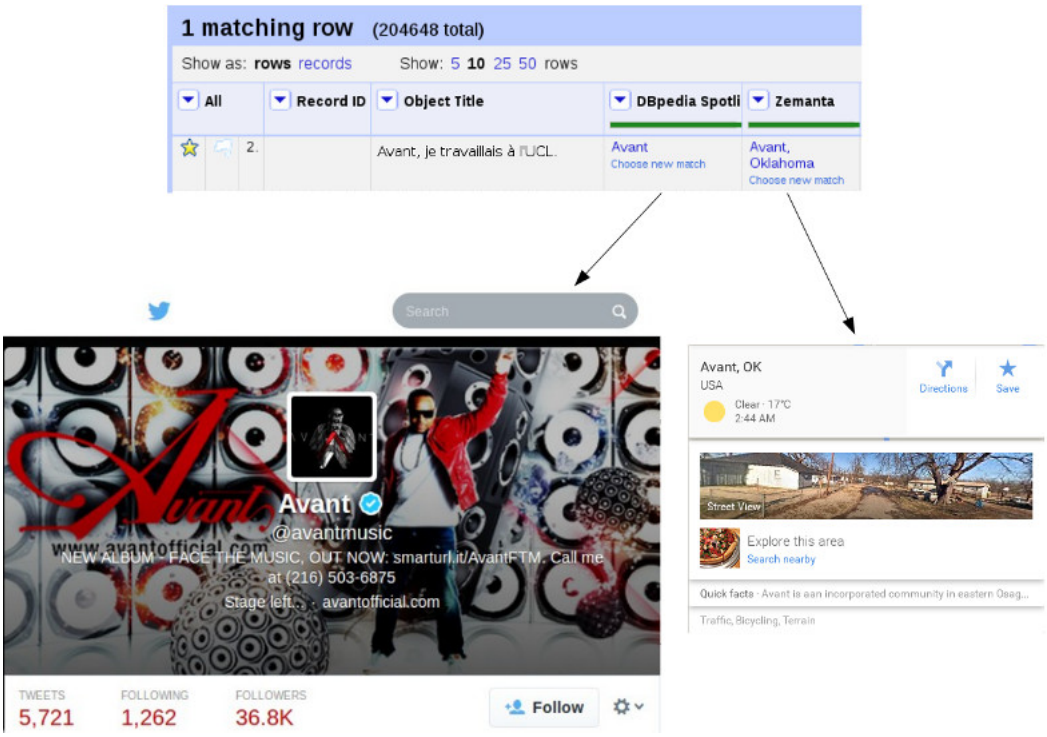


Figure IV.3: Problematic NER due to multilingual ambiguity

In some cases, achieving English-grade results in other languages only requires a limited amount of adaptation, e.g. the addition of a few language-specific rules. Richman and Schone (2008), for instance, exploit the (asymmetric) multilingual structure of Wikipedia to extend NER coverage to other languages including French, Polish, Portuguese, Russian, Spanish and Ukrainian, bootstrapping[25] their system with English data.

Nonetheless, as Palmer and Day (1997) pointed out a decade earlier, "it is not [always] clear what resources are required to adapt systems to new languages". As a matter of fact, systems that were conceived with only English in mind and achieved state-of-the-art performance on English text have been shown to score rather poorly on other languages, including closely related Germanic languages, even after some amount of tedious adaptation work.[26]

[25] Bootstrapping is a machine learning technique consisting in the iterative improvement of a system's performance by feeding it initial data and letting it discover further relevant material.

[26] The system of Chieu and Ng (2003), for instance, which participated in the CoNLL-2003 shared task, ranked amongst the best on English data by achieving over 88% F-score, while it lagged behind on German data with barely 65% F-score, although the authors experimented with lots of local and global features for their maximum entropy tagger.

Overcoming this bias entails planning the conception of a system in a language-independent way right from the start. In the words of Hallot (2005), "the multilingual dimension of applications implies some data structuring needs that should be directly tackled since the inception of the application". Concretely, this requires some basic linguistic knowledge, or at the very least, as Bender (2011) humorously remarks, to know some helpful linguists.

With the emergence of refined machine learning algorithms, statistical NER systems have increasingly been supplanting rule-based techniques, achieving remarkable results on multiple languages, although not yet rivalling human performance. In order to be successful, these systems need to be designed from the start to be maximally language-independent, as noted by Cucerzan and Yarowsky (1999).

De Meulder and Daelemans (2003) experiment with English and German NER, using a memory-based learner (a type of supervised machine learning)[27] and unannotated data (also see Hendrickx and Van Den Bosch (2003)). They note that while gazetteers are not directly beneficial for English data, they lead to substantial improvement on German, a language that is reputed resistant to traditional NER techniques due to its lack of discrimination between common and proper nouns, all substantives being capitalised. This general idea will be exploited in Chapter V by using dictionaries of labels extracted from KBs.

When working with languages such as Russian, Chinese or Arabic that do not make use of Latin script, the handling of a Unicode character encoding such as UTF-8 has to be supported. Failure to do so may cause unexpected results or even a complete break-up of algorithms. Erroneous transliteration can also result in inaccurate disambiguation of acronyms:

Example 25. Генсек НАТО осудил российскую гуманитарную миссию[28]

While НАТО in fact is the Russian spelling for NATO, a literal reading could seem to refer to the Highways Agency Traffic Officers for instance, altering the meaning of the sentence. A Unicode-aware system would be able to avoid this mistake,[29] which is confusing even for humans (English speakers generally pronounce CCCP literally although it should be transliterated to SSSR). This quick round-up amply illustrates the various types of problems that can arise when dealing with named entities in a multilingual context, even though they are commonly considered fairly language-independent textual units.

[27] In the words of Daelemans et al. (1999), "the unique property of memory-based approaches which sets them apart from other learning methods is the fact that they are *lazy learners*: they keep all training data available for extrapolation" (italics theirs).

[28] NATO Secretary General condemned Russian humanitarian mission.

[29] The Latin capital H corresponding to U+0048, while the Cyrillic capital "En" is rendered as U+041D.

2.1.2 Other cross-lingual applications

While most of the work in language-independent natural language processing has been focused on named-entity recognition due to the relatively stable lexicalisation of named entities from one language to another, some studies have also experimented with more ambitious claims to language independence involving common semantic roles or even syntactic structures.

Mehdad et al. (2010), for instance, introduce the task of cross-lingual textual entailment (CLTE) as a semantic relation between two text portions in different languages. Given a text T written in English and a hypothesis H in French, for instance, it can be said that T entails H if a human being would be able to infer that H after reading T:

Example 26. T: Art Nouveau architect Victor Horta is a native of Ghent.
H: Horta est né à Gand

CLTE was subsequently adopted as a task in the 2012 SemEval campaign. The added value of integrating textual entailment components into NLP systems has been widely acknowledged, including for information extraction (Romano et al., 2006).

Cross-lingual approaches were also developed for other information extraction tasks such as relation detection (Kim et al., 2010) and event mining (Vossen et al., 2012). Kim et al. (2010) propose a method that "leverages parallel corpora to bootstrap a relation detector without significant annotation efforts for a resource-poor language". This allows them to achieve significant results for detecting relations in the Korean language, without adding much language-specific components since their system also works for English.

Using a methodology closer to our own, Vossen et al. (2012) propose a concept-based event mining system relying on parallel wordnets as a shared knowledge interface for multiple languages.[30] With a single semantic model and using cross-wordnet links, the authors report transferring their English event-mining patterns to Dutch in less than a day's work – replacing English prepositions by Dutch equivalents, adapting the word order, etc. – while keeping the semantic backbone.

Such complex tasks, however, obviously involve more ingenuity than lower-level ones such as NER. Taking the intricacies of individual languages into account can only be performed at a cost for the level of language-independence, and an arbitration between the complexity of the model and the degree of language-specialisation will therefore become necessary.

[30] See `http://compling.hss.ntu.edu.sg/omw/` for a complete list of available avatars based on the original WordNet.

2.2 The Semantic Web in other languages

Linking online datasets in multiple languages requires to make their semantics explicit, which can be achieved through the use of linguistic metadata. For unstructured documents, however, this is a task that is far from obvious. As Zhang et al. (2015) remark, "many multilingual documents on the Web have no complete metadata. Consequently, text analysis based on natural language processing becomes the alternative way to establish the semantic links based on the full texts of documents".

Recently, Buitelaar and Cimiano (2014) edited a comprehensive volume dedicated to the multilingual Semantic Web, with contributions focusing on a large variety of topics ranging from resource and terminology management to Linked Data publishing, ontology engineering and complex applications such as the interoperability of multilingual Web services. Much remains to be done, however, in order to build a truly universal Web of Knowledge, as emphasised for instance by the fourth chapter on under-resourced South African languages. Specifically, the claim to multilingual comprehensiveness of Wikipedia and related knowledge bases deserves a closer look.

In many tech-related domains, English resources predate the apparition of multilingual ones, as we have already seen for entity extraction for instance. Knowledge bases are no exception and the articulation of (often more comprehensive) English-only resources with newer, more modest ones in various languages – complicated by the potential absence of a one-to-one correspondence between concepts from a language to another – raises issues in terms of coherence and uniqueness. Cabrio et al. (2014, p. 137) sum up this situation:

> "In order to publish information extracted from language-specific pages of Wikipedia in a structured way, the Semantic Web community has started an effort of internationalization of DBpedia. Language-specific DBpedia chapters can contain very different information from one language to another; in particular, they provide more details on certain topics or fill information gaps. Language-specific DBpedia chapters are well connected through instance interlinking, extracted from Wikipedia. An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms in Wikipedia to a common ontology, enabling the exploitation of [...] language-specific DBpedia chapters. However, the mapping process is currently incomplete, it is time-consuming as it is performed manually, and it may lead to the introduction of redundant terms in the ontology."

The state of affairs can vary a lot from one KB to the other. DBpedia mimics its structure on Wikipedia and thus offers a complete semantic network for each language, which somewhat contradicts its ambition to map the sum of knowledge in a single graph-based structure (see Chapter II). The concept of the planet Earth, for instance, can alternatively be referred to as `http://dbpedia.org/resource/Earth` or as `http://fr.dbpedia.org/resource/Terre` or even `http://ru.dbpedia.org/resource/Земля`.

Since each language chapter[31] has its own properties, all speakers are not equal before the access to knowledge. Admittedly, the multiple URIs identifying a resource in various languages are well interlinked via the `owl:sameAs` property, but we have seen the limits of this mechanism in Section 1.3 and it remains at best a dubious way to proceed if we are sincere about language independence. Moreover, as already mentioned in Chapter II with the example of the philosopher Guido Calogero (whose page in only available in Italian, Basque and Polish), some resources are missing altogether from the main DBpedia which is thus far less comprehensive without taking its multilingual addenda into account.

The case of Wikidata is very different, since its makers chose to represent concepts with numbers rather than lexicalised strings, in an effort towards universality. The URI for Earth, for instance, is `https://www.wikidata.org/wiki/Q2` which is less explicit but can be used indifferently by English, Russian or Chinese natives. To accommodate these users with documentation in their own languages, Wikidata does not use alternative subdomains but rather adds an extra GET parameter in the URL in this form: `https://www.wikidata.org/wiki/Q2?uselang=fr`.[32]

However, none of these various technical implementations does satisfactorily represent the fact that there is no necessary one-to-one correspondence of concepts across languages. The concept of “future”, for instance, can be expressed in French both as “futur” and “avenir”, while conversely the Dutch word “koninklijk” has three English equivalents – “kingly”, “royal”, and “regal” – with subtle meaning variations and fixed collocations.

To sum up, most knowledge bases include a multilingual dimension, but all of them simplistically assume that the structure of language can be more or less adequately represented by establishing links between equivalent concepts. While this assumption appears to hold true most of the time, this model suffers from obvious shortcomings, especially when dealing with concepts evolving over time, as will be discussed in Section 3.

[31] `http://wiki.dbpedia.org/about/language-chapters/`

[32] Note that leaving this attribute out unsurprisingly defaults to English.

2.3 Multilingual corpora

While monolingual corpora – mainly English ones such as the Brown Corpus (Francis, 1964) and the Penn Treebank (Marcus et al., 1993) – have dominated the field of corpus linguistics for a few decades, this trend is changing fast with information globalisation, and corpora become more and more multilingual (Zhang et al., 2015). The Web itself has been considered by some researchers to be a huge, heterogeneous, multilingual corpus (Liu and Curran, 2006).

The Multilingual Corpus I, collected by the European Corpus Initiative[33] (ECI/MCI) and distributed on CD-ROM since 1994, was a first step in this new direction. It was followed by the Reuters corpus,[34] initially released in 2000, the second volume of which is dedicated to multilingual news stories. Other major initiatives include the Europarl[35] parallel corpus (Koehn, 2005) and more recently the Pentaglossal corpus (Forsyth and Sharoff, 2014) which will be introduced in more detail in Chapter V for generalisation purposes.

The *Historische Kranten* corpus used in this dissertation is multilingual in a moderate sense: it covers three languages (Dutch, French and English) from two branches (Germanic and Romance) of a single family (Indo-European). However, Indo-European languages account for 45% of speakers worldwide, and our corpus is at least representative of some of the linguistic diversity found in Western Europe.

Handling a multilingual corpus can be achieved in either a language-specific or a language-independent way. A language-specific approach would involve a preliminary language detection module prior to any linguistic processing, except if the language of each item has been encoded. We have seen in Chapter III that although such a `<language>` tag exists in the XML files of the *Historische Kranten* corpus, its misuse makes it utterly useless. The alternative is a language-independent approach which will be favoured here, relying on the multilingual structure of knowledge bases described above.

Even subtler is the multilingualism involved in the juxtaposition of Flemish dialects, sometimes radically diverging from the standard Dutch language, and their evolution over time. Even for a relatively codified languages such as French, significant changes can occur over the course of a century and a half, and the French spoken today is not quite the same as the one spoken by the journalists of young Belgium in 1830. The next section is dedicated to this important issue.

[33] `http://www.elsnet.org/eci.html`
[34] `http://trec.nist.gov/data/reuters/reuters.html`
[35] `http://www.statmt.org/europarl/`

3 Language Evolution

"Le temps altère toutes choses : il n'y aucune raison
pour que la langue échappe à cette loi universelle."
Ferdinand de Saussure (1916)

Language is, by nature, in constant evolution. Deutscher (2005, p. 1) starts his book on the origins of language with this apparently contradictory statement: "Language is mankind's greatest invention – except, of course, that it was never invented". Instead, Deutscher prefers to talk about the *unfolding* of language, an unstoppable phenomenon that defies all regulations, following its course unshakably century after century, and surviving all changes.

In contrast, human attempts to capture the essence of language have almost always reduced it to a fixed, static object. Whether at the lexical (dictionaries), syntactic (grammars) or semantic (thesauri) level, linguists rely on a *snapshot* of language of which they are seldom able (or willing) to distance themselves. Research in NLP has been no exception to this synchronic use of language, with very few systems taking its temporal dimension into account.

3.1 The generative lexicon

One of the first computational linguists to challenge this state of affairs was Pustejovsky (1991) with his idea of a generative lexicon – a lexicon in constant evolution that does not need to be set in stone once and for all.[36] This open model allows to *generate* meaning (Pustejovsky and Boguraev, 1993, p. 194):

> "The traditional organization of lexicons in natural language processing (NLP) systems assumes that word meaning can be exhaustively defined by an enumerable set of senses per word. Computational lexicons, to date, generally tend to follow this organization. [...] One disadvantage of such a design follows from the need to specify, ahead of time, the contexts in which a word might appear; failure to do so results in incomplete coverage. [...] Rather than taking a "snapshot" of language at any moment of time and freezing it into lists of word sense specifications, the model of the lexicon proposed here does not preclude extensibility: it is open-ended in nature and accounts for the novel, creative, uses of words in a variety of contexts by positing procedures for generating semantic expressions for words on the basis of particular contexts."

[36] Originally developed for the English language only, this model was partially extended to French by Pierrette Bouillon (1997), with a special focus on adjectives.

Pustejovsky and Boguraev (1993, p. 195) go on highlighting the advantages of their model for designing better NLP tools:

> “ Adopting such a model presents a number of benefits. [...] From the point of view of a natural language processing system, it can offer improvements in robustness of coverage. [Moreover,] some classically difficult problems in lexical ambiguity are resolved by viewing them from a different perspective. ”

This new perspective amounts to a “dynamic interpretation of a word in context”, in contrast to approaches enumerating word senses and subsequently forcing an algorithm to select one of these predefined senses. Static approaches fail to account for the diversity and intrinsic creativity of language, since “external contextual factors alone are not sufficient for precise selection of a word sense; additionally, often the lexical entry does not provide enough reliable pointers to critically discriminate between word senses” (Pustejovsky and Boguraev, 1993). Another consequence of predefining word senses is their multiplication for every minor variation of meaning.[37] In general, a dynamic model is better suited to represent the shifting nature of reality, where different levels of understanding constantly interact.

3.2 Stratified timescales

In order to account for the multiplicity of meanings across time, a robust model is needed to “substituer à la continuité insaisissable du temps une structure signifiante” (Prost, 1996, p. 115). Braudel (1949) introduces the framework of stratified timescales (*temporalités étagées*) which distinguishes between the long-term duration of geographical structures, the medium term of economical conjunctures and the short term of political events.

These different layers of time are not independent but interact with one another in more or less obvious ways: the geo-strategic position of the Mediterranean sea influenced the economic prosperity of the region, which in turn had an effect on everyday’s political life. But conversely, daily events had, in the longer run, an influence on the economy and ultimately on the geography of the Mediterranean world. This reverse effect is not acknowledged in Braudel’s original work but can be made explicit thanks to Norbert Elias (1996) and his explanation of all temporal phenomena as evolving continuums (*Wandlungskontinua*) interacting with one another in incessant fluxes.

[37] WordNet is notoriously guilty of this kind of over-segmentation, with 44 distinct senses listed for the verb “to give” for instance.

In this section, we first see how Isabelle Boydens applied this framework to the domain of administrative databases and showed how it could be generalised to all empirical databases in order to get operational results measurable in terms of costs and benefits, notably by reducing the number of formal anomalies arising over time thanks to a better understanding of the interaction between the different timescales involved in the management of these databases. We will then transpose the model of stratified timescales to our own field of interest and show how it applies to language evolution as a whole.

3.2.1 Application to empirical databases

Boydens (1999) generalised and extended Braudel's paradigm by applying it to the domain of databases containing strategic empirical information in the context of Belgian social security. She described the interaction between three levels of temporality (Boydens, 2011, p. 119):

> “ Three levels of transformation are interacting within the information system: the evolution of jurisprudence, the changes made within databases, and the categories observable in the field. These three levels of reality [...] operate, according to their nature, on different timescales. Thus we have the long-term for legal rules, renewed from one quarter or one year to the next, the medium-term for the management of databases, and the short-term for the observable reality, that is, that of the citizens or companies subject to administration, which is continuously evolving. ”

These three levels of reality are deeply interlinked but evolve in an asynchronous manner: a sudden change in the reality of a field will produce anomalies in the corresponding information system if the laws regulating the system had not foreseen the change. After legal interpretation, an adaptation of the legislation can trigger a restructuring of the database, with direct effects on the underlying reality.

This framework has thus very concrete operational consequences for database systems and people affected by them. Boydens (2011, p. 120) takes the example of private nursing homes which were originally excluded from the non-market sector due to their profitable nature. The declaration of such institutions as non-market was regarded erroneous in databases until legal interpretation eventually included them in this sector, forcing a re-engineering of the corresponding database schema:

> “ This restructuring was the result of a human decision aimed at bringing the model temporarily into line with the new observations. This phenomenon of transformation corresponds to the so-called “strange loop” mechanism defined by [Hofstadter (1980)]. In the absence of such an intervention, the gap between the database and the reality widens. ”

The concept of a *strange loop* therefore refers to the adaptation of a model due to contradictory observations, where the model should normally affect the perception of reality itself. Failure to account for the evolution of reality often results in a so-called *ghost factory*: a company or government body may invest important resources (in time, money, and human power) to produce anomalies and subsequently correcting them.

3.2.2 Application to language evolution

We have seen that Braudel's framework of stratified timescales is a productive environment that can be applied to heterogeneous domains in order to account for the interaction between different temporalities. Pushing the analogy further, we consider that language[38] materialises in this stratified time, ceaselessly evolving on three different layers interacting with one another:

- the long term of language appearance and extinction, or rather slow transformation (e.g. from Latin to Romance languages) over centuries
- the medium term of common usage, of dictionaries and grammar rules that evolve over years or decades
- the short term of everyday speech, of creative word uses and neologisms that appear on a daily basis

We see that the long term has an influence on the medium term (Latin declensions eroded and disappeared but left traces in the lexicon of modern French) and the medium on the short (prescriptive grammar influences the way speakers express themselves in a given context), but also, perhaps less obviously, that the short term affects the medium term (descriptive grammar tries to keep up with the way people actually speak, resulting in new hype words entering dictionaries) and the medium the long (the standardisation of the Italian language was partly based on literary classics). See Figure IV.4 for an illustration of the interactions between these three layers of temporality.

[38] As earlier (Section 2), the English term is used here in the broader sense of French *langage* (i.e. human system of communication), as opposed to particular languages (French *langues*).

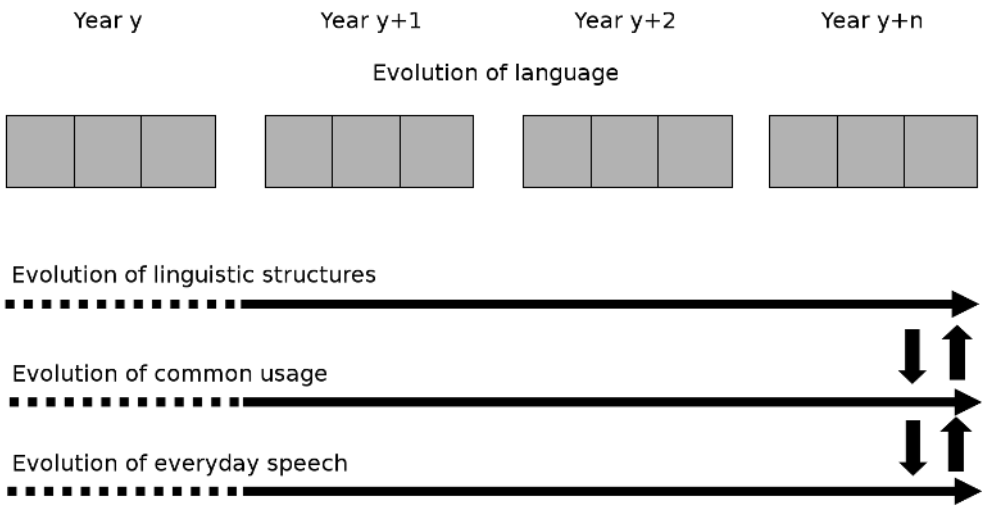


Figure IV.4: Evolution of language, adapted from Boydens (2011)

In the context of the semantic enrichment of our historical newspapers, all three levels are important. The proliferation of Dutch varieties versus the economic dominance of *Standaardnederlands* is an example of how long-term evolution affects language: are Flemish dialects from the early 19th century still sufficiently related to present-day Dutch to be considered the same language? Moreover, new realities appear and are expressed either with new words or with existing ones, which thereby acquire new meanings.

Archaic sentence constructions and words from centuries past constitute the focus of medium-term language evolution, sometimes requiring ancient dictionaries to retrieve word meanings. Finally, short-term lexical usage is the harder level for information extraction systems, not only for distant historical sources but for virtually every kind of content: the task of linking mentions in tweets to a stable external reference, for instance, can prove very challenging (Derczynski et al., 2015), especially in a multilingual context.

3.3 Concept drift

Shifts in the meaning of words have been widely studied. In the domain of machine learning, for instance, the fact that "real world concepts are often not stable but change with time" (Tsymbal, 2004) has long been identified as a potential issue affecting systems. Widmer and Kubat (1996) have called this problem *concept drift* and linked it to the related notion of *hidden contexts*, stressing the fact that the cause for the change is often not known a priori.

Similarly, but in relation to history, Prost (1996, p. 63) notes that "les concepts ont beaucoup changé de sens, et ceux qui nous paraissent transparents sont les plus dangereux. [...] Aussi pourrait-on ériger l'histoire des concepts en préalable de toute autre histoire". Tracking these shifts is therefore a laudable goal in order to understand how languages evolve over time, and a critical component that could be integrated into NLP systems to improve specific modules such as coreference resolution and entity disambiguation.

To convince ourselves of the pervasiveness of this phenomenon, let us focus on the term *bourgeois*. Initially (and etymologically), it was a neutral word referring to a town-dweller (from French *bourg*, Late Latin *burgus*, Old High German *burg*), as in Example 27. Later, it became more closely associated with the *bourgeoisie* and upper-middle-class people, as seen in Example 28. Today, the term is rarely used in either sense and appears mainly as a family name (Example 29) or in some codified fixed forms (Example 30).

Example 27. il fit distribuer 17 000 pains aux malheureux bourgeois d'Ypres

Example 28. un bourgeois dont le gros sac fait concurrence au gros ventre

Example 29. André Bourgeois in Raad van Advies voor Vreemdelingen

Example 30. Château La Montagne Cru Bourgeois Medoc

Tsymbal (2004) distinguishes between sudden and gradual concept drift, and also refers to virtual concept drift when it is not an actual concept that is changing but rather the underlying data distribution. On a practical level, however, the author notes that "it is not important, what kind of concept drift occurs, real of virtual, or both. In all cases the current model needs to be changed". Also see Masud et al. (2010) for a more technical discussion of concept evolution in data streams.

According to Derczynski et al. (2015), "systems trained on one dataset may perform well on that dataset and other datasets gathered at the time, but not actually generalise well in the long run". This observation has important consequences for corpora spread over a broad time span such as the *Historische Kranten*. Working with KBs implies that we have access to the sum of knowledge at a given moment. While the Web of Documents is very dynamic and in constant evolution, Linked Data tend to be more static, sometimes lacking in actuality with respects to recent trends. In other words, the Web (especially the social Web) evolves in the short term, whereas the Semantic Web only gets updated in the medium term, accumulating encyclopaedic knowledge for the long term.

This is especially true when dealing with large dumps of KBs that take a long time to backup and to download: DBpedia dumps, for instance, are generated only at the rate of one per year on average.[39] Even if dumps were more regular, it would be very difficult to keep up with the current state of knowledge, since Wikipedia articles get updated every second. Data stored in KBs can quickly become outdated, and Wikipedia articles then need to be re-extracted before becoming available to users. To remedy this situation, DBpedia Live[40] was set up to allow a continuous synchronisation with Wikipedia.

Detecting concept drift automatically remains a challenge to be tackled by semantic annotation tools if their output is to remain relevant over time. As Tsymbal (2004) notes for data mining, "an important problem with most of the real-world datasets in existing experimental investigations is that there is little concept drift in them, or concept drift is introduced artificially". Without offering a full-scale solution to this problem, we can at least bear in mind that concepts extracted from KBs are never fixed once and for all but always subject to future revision and complementation. We will now apply the notion of concept drift to place names and concepts to understand their evolution.

3.3.1 Application to place names

As seen at the end of Chapter III, locations are of special interest to users of the *Historische Kranten* corpus. The unique context of the city of Ypres and its surroundings[41] makes linguistic variation of place names in the region particularly productive, especially from a diachronic perspective. The toponym Passendale, for instance, appears over 10 000 times in the corpus in various written forms. The old Dutch form *Passchendaele* (and its variant *Passchendale*) was retained in the English war literature, despite the fact that the new spelling became effective in Belgium and the Netherlands around 1946–1947. Table IV.2 illustrates this shift in *Het Ypersch nieuws*:

Spelling	**1920s**	**1930s**	**1940s**	**1950s**	**1960s**	**1970s**
Passchendaele	30	406	82	-	2	-
Passchendale	28	1 032	891	13	-	-
Passendale	-	-	1	691	879	137

Table IV.2: Spelling shift from Passchendaele to Passendale

[39] `http://wiki.dbpedia.org/datasets/`

[40] `http://live.dbpedia.org/`

[41] In the Westhoek, close to France, ground to several battles involving English troops, etc.

Admittedly, spelling shift is not concept drift, but the two are interlinked since the old spelling *Passchendaele* has now acquired a new, war-related meaning and it not used in other contexts any longer. The mechanism of Wikipedia (and DBpedia) redirections allows to capture these variations and to correctly disambiguate both forms with a single URI. In fact, `dbr:Passchendale` is automatically redirected to `dbr:Passendale` thanks to the `dbo:wikiPageRedirects` OWL property[42] (`dbr:Passchendaele` being a separate resource). Whereas a plain NER tool could possibly recognize Passendale and Passchendaele as locations, they would fail to identify both as the same place without the help of an additional coreference module. This demonstrates the added value of an LOD-based approach to full disambiguation and temporal tracking of concepts, which will be further developed in our research perspectives.

3.3.2 Emergence and salience of concepts

NLP techniques can be leveraged to monitor the apparition of new concepts or their relative importance over time. Bird et al. (2009) show how to use the Natural Language Toolkit (see Chapter V) to track the use of terms. An example of this on US inaugural addresses from 1789 to 2009 is shown in Figure IV.5.

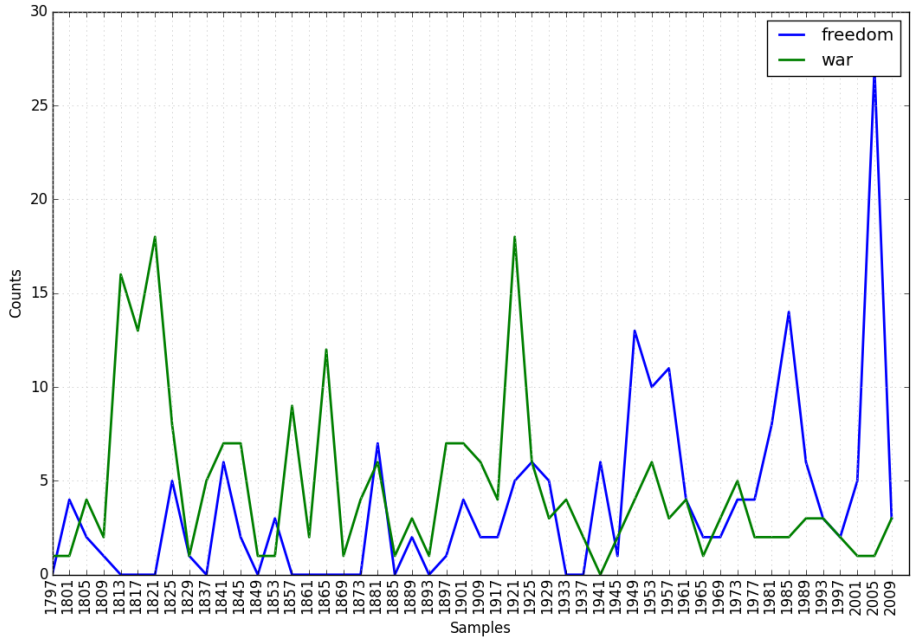


Figure IV.5: Relative salience of terms in US inaugural addresses

[42] In the way non-descriptors refer to corresponding descriptors (standard froms) in thesauri.

Similarly, we can automatically detect the "hot topics" of any given year in the *Historische Kranten* corpus by computing the relative frequency of terms compared to their general distribution. Table IV.3 shows an example of this for salient terms used in the Ypres Times from 1922 to 1935, along with their scores based on the term frequency–inverse document frequency statistic.[43]

Year	Term	Score
1922	Hartley	0.8
	Teddington	1.0
	Cavendish	1.0
1923	Sage	1.0
	Communal	0.769230769231
1924	Bellew	1.0
	Sclater	0.733333333333
	Ramsdale	0.730769230769
1925	Bowley	0.727272727273
	Managers	0.916666666667
1926	Ginchy	0.785714285714
	Rummy	1.0
1928	Portsmouth	0.882352941176
	Fokker	0.727272727273
1929	monitors	0.866666666667
1930	Blunden	0.818181818182
	Morgan	0.9
	Dorrien	0.8
1931	Danny	1.0
	Seamen	0.727272727273
1932	Thiépval	1.0
1933	Feathers	0.75
	Morval	0.862068965517
1934	Gordons	0.75
1935	starvation	0.714285714286
	Jubilee	0.9

Table IV.3: Hot topics from the Ypres Times

Although some trends appear from the data, the tracking of the evolution of concepts over time would benefit from a more comprehensive approach which is beyond the scope of the present dissertation. This will encourage us to formulate a number of lines of enquiries in our conclusions, when anticipating future research perspectives and potential applications.

[43] The tf–idf score reflects the salience of a word in a corpus (Rajaraman and Ullman, 2011).

Summary

In this fourth chapter, we have dealt with transversal subjects that are inherent to many collections of data from the digital humanities and elsewhere but are seldom addressed as real issues, or only in a peripheral way. The three central aspects we proposed to cover were the quality of content, the handling of multilingualism and the evolution of language and concepts over time.

Accounting for the quality of data is particularly important when dealing with uncurated collections and information sources, but we have seen that quality is never an absolute factor that can be measured objectively for all purposes. On the contrary, it always depends on the needs of users and is therefore relative to usage. In particular, processing digitised collections raises the question of OCR quality and of its impact on further processing such as semantic enrichment, while the data sources used for this enrichment are themselves subject to variable quality.

Another central issue is the question of multilingualism and language-independence. Many NLP tools have been developed for English and their adaptation to other languages can prove harder than expected. In the context of the Web, designing systems that are portable to any language without too much customisation is a key challenge in order to access the sum of knowledge available online in an automated way. Currently, however, this goal remains largely unattained, thereby threatening to confine the Web in the position of a digital Babel of multiple individual linguistic crystallisations with loose interdependence. On a smaller scale, the *Historische Kranten* project contains newspapers in three languages and therefore needs approaches to information extraction suited to account for this reality. Cross-lingual IE tools are clearly in demand for all kinds of semantic enrichment projects, as shown by the success enjoyed by multilingual digital corpora worldwide.

Finally, the evolution of language over time needs to be taken into account. As emphasised by Isabelle Boydens, information is not given once and for all but progressively constructed, and concepts must be seen as dynamic objects rather than static ones. Braudel's framework of stratified timescales, coupled with Elias's model of evolving continuums, allow us to represent the interactions between different layers of time in order to better apprehend the underlying reality. Applying this framework to language, we showed how concept drift affects IE systems in often unacknowledged ways.

Our hypothesis regarding the practical consequences of quality, language, and time on information extraction and semantic enrichment will now be tried and tested against the empirical data of the *Historische Kranten*.

Chapter V

Knowledge Discovery

Outline

This final chapter builds upon the material discussed in the previous four in order to achieve knowledge discovery, i.e. to provide users not only with a structured answer to their requests but also with additional facts relevant to their queries. Knowledge discovery can indeed be considered the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al., 1992).

Section 1 introduces MERCKX, a knowledge extractor developed for this thesis. Related tools are reviewed in a comparative perspective, emphasising their limitations and focusing on potential improvements that can be achieved with Linked Data to handle multiple languages efficiently. The components of our system are then detailed before describing our workflow in detail, showing the originality of our approach.

The performance of entity linking is evaluated in Section 2 in order to compare the results we obtained with state-of-the-art systems. We first look at some tools from Section 1 in the light of the SQuaRE ISO standard, before defining the methodology we followed for the evaluation – including the use of a gold-standard corpus based on the *Historische Kranten* corpus working as an artificial real-world referent – and discussing our results.

Finally, a validation of our findings is proposed in Section 3 through a proof-of-concept implementation for the Ypres City Archive. We focus on ways to go beyond the limitations of traditional search engines and provide insight for creative applications based on semantic enrichment and Linked Open Data, before generalising our approach to other languages, application domains, and types of entities.

Contents

1 MERCKX: A Knowledge Extractor

Whereas NLP technologies of the first generation were in general commercial solutions, developed inside companies and sold on physical supports such as CDs, the increased popularity of the Web has allowed a proliferation of online tools available under various licences. But what does constitute a useful tool? Kent (1978, p. 194) offers the following insight:

> “ Useful tools have well-defined parts, and predictable behavior. They lend themselves to solving problems we consider important, by any means we can contrive. We often solve a problem using a tool that wasn't designed for it. Tools are available to be used, don't cost too much, don't work too slowly, don't break too often, don't need too much maintenance, don't need too much training in their use, don't become obsolete too fast or too often, are profitable to the toolmaker, and preferably come with some guarantee, from a reliable toolmaker. Tools don't share many of the characteristics of theories. Completeness and generality only matter to the extent that a few tools can economically solve many of the problems we care about. ”

We gather from this definition that a tool is essentially something that *works*, i.e. that fits a certain use. This is clearly in line with the framework of data quality defined in Chapter IV. Quality being a relative concept, it entails that the perfect tool does not exist: a good-enough tool for a given purpose always constitutes a trade-off between these various criteria, and many more besides.

In this section, we present MERCKX (Multilingual Entity/Resource Combiner & Knowledge eXtractor), a tool that we designed in the context of our thesis in order to extract entity mentions from documents and to link them to DBpedia (Bizer et al., 2009b). For example, a textual mention of “Rabelais” could be disambiguated with `dbr:François_Rabelais`, which includes the alternative label “Alcofribas Nasier” (one of his anagrammatic pseudonyms) but excludes information about American composer Akira Rabelais (who has his own unique URI: `dbr:Akira_Rabelais`) or the main-belt asteroid named after the French humanist (`dbr:5666_Rabelais`).

In order to underline the originality of our approach, we will first review a few related tools in Section 1.1, before detailing the components of MERCKX in Section 1.2 and explaining its inner workings in Section 1.3.

1.1 Similar tools

This section opens with a panel of existing tools, detailing their functionalities but also their limits, especially in terms of multilingual coverage. It brings together two types of tools capable of performing entity linking: named-entity recognisers and semantic annotators, which we chose to group under a single heading since they achieve the same results nowadays. In our rapidly evolving information economy, a presentation of existing solutions, both open-source and commercial, is necessarily biased and incomplete. Acknowledging the fact that several tools have certainly escaped our attention, we nevertheless attempt to draw a good picture of the services available in the broad field of semantic enrichment technologies.

In Chapter I, we tried to define what constitutes a valid named entity. The distinction between a term (common noun) and entity (proper noun) is sometimes blurred by tools that do not establish a formal distinction between the two. Practically speaking, a named entity could even be defined as something retrieved by a NER tool, which is admittedly circular but would be sufficient for some purposes. The relevance of the output for users is indeed more of a concern than the linguistic category used to label it. Semantic annotation, on the other hand, is the process of attaching supplementary information from the Semantic Web to the content of documents (see Chapter II, Section 3.2). In contrast to some NER tools, semantic annotators explicitly integrate a linking component and do not limit themselves to entities: any type of content is liable to be enriched with external facts from the Web of Data.

The NERD platform[1] (Rizzo and Troncy, 2011) allows to experiment with these different services (and some others not covered here) within a single graphical user interface (see Figure V.1). Another effort to integrate various services is the NER extension[2] of OpenRefine[3] (Verborgh and De Wilde, 2013; van Hooland et al., 2015). Originally relying on three third-party NER APIs (AlchemyAPI, DBpedia Spotlight, and Zemanta), this extension has now been extended to other services, such as dataTXT.[4]

In the remainder of this section, we propose a quick overview of some of the most popular entity linking tools. This survey starts with Spotlight, which is totally open source, in order to study the general workings of such tools, and goes on with services that operate as black boxes but achieve better results.

[1] http://nerd.eurecom.fr/
[2] https://github.com/RubenVerborgh/Refine-NER-Extension/
[3] http://openrefine.org/
[4] https://dandelion.eu/datatxt/

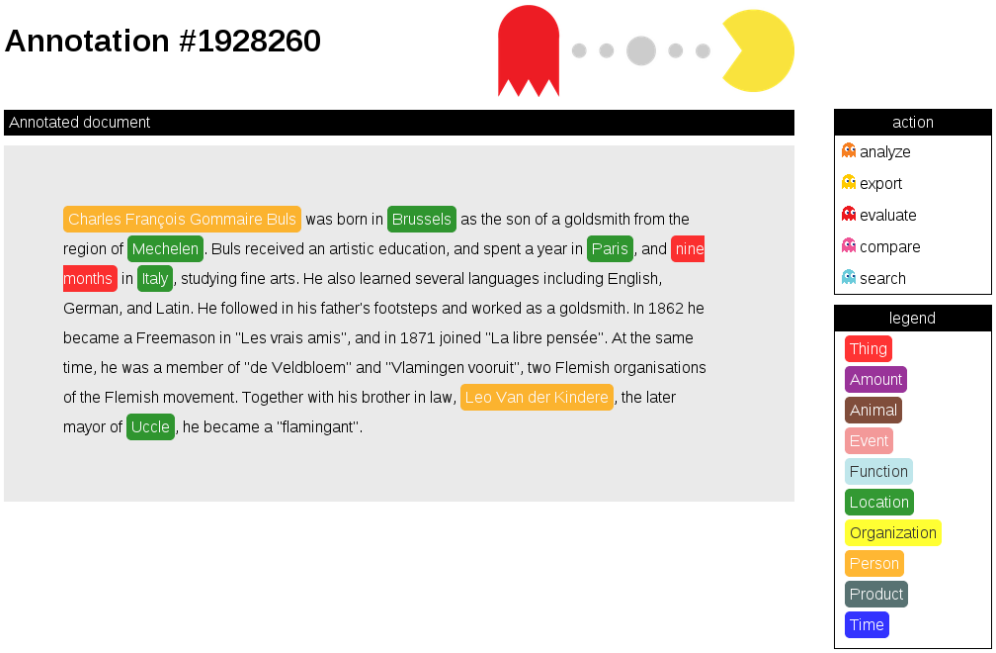


Figure V.1: NERD platform

1.1.1 DBpedia Spotlight

DBpedia Spotlight[5] 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 (Mendes et al., 2011, italics theirs):

> "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 [...]."

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

However, the original Spotlight was designed for English only, as it is the case for many tools. To counterbalance this limitation, Daiber et al. (2013) developed a new multilingual version of DBpedia Spotlight, which they claim is faster, more accurate, and easier to configure. This statistical version has been adopted for the online demo.[6] 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%.

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.[7] Surprisingly, the language-dependent implementation does not improve the results significantly: it only outperforms the language-independent implementation by less than a percentage point.

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 (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. As a fully transparent tool, the multilingual version of DBpedia Spotlight will be used as a baseline in our evaluation process in Section 2.

1.1.2 OpenCalais

Powered by Thomson Reuters, OpenCalais provides several services such as NER, topic modelling, relation detection and event extraction. The Calais Viewer[8] allows to experiment with its technology (see figure V.2), while prolonged use requires an API key (free for the basic service, with paid upgrades to be purchased if needed). Although NER has a limited support for French and Spanish, other services are restricted to the English language, with no plans for additional language support.[9] Contrary to most NER services, OpenCalais uses exclusively its own knowledge base – called Open PermID[10] – which makes interoperability of the extracted entities difficult.

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

[7] `https://opennlp.apache.org/` ; see Chapter I (Section 2.2.2) for the distinction between ruled-based and data-driven approaches.

[8] `http://viewer.opencalais.com/`

[9] `http://www.opencalais.com/forums/calais-initiative/language-support`

[10] Beta version available at `https://permid.org/` (as of October 20, 2015).

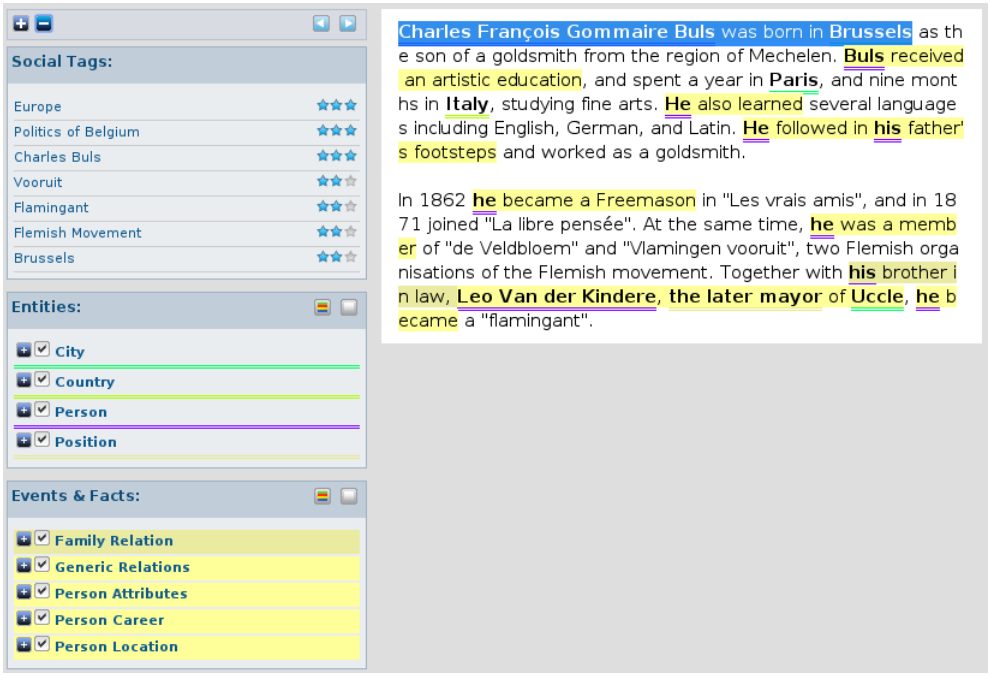


Figure V.2: Calais Viewer

1.1.3 AlchemyAPI

AlchemyAPI, a company acquired by IBM in March 2015, provides a NER service through an application programming interface (API), as its name suggests. IBM integrated Alchemy in its Watson computer,[11] which won the Jeopardy! TV game show in 2011.[12] Alchemy is a commercial application charging customers per API call, but free API keys can be obtained for testing purposes with a daily limit of 1 000 transactions.

As of April 2015, AlchemyAPI supports NER in eight languages: English, French, German, Italian, Portuguese, Russian, Spanish, and Swedish. However, the quality of the extraction is unequal from one language to another, with full disambiguation working much better for English (Hengchen et al., 2015). It also includes several other components which can be tested in an online demo,[13] – such as relation detection and sentiment analysis – but again these do not live up to expectations for languages outside English.

[11] http://www.alchemyapi.com/company/press/ibm-acquires-alchemyapi-enhancing-watson%E2%80%99s-deep-learning-capabilities

[12] http://techcrunch.com/2013/02/07/alchemy-api-raises-2-million-for-neural-net-analysis-tech-on-par-with-ibm-watson-google/

[13] http://www.alchemyapi.com/products/demo/alchemylanguage

1.1.4 Stanford NER

The University of Stanford offers a statistical NER service[14] based on conditional random fields (CRFs)[15] (Finkel et al., 2005). Stanford NER can either be downloaded under a GPL license, or used online as a demo version.[16] It is a Java implementation of a semi-supervised approach to NER, provided with models trained on English, German and Chinese. Users can theoretically also train their own models for use on other languages or specific application domains, although this requires large-scale human-annotated data, which is a *de facto* limitation for usage with languages not listed above. Stanford NER only recognises and classifies named entities in the traditional sense defined in Chapter I, providing 7 entity types: LOCATION, ORGANIZATION, DATE, MONEY, PERSON, PERCENT, and TIME. No disambiguation is proposed.

1.1.5 AIDA

AIDA[17] is a framework developed by the Max Planck Institute for Informatics. It is based on YAGO and intended for entity detection and disambiguation in text (Hoffart et al., 2011). The idea behind AIDA is to use a graph-based approach (see Chapter II) to perform a collective mapping of textual mentions to entities. For instance the mention of "Kashmir" in a text could refer to at least two distinct entities depending on the context: the Asian region or a song by Led Zeppelin. Both mentions and entities can be used as vertices (nodes) in the graph, while edges (links) between them are of two types:

- *mention–entity* edges capture the similarity between the context of a mention and a candidate;
- *entity–entity* edges capture the semantic relatedness between entities.

In other words, if a textual mention has two potential entity candidates for disambiguation, the one with the most similar context (which will be closer in the graph) will be chosen. Similarly, close entities are bound to be semantically related. Although Hoffart et al. (2011) and Yosef et al. (2011) mention that AIDA handles "natural language" without any specifics (see Bender (2011) for this common bias), the GitHub page of the project[18] makes it explicit that only English text is supported.

[14] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[15] CRFs typically take more context into account than ordinary classifiers.
[16] `http://nlp.stanford.edu:8080/ner/`
[17] `http://www.mpi-inf.mpg.de/yago-naga/aida/`
[18] `https://github.com/yago-naga/aida`

1.1.6 Zemanta

Developed as a Web content enrichment platform, Zemanta[19] 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,[20] 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.

Zemanta requires an API key in order to use its services,[21] 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). 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. The good results achieved in former experiments made us select it as a candidate for evaluation in Section 2.

1.1.7 Babelfy

Moro et al. (2014) introduce Babelfy,[22] a system bridging entity linking and word sense disambiguation and based on the BabelNet[23] multilingual encyclopaedic dictionary and semantic network which is constructed as a mash-up of Wikipedia and WordNet. 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: Figure V.3 shows two football players successfully identified on the basis of their first names only, with a limited amount of additional context. Relying on a large-scale multilingual network, Babelfy officially supports 267 languages, in addition to a language-agnostic option. This unusually broad linguistic coverage, coupled with its original approach to entity disambiguation, justifies the fact that Babelfy is also among the tools that will be evaluated in Section 2.

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

[20] 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/.

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

[22] http://babelfy.org/

[23] http://babelnet.org/

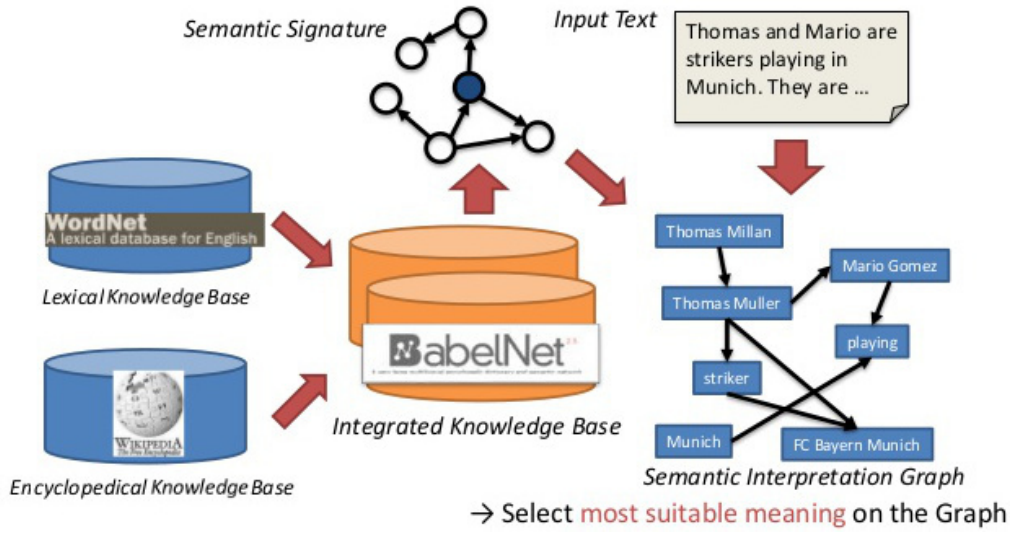


Figure V.3: Babelfy, adapted from slides based on Moro et al. (2014)

1.2 Components

The previous section presented a number of NER and semantic annotation applications able to link and disambiguate entities with URIs. These tools, however, suffer from a number of shortcomings, as will be demonstrated in Section 2: they work unequally well on different languages, suffer from cross-lingual ambiguity, do not handle OCR output very well, and are only portable to a limited extent.

In order to overcome these various limitations, we chose to build our own knowledge extractor, based on personal insights and on the integration of a few external components. Our main experiment focuses on locations, but the system could be adapted to other languages, domains, and types of content, as will be shown in Section 3. For full transparency, the scripts of the different parts of MERCKX are openly maintained on GitHub.[24] The source code of the main program is also provided in Appendix A.

In what follows, we review the key components of our system: the Python language and its natural language toolkit, the architecture of named-entity extractor X-Link, and the use of DBpedia dumps. The workflow used by this tool to process documents from our corpus and link them to the information contained in knowledge bases will then be put forward in Section 1.3.

[24] `https://github.com/ulbstic/ypres`

1.2.1 Python and NLTK

We chose to implement our tool in Python[25] since its simple, object-oriented syntax is well-suited to handle natural language, and because it is widely used across the NLP community (Bird et al., 2009). Its comprehensive natural language toolkit (NLTK)[26] (Garrette and Klein, 2009) makes the choice even more obvious. NLTK provides several ready-to-use NLP components such as tokenisers, part-of-speech taggers, chunkers, and syntactic parsers. A basic pipeline using NLTK is displayed in Figure V.4. In this example, a raw HTML file is parsed into a string of text, then further processed to obtain individual word forms (lemmas) and build a vocabulary.

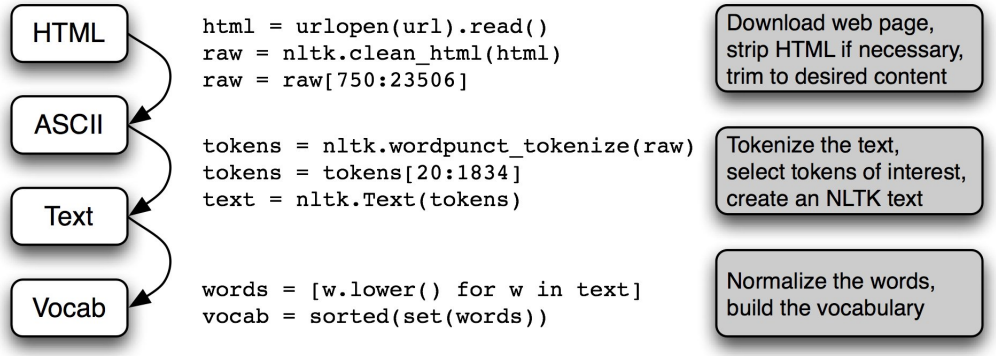


Figure V.4: NLTK pipeline, reproduced from Bird et al. (2009, p. 86)

The toolkit also comes with many linguistic resources and corpora, including lists of basic stopwords in 14 languages.[27] Handling stopwords in a multilingual content is particularly tricky: in a retrieval application ignoring diacritics, for instance, excluding the common English article "the" could make it difficult to retrieve results about "thé" (tea) in French for instance. Moreover, lists of stopwords are not deterministic and can vary with the domain or the interests of the users. A good example in the biomedical context is the word "cell", which is by no means empty (it has a very concrete meaning) and has an entry in the Unified Medical Language System (UMLS),[28] but is far too general in itself to provide relevant information for users. For a detailed discussion of stopwords (*mots vides* in French), see Deviaene (2008).

[25] `https://www.python.org/`

[26] `http://www.nltk.org/`

[27] Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish, and Turkish.

[28] `http://www.nlm.nih.gov/research/umls/`

1.2.2 X-Link

X-Link[29] is a fully configurable, Linked Data-based, Named Entity Extraction (NEE)[30] tool (Fafalios et al., 2014). To our best knowledge, this is the system which comes closest to the work undertaken in this dissertation as a whole, with lots of shared intuitions and insights. The fact that it was developed quite recently and is used in production for several European projects encourages us to think that there is a real need for such applications today.

As shown in Figure V.5, the architecture of X-Link is based in part on the GATE ANNIE system,[31] but it can use any other NER system that takes text as input and returns a list of named entities. The system relies on dynamic gazetteers of DBpedia multilingual labels that are extracted in advance and kept in memory for the subsequent semantic enrichment phase. Central to the approach is configurability: the user can modify any component of the system directly in the Web interface, which makes it extremely portable.

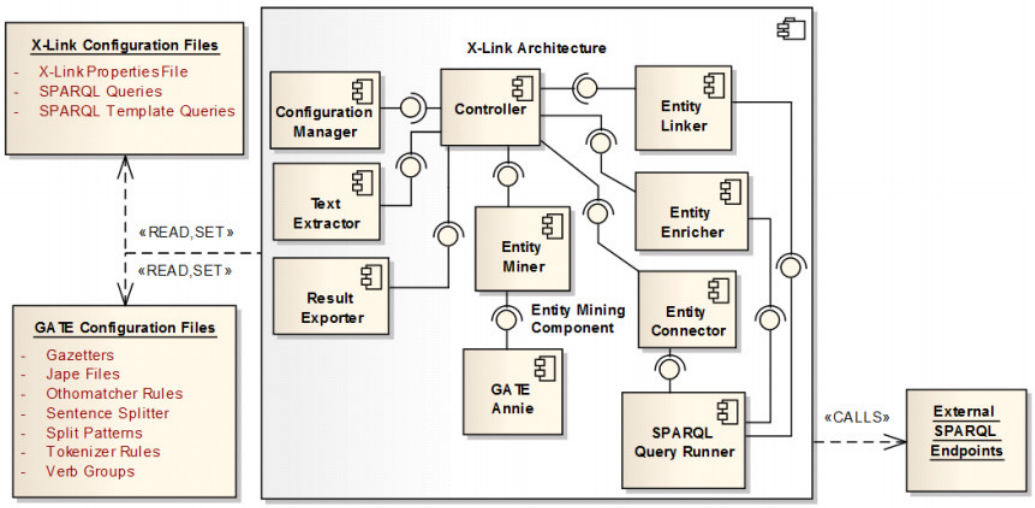


Figure V.5: X-Link, reproduced from Fafalios et al. (2014)

Fafalios et al. (2015) observe that "since a lot of information about *named entities* is already available as Linked Open Data (LOD), the exploitation of LOD by a NEE system could bring wide coverage and fresh information". They also note that while tools such as DBpedia Spotlight play this role up to a point, they fail to fully exploit the dynamic and distributed nature of LOD.

[29] `http://ics.forth.gr/isl/X-Link/`

[30] The creators of XLink consider NEE as the combination of NER and entity linking.

[31] `http://services.gate.ac.uk/annie/`

To illustrate their point, Fafalios et al. (2014) provide a real scenario (based on the iMarine[32] project) which looks surprisingly similar to our own use case, except for fish species which should be replaced by locations:

> “ For giving users an overview of the search results and allowing them to explore them in a faceted way, you want to use a NEE tool for identifying (at real time) Fish Species in the snippets or the full contents of the top results. You think that it would also be useful to link (on demand) the identified species with related semantic resources, as well as retrieve more information (e.g. a short description of the species, an image, its taxonomy, etc.) by querying (at real-time) online Semantic Knowledge Bases. ”

For all these reasons, we chose to implement parts of this architecture into our own system, since relying on the online version could prove problematic in the case of a server downtime. The interesting components of X-Link were therefore transposed to our needs and seamlessly integrated in our workflow.

1.2.3 DBpedia dump

The third component of MERCKX is its reliance on Linked Open Data. As mentioned in Chapter II, DBpedia can be queried through its SPARQL endpoint.[33] For instance, listing all species of fish present in the KB can be achieved with:

```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Fish
}
```

A sample of the results is shown in Table V.1 overleaf. Additionally, the *labels* corresponding to these URIs (i.e. the lexicalised forms of the concepts in a given language) can be accessed with the following query:

```
SELECT DISTINCT ?label WHERE {
  ?uri a dbo:Fish .
  ?uri rdfs:label ?label
}
```

[32] http://www.i-marine.eu/
[33] http://dbpedia.org/sparql

uri
http://dbpedia.org/resource/Astroscopus_guttatus
http://dbpedia.org/resource/Barbus
http://dbpedia.org/resource/Black_ruby_barb
http://dbpedia.org/resource/Channichthyidae
http://dbpedia.org/resource/Checker_barb
http://dbpedia.org/resource/Cherry_barb
http://dbpedia.org/resource/Cow_shark

Table V.1: URIs of fish species

This second command returns a list of fishes as they are likely to appear in text, along with the language of expression:

label
"Astroscopus guttatus"@en
"Nördlicher Elektrischer Sterngucker"@de
"Astroscopus guttatus"@nl
"Skaber amerykański"@pl
"Североамериканский звездочёт"@ru

Table V.2: Labels of fish species

However, querying the online endpoint suffers from at least three disadvantages: it is slow (several minutes for long texts), unreliable (the server being frequently overloaded or in maintenance), and incomplete (limited to 10 000 results at a time). To make up for these issues, we chose to work with a local dump of DBpedia. We selected the dump of August 2014[34] which was the most recent at the beginning of our experiments. MERCKX requires two files:

- instance types[35] to select all URIs from a category (places, for instance)
- English labels[36] to map these URIs to their lexicalised forms

Additionally, labels from other languages can be loaded as well to increase recall. In our experiment described in Section 2, we also used French and Dutch labels. These gazetteers are the only language-specific resources of our system, and they will be combined in a single multilingual dictionary.

[34] http://wiki.dbpedia.org/Downloads2014

[35] http://downloads.dbpedia.org/2014/en/instance_types_en.nt.bz2 (122 MB)

[36] http://downloads.dbpedia.org/2014/en/labels_en.nt.bz2 (163 MB)

1.3 Workflow

The workflow of MERCKX for the extraction and disambiguation of entities consists of five phases: downloading resources, building the dictionary, tokenising the text, spotting entity mentions, and annotating mentions with positions and URIs. The first two steps can be time-consuming depending on the chosen entity type and additional languages (about 5 minutes and one minute respectively in our experiment), but they need to be performed only once.

1.3.1 Download

In order to simplify the download and decompression of the DBpedia dump, we provide a shell script doing this automatically.[37] This script invokes another one written in Python[38] 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://dbpedia.org/resource/Alabama>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<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>

<http://dbpedia.org/resource/Abraham\_Lincoln>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/OfficeHolder>
```

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

[37] https://github.com/ulbstic/ypres/blob/master/merckx-init.sh
[38] https://github.com/ulbstic/ypres/blob/master/merckx-init.py

1.3.2 Dictionary

In the second phase, MERCKX maps all the URIs to their corresponding labels 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
```

To make this more legible and reduce the size of the file, thereby saving time, the `merckx-init.py` 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 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
```

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 lesser-known (and thus less frequent) French municipality of Le Liège and to the Liège metro station in Paris. If the languages are combined in that order, the conflict between URIs will probably 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).

Table V.3 shows a summary of the number of places extracted (URIs and labels by language, plus the combined labels).

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

Table V.3: Summary of the extracted places

In total, 735 062 unique locations were found in the DBpedia dump of August 2014.[39] 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.

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

[39] 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. This point will be discussed in more detail in our conclusions.

for the benefit of the English central version. This state of affairs constitutes one of the major downsides of the current structure of DBpedia, which is simply replicated from Wikipedia rather than organised in a language-independent manner (see Chapter II). 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 (1 106 and 4 477 respectively).

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

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

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

1.3.3 Tokenisation, spotting, and annotation

The next step is to tokenise the documents we want to enrich with the NLTK WordPunctTokenizer and to perform a simple greedy lookup[40] 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.

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 (assuming that the dictionary mapping is correct in the first place). This corresponds to the format of the Entity Discovery and Linking track[41] at the Text Analysis Conference,[42] which will also be used to annotate our gold-standard corpus in Section 2.2.3. 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. Potential applications for this will be presented in Section 3.

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

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

[42] `http://www.nist.gov/tac/`

2 Evaluation

Evaluation of IE systems is crucial in order to understand how they compare to one another and to a common baseline. Several resources have been made available for evaluating such tools. SemEval (Semantic Evaluation, formerly SensEval), for instance, is a series of evaluation campaigns for computational semantic analysis, which started in 1998 and is still ongoing. It is considered the reference for state-of-the-art semantic systems, with an evaluation workflow widely accepted by the NLP community (see Figure V.6). For entity linking in particular, the AGDISTIS framework (Speck and Ngomo, 2014) provides annotated corpora for benchmarking purposes.

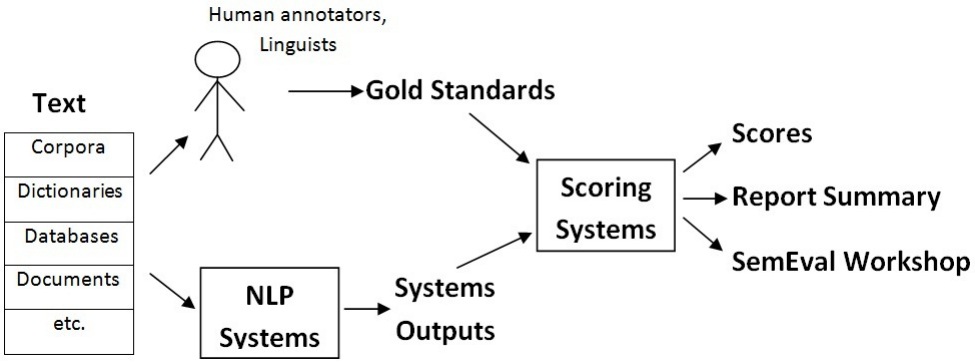


Figure V.6: SemEval workflow, reproduced from Liling Tan (CC BY-SA)

IE systems are often evaluated in terms of precision, recall, and F-score. The performance of human annotators on NER is known to be relatively high, with F-scores over 96% on most entity types (Sundheim, 1995). By comparison, state-of-the-art systems can reach F-scores of 92% for PER and LOC, with lower scores of about 85% for ORG (Tjong Kim Sang and De Meulder, 2003). Palmer and Day (1997) note that "incremental advances above the baseline can be arduous and very language specific". We intend to find out if this also holds true for the more advanced task of entity linking.

In this section, we will give importance to evaluation practices and conduct an experiment of our own. Section 2.1 offers a preliminary assessment based on the comparison of the linguistic coverage of the evaluated tools, followed by a black-box analysis with the SQuaRE ISO standard. Section 2.2 introduces the methodology used for the glass-box evaluation of the extraction component – defining its objectives and including metrics along with our corpus – while Section 2.3 discusses the results obtained for the different systems.

2.1 Preliminary assessment

The list of NER services presented in Section 1.1 could go on but we chose to limit it to some of the biggest players in the field. Rodriquez et al. (2012), for instance, used the NER module of OpenNLP[43] for their evaluation, but they reported very poor results even for English, so we chose not to include this tool in our review.[44] For a good synthetic view of the features of these services, we refer the reader to Derczynski et al. (2015, p. 34).

2.1.1 Linguistic coverage

Table V.4 compares the language support of the tools we presented earlier. Despite some efforts to increase the linguistic coverage of semantic technologies, English remains to this day the only language to be fully supported by all seven NER tools, although Babelfy can boast a broad coverage. This unfortunate situation calls for more massive investment into language-independent approaches that do not require substantial adaptation work nor linguistic resources.

Language	AIDA	ALCH	BABE	CALA	SPOT	STAN	ZEMA
Arabic			✓				
Bulgarian			✓				
Cebuano			✓				
Chinese			✓			✓	
Danish			✓		✓		
Dutch			✓				(✓)
English	✓	✓	✓	✓	✓	✓	✓
French		✓	✓	✓	✓		(✓)
German		✓	✓		✓	✓	
Hindi			✓				
Hungarian			✓		✓		
Italian		✓	✓		✓		
Portuguese		✓	✓				
Romanian			✓				
Russian		✓	✓		✓		
Spanish		✓	✓	✓	✓		
Swedish		✓	✓				

Table V.4: Linguistic coverage of NER tools

[43] `https://opennlp.apache.org/`

[44] Moreover, the project seems to have come to a halt, with no new release over the past two years (version 1.5.3 having been released in April 2013).

2.1.2 SQuaRE analysis

Systems and software Quality Requirements and Evaluation (SQuaRE) is a family of international standards designed to evaluate the quality of software (ISO, 2011b). ISO/IEC 25010:2011 offers a typology of 8 quality characteristics further divided into 31 sub-characteristics, detailed in Table V.5. This framework has notably been used to evaluate the output of machine translation.[45]

Characteristic	**Sub-characteristic**
Functional suitability	Functional completeness
	Functional correctness
	Functional appropriateness
Performance efficiency	Time behaviour
	Resource utilization
	Capacity
Compatibility	Co-existence
	Interoperability
Usability	Appropriateness recognizability
	Learnability
	Operability
	User error protection
	User interface aesthetics
	Accessibility
Reliability	Maturity
	Availability
	Fault tolerance
	Recoverability
Security	Confidentiality
	Integrity
	Non-repudiation
	Accountability
	Authenticity
Maintainability	Modularity
	Reusability
	Analysability
	Modifiability
	Testability
Portability	Adaptability
	Installability
	Replaceability

Table V.5: Quality characteristics of SQuaRE

[45] See the portal of the FEMTI initiative: `http://www.isi.edu/natural-language/mteval/`.

As we mentioned in Section 1.1, there is necessarily a trade-off between these ideals, and favouring one at the expense of others highly depends on the actual needs of users. In Section 2.3, four systems will be evaluated: DBpedia Spotlight, Zemanta, Babelfy, and our MERCKX tool. Before conducting the quantitative evaluation on the output of these systems, Table V.6 provides a (necessarily subjective) qualitative assessment of their advantages and downsides along the lines of the eight main characteristics of SQuaRE.

Characteristic	**Spotlight**	**Zemanta**	**Babelfy**	**MERCKX**
Functional suitability	-	+	+/-	+
Performance efficiency	-	--	+/-	++
Compatibility	+/-	-	+	+
Usability	+/-	+	++	+/-
Reliability	+	+/-	+	+/-
Security	+/-	+/-	+/-	+/-
Maintainability	+	--	+/-	+
Portability	+/-	+/-	+	++

Table V.6: Evaluation of systems with SQuaRE

A 5-point scale is used to evaluate how the systems fare on the eight characteristics: ++ (excellent), + (good), +/- (average), - (poor), -- (terrible). Compatibility, usability, security, and maintainability are less critical since the end users would not have to operate the tools directly. In contrast, functional suitability, performance efficiency, reliability and portability are essential features in order to provide the users with quick, accurate results across different application domains. Table V.6 confirms that no system is perfect, although Babelfy does a pretty good job.

DBpedia Spotlight is a quite reliable tool which works transparently, making its components easily reusable. Although it suffers from poor results and efficiency, its average performance makes it a good candidate for establishing a baseline. Zemanta has a good functional correctness and is easy to use, but has important downsides in terms of efficiency (the user has to wait several seconds between each request) and its black-box technology means it cannot be properly analysed or modified at all. Babelfy offers an excellent interface, good interoperability thanks to a clean output format, and decent robustness and portability. Finally, the main advantages of our MERCKX system compared to the three other tools are its efficiency and portability, although its reliability is not yet optimal due to its low degree of maturity.

2.2 Methodology

A method is "un ensemble défini de procédures intellectuelles tel que quiconque, respectant ces procédures et posant la même question aux mêmes sources, aboutisse nécessairement aux mêmes conclusions" (Prost, 1996, p. 290). In this section, we present the objective of the evaluation process along with the metrics and corpus used, with a view towards reproducibility.

2.2.1 Objective

The aim of our evaluation is to answer the following three questions:

- How do entity linking systems score on the *Historische Kranten* corpus?
- Does MERCKX improve on the established baseline?
- Is information extraction heavily dependent on the quality of the text material or does it also work on poor OCR output?

To fully address these questions, two complementary types of evaluation should ideally be carried out: intrinsic (evaluation of the *correctness* of named entities) and extrinsic (evaluation of the *relevance* of the named entities to end users). The degree of correctness of an algorithm is measured with respect to certain specifications, which brings us back to the concept of *fitness for use* defined in the context of data quality (see Chapter IV, Section 1). In fact, "*selon le type d'information envisagé*, la question de la *correction* fait place à celle de l'*interprétation*" (Boydens, 1999, p. 129, italics hers).

However, the extrinsic evaluation requires an implementation that is not completed yet. We will therefore limit ourselves to the intrinsic evaluation for the time being, momentarily leaving aside the validation by end users but only to come back to it in the research perspectives described in our conclusions. Both types of evaluation are indeed necessary to get a comprehensive picture of the usefulness of an IE system.

For the present experiment, we focus on mentions of places, since Google Analytics statistics of the *Historische Kranten* website showed that they were especially favoured by users (see Chapter III): locations represent 60% of queries, compared to 6% for persons and 4% for organisations.[46] This tendency is consistent with the observations of Blanke et al. (2012) who note that "place names are often mentioned in the archival descriptions; researchers would like to be able to search for these locations, and place name extraction from the descriptions can help here".

[46]The remaining 30% being accounted for by concepts and rarer entity types.

In addition to our MERCKX system, we will evaluate three related tools already presented in Section 1.1: DBpedia Spotlight, Zemanta, and Babelfy. Spotlight is an obvious choice because it is one of the few services that does not operate as a black box: its inner workings are well-documented. It will provide us a baseline to measure other systems against. Zemanta, in contrast, is a closed-source NER tool, but we have shown in previous work (van Hooland et al., 2015) that it outperforms tools such as AlchemyAPI and Spotlight on English text. This experiment will allow to evaluate how well it adapts to other languages. Finally, Babelfy is a promising new player in the entity linking field. As a semantic annotation tool, it is not limited to named entities, but places are included in its output among several other types of content.

2.2.2 Metrics

Different measures are used in our experiment to evaluate the output of information extraction systems. We present those that were used, starting with precision, recall and F-score, before considering an alternative metric: slot error rate. Additionally, we discuss two different ways to apply these measures : simple entity match and strong annotation match.

The choice of an evaluation metric over another is not trivial: although quantitative analysis carries an aura of objectivity, numbers can be used to make empirical data appear deterministic, as emphasised in Chapter III. However, common denominators remain indispensable in order to compare systems efficiently. Metrics are therefore better understood as relative references than as absolute ones really describing the performance of a system.

Precision is the percentage of correct entities (*true positives*) among all the entities retrieved by the system. Entities mistakenly identified are called *false positives* and introduce *noise*:

$$\text{precision} = \frac{|\{\text{true positives}\}|}{|\{\text{true positives}\} \cup \{\text{false positives}\}|} \tag{V.1}$$

Recall is the percentage of correct entities retrieved among all those that should have been. Non-identified entities are called *false negatives* and cause *silence*:

$$\text{recall} = \frac{|\{\text{true positives}\}|}{|\{\text{true positives}\} \cup \{\text{false negatives}\}|} \tag{V.2}$$

When there is a risk of very low numbers of entities in a text (data sparsity), 1 can be added to the numerator and denominator (additive smoothing) to avoid an undefined division by 0.

The F-score or F-measure (van Rijsbergen, 1975) is a way to balance both measures, because obtaining either a very good precision (on a single entity retrieved) or an excellent recall (with lots of errors) would be meaningless. This harmonic mean includes a β variable that allows to give preponderance to precision or recall depending on one's needs, and is defined as:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{V.3}$$

In general we will want to give equal weight to precision and recall by considering that $\beta = 1$. The resulting formula is called F_1:

$$F_1 = \frac{2PR}{P + R} \tag{V.4}$$

Slot error rate Makhoul et al. (1999) acknowledged the usefulness of precision and recall as IE metrics, but showed that the F-score under-represents the overall error rate by about 30% by deweighting insertions (strict false positives) and deletions (strict false negatives) by a factor of two, thereby making systems look better than they are in reality.[47] To overcome this problem, they introduced a new metric called the *slot error rate* (SER), representing the true cost to the user in having a given system making errors. SER is defined as the ratio of the total number of slot errors (insertions, deletions, and substitutions) by the total number of slots in the reference:

$$\text{slot error rate} = \frac{|\{\text{insertions}\} \cup \{\text{deletions}\} \cup \{\text{substitutions}\}|}{|\{\text{slots}\}|} \tag{V.5}$$

Since the F-score is almost universally used in evaluation campaigns without being questioned as a valid metric, we found this critique relevant and original, and chose to include SER in our evaluation in Section 2.

[47] The third type of error being substitutions which are false positives and false negatives at the same time.

Simple entity match vs. strong annotation match According to Ruiz and Poibeau (2015), "the E[ntity] L[inking] literature has stressed the importance of evaluating systems on more than one measure". Thanks to the evaluation framework made available by the authors,[48] we became acquainted with the distinction between simple entity match (ENT), i.e. without alignment, and strong annotation match (SAM) which is stricter on entity boundaries (Cornolti et al., 2013):

> « 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. »

Both measures will be used to evaluate our results in Section 2.3 in order to get a more nuanced picture of what can be achieved by entity linking tools.

2.2.3 Corpus

As discussed in detail in Chapter III, the very nature of empirical data prevents us from drawing a bijection between the real object and its representation. In contrast to deterministic data where a model can be used as a reliable referent, no such referent can be used for interpretable content whose reality varies over time (Boydens, 1999, p. 144):

> « afin de vérifier la correction d'une valeur, il faut disposer d'un référentiel normatif. Or, dans un domaine d'application empirique, ce référentiel n'existe pas. En d'autres termes, il n'existe jamais de projection biunivoque nécessaire entre une représentation informatique et le réel observable correspondant. »

To overcome this limitation, a common practice is to create an artificial referent which is not quite similar to the real world, not quite perfect, but agreed upon by researchers as the next best thing in order to offer a stable base for evaluation. Such a pseudo-deterministic referent is called a *gold-standard corpus* (GSC), as already introduced in Section 2.1 of Chapter I.

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. (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 newspapers archives.

[48] `https://sites.google.com/site/entitylinking1`

Sample selection Since the *Historische Kranten* corpus contains 1 028 555 articles, we calculated with the help of an online tool[49] 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.[50] 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 disambiguated them with their corresponding DBpedia URIs, yielding a total of 662 locations in the following format:

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

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.

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:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{V.6}$$

[49] `http://www.surveysystem.com/sscalc.htm`

[50] Documents in the sample contain 1430 characters on average, which is comparable to the subset of the AQUAINT corpus used by Milne and Witten (2008).

$Pr(a)$ stands for the relative agreement between two raters and $Pr(e)$ for the probability of random agreement. 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 V.7 presents the raw annotation counts, along with the kappa by language.

Lang.	**Both**	**A**	**B**	**None**	**Tot**	**Pr(a)**	**Pr(e)**	κ
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 V.7: Cohen's kappa for our GSC

The average kappa of .911 shows a high agreement that is largely sufficient to consider the GSC reliable. This unusually 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 (supplied in Appendix B). 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       dbr:Bouvancourt
gsc3.txt       199       205       dbr:Fismes
gsc3.txt       561       565       dbr:Pévy
gsc3.txt       626       640       dbr:East_Riding_of_Yorkshire
gsc3.txt      1076      1082       dbr:Trigny
gsc3.txt      1145      1151       dbr:Muizon
gsc3.txt      1200      1205       dbr:Vesle
```

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 available on GitHub at `https://github.com/ulbstic/ypres/tree/master/gsc`.

2.3 Results and discussion

"Machines take me by surprise with great frequency."
Alan Turing (1950)

In order to compute the precision, recall and F-score of MERCKX against our GSC and to compare it to related systems, we used the *neleval* tool[51] which is a collection of Python evaluation scripts for the TAC[52] 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 (ENT versus SAM) to compare. The parametrised `scores.sh` shell script is shown below:

```
for measure in ent sam ; do
  if [ $measure = sam ] ; then
    lookfor="strong_link_match"
  else lookfor="entity_match"
  fi
  for corpus in raw corr ; do
    echo "** ${corpus} [${measure}] **"
    for system in Spotlight Zemanta Babelfy MERCKX; do
      echo "== ${system} =="
      ./nel evaluate -g ../gsc/res/ref/${corpus}-gold.mapped \
      ../gsc/res/sys/${corpus}-${system}-${measure}.mapped \
      | grep $lookfor$
    done
  done
done
```

Tables V.8 and V.9 present the results for simple entity match (ENT) and strong annotation match (SAM) respectively (see Section 2.2.2), with the best figures indicated in bold.[53] MERCKX outperforms the three other systems evaluated, except for precision where Zemanta scores best.[54] The columns marked "Raw" show the results obtained on the original GSC, while those marked "Corr" indicate scores obtained on corrected OCR. A paper describing preliminary results of this glass-box evaluation of the extraction component of entity linking systems has been accepted for publication (De Wilde, 2015).

[51] `https://github.com/wikilinks/neleval`
[52] `http://www.nist.gov/tac/`
[53] Since E and SER are error rates, the best scores for these metrics are the lowest.
[54] Consistently with results reported by Rizzo and Troncy (2011).

System	Precision		Recall		F_1-score		E = 1 - F_1		SER	
	Raw	Corr	Raw	Corr	Raw	Corr	Raw	Corr	Raw	Corr
Spotlight	.466	.468	.192	.207	.272	.287	.728	.713	1.028	1.028
Zemanta	**.887**	**.898**	.333	.371	.485	.525	.515	.475	**.709**	.671
Babelfy	.656	.688	.376	.446	.478	.541	.522	.459	.822	.756
MERCKX	.712	.744	**.488**	**.559**	**.579**	**.638**	**.421**	**.362**	**.709**	**.634**

Table V.8: Simple entity match (ENT)

System	Precision		Recall		F_1-score		E = 1 - F_1		SER	
	Raw	Corr	Raw	Corr	Raw	Corr	Raw	Corr	Raw	Corr
Spotlight	.235	.287	.190	.251	.210	.268	.790	.732	1.216	1.148
Zemanta	**.867**	**.888**	.278	.362	.421	.515	.579	.485	.766	.685
Babelfy	.662	.711	.321	.399	.433	.511	.657	.489	.852	.771
MERCKX	.782	.805	**.443**	**.517**	**.566**	**.629**	**.434**	**.371**	**.680**	**.610**

Table V.9: Strong annotation match (SAM)

2.3.1 Quantitative analysis

Precision is consistently ahead of recall, with Zemanta reaching scores between 85% and 90%. The harder task of strong annotation match (which takes 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. 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.

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 similar F-scores under 50%. Spotlight fares disappointingly, with F-scores around the 25% mark. The basic error rate E is calculated by substracting the F-score from 1: since MERCKX has a 58% F-score, the remaining 42% are mistakes. As discussed in Section 2.2.2, however, a better approximation of the true cost for users can be attained with the slot error rate metric.

Contrary to the other systems, MERCKX reduces its error rate on the SAM task. Spotlight reaches error rates over 1: Makhoul et al. (1999) concede that "Some may feel uncomfortable with the notion of an error rate that is greater than 100%, but this possibility is not as unreasonable as it might appear at first glance." In fact, a system that produces nothing is bad enough, but a system that produces only false positives (insertions) is arguably even worse. Results on corrected OCR (columns marked "Corr") will be discussed separately in Section 2.3.3.

2.3.2 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. (1999): insertions, deletions, and substitutions.

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 `http://fr.dbpedia.org/resource/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 (except as an ancient way to go there), 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.

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.[55]

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 tokenised separately) can cause havoc.

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: "Vitry-le-François" instead of "Vitry-sur-Seine".

2.3.3 Impact of OCR

In similar work on Holocaust testimonies, Rodriquez et al. (2012) found that "manual correction of OCR output does not significantly improve the performance of named-entity extraction". 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.

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. Slot error rates also regularly decrease by up to 8% (although it does not seem to affect the SER of Spotlight on the ENT task).

[55] 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. While helping to track evolution of place names over time, the danger of confusion with the French determiner *cette* is obvious. This question will be further discussed in the research perspectives put forward at the end of the thesis.

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 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% score. A new testing on the second OCR performed by Picturae with version 10 of ABBYY FineReader (see Chapter IV, Section 1.2) would be necessary to determine if about 90% word accuracy (instead of 82%) would be sufficient to achieve better entity linking.

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 a quadratic time complexity.

3 Validation

In this section, we will start by looking at the limits of existing search capabilities (Section 3.1) in order to improve on them and to offer our own concrete applications (Section 3.2), before transposing our model to other languages, domains and entity types (Section 3.3) to test its portability.

3.1 Beyond search engines

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:

Example 31. Links zien wij Christiaan Huygens die met zijn slingeruurwerk de oplossing bracht voor het meten van de tijd

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

Example 32. 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

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

Example 33. [...] en op het uurwerk toegepast door den Hollander Huyghens (1629-1695).

A correct disambiguation with DBpedia URI `dbr:Christiaan_Huygens` would include 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)`. Disambiguating results is something out of scope for most search engines, as noted by Bade (2008, p. 34):

> "The chief difference between the library catalogue and Google is that the material described in a library catalogue is a deliberately limited set of materials which have been selected by subject specialists to be included in the library because of their value for research and (hopefully) described by persons sharing the intellectual commitments and scholarly vocabulary of the authors of those materials. None of this can be said of the items retrieved via a Google search. The Google search permits no filter (other than the hidden algorithms!) between the information and the user, and hence 426,000 responses to a query."

Providing these filters, in the form of semantically-enriched content to be browsed and faceted by users, should therefore be the leitmotiv when building advanced knowledge discovery applications making up for the shortcomings of full-text search. Transparency is also important in order not to reproduce the black-box technology of commercial search engines. Figure V.7 illustrates the path from a corpus of documents to discovered knowledge.

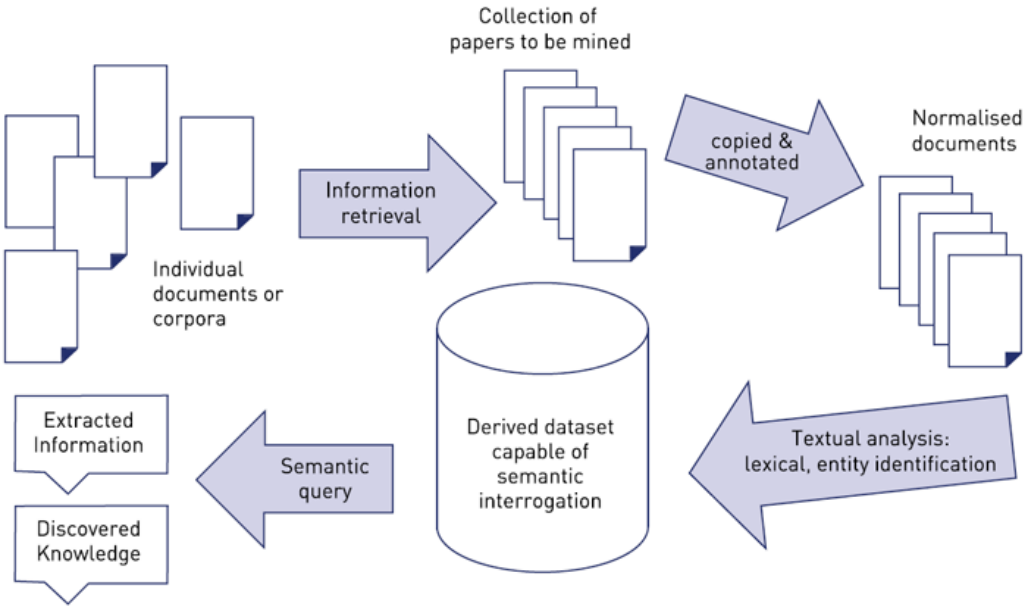


Figure V.7: From corpus to knowledge, reproduced from JISC (CC BY-NC-ND)

This structural shift raises the question of the role of traditional documents in a knowledge-driven society. In our introduction, we have defined knowledge as "a fluid mix of framed experience, values, contextual information, expert insight and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information" (Davenport and Prusak, 1998, p. 5). This practical definition allows us to include all kinds of facts contained in a knowledge base.

Defining what constitutes a document is trickier. In a survey on this question, Buckland (1997) quotes an early attempt at standardisation by the International Institute of Intellectual Cooperation in 1937 defining a document as "any source of information, in material form, capable of being used for reference or study or as an authority". Interestingly, "source of information" is rendered as "base de connaissance" in the French translation provided by the Union Française des Organismes de Documentation.

Briet (1951) had devised an even broader definition by considering that a document is "tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène physique ou intellectuel". But the key feature of documents, as Havelange (2014) reminds us, is to convey an educational value (from Latin *docere, doceo*, I teach). In order to teach us something, however, a document must be recognised as such: "un document [...] n'existe pas 'en soi', mais dans le cadre, seulement, du dispositif de savoir ou d'expression qui le convoque en cette qualité de document" (Havelange, 2014).

Anything, in other words, can become a document if used in a way to suit a given purpose. While information extraction progressively ousted the document as the primary informational unit in favour of smaller knowledge items such as entities and facts (materialised on the Web of Data in the form of RDF triples), documents in turn became abstract entities with their associated metadata, allowing them to be manipulated, automatically summarised and linked to other documents in a semantically meaningful way.

For Stern (2013, p. 38), "l'Annotation Sémantique promeut le document et son contenu textuel comme objet central dans le Web Sémantique, et se place ainsi dans la lignée du T[raitement] A[utomatique des] L[angues], qui s'intéresse de façon primordiale à ces objets". For Blanke and Kristel (2013) however, "the traditional distinction between collection-level and document-level documentation is disappearing fast in a digital environment". In a Big Data context, the metadata of field practitioners become the raw data of computer scientists, and the distinction between the two is blurred.

But is there an intrinsic difference between a traditional document and a digital document? Buckland (1997) considers that "an emphasis on the technology of digital documents has impeded our understanding of digital documents as documents". This observation validates the reflections on the Hype cycle already put forward in Chapter III, and strengthens the premise of van Hooland (2009, p. 2) stating that "innovative technologies offer new possibilities for metadata creation and management, but can also have a negative impact upon the quality of metadata".

3.2 Applications

In this section, we present two concrete applications of knowledge discovery: search suggestions based on semantic proximity on the one hand, and the automated exploitation of related resources on the other hand. Both applications have the potential to enrich the *Historische Kranten* website by providing the end users with a more comprehensive picture of the content available.

3.2.1 Search suggestions

The search engine used by the *Historische Kranten* website relies on Apache Lucene.[56] Its Solr implementation provides a module called MoreLikeThis which constructs a lucene query based on terms vectors within a document, in order to provide users with related content suggestions. However, this module fared very poorly on the website and was subsequently removed by Picturae developers due to performance issues.

We propose to use an alternative approach based on semantic similarity measures introduced in Chapter II. The DBpedia FindRelated service[57] is an example of such a tool allowing to compute symmetric or asymmetric distances between DBpedia resources with a cut-off threshold to limit the number of results. For instance, a search for resources related to Poperinge using the symmetric model and a threshold of 0.2 yields the following results in XML format:

```
<results>
 <resource>http://dbpedia.org/resource/Poperinge</resource>
 <model>symmetric</model>
 <threshold>0.2</threshold>
 <result>
  <resource>http://dbpedia.org/resource/Anne_Provoost</resource>
  <distance>0.16370000505196158</distance>
 </result>
 <result>
  <resource>http://dbpedia.org/resource/Jef_Planckaert</resource>
  <distance>0.18771837736522878</distance>
 </result>
 <result>
  <resource>http://dbpedia.org/resource/Proven</resource>
  <distance>0.18771837736522878</distance>
 </result>
 <result>
  <resource>http://dbpedia.org/resource/Reningelst</resource>
  <distance>0.18771837736522878</distance>
 </result>
</results>
```

[56] http://lucene.apache.org/

[57] http://wiki.dbpedia.org/services-resources/find-related

These results fall into two categories: people and places. People include writer Anne Provoost and cyclist Jef Planckaert who were both born in Poperinge, while places Proven and Reningelst are subdivisions (*deelgemeenten*) of the municipality. Including these related resources in the form of search suggestions allows to broaden the range of documents retrieved, since newspaper articles about Reningelst are bound to interest users querying the website about Poperinge.

Another such tool is DBpedia Spotlight's Rel8[58] which relies on distributional similarity but does not provide customisation parameters like FindRelated. Unfortunately, the web demo appears to have been down for some time, but Mendes and Jakob (2013) provide an overview of results that could be obtained when searching for resources related to the Scala programming language, for instance:

```
[
  {"Clojure": 1.447},
  {"Groovy_(programming_language)": 1.306},
  {"Objective_Caml": 1.281},
  {"Go_(programming_language)": 1.233},
  {"Erlang_(programming_language)": 1.229},
  {"D_(programming_language)": 1.147},
  {"Racket_(programming_language)": 1.127},
  {"F_Sharp_(programming_language)": 1.121},
  {"Ruby_(programming_language)": 1.07}
]
```

Their relatedness scores are a combination of cosine similarity between resources, represented as vectors in a Vector Space Model, and of a neighbourhood measure based on the *wikiPageLinks* dataset (Mendes and Jakob, 2013):

> “ In this case, a resource $r1$ is in the neighborhood of $r2$ if there is a property connecting $r1$ and $r2$. After aggregating all resources in the neighborhood of a set of query terms, resources that are more related to all terms should appear more often. ”

Search suggestions based on distance and relatedness would allow users of the *Historische Kranten* website to discover more about their topics of predilection and to enrich their queries seamlessly with semantic content.

[58] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Rel8`

3.2.2 Related resources and data visualisation

In order to demonstrate the practicability of the approach proposed in this dissertation, a prototype is provided at `http://mastic.ulb.ac.be/ypres/`. This portal includes two proof-of-concept applications offering data visualisation and knowledge discovery related to the *Historische Kranten* corpus.

The Place Browser allows to visualise on a map the prominent locations used in the corpus and to learn more about them. As seen on Figure V.8, most of these locations are cluttered in the Westhoek region.[59] Dot sizes are proportional to the prominence of the locations in the corpus. By hovering the mouse over a dot, the user can see the number of potential documents about the associated place and click it to access relevant information. This application is based on the Google Geochart visualisation technology.[60]

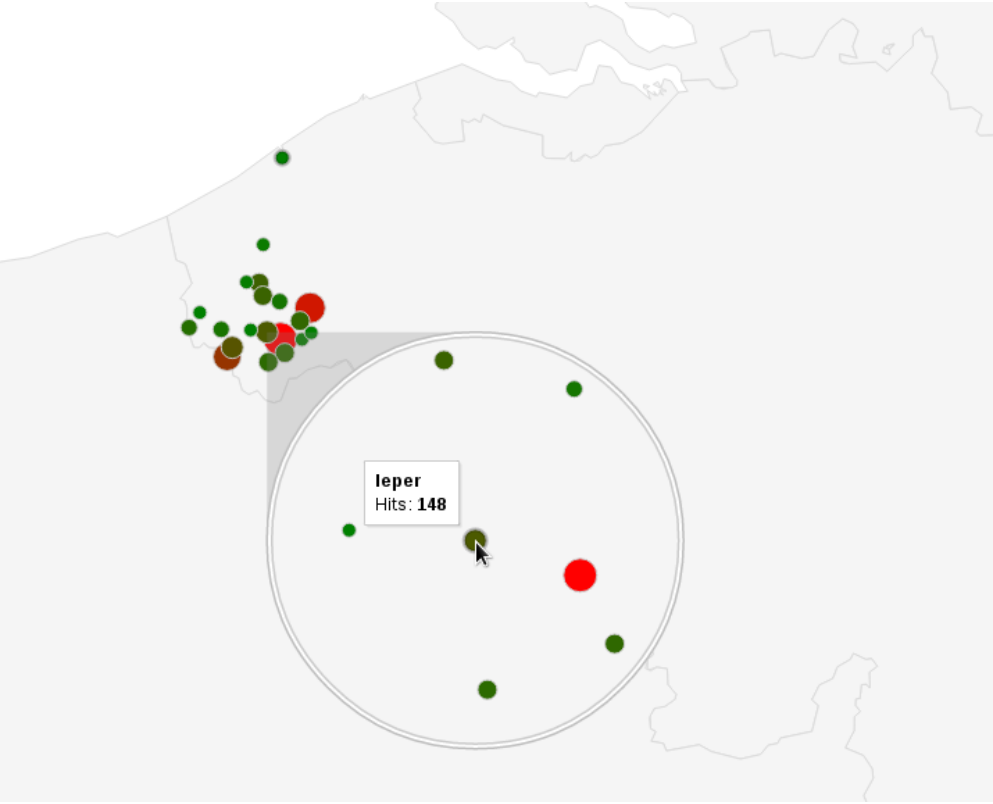


Figure V.8: Place Browser

[59] See `http://www.westhoekverbeeldt.be/`.
[60] `https://developers.google.com/chart/interactive/docs/gallery/geochart`

The markers for each location are automatically plotted on the map from their coordinates, extracted from DBpedia once they have been disambiguated by MERCKX. A sample of the resulting JavaScript code is provided below:

```
var data = google.visualization.arrayToDataTable([
    ["Lat", "Lng", "Place", "Hits"],
    [50.8345035, 2.9225399, "Zillebeke", 398],
    [50.9002575, 3.0207328, "Passendale", 333],
    [50.7971484, 2.7464372, "Westouter", 252],
    [50.8170545, 2.7634047, "Reningelst", 163],
    [50.8492265, 2.8779465, "Ieper", 148],
    ...
]);
```

The demo app does not yet use the total numbers of occurrences from the corpus, but rather the hit counts from the locations most commonly searched by users (see Chapter III). The links behind the dots currently point to the Dutch Wikipedia, but they could easily be replaced by other resources once the tool is implemented on the *Historische Kranten* website.

The People Finder (Figure V.9) proposes to learn about famous people born in the Westhoek (or any other place for that matter) in order to arouse interest in important historical figures mentioned in the periodicals.

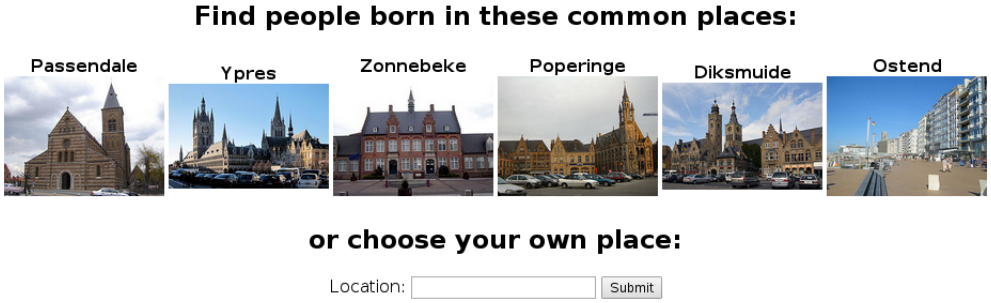


Figure V.9: People Finder

This tool, coded in PHP, dynamically queries DBpedia in order to find all persons linked to a given place and retrieves additional information about them, along with their pictures. The full code is available on GitHub,[61] and a sample result for Ypres is shown in Figure V.10.

[61] https://github.com/ulbstic/ypres/tree/master/people

Figure V.10: Discovering related entities

3.3 Generalisation

Portability is the last, but not the least, of the eight criteria listed by the SQuaRE ISO standard (see Section 2.1.2). It is indeed an important dimension to demonstrate the capacity of a tool to be used across a wide range of contexts. But if, as we have suggested in chapters III and IV, empirical objects escape any form of deterministic formalisation, then how can our approach be properly generalised? In other words, how can we be sure that the method and tools we have proposed can be adapted successfully to other uses, and that the results obtained can be replicated? Hermeneutics, the science of interpretation, helps us to break this deadlock (Boydens, 1999, p. 470):

> “ L'herméneutique nous enseigne que si les réalités d'ordre humain ou social ne peuvent faire l'objet d'une approche déterministe, il est néanmoins possible de les appréhender dans une perspective généralisante. [...] [C]es enseignements révèlent qu'en l'absence de critère de validation déterministe, une approche interprétative permet d'obtenir des résultats opérationnels [...] ”

For Elias (1996, p. 99), "les problèmes du temps ne se laissent pas ranger dans des cases correspondant à la répartition des disciplines scientifiques actuellement prévalente et à la compartimentalisation qui en découle de notre appareil conceptuel". Our field of application remains within the humanities, but can it be generalised to other collections of data? The danger of in-depth analysis of empirical objects is in fact to be so restrictive as to become completely unrepresentative of the field as a whole.

In this last section, we will test the portability of our methodology in three ways. We first evaluate MERCKX on other languages (Section 3.3.1), then on various application domains (Section 3.3.2), and finally on different entity types (Section 3.3.3). To meet these goals, we will use two additional corpora: the Pentaglossal corpus (mixed collections of documents in five languages including Russian and Chinese) and the Perelman archive (French and English correspondence), which will be described in beside the results obtained.

3.3.1 Other languages

In our experiment, we focused on three languages: English, French, and Dutch. Since two of them are Germanic languages and all three of them Western European languages, it is still unclear how well MERCKX would generalise to content from outside these linguistic groups. The multilingual structure of DBpedia should in theory ensure that our approach is portable to any of the 128 languages it covers, but this hypothesis has yet to be tested.

To do so, we introduce the Pentaglossal corpus,[62] a "parallel corpus comprising 113 texts in five languages, namely, English, French, German, Russian, and Chinese" (Forsyth and Sharoff, 2014). The documents range from the very ancient (Old Testament extracts) to the very new (recent news items) and thus cover a broad period of time, exemplifying language evolution.

The main advantage of this corpus is to include non-Western languages – Russian and Chinese – thereby allowing to generalise the findings made for the three languages used in the Ypres corpus. The parallel structure of the Pentaglossal corpus also makes it easy to check the validity of the information extracted without mastering these languages. Texts are classified into thirteen categories, as shown in Table V.10, each text having a translation equivalent in the other four languages. The corpus also has the advantage of being very diverse, contrary to other parallel corpora such as Europarl (Koehn, 2005), which is "homogeneous in terms of its topics and genres, [making] it difficult to generalize any results obtained from it" (Forsyth and Sharoff, 2014).

[62] `http://corpus.leeds.ac.uk/tools/`

Code	Docs	Tokens	Description
Bib1	5	5 503	Bible, Old Testament extracts
Bib2	6	10 140	Bible, New testament extracts
Corp	6	5 074	Corporate statements of self-promotion
Fict	30	138 704	Fiction: novel chapters or short stories
Marx	5	31 499	Marxist documents
News	10	7 078	News articles
Opac	3	3 766	Open access declarations
Tedi	11	22 758	Transcripts from Ted.com initiative
Tele	14	44 856	Telematics, engineering
Teli	1	2 733	Telematics, instructions
Tels	15	8 974	Telematics, software
Unit	4	19 205	United Nations documents
Wind	3	7 417	Wind energy articles

Table V.10: Pentaglossal corpus, adapted from Forsyth and Sharoff (2014)

In order to isolate the language variable, we evaluated MERCKX on a similar domain (News) and with the same type of entities (places). Since the Pentaglossal corpus does not provide a gold-standard corpus for entity linking, we used the output of MERCKX on the ten English texts as a reference, and compared the four other languages against it, exploiting the exact parallel between file names. Table V.11 shows the results obtained.

Language	Precision	Recall	F_1-score
French	.723	.667	.694
German	.532	.647	.584
Russian	.600	.294	.395
Chinese	.194	.255	.220

Table V.11: Generalisation to other languages

While the absolute scores do not necessarily reflect reality since no human annotation was performed upstream, proceeding in this way still allows for an objective comparison of languages. A similar methodology was used in the Collaborative Annotation of a Large Biomedical Corpus (CALBC) challenge[63] for instance, where a "silver" standard was computed in retrospect from the common output of participating systems rather than produced by annotators.

[63] http://www.ebi.ac.uk/Rebholz-srv/CALBC/

We observe an inverse correlation between F-score and linguistic complexity. While the extraction of places with MERCKX reaches almost 70% on French news, it stays around 60% for German (with lower precision due to the capitalisation of common nouns), 40% for Russian (handling the Cyrillic characters efficiently at 60% precision but lacking in recall due to the lack of coverage of the Russian DBpedia) and only just over 20% on Chinese (which is challenging in terms of tokenisation due to the absence of word boundaries).

In a nutshell, we can say that MERCKX is language-agnostic since it does not include any language-specific component, but not yet truly language-independent in the sense of Bender (2011, see Chapter IV), since the language variable still affects the results significantly. This can be partly explained by the discrepancies between the various linguistic chapters of DBpedia.

3.3.2 Other domains

In the words of Maturana et al. (2013), "Linked Data expresses the mechanical possibilities of discovering knowledge related to almost every domain of human interest". Exploiting these possibilities is essential to attain a higher degree of understanding of the vast collections of data present on the Web.

Although the Pentaglossal Corpus was mainly designed to compare the performance of systems on different languages, its topical diversity also makes it a useful referent to evaluate domain-independence. Table V.12 shows the results of MERCKX on the thirteen domains.

Domain	**Precision**	**Recall**	F_1**-score**
Bib1	.467	.333	.389
Bib2	.407	.550	.468
Corp	.444	**1**	.615
Fict	.551	.667	.603
Marx	.385	.588	.465
News	.766	.692	.727
Opac	.400	.667	.500
Tedi	.708	.872	**.782**
Tele	.111	.200	.143
Teli	.429	**1**	.600
Tels	.333	.364	.348
Unit	.294	.556	.385
Wind	**.769**	.769	.769

Table V.12: Generalisation to other domains

To isolate the domain variable, the evaluation was performed on a language from the *Historische Kranten* corpus: French (Dutch being absent from the Pentaglossal corpus and English used as a reference again), with the place entity type. Surprisingly, the best results are not achieved on news: precision peaks at 77% for wind energy articles and recall reaches 100% for corporate statements and telematics instructions, which offer very different content (although the number of locations mentioned is admittedly small: 4 and 3 respectively), while TED talks yield the best F-score just over 78%.

To further validate our findings and evaluate the relevance of extracted places, an additional experiment was conducted on yet another corpus: the correspondence of Belgian logician and philosopher Chaïm Perelman. The Perelman archive, maintained by the Research Group in Rhetoric and Argumentation of the Université libre de Bruxelles, consists of 42 boxes, representing about 30 000 to 40 000 sheets of paper.[64]

Just under 12 000 of these represent carbon copies of the total outgoing correspondence of Perelman with academics and the wider intellectual world, mixing administrative and scientific letters. The collection roughly covers a quarter of a century, with the oldest letters dating back to 1960 while the newest were produced just a month before Perelman's death on 22 January 1984. The details of the correspondence figures can be found in Table V.13.

Year	**# pages**	**Year**	**# pages**
1960	565	1972	543
1961	5	1973	707
1962	345	1974	552
1963	547	1975	577
1964	239	1976	499
1965	452	1977	591
1966	458	1978	544
1967	646	1979	393
1968	611	1980	565
1969	859	1981	468
1970	671	1982	0
1971	647	1983	395
TOTAL			**11 879**

Table V.13: Perelman's correspondence volume

[64] `http://gral.ulb.ac.be/archives-chaim-perelman`

The corpus consists of one PDF per year, with letters in reverse order. OCR was performed with ABBYY FineReader 11 at the ULB library of social sciences. However, the quality of the carbon paper and a lack of parametrisation led to even poorer OCR output (word accuracy of 57.5% and character accuracy of 68.8% on a tested sample) than in the case of the *Historische Kranten* corpus. The name "Perelman", for instance, appears in 740 different spellings, although some are much more common than other. Table V.14 shows the 10 more frequent ones with their number of occurrences, and percentage of the total.

spelling	**#**	**%**
Perelman	3100	43.7
Perelaan	701	9.9
Paralaan	483	6.8
Perelnan	310	4.4
Paralman	132	1.9
Paralnan	115	1.6
Perelraan	114	1.6
Ferelman	110	1.5
Pereloan	76	1.1
Parelaan	62	0.9

Table V.14: Variations of Perelman's surname due to OCR errors

MERCKX extracted 12 586 mentions of 951 different places from the 11 879 pages in the correspondence. The most common ones are listed in Table V.15, sorted by order of decreasing frequency. Unfortunately, there is no structured reference of the places of residence of Perelman's correspondents against which to check the validity of the extraction process, but most of these locations are present in the summary of the manual inventory of the collection.[65]

A thorough qualitative analysis nonetheless shows that the results contain some errors. One of the obvious missing entities (false negatives) is "Brussels" which appears almost 3 700 times in the letters and should therefore precede Paris in the table. This striking absence is due to a technicality in DBpedia: the French label for `dbr:Brussels` is not "Bruxelles" but "Région de Bruxelles-Capitale", which of course never appears in its full form in the correspondence of Perelman. As already mentioned in Section 2.3, including redirections in addition to labels would solve this issue pretty straightforwardly.

[65] `http://perelman.ulb.be/sites/default/files/inventaire89pp_final.pdf`

Place	#	Place	#
Paris	1073	Canada	147
France	454	Sydney	146
Belgium	406	Switzerland	126
Jerusalem	345	Mexico City	125
New York City	342	Chicago	105
Israel	241	Liège	101
Leuven	233	Europe	93
Tours	214	Geneva	90
Italy	188	Germany	89
Poland	167	Turin	89
Montreal	148	Oxford	82

Table V.15: Most frequent places in Perelman's letters

There is also some amount of noise in the results (false positives). Bruxelles is often abbreviated "Brux." by Perelman, but Brux happens to be a French commune in the Vienne department, causing mentions of it to be mistakenly disambiguated to `dbr:Brux`. Notwithstanding these few errors, most locations extracted look correct and could be used to draw a map of Perelman's scientific collaborations. This illustrates the fact that MERCKX is not limited to newspaper articles but can be extended to any type of content.

3.3.3 Other entities

Finally, we should evaluate how well MERCKX adapts to other entity types such as Persons and Organisations. In theory, any class from the DBpedia ontology[66] could be used as a valid category in order to filter entities. In practice, however, some of these classes are not sufficiently populated to achieve decent recall. For instance, the `dbo:PhilosophicalConcept` class would be of great interest to the users of the Perelman archive, but it appears to be completely empty. Even its parent class `dbo:TopicalConcept` only contains 3 466 elements, compared with 725 546 places for instance.

It appears that our sample of 100 texts from Ypres contains very few organisations. MERCKX is able to disambiguate a mention of "Red Cross" to `dbr:International_Red_Cross_and_Red_Crescent_Movement`, for example, but the number of entities extracted is not high enough to compute precision and recall reliably.

[66] `http://mappings.dbpedia.org/server/ontology/classes/`

More organisations can be found in the News subset of the Pentaglossal corpus. As already done for languages and domains, we used the English as reference and evaluated the output of the extraction on French. MERCKX achieved a precision of 79% and a recall of 56%, yielding a F-score of 65.5% on a total of 34 mentions from 10 texts (compared with an F-score of 72.7% for places, see Section 3.3.2). Table V.16 illustrates the diversity of the organisations extracted.

Organisation	**# mentions**
FC Barcelona	12
Fidesz	4
European Parliament	3
Goldman Sachs	2
State Duma	2
Twitter	2

Table V.16: Organisations from the Pentaglossal News

Famous persons are tantamount to absent from both the Ypres sample and the Pentaglossal News, but they appear quite often in Perelman's letters. The correspondents of Perelman include many different people: the *index nominorum* manually built by researchers identifies over 2 000 of them. Only a fraction of them have a DBpedia page, though, which means that recall will necessary be quite low. In its current form, MERCKX was able to extract 90 mentions of 39 people. Table V.17 displays the most frequent ones.

Person	**# mentions**
Eugène Dupréel	18
Max Black	7
Nicholas Rescher	5
Philippe Devaux	5
Marvin Farber	4
Abraham Kaplan	3
Michel Meyer	3
Raymond Aron	3
Ruth Barcan Marcus	3
Wayne C. Booth	3

Table V.17: People mentioned in Perelman's contacts

A concrete application,[67] using Geochart again, is the visualisation of the geographical dispersion of Perelman's correspondents. Figure V.11 shows their concentration across Europe and the East Coast of the United States:

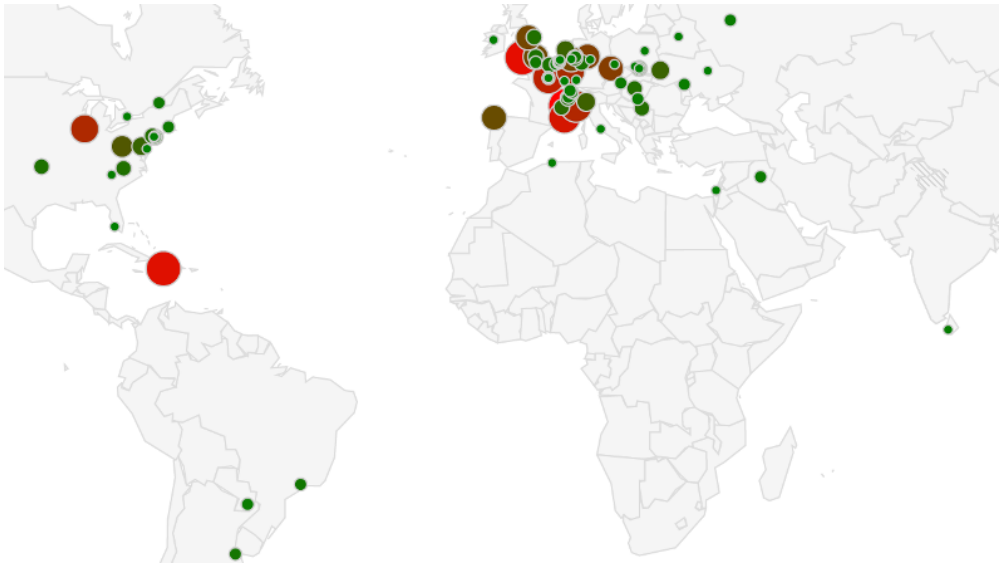


Figure V.11: Geographical dispersion of Perelman's correspondents

Finally, MERCKX also allows to look for all DBpedia resources without filtering them by type. While this is very difficult to evaluate with a GSC since virtually every word would have to be annotated with a URI, the field survey from Chapter III showed that users of the *Historische Kranten* website were interested in concepts such as "war" and "murder". The extension to any type of content, illustrated in Table V.18, truly demonstrates the portability of our approach.

Concept	**# mentions**
Civil law notary	17
Profession	15
Pharmacy	14
Price	14
Time	14
Province	11
War	4

Table V.18: Concepts extracted from the Ypres sample

[67] Demo accessible at `http://mastic.ulb.ac.be/perelman/`.

Summary

In this last chapter, we came full circle by redefining information extraction with semantic enrichment as knowledge discovery. The material introduced gradually in the previous chapters was here eventually connected and structured in a meaningful way, producing a clear picture of the achievements of this dissertation.

Several tools were surveyed in order to grasp their functionalities, but also their limits, especially in handling multiple languages in an efficient manner. Named-entity recognition systems, first introduced in Chapter I, traditionally stopped after the recognition of an entity's boundaries and its classification into broad semantic categories. Although some NER tools now go further by providing full disambiguation with LOD resources (entity linking), this evolution remains largely ignored by mainstream computational linguistics, which also continues to regard entities and terms as irreducibly different language units. Building on more advanced semantic annotation tools, we therefore proposed a system of our own called MERCKX, based on the intuitions behind our four research questions delineated in the introduction.

In order to compare the results we obtained with state-of-the-art entity linking systems, we first performed a preliminary assessment of four tools (DBpedia Spotlight, Zemanta, Babelfy, and our own MERCKX) in the light of linguistic coverage and the SQuaRE ISO standard criteria, before defining our methodology in terms of objectives, metrics and corpora. The results show that MERCKX outperforms existing systems for recall and F-score on both the tasks of simple entity match (without alignment) and strong annotation match (with strict character alignment), although Zemanta offers a better precision. Manual correction of the OCR output was also shown to improve the entity linking process by about 5%, contrary to expectations from the literature.

After reminding the stakes of renewing the old information retrieval model in the face of information overload, we proceeded to imagine proof-of-concept applications that could be implemented in the context of the *Historische Kranten* project, but could also be generalised to other domains where empirical and multilingual content is prominent. In doing so, we strived to provide collection owners not only with theoretical recommendations as to how value can be added to digitised content in an automated way, but also with practical, semantic enrichment code that can be easily adapted to meet a whole range of unforeseen requirements.

Conclusions

Outline

To conclude this dissertation, we focus on the past, the present, and the future. The past looks in retrospect at our journey which is now almost completed; the present takes stock of our various achievements and of their limitations; the future foresees what remains to be done and how we could accomplish it. These levels of temporality materialise in three complementary sections.

We start in Section 1 with a detailed summary of the five chapters in order to emphasise the progression from our starting point up to the finishing line. Since several discoveries were made along the way, we dwell on each of them to sketch a comprehensive overview of the contents of the thesis and of its operational consequences.

Section 2 recapitulates the principal outcomes, starting with key findings and tentative answers to the four research questions raised in introduction. Crucially, we emphasise the limitations of our work and justify several aspects that we deliberately chose not to take into account. We also offer a number of recommendations for projects contemplating the semantic enrichment of their content, in the humanities or elsewhere.

Finally, Section 3 paves the way for further research and identifies issues that remain to be tackled in the future, ending with some open questions. These perspectives include the practical implementation of a semantic search engine for the *Historische Kranten* project, but also new ideas of application domains that could benefit from the approach developed in this dissertation, along with an extension of our temporal framework to tackle the evolution of concepts.

Contents

1 Overview

In this first section, we offer a summary of the material covered in the five chapters of the present work. This dissertation can be seen as a journey from information extraction to knowledge discovery through the twists and turns of the Semantic Web and entity linking, decoding the map of empirical data and the digital humanities, while overcoming the obstacles of data quality, multilingualism and language evolution.

Every step brought valuable insight in support of our main thesis, which advocated generalisation over specialisation as a sound, productive approach to semantic enrichment and knowledge discovery, thereby arguing in favour of the interconnection of disciplines and against the development of information extraction techniques specialised for a single language, domain, or type of content. The contribution of each chapter is detailed below.

Information Extraction

Our journey started in Chapter I with a historical glance at the achievements of the last few decades in the automated processing of unstructured content. Because natural languages are intrinsically ambiguous, getting machines to "understand" them has been a major challenge ever since the birth of artificial intelligence. Natural language processing – and information extraction in particular – developed into independent fields of research with the declared goal of tackling this issue.

The refinement of information extraction techniques, mainly named-entity recognition (NER) but also other tasks such as relation detection and event extraction, has allowed to represent the content of documents in an increasingly satisfying manner, although much work has been invested in English to the detriment of other languages, and full entity disambiguation has remained elusive. To overcome these limitations, fresh input from elsewhere was required, which set us on an epistemological quest.

In its historical meaning, epistemology refers to a branch of philosophy studying the nature of knowledge and ways to acquire it in a useful manner. Investigating knowledge acquisition techniques from different perspectives helped us to overtake the somewhat technocentrist approach of traditional information extraction, and gave a new breadth to the disambiguation task. But in order to achieve this ideal, the resort to comprehensive multilingual resources was required.

Semantic Enrichment with Linked Data

Chapter II introduced the Semantic Web and Linked Data as good practices to transform the loose Web of documents into a new Web of structured data. Building on the work of visionaries such as Paul Otlet, Tim Berners-Lee and others imagined a second generation of technologies that could be used by machines in order to process the contents of the Web without any amount of human intervention.

Relying on W3C standards such as XML, RDF, OWL, and SKOS, the Semantic Web unfortunately did not live up to Berners-Lee's expectations, but nevertheless resulted in an myriad of new semantic resources, from knowledge bases to ontologies. One of the specificities of these references is to identify each object and property with a uniform resource identifier, allowing in theory to represent them unequivocally, although quality issues can (and will) arise.

By exploiting these resources, we were able to overcome the limitations of a classic approach to information extraction: the task of entity linking goes one step further than NER and provides a full disambiguation of entities, along with links to related information present in the Linked Open Data cloud, allowing to retrieve additional knowledge. Newly equipped with these techniques, we proceeded to confront them against real-world material.

The Humanities and Empirical Content

In Chapter III, we focused on the specificities of empirical data, as opposed to deterministic facts. The empirical nature of the humanities implies that their objects of study are necessarily subjective, despite claims to the contrary by positivist scholarly trends, of which the digital humanities are but the latest embodiment.

The enthusiasm for quantitative approaches in the humanities is not new, but it reached a new high with the apparition of the concept of distant reading and the forecast about the end of theory in a world ruled by Big Data. While guarding ourselves from the dangers of using computing techniques at all costs, we assessed what could be gained by leveraging existing resources from the Web of Data through the mechanism of semantic enrichment.

To confront theory with practice, we applied this intuition to a real-world case study: the *Historische Kranten* corpus – a million-document, trilingual (Dutch/French/English) archive. A special attention was paid to the search interests of users, ensuring that our methods and tools were never disconnected from the actual needs of the field, while providing constant feedback through the "boomerang of reality" mechanism.

Quality, Language, and Time

Chapter IV dealt with three particularly pervasive issues, related to the quality of data and resources, the development of multilingual approaches, and the evolution of concepts over time. We saw that data quality can never be defined in absolute terms but is relative to usage, in agreement with the *fitness for use* principle. Applied to uncurated collections and knowledge bases, this conception of quality allowed us to focus on the implications of adopting Linked Open Data for cultural institutions, in a cost-benefit perspective.

The proliferation of languages is another pitfall affecting many applications, designed for English only whereas our globalised world in general – and the Web in particular – is increasingly multilingual. Mindful not to lock ourselves into an over-specialised approach that would bear little significance outside this specific project, we stressed the importance of portability to other languages, contemplating the possibility of generalising our findings.

Furthermore, natural language is not static but constantly evolves over time, with terms changing their meaning and new concepts appearing continuously. This makes ambiguities even more tricky to manage, especially in the context of empirical content scattered over a long time period. Possible solutions to this problem will be envisioned in Section 3.3.

Knowledge Discovery

Taking the previous elements into consideration, Chapter V strived to converge towards an integrated solution, going beyond information extraction in its limited sense in favour of knowledge discovery, i.e. the automated exploitation of new facts related to a user's query thanks to the semantic enrichment of documents with Linked Data. After surveying several tools, we proposed our own proof-of-concept system called MERCKX (Multilingual Entity/Resource Combiner and Knowledge eXtractor) to extract and disambiguate relevant terms and entities from text and enrich content automatically with additional information instantly retrieved about these resources, thanks to ontology properties and semantic relatedness heuristics.

Importance was given to evaluation standards and practices in order to compare our system with state-of-the-art applications. The results on both raw and corrected OCR showed that MERCKX outperforms existing entity linking tools by about ten percentage points, although much remains to be done to improve its precision (through clustering algorithms) and recall (through the exploitation of more LOD relations) on the one hand, and to achieve a better robustness on poor OCR output on the other hand.

Finally, this work did not stay at the theoretical level. To demonstrate the operational character of our approach, we presented a few applications that could be implemented on the *Historische Kranten* website, but also generalised to any other project sharing its core characteristics: a vast collection of documents, digitised empirical material (with or without OCR), multiple languages intermixing, and a broad historical period to take into account.

2 Outcomes

"Experience [...] was merely the name men gave to their mistakes."
Oscar Wilde[1]

In this section, we build upon the summary provided above in order to gather elements of answer to the questions raised at the onset of this work (Section 2.1). Some aspects obviously had to be left out, or simply failed to yield the results we expected, so we highlight these limitations to put the scope of our thesis into perspective (Section 2.2). Notwithstanding some important open issues, we close up with operational recommendations for libraries, archives, and museums wishing to put into practice the methods described in this dissertation (Section 2.3).

2.1 Main findings

In order to recapitulate what has come out of our work, let us go back to the four research questions outlined in the introduction of this thesis. Although these questions were linked specifically to chapters I to IV, the underlying themes were present all along the way and constantly intertwined. In what follows, we provide tentative answers in the light of the theoretical findings from these first four chapters, while qualifying our conclusions in the light of the concrete results obtained in the last one.

Question 1 questioned the legitimacy of the distinction between terms and entities from the practical perspective of the end users of extraction tools.

We have seen that, while this artificial separation may still be justified in some cases – including for building a gold-standard corpus with predefined entity types in order to perform a qualitative evaluation of a set of systems –, most semantic enrichment tools do not draw a dividing line between the common and proper nouns they are extracting any longer. In particular, entity linking systems relying on knowledge bases and other semantic resources can indifferently serve pages about terms and entities, since there is no formal criterion to distinguish between the two. Whereas older tools focused either on named-entity recognition or on terminology extraction, this distinction is now blurred to the point that we can see no practical reason to maintain for knowledge discovery, although the distinction may remain valid in other contexts where no external resources are available.

[1] *The Picture of Dorian Gray*, Chapter 4.

Question 2 pondered over the added value of interdisciplinary input for information extraction (IE).

After combining findings from several research domains, from linguistics to history and from computer sciences to philosophy, we confirm that this synergy broadened our horizons and that seemingly irreconcilable disciplines reinforced one another, enriching the theoretical framework used to address our object of study. Specifically, research on the Semantic Web and Linked Data allowed us to overcome some limitations faced by IE, while Linked Data resources were in turn studied critically in the light of data quality research and hermeneutics. This is not to say that interdisciplinarity is beneficial in all scenarios, but in our own it definitely was.

Question 3 wondered if decompartmentalisation could improve IE systems.

While the benefits of domain-specific IE cannot be denied in a number of cases (Chiticariu et al., 2010; Tang et al., 2015), ad hoc approaches developed for a given use are relatively expensive and challenging to maintain over time. In empirical domains where the objects of study are constantly evolving, and particularly in the humanities and the cultural heritage where budgets are often limited, the added value of generic IE systems is obvious (Albanese and Subrahmanian, 2007). General-domain knowledge bases will sometimes show their limits with texts containing a high proportion of undocumented entities. Nevertheless, they often outperform specialist systems in terms of coverage (Fafalios et al., 2015), and therefore help to improve recall scores.

Question 4 weighted the pros and cons of language-independent IE.

Increasingly, the focus of natural language processing has been on cross-lingual approaches allowing to handle the growing amount of multilingual content available in digital form. The shift from linguistic to data-driven and hybrid methods also made systems less reliant on language-specific symbolic rules, encouraging the development of more language-independent methods. Although some NLP components remain necessarily language-dependent – such as deep parsing and stopwords processing, for instance – shallow analysis of multilingual text has become commonplace, with state-of-the-art systems achieving comparable performance on several languages. In the context of the *Historische Kranten* project, we showed that an application (MERCKX) relying on a multilingual knowledge base could efficiently handle three languages, and be extended to others with some adaptation effort.

2.2 Limitations

As mentioned in our introduction, searching for the perfect IE system that will change the lives of end users is necessarily a quixotic quest. Working in a fundamentally interdisciplinary environment is a rewarding experience, but it can at times become frustrating to wander ceaselessly from one theoretical framework to the other, without ever being able to get to the bottom of things.

Although this dissertation may occasionally have sounded quite assertive about the practicability of a fully generalised approach to entity linking, Chapter V put things into perspective by showing that, in practice, things are not always as clear-cut as we would have liked them to be. In particular, the last section about the generalisation of our approach to other languages, domains, and types of entities revealed a mixed picture in terms of practical portability. MERCKX is indeed language-agnostic since it does not include any language-specific component, but not yet truly language-independent because the language variable still affects the results significantly.

This can be explained on the one hand by the discrepancies between the various linguistic chapters of DBpedia, and on the other hand by the complexity of some languages, such as the absence of word boundaries in Chinese or the right-to-left writing system of Arabic. A fully language-independent system would obviously need to incorporate labels from more languages, relying on a fallback mechanism when an entity does not exist in a specific tongue.

Similarly, all domains are not equal before the extraction process, and the quality of the knowledge discovered varies widely from one type of entity to the other. Including miscellaneous concepts (common nouns) requires to consider all lowercase words as valid candidates – driving up the amount of noise – and filtering them with stopwords would imply compromising with our ideal of language-independence. As a whole, the question of the quality of Linked Open Data remains far from solved, and using these uncurated resources in strategic domains remains problematic.

Another issue we did not explicitly take into account in this work but which badly affects information accessibility is the incorrect indexation of documents, making them virtually untraceable. Bade (2004) addresses the impact of bibliographic errors on information retrieval and gives a damning report of encoding practices in the information age.

Finally, Boydens (1999, p. 129, italics hers) notes that “la question de l’adéquation d’une représentation informatique à son objet, *en l’absence de référentiel*, demeure ouverte”. While entity linking allows to disambiguate between equivocal terms thanks to the resort to different URIs, the relationship between the sense and the reference remains as elusive as ever.

2.3 Operational recommendations

Building upon the experience acquired in the context of this dissertation, we can formulate a number of practical recommendations for cultural heritage players eager to make the best of semantic enrichment applications to add value to their collections. Far from theoretical musings, these pieces of advice are very down-to-earth and take into account the reality of the cultural sector, including financial and situational constraints.

Take advantage of what is there

Documents are seldom so special that they require to be processed in a completely specific way. One should beware of not reinventing the wheel. Lots of facts about the world are available out there in knowledge bases, and the tools to exploit them are largely free to use. Before investing time and money into the development of ad hoc resources and applications, one should take the time to assess the needs of users and survey the existing tools likely to meet them. In most cases, building upon open source material will allow to achieve significant benefits at a reduced cost.

Imperfect knowledge is (often) better than none

Being conscious of the bad quality of data does not imply to be deterred by it. No data are ever perfect, and acceptable quality is always more a matter of balancing conflicting aspects rather than aiming for an elusive absolute. Except in application domains where a trifling error can have devastating consequences in terms of money losses (e.g. trading), trials (e.g. law) or human lives (e.g. medicine), *good enough* semantic content will always be of a higher value than none at all for end users, given the financial constraints.

Go multilingual

In an economy of knowledge largely dominated by the English language, the temptation is great to concentrate all development work on the processing of this natural language only in the hope that subsequent adaptation to other languages will then be seamless. We have shown, however, that this is a vain hope, and this position is increasingly indefensible in the light of recent developments of language-independent systems. Engaging from the start with the multilingual dimension of global knowledge allows to adopt a comprehensive point of view, while anticipating further developments.

Time matters

Language is not static, and nor is our environment. Concepts appear, evolve, and die out all the time: acknowledging this constantly moving reality requires an extraction model taking the temporal dimension into account. While the Web is dynamic by nature, Linked Open Data are often more rigid and exploited in an artificially fixed form, materialised by database dumps exported at a given time. Being conscious of this crucial limitation and attempting to remedy it by designing mechanisms able to track the evolution of concepts over time is an essential part of any information extraction system, as will be further demonstrated in Section 3.3.

Graphs are good

Blanke and Kristel (2013) investigated the added value of graph databases for the humanities. They conclude that, "with their emphasis on relationships, graph databases are particularly well suited for historical research in particular and humanities research in general". Since the content emanating from the humanities is often complex, the authors show "how graph databases integrate with traditional ways of searching and browsing historical collections [to] support more advanced means of access to facts in the documents and enable deep semantically meaningful access to the documents". The focus on relationships is essential. Whereas links in relational databases can only be represented through computationally expensive join operations between tables, "relationships are first-class citizens in graph databases". Graph databases are therefore good candidates for the organisation of knowledge and offer very efficient traversal algorithms, enabling the discovery of semantically-related information with a performance superior to most traditional databases.

Above all, experiment

With the democratisation of NLP tools and the proliferation of open source solutions available online, experimenting with semantic enrichment does not require a degree in linguistics or IT any longer. With a hands-on mindset, any motivated individual can get to grips with the basics of entity recognition and linking, and understand how to gain advantage of these technologies in order to improve the daily experience of end users. The *Free Your Metadata* project[2] contributed to promote this attitude of open-mindedness towards computational techniques in the cultural heritage sector.

[2] `http://freeyourmetadata.org/`

3 Perspectives

"We can only see a short distance ahead,
but we can see plenty there that needs to be done."
Alan Turing (1950)

This final section brings together some issues that were either outside the scope of this dissertation, or that could not be tackled within its timeframe due to practical limits. First and foremost, MERCKX remains to date in a proof-of-concept state, and has not yet been implemented into the *Historische Kranten* website, although we have plans to do so in a near future (Section 3.1).

The obvious step following this implementation is the evaluation of the relevance of this system for end users, which is discussed in Section 3.2. Lastly, we propose in Section 3.3 to extend our model to application domains from outside the humanities, and imagine how this could prove useful in totally different contexts, with a special emphasis on temporal issues.

3.1 Implementation

We are now looking forward to integrating the MERCKX algorithms into the *Historische Kranten* project's Web interface, in order to watch it improve the search experience of end users. Following a conclusive meeting in Ypres on 9 June, 2015, we were invited to travel to the Picturae headquarters in Heiloo (near Amsterdam) on 21 September, 2015 to give a full presentation to the CEO, and have an in-depth talk with developers (see Appendix C for a follow-up of this meeting).

Before materialising into a full-fledged application, MERCKX still needs development work to gain in maturity and robustness, although the fundamental components are already in place. Following the preliminary evaluation performed in Chapter V, the source code could benefit from a few improvements before the workflow is launched on the whole million-document collection. Reflections on what remains to be done are provided below.

To address the issue of low recall, we could leverage other links in addition to the `rdfs:label` explicit relationship between a concept and the terms to express it in various languages. For instance, the `dbo:wikiPageRedirects` property references redirection pages, which in turn contain some extra labels. Similarly, the `owl:sameAs` property could be used to exploit resources from the other language chapters of DBpedia, but also from external ontologies containing their own types: `yago:YagoGeoEntity`, `wikidata:Q486972` (human settlement) or `http://schema.org/Place` for instance.

To boost precision, we plan to limit the negative impact of OCR errors by implementing the Levenshtein distance clustering algorithm. To cross-check the validity of entities, we could experiment with the combination of several knowledge bases instead of DBpedia only: for place names, the aggregation of GeoNames[3] and GeoVocab[4] looks promising. Working with the new digitised version of the corpus instead of the original one we used in this thesis should also improve the accuracy of the extraction. Finally, learning from the positive aspects of Zemanta and Babelfy, a better-informed disambiguation process could be set up with a graph-based approach taking into account the context and the semantic proximity of resources with related entities.

3.2 Extrinsic evaluation

A limitation we did not mention yet is the absence of an external assessment. In Chapter V, we focused on the intrinsic evaluation of entities, testing their formal correctness against a manually annotated gold-standard corpus. In addition, an extrinsic evaluation should be performed in order to assess the relevance of the entities retrieved for end users. This second type of evaluation is crucial because a beautiful theoretical model or tool can prove totally at odds with the needs of users, or the reality of fieldwork. Indeed, the fact that an entity is correct does not necessarily mean that it is useful. In other words, "la cohérence formelle d'une représentation informatique n'implique nullement que celle-ci soit utile dans la pratique" (Boydens, 1999, p. 488).

The relevance of entities is highly context-dependent: for instance, a mention of "Brussels", correctly disambiguated, could be of interest if found in American literature, but would be quite insignificant in the minutes of the European Parliament. It is also user-dependent, as different kinds of people will have different interests in a given collection of documents. Buckland (1997) remarked that relevance "is now generally considered to be situational and ascribed by the viewer". It is therefore important to analyse the search behaviour of users in order to have realistic expectations about their needs.

After the implementation phase, we thus plan to interact with the users to get feedback about the relevance and usefulness of entities extracted, and of automatic search suggestions based on semantic relatedness. This empirical survey would involve asking evaluators to query the *Historische Kranten* collection with the old Lucene-based search engine and with a new one based on MERCKX, in order to see the difference and compare the results obtained, using the evaluation grid of the SQuaRE ISO standard introduced in Chapter V.

[3] `http://www.geonames.org/`
[4] `http://geovocab.org/`

In a similar experiment, Miliaraki et al. (2015) performed a large-scale analysis of the usefulness of the Yahoo! Spark system which allows exploratory entity search by providing users with related entity suggestions based on their query and exploiting the Semantic Web. We hope that this kind of analysis will help to further demonstrate the significance of our work.

3.3 Other applications

As emphasised in Chapter III, the present approach is not limited to cultural heritage content or the humanities but can be extended to any empirical domain. The nature of our case study made that much effort was devoted to the extraction of places from historical periodicals, but this limitation of scope is by no means inherent to the method used. To illustrate this claim, we review a few potential application domains from outside the humanities that could benefit from entity linking, semantic enrichment, and knowledge discovery.

Politicisation of immigration

The SOM project,[5] funded by the European FP7 programme, aimed to identify trends in the support and opposition to migration by analysing political claims from the national press of seven European countries. The methodology used by the project partners was to manually read, cut out, and index relevant news.

Although the project ran over three years, only 971 newspaper articles were used in total, casting doubt on the representativeness of the results. A second phase could involve the validation of these results with the help of a larger-scale, automated extraction process. However, identifying claims is more complex than for entities, especially in a multilingual context (the SOM corpus covering Dutch, English, French, German, Spanish). A preliminary study is currently in progress in order to evaluate the viability of this alternative approach.

Tracking disease outbreaks

The recent H1N1 influenza and Ebola pandemics emphasised the need for very quick responses from the World Health Organisation and other public health agencies in the event of a major outbreak. Projects such as HealthMap[6] and GermTraX[7] recognise this necessity and exploit the vast amount of data

[5] `http://www.som-project.eu/`
[6] `http://www.healthmap.org/`
[7] `http://www.germtrax.com/`

available online, especially on social media, to detect emergencies and allow people in charge to react swiftly. The integration, into such tools, of biomedical knowledge discovery systems like BioGraph[8] (Liekens et al., 2011) could help to improve the tracking of previously unknown viruses and to anticipate the secondary effects of health policies.

Natural disasters

Likewise, the unexpected eruption of the Cotopaxi volcano near Quito (Ecuador) calls for well-coordinated responses within a short period of time. The Global Disaster Alert and Coordination System,[9] for instance, is "a cooperation framework between the United Nations, the European Commission and disaster managers worldwide to improve alerts, information exchange and coordination in the first phase after major sudden-onset disasters".

However, this website is only updated in retrospect, when a major disaster is already under way. Initiatives such as Earth Alerts[10] aim to rectify this shortcoming by collecting all signs of potential disasters in an open interface. Tracking seemingly insignificant facts posted online with information extraction techniques is also a promising development of the work presented here, especially when taking the temporal dimension into account.

Interactions between timescales

In Chapter IV, we explored the influence of time on linguistic phenomena and showed, after Boydens (1999), how the Braudelian framework of stratified timescales, enriched with the evolving continuums of Elias, allowed us to account for the asynchronous evolution of language on three temporal layers dynamically interacting with one another: the long term of language change, the medium term of common usage, and the short term of everyday speech.

Tracking the emergence, disappearance, and evolution of concepts is a daunting task. But since concepts are progressively constructed over time, a knowledge discovery system implementing this framework would in theory be able to keep up with the underlying reality in an operational perspective. Like for named-entity recognition and entity linking, knowledge bases can be instrumental to the disambiguation of such concepts, even when the terms used to refer to them are constantly evolving. Although we did not develop this system, we will offer insights as to how it could be achieved in the future.

[8] http://www.clips.uantwerpen.be/BiographTA/

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

[10] http://earthalerts.manyjourneys.com/web/

A first possibility would be to monitor the evolutions of knowledge bases themselves. This could be done either continuously by automatically querying live KBs for new concepts,[11] or episodically on discrete dumps downloaded every few months or years.[12] Both approaches have advantages and drawbacks. The dynamic version is more up-to-date and allows to detect changes the second they happen,[13] but it is also slower and requires constant resources, in addition to being dependent on the availability of the server.

In contrast, the static version is more liable to be outdated, but is also more reliable to see evolutions in the long run. Between the English DBpedia dump of August 2014 and the one of April 2015, that is to say over a mere eight-month period, 594 381 labels were incorporated in the knowledge base and 23 333 of them disappeared. Monitoring these labels would reflect on the fluctuating nature of knowledge representation, but also teach us something about the underlying, evolving reality.

Another promising track would be to analyse a historical corpus, such as the *Historische Kranten*, in an explicitly diachronic perspective, to show the interactions between the levels of temporality inside a specific timeframe, using external linguistic resources. For instance, a new concept appearing in the corpus (short term) could be tracked to see if it becomes at some point formalised in a dictionary (medium term) and thereby contributes to the evolution of the language in question (long term). Inversely, the disappearance or shift in meaning of concepts in prescriptive grammars could have an impact on the reporting of events in the press, and ultimately on the way people discuss them.

This could be performed in a language-independent manner by applying probabilistic topic modelling[14] on a collection of documents and comparing the results obtained for each language, in order to determine if the evolutions are synchronous from one language to the other. Since there is a risk of an anachronistic bias inherent to the use of contemporary knowledge bases for the analysis of historical content, evaluating the effectiveness of this approach would require a new gold-standard corpus, more representative of the temporal dispersion of the collection. Research along these lines is planned to be conducted in the next coming months.

At the end of this journey, we are left with more questions than answers about the best itinerary to reach the promised land of knowledge discovery, but is that not precisely the way it is meant to be?

[11] `http://live.dbpedia.org/` for instance.
[12] Like those made available at `http://wiki.dbpedia.org/datasets/`.
[13] See Wikipedia Live Monitor: `http://wikipedia-live-monitor.herokuapp.com/`.
[14] See `https://github.com/ulbstic/topic-modeling-tool-FR`.

Appendix A

Source Code

The main program used by our MERCKX system (`merckx.py`) *is shown below. More scripts written for the* Historische Kranten *project can be found online at* `https://github.com/ulbstic/ypres`*, along with our gold-standard corpus and material from the test application at* `http://mastic.ulb.ac.be/ypres/`.

```
import os,sys,urllib,urllib2
from urllib import unquote_plus as up
from nltk.tokenize import WordPunctTokenizer as tokenizer

# check parameters
txtType = sys.argv[1].upper() if len(sys.argv) > 1 else ""
txtFile = sys.argv[2] if len(sys.argv) > 2 else ""
entityTypes = {"EVE":"Event", "LOC":"Place",
               "PER":"Person", "ORG":"Organisation"}
if txtType not in entityTypes or not os.path.isfile(txtFile):
  print "Syntax: merckx.py <type> <filename>"
  print "where <type> is the entity type: EVE LOC PER ORG"
  print "and <filename> is the name of text file to analyze"
  print
  sys.exit(0)

# load labels of entities
labels = {}
entityType = entityTypes[txtType]
filename = "data/labels_"+entityType+".lst"
with open(filename) as inFile:
  lines = inFile.read().decode("utf8").strip().split("\n")
  for line in lines:
    label,uri = line.split("\t")
    labels[label] = uri
```

```
# load + tokenize + extract entities from text file
source = os.path.basename(txtFile)
text = open(txtFile).read().decode("utf8")
tokens = tokenizer().tokenize(text) # list of tokens
pos = list(tokenizer().span_tokenize(text)) # list of positions
sep = " " # word separator; may be language-dependent
lastpos = 0 # lastpos position

for i,w in enumerate(tokens): # for every token
  if len(w) >= 3: # ignore less than 3 chars
    w3 = sep.join(tokens[i:i+3])
    if i+1<len(tokens) and tokens[i+1] == "-": # e.g. Pays-Bas
      w3 = "".join(tokens[i:i+3])
    w2 = sep.join(tokens[i:i+2])
    w1 = w
    # NEW YORK CITY => New York City
    w3 = w3.title() if w3.isupper() else w3
    w2 = w2.title() if w2.isupper() else w2
    w1 = w1.title() if w1.isupper() else w1
    comp = ["",w1,w2,w3]
    if i+2<len(tokens) and comp[3] in labels: # 3 words
      label = w3
      start = pos[i][0]
      end = pos[i+2][1]
    elif i+1<len(tokens) and comp[2] in labels: # 2 words
      label = w2
      start = pos[i][0]
      end = pos[i+1][1]
    elif comp[1] in labels: # 1 word
      label = w1
      start = pos[i][0]
      end = pos[i][1]
    else:
      continue
    if start < lastpos: # skip words already processed
      continue
    lastpos = end
    uri = up(str(labels[label]))

    # display results
    print("{0}\t{1}\t{2}\t{3}".format(source,start,end,uri))
```

Appendix B

Guidelines for Annotators

We here transcribe the detailed guidelines provided to external annotators for the construction of our gold-standard corpus based on the identification of place mentions in a sample of 100 texts from the Historische Kranten *corpus:*

- Identify all individual mentions of places/locations in the sample
- Additionally, write down the positions of the first and last character
- For instance, the sentence "I work in Brussels and live in Mons." should result in the following annotations:
  - `Brussels 10 18`
  - `Mons 31 35`
- Place mentions should include (but are not limited to):
  - countries, states & provinces
  - cities, towns & municipalities
  - villages, hamlets & boroughs (*deelgemeenten*)
  - mountains & rivers
- Include places regardless of case, i.e. even when in UPPERCASE
- Do not include street names, nor the places in them ("rue de Lille")
- Do not include adjectives ("Belgian") nor genitive forms ("te Yperen")
- If a place appears several times in a row, annotate all mentions
- When OCR errors are present, try to guess the original place name
- When in doubt, check the original on the *Historische Kranten* website

Appendix C

Follow-up

On September 21, we travelled to Heiloo in order to discuss implementation of MERCKX with Picturae. This meeting was very productive, as we exchanged ideas for hours with developers who were definitely interested in potential applications for this website and others they are maintaining. At the end of the day, Mr Tiessen announced that he could secure a budget for the development of a proof-of-concept semantic search engine based on MERCKX.

Concretely, the outcomes of this meeting can be summarised as follows:

- The people from Picturae were not aware of the importance of locations for users (which we demonstrated based on the analytics of the website) and are keen to upgrade the current Solr engine with a specific index of place names, along with a visualisation tool of places on a map.
- The relevance of multilingual tools has been mentioned repeatedly, since classic tools such as OpenCalais and AlchemyAPI do not handle Dutch at all.
- They are also interested in other NLP applications such as topic modelling (a research trend explored by our colleague Simon Hengchen in his PhD thesis) and are eager to establish links between their collections (full-text descriptions and structured metadata) through Linked Data.

When the implementation materialises, it will pave the way for an external validation of our results by end users of the *Historische Kranten* website, who could compare the output of MERCKX with the type and amount of information that they obtained with the former search engine. Since the system is open, it is challenging to track users doing regular research, but people using computers in the archive building could be personally contacted and the CO7 heritage cell is willing to put a user group together in order to conduct an online survey.

Detailed Contents

Bibliography

(Ackoff, 1989) Russell L. Ackoff. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16:3–9, 1989.

(Agirre et al., 2012) Eneko Agirre, Ander Barrena, Oier Lopez De Lacalle, Aitor Soroa, Samuel Fernando, and Mark Stevenson. Matching Cultural Heritage Items to Wikipedia. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 1729–1735, Istanbul, Turkey, 2012.

(Ahn, 2006) David Ahn. The Stages of Event Extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events (ARTE)*, pages 1–8, Sydney, 2006.

(Akbik and Broß, 2009) Alan Akbik and Jügen Broß. Wanderlust: Extracting Semantic Relations from Natural Language Text using Dependency Grammar Patterns. In *Proceedings of the Semantic Search Workshop at the 18th International World Wide Web Conference*, Madrid, 2009.

(Albanese and Subrahmanian, 2007) Massimiliano Albanese and V. S. Subrahmanian. T-REX: A Domain-Independent System for Automated Cultural Information Extraction. In *Proceedings of the 1st International Conference on Computational Cultural Dynamics (ICCCD)*, University of Maryland, 2007.

(Alex et al., 2012) Bea Alex, Claire Grover, Ewan Klein, and Richard Tobin. Digitised Historical Text: Does it have to be mediOCRe? In *Proceedings of KONVENS*, pages 401–409, Vienna, 2012.

(Alexopoulos et al., 2015) Panos Alexopoulos, Ronald Denaux, and Jose Manuel Gomez-Perez. Troubleshooting and Optimizing Named Entity Resolution Systems in the Industry. In *The Semantic Web. Latest Advances and New Domains*, volume 9088 of *Lecture Notes in Computer Science*, pages 559–574. Springer, 2015.

(Alfonseca and Manandhar, 2002) Enrique Alfonseca and Suresh Manandhar. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In *Proceedings of the 1st International Conference on General Word-Net*, Mysore, India, 2002.

(Allan, 1986) Keith Allan. *Linguistic Meaning*. Routledge, London, 1986.

(Ananiadou and McNaught, 2006) Sophia Ananiadou and John McNaught, editors. *Text Mining for Biology and Biomedicine*. Artech House, London, 2006.

(Anderson, 2008) Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, 16(07), 2008.

(Aone and Ramos-Santacruz, 2000) Chinatsu Aone and Mila Ramos-Santacruz. REES: A Large-Scale Relation and Event Extraction System. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 76–83, Seattle, WA, USA, 2000. ACL.

(Ariès, 1986) Philippe Ariès. *Le temps de l'histoire*. Éditions du Seuil, Paris, 1986.

(Auer and Lehmann, 2007) Sören Auer and Jens Lehmann. What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In *The Semantic Web: Research and Applications*, pages 503–517. Springer, 2007.

(Augenstein et al., 2014) Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Relation Extraction from the Web using Distant Supervision. In *Knowledge Engineering and Knowledge Management*, pages 26–41. Springer, 2014.

(Babych and Hartley, 2003) Bogdan Babych and Anthony Hartley. Improving Machine Translation Quality with Automatic Named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 1–8, Stroudsburg, PA, USA, 2003. ACL.

(Bade, 2004) David Bade. *The Theory and Practice of Bibliographic Failure, or, Misinformation in the Information Society*. Chuluunbat, Ulan Bator, 2004.

(Bade, 2008) David Bade. *Responsible Librarianship: Library Policies for Unreliable Systems*. Library Juice Press, Litwin Books, Sacramento, CA, USA, 2008.

(Banko and Brill, 2001) Michele Banko and Eric Brill. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting on ACL*, ACL '01, pages 26–33, Stroudsburg, PA, USA, 2001.

(Batini and Scannapieco, 2006) Carlo Batini and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Springer, New York, 2006.

(Becker et al., 2002) Marcus Becker, Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. SProUT – Shallow Processing with Unification and Typed Feature Structures. In *Proceedings of the International Conference on Natural Language Processing (ICON)*, Kanazawa, Japan, 2002.

(Bender, 2011) Emily M. Bender. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26, 2011.

(Bender et al., 2003) Oliver Bender, Franz Josef Och, and Hermann Ney. Maximum Entropy Models for Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT–NAACL*, volume 4, pages 148–151. ACL, 2003.

(Berners-Lee, 2000) Tim Berners-Lee. *Weaving the Web*. HarperCollins, New York, 2000.

(Berners-Lee et al., 2001) Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):28–37, 2001.

(Bingel and Haider, 2014) Joachim Bingel and Thomas Haider. 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, 2014.

(Bird et al., 2009) Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, USA, 2009.

(Bizer et al., 2009a) Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009a.

(Bizer et al., 2009b) Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – A Crystallization Point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009b.

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

(Blanke et al., 2012) Tobias Blanke, Michael Bryant, and Mark Hedges. Open Source Optical Character Recognition for Historical Research. *Journal of Documentation*, 68(5):659–683, 2012.

(Bledsoe and Browning, 1959) Woodrow Wilson Bledsoe and Iben Browning. Pattern Recognition and Reading by Machine. In *Proceedings of the 9th Eastern Joint Computer Conference*, pages 225–232, Boston, MA, USA, 1959.

(Bouillon, 1997) Pierrette Bouillon. *Polymorphie et sémantique lexicale: le cas des adjectifs*. PhD thesis, Paris 7, 1997.

(Bouillon et al., 2000) Pierrette Bouillon, Cécile Fabre, Pascale Sébillot, and Laurence Jacqmin. Apprentissage de ressources lexicales pour l'extension de requêtes. *Traitement automatique des langues*, 41(2):367–393, 2000.

(Bouillon et al., 2001) Pierrette Bouillon, Vincent Claveau, Cécile Fabre, and Pascale Sébillot. Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements. In *Proceedings of 1st International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland, 2001.

(Bourigault et al., 1996) Didier Bourigault, Isabelle Gonzalez-Mullier, and Cécile Gros. LEXTER, a Natural Language Processing Tool for Terminology Extraction. In *Proceedings of the 7th Euralex International Congress*, pages 771–779, Göteborg, Sweden, 1996.

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

(Boydens, 2011) Isabelle Boydens. Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium. In Saïd Assar, Imed Boughzala, and Isabelle Boydens, editors, *Practical Studies in E-Government: Best Practices from Around the World*, pages 113–130. Springer, 2011.

(Braudel, 1949) Fernand Braudel. *La Méditerranée et le monde méditerranéen à l'époque de Philippe II*. Armand Colin, Paris, 1949.

(Bresnan and Kaplan, 1982) Joan Wanda Bresnan and Ronald M. Kaplan. Introduction: Grammars as Mental Representations of Language. In Joan Wanda Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, USA, 1982.

(Briet, 1951) Suzanne Briet. *Qu'est-ce que la documentation ?* Éditions documentaires, industrielles et techniques, Paris, 1951.

(Brooks, 1947) Cleanth Brooks. *The Well Wrought Urn. Studies in the Structure of Poetry*. Reynal & Hitchcock, 1947.

(Brun et al., 2007) Caroline Brun, Maud Ehrmann, and Guillaume Jacquet. XRCE-M: A Hybrid System for Named Entity Metonymy Resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 488–491, Prague, 2007. ACL.

(Brun et al., 2009) Caroline Brun, Maud Ehrmann, and Guillaume Jacquet. Résolution de métonymie des entités nommées: proposition d'une méthode hybride. *Traitement Automatique des Langues (TAL)*, 50:87–110, 2009.

(Buckland, 1997) Michael K. Buckland. What is a "Document"? *Journal of the American Society for Information Science*, 48(9):804–809, 1997.

(Buitelaar and Cimiano, 2014) Paul Buitelaar and Philipp Cimiano, editors. *Towards the Multilingual Semantic Web*. Springer, Berlin, 2014.

(Buitelaar et al., 2005) Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology Learning from Text: An Overview. In *Ontology Learning from Text: Methods, Applications and Evaluation*, pages 3–12. IOS Press, 2005.

(Buitinck and Marx, 2012) Lars Buitinck and Maarten Marx. Two-Stage Named-Entity Recognition using Averaged Perceptrons. In *Proceedings of the 17th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 7337 of *Lecture Notes in Computer Science*, pages 171–176, Groningen, The Netherlands, 2012. Springer.

(Bunescu and Pasca, 2006) Razvan C. Bunescu and Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *EACL*, volume 6, pages 9–16, 2006.

(Busa, 1980) Roberto Busa. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, 14(2):83–90, 1980.

(Busemann and Krieger, 2004) Stephan Busemann and Hans-Ulrich Krieger. Resources and Techniques for Multilingual Information Extraction. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

(Byrd and Ravin, 1999) Roy J. Byrd and Yael Ravin. Identifying and Extracting Relations in Text. Technical report, T.J. Watson IBM Research Center, Yorktown Heights, NY, USA, 1999.

(Cabrio et al., 2014) Elena Cabrio, Julien Cojan, and Fabien Gandon. Mind the Cultural Gap: Bridging Language-Specific DBpedia Chapters for Question Answering. In Paul Buitelaar and Philipp Cimiano, editors, *Towards the Multilingual Semantic Web*, pages 137–154. Springer, 2014.

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

(Carreras et al., 2002) Xavier Carreras, Lluis Marquez, and Lluís Padró. Named Entity Extraction using Adaboost. In *Proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. ACL, 2002.

(Cavnar and Trenkle, 1994) William B. Cavnar and John M. Trenkle. N-gram-Based Text Categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, pages 161–175, Las Vegas, NV, USA, 1994.

(Charton and Torres-Moreno, 2009) Eric Charton and Juan Manuel Torres-Moreno. Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, 2009.

(Chieu and Ng, 2002) Hai Leong Chieu and Hwee Tou Ng. Named Entity Recognition: A Maximum Entropy Approach using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1, pages 1–7. ACL, 2002.

(Chieu and Ng, 2003) Hai Leong Chieu and Hwee Tou Ng. Named Entity Recognition with a Maximum Entropy Approach. In *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL)*, pages 160–163, Edmonton, Canada, 2003.

(Chiticariu et al., 2010) Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012. MIT, MA, USA, 2010.

(Chomsky, 1956) Noam Chomsky. Three Models for the Description of Language. *IRE Transactions on Information Theory*, 2:113–124, 1956.

(Ciravegna, 2001) Fabio Ciravegna. Adaptive Information Extraction from Text by Rule Induction and Generalisation. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1251–1256, Seattle, WA, USA, 2001.

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

(Cornolti et al., 2013) Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A Framework for Benchmarking Entity-Annotation Systems. In *Proceedings of the 22nd International Conference on the World Wide Web*, pages 249–260, Rio de Janeiro, Brasil, 2013.

(Cowie and Lehnert, 1996) Jim Cowie and Wendy Lehnert. Information Extraction. *Commun. ACM*, 39(1):80–91, 1996.

(Coyle, 2006) Karen Coyle. Mass Digitization of Books. *The Journal of Academic Librarianship*, 32(6):641–645, 2006.

(Crombez, 2015) Thomas Crombez. The Document as Event: Assessing the Value of Digital Collections of Theatrical Heritage. In Bruno Forment and Christel Stalpaert, editors, *Theatrical Heritage: Challenges and Opportunities*, pages 195–205. Leuven University Press, 2015.

(Cucerzan and Yarowsky, 1999) Silviu Cucerzan and David Yarowsky. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*, pages 90–99, College Park, MD, USA, 1999.

(Curran and Clark, 2003) James R. Curran and Stephen Clark. Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of the 7th Conference on Natural Language Learning at HLT–NAACL*, volume 4, pages 164–167. ACL, 2003.

(Daelemans and Hoste, 2002) Walter Daelemans and Véronique Hoste. Evaluation of Machine Learning Methods for Natural Language Processing Tasks. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, 2002.

(Daelemans et al., 1999) Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. Memory-Based Shallow Parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 53–60, 1999.

(Daiber et al., 2013) Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124, Graz, Austria, 2013. ACM.

(Davenport and Prusak, 1998) Thomas H. Davenport and Lawrence Prusak. *Working Knowledge: How Organizations Manage What They Know*. Harvard Business Press, 1998.

(De Meulder and Daelemans, 2003) Fien De Meulder and Walter Daelemans. Memory-Based Named Entity Recognition using Unannotated Data. In *Proceedings of the 7th Conference on Natural Language Learning at HLT–NAACL*, pages 208–211, 2003.

(de Saussure, 1916) Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Paris, 1916.

(De Wilde, 2009) Max De Wilde. La réalisation linguistique des actes de langage directifs. Typologie – Sémantique – Pragmatique. Master's thesis, Université libre de Bruxelles (ULB), 2009.

(De Wilde, 2015) Max De Wilde. Improving Retrieval of Historical Content with Entity Linking. In *Proceedings of the 1st International Workshop on Semantic Web for Cultural Heritage*, volume 539 of *Communications in Computer and Information Science*, Poitiers, France, 2015. Springer.

(Derczynski et al., 2015) Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of Named Entity Recognition and Linking for Tweets. *Information Processing & Management*, 51(2):32–49, 2015.

(Deutscher, 2005) Guy Deutscher. *The Unfolding of Language*. Henry Holt and Company, New York, 2005.

(Deviaene, 2008) Laure Deviaene. Les mots vides : des origines linguistiques au contexte documentaire. Essai d'état de l'art et d'expérimentation à des fins méthodologiques et opérationnelles. Master's thesis, Université libre de Bruxelles (ULB), 2008.

(Diller, 1996) Karl C. Diller. How Human Languages Cohere: Languages Seen as Artificial Life. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, San Diego, CA, USA, 1996. Psychology Press.

(Doddington et al., 2004) George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 837–840, 2004.

(Dreyfus, 1972) Hubert Dreyfus. *What Computers Can't Do*. MIT Press, New York, 1972.

(Dreyfus and Dreyfus, 1986) Hubert Dreyfus and Stuart Dreyfus. *Mind Over Machine*. Blackwell, Oxford, 1986.

(Drucker, 2012) Johanna Drucker. Humanistic Theory and Digital Scholarship. In Matthew K. Gold, editor, *Debates in the Digital Humanities*, pages 85–95. University of Minnesota Press, 2012.

(Ehrmann, 2008) Maud Ehrmann. *Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris Diderot – Paris VII, 2008.

(Einstein, 1922) Albert Einstein. *Geometry and Experience*. Methuen & Co., London, 1922.

(Elias, 1996) Norbert Elias. *Du temps*. Fayard, 1996.

(Enache and Angelov, 2010) Ramona Enache and Krasimir Angelov. Typeful Ontologies with Direct Multilingual Verbalization. In *Proceedings of the 2nd Workshop on Controlled Natural Languages (CNL)*, Marettimo Island, Sicily, Italy, 2010.

(Exner and Nugues, 2012) Peter Exner and Pierre Nugues. Entity Extraction: From Unstructured Text to DBpedia RDF Triples. In *Proceedings of the Web of Linked Entities Workshop (WoLE)*, pages 58–69, Boston, MA, USA, 2012.

(Fafalios et al., 2014) Pavlos Fafalios, Manolis Baritakis, and Yannis Tzitzikas. Configuring Named Entity Extraction through Real-Time Exploitation of Linked Data. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS)*, Thessaloniki, Greece, 2014.

(Fafalios et al., 2015) Pavlos Fafalios, Manolis Baritakis, and Yannis Tzitzikas. Exploiting Linked Data for Open and Configurable Named Entity Extraction. *International Journal on Artificial Intelligence Tools (IJAIT)*, 24(02), April 2015.

(Farmakiotou et al., 2000) Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. Rule-Based Named Entity Recognition for Greek Financial Texts. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX)*, pages 75–78, Patras, Greece, 2000.

(Fayyad et al., 1996) Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3): 37–54, 1996.

(Feldman and Sanger, 2007) Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

(Fellbaum, 1998) Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

(Fenn and Raskino, 2008) Jackie Fenn and Mark Raskino. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. Harvard Business Press, 2008.

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

(Finkel et al., 2005) Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 363–370, Ann Arbor, MI, USA, 2005.

(Florian, 2002) Radu Florian. Named Entity Recognition as a House of Cards: Classifier Stacking. In *Proceedings of the 6th Conference on Natural Language Learning*, volume 20, pages 1–4. ACL, 2002.

(Florian et al., 2003) Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named Entity Recognition through Classifier Combination. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL)*, volume 4, pages 168–171. ACL, 2003.

(Forsyth and Sharoff, 2014) Richard S. Forsyth and Serge Sharoff. Document Dissimilarity Within and Across Languages: A Benchmarking Study. *Literary and Linguistic Computing*, 29(1):6–22, 2014.

(Francis, 1964) Winthrop Nelson Francis. A Standard Sample of Present-Day English for Use with Digital Computers. Report to the U.S. Office of Education on Cooperative Research Project No. E-007, 1964.

(Frawley et al., 1992) William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge Discovery in Databases: An Overview. *AI Magazine*, 13(3):57–70, 1992.

(Frege, 1960) Gottlob Frege. On Sense and Reference. In Peter Geach and Max Black, editors, *Translations from the Philosophical Writings of Gottlob Frege*, pages 56–78. Blackwell, Oxford, 2nd edition, 1960.

(Frontini et al., 2015) Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. Semantic Web Based Named 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, 2015.

(Futrell et al., 2015) Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-Scale Evidence of Dependency Length Minimization in 37 Languages. In *Proceedings of the National Academy of Sciences of the United States of America*, 2015.

(Gabrilovich and Markovitch, 2007) Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611, Hyderabad, India, 2007.

(Garrette and Klein, 2009) Dan Garrette and Ewan Klein. An Extensible Toolkit for Computational Semantics. In *Proceedings of the 8th International Conference on Computational Semantics*, pages 116–127, Tilburgs, The Netherlands, 2009. ACL.

(Gillam et al., 2009) Michael Gillam, Craig Feied, Jonathan Handler, Eliza Moody, Ben Shneiderman, Catherine Plaisant, Mark Smith, and John Dickason. The Healthcare Singularity and the Age of Semantic Medicine. The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

(Ginzburg, 1989) Carlo Ginzburg. *Mythes, emblèmes, traces: morphologie et histoire*. Flammarion, Paris, 1989.

(Ginzburg, 2002) Carlo Ginzburg. *Wooden Eyes: Nine Reflections on Distance*. Verso Books, 2002.

(Goutte et al., 2014) Cyril Goutte, Serge Léger, and Marine Carpuat. The NRC system for discriminating similar languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145, Dublin, Ireland, 2014.

(Green and Bide, 1996) Brian Green and Mark Bide. Unique Identifiers: A Brief Introduction. Book Industry Communication, 1996.

(Greenberg and Méndez, 2007) Jane Greenberg and Eva Méndez. *Knitting the Semantic Web*. The Haworth Information Press, Binghamton, NY, USA, 2007.

(Grishman and Sundheim, 1996) Ralph Grishman and Beth Sundheim. Message Understanding Conference–6: A Brief History. In *Proceedings of the 16th International Conference on Computational Lingusitics (COLING)*, pages 466–471, 1996.

(Gruber, 1995) Thomas R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human–Computer Studies*, 43(5):907–928, 1995.

(Guasch and Sola, 1998) Jaume Guasch and Joan Sola. Implications on the Supersymmetric Higgs Sector from Top Quark Decays at the Tevatron. *Physics Letters B*, 416(3):353–360, 1998.

(Gupta et al., 2005) Suhit Gupta, Gail E. Kaiser, Peter Grimm, Michael F. Chiang, and Justin Starren. Automating Content Extraction of HTML Documents. *World Wide Web Journal*, 8(2):179–224, 2005.

(Hahn, 2008) Trudi Bellardo Hahn. Mass Digitization. *Library Resources & Technical Services*, 52(1):18–26, 2008.

(Hallot, 2005) Frédéric Hallot. Multilingual Semantic Web Services. In *Proceedings of the OTM Workshop: On the Move to Meaningful Internet Systems*, volume 3762 of *Lecture Notes on Computer Science*, pages 771–779. Springer, 2005.

(Halpin et al., 2010) Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. When `owl:sameAs` Isn't the Same: An Analysis of Identity in Linked Data. In *Proceedings of the 9th International Semantic Web Conference (ISWC2010)*, pages 305–320, Shanghai, China, 2010.

(Han and Sun, 2011) Xianpei Han and Le Sun. 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, 2011.

(Harris, 1962) Zellig S. Harris. *String Analysis of Sentence Structure*. Mouton, The Hague, 1962.

(Hartig and Zhao, 2010) Olaf Hartig and Jun Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *Provenance and Annotation of Data and Processes*, pages 78–90. Springer, 2010.

(Havasi et al., 2007) Catherine Havasi, Robert Speer, and Jason B. Alonso. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*, pages 27–29, Borovets, Bulgaria, 2007.

(Havelange, 1993) Carl Havelange. La critique historique aujourd'hui. Réflexions critiques à propos du document et du sens. *Revue du cercle des alumni de la Fondation Universitaire*, LXIV, 1993.

(Havelange, 2014) Carl Havelange. La condition documentaire. Libres propositions pour une intelligence plurielle du document et de ses usages. *La Licorne*, 109, 2014.

(Hearst, 1999) Marti A. Hearst. Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 3–10, College Park, MD, USA, 1999.

(Hendrickx and Van Den Bosch, 2003) Iris Hendrickx and Antal Van Den Bosch. Memory-Based One-Step Named-Entity Recognition: Effects of Seed List Features, Classifier Stacking, and Unannotated Data. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL)*, volume 4, pages 176–179, 2003.

(Hengchen et al., 2015) Simon Hengchen, Seth van Hooland, Ruben Verborgh, and Max De Wilde. L'extraction d'entités nommées: une opportunité pour le secteur culturel? *Information, données & documents*, 52(2):70–79, 2015.

(Heylighen, 2014) Francis Heylighen. Challenge Propagation: Towards a Theory of Distributed Intelligence and the Global Brain. *Spanda Journal*, V(2):51–63, 2014.

(Hinzen, 2012) Wolfram Hinzen. The philosophical significance of Universal Grammar. *Language Sciences*, 34(5):635–649, 2012.

(Hitzler et al., 2009) Pascal Hitzler, Markus Krotzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. CRC Press, Boca Raton, FL, USA, 2009.

(Hoffart et al., 2011) Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing*, pages 782–792, 2011.

(Hofstadter, 1980) Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid. A Metaphorical Fugue on Minds and Machines in the Spirit of Lewis Carroll*. Random House, New York, 1980.

(Hogan et al., 2010) Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the Pedantic Web. In *Proceedings of the 3rd International Workshop on Linked Data on the Web*, Raleigh, NC, USA, 2010.

(Hogenboom et al., 2011) Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. An Overview of Event Extraction from Text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at 10th International Semantic Web Conference (ISWC)*, volume 779, pages 48–57, Bonn, Germany, 2011.

(Holley, 2009) Rose Holley. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, 15(3/4), 2009.

(Hoste, 2009) Veronique Hoste. Name Recognition: NER and WEPS. Unpublished lecture slides, 2009.

(ISO, 2011a) ISO. Information and Documentation – Thesauri and Interoperability with Other Vocabularies – Part 1: Thesauri for Information Retrieval. ISO 25964–1:2011, International Organization for Standardization, Geneva, Switzerland, 2011a.

(ISO, 2011b) ISO. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. ISO 25010:2011, International Organization for Standardization, Geneva, Switzerland, 2011b.

(Jankowski, 2009) Nicholas W. Jankowski. *E-Research: Transformation in Scholarly Practice*. Routledge, New York, 2009.

(Ji and Grishman, 2006) Heng Ji and Ralph Grishman. Data Selection in Semi-Supervised Learning for Name Tagging. In *Proceedings of the Workshop on Information Extraction Beyond the Document*, pages 48–55, Sydney, Australia, 2006.

(Ji et al., 2014) Heng Ji, Joel Nothman, and Ben Hachey. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proceedings of the Text Analysis Conference (TAC)*, 2014.

(Jurafsky and Martin, 2009) Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2009.

(Juran, 1951) Joseph M. Juran. *Quality Control Handbook*. McGraw–Hill, New York, NY, USA, 1951.

(Kent, 1978) William Kent. *Data and Reality: Basic Assumptions in Data Processing Reconsidered*. North Holland, Amsterdam, 1978.

(Kim et al., 2009) Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, Boulder, CO, USA, 2009.

(Kim et al., 2010) Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. A Cross-Lingual Annotation Projection Approach for Relation Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 564–571, Beijing, China, 2010.

(Knuth, 2014) Magnus Knuth. Linked Data Cleansing and Change Management. In *Knowledge Engineering and Knowledge Management*, pages 201–208. Springer, 2014.

(Koehn, 2005) Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005.

(Kozareva et al., 2007) Zornitsa Kozareva, Óscar Ferrández, Andrés Montoyo, Rafael Muñoz, Armando Suárez, and Jaime Gómez. Combining Data-Driven Systems for Improving Named Entity Recognition. *Data & Knowledge Engineering*, 61 (3):449–466, 2007.

(Kramdi et al., 2009) Seif Eddine Kramdi, Ollivier Haemmerlé, and Nathalie Hernandez. Approche générique pour l'extraction de relations à partir de textes. In *Actes des Journées Francophones d'Ingénierie des Connaissances*, pages 97–108. Presses Universitaires de Grenoble, 2009.

(Kripke, 1980) Saul Kripke. *Naming and Necessity*. Harvard University Press, Cambridge, MA, USA, 1980.

(Kulkarni et al., 2009) Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining*, pages 457–466, Paris, 2009.

(Kupietz et al., 2010) Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 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, 2010.

(Ladrière, 1984) Jean Ladrière. *L'articulation du sens: Discours scientifique et parole de la foi*, volume 1. Éditions du Cerf, Paris, 1984.

(Le Roy Ladurie, 1973) Emmanuel Le Roy Ladurie. *Le territoire de l'historien*. Gallimard, Paris, 1973.

(Leal et al., 2012) José Paulo Leal, Vânia Rodrigues, and Ricardo Queirós. Computing Semantic Relatedness using DBpedia. In *Proceedings of the 1st Symposium on Languages, Applications and Technologies (SLATE)*, pages 133–147. Schloss Dagstuhl, 2012.

(Lefever et al., 2009) Els Lefever, Lieve Macken, and Veronique Hoste. Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 496–504, Athens, 2009.

(Lehmann et al., 2015) Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, and Sören Auer. DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

(Liekens et al., 2011) Anthony M. Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, and Jurgen Del-Favero. BioGraph: Unsupervised Biomedical Knowledge Discovery via Automated Hypothesis Generation. *Genome Biology*, 12(6), 2011.

(Lin et al., 2010) Yiling Lin, Jae-Wook Ahn, Peter Brusilovsky, Daqing He, and William Real. 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, 2010.

(Liu and Curran, 2006) Vinci Liu and James R. Curran. Web Text Corpus for Natural Language Processing. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, Trento, Italy, 2006.

(Lonsdale et al., 2010) Deryle W. Lonsdale, David W Embley, and Stephen W Liddle. Ontologies for Multilingual Extraction. In *Proceedings of the 1st International Workshop on the Multilingual Semantic Web*, pages 1–4, Raleigh, NC, USA, 2010.

(Lui and Baldwin, 2012) Marco Lui and Timothy Baldwin. langid.py: An Off-the-Shelf Language Identification Tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, Jeju, South Korea, 2012. ACL.

(Maedche and Staab, 2001) Alexander Maedche and Steffen Staab. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

(Mahdisoltani et al., 2015) Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*, Asilomar, CA, USA, 2015.

(Makhoul et al., 1999) John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance Measures for Information Extraction. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 249–252, Gaithersburg, MD, USA, 1999.

(Marcus et al., 1993) Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

(Markert and Nissim, 2006) Katja Markert and Malvina Nissim. Metonymic Proper Names: A Corpus-Based Account. In Anatol Stefanowitsch and Stefan Th. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, pages 152–174. De Gruyter Mouton, 2006.

(Masud et al., 2010) Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han, and Bhavani Thuraisingham. Addressing Concept-Evolution in Concept-Drifting Data Streams. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pages 929–934, Sydney, Australia, 2010.

(Maturana et al., 2013) Ricardo Alonso Maturana, María Ortega, María Elena Alvarado, Susana López-Sola, and María José Ibáñez. Mismuseos.net: Art After Technology. Putting Cultural Data to Work in a Linked Data Platform. LinkedUp Veni Challenge, 2013.

(Mazlack and Feinauer, 1980) Lawrence J. Mazlack and Richard A. Feinauer. Establishing a Basis for Mapping Natural-Language Statements onto a Database Query Language. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 192–202, Kent, UK, 1980.

(McCallum, 2005) Andrew McCallum. Information Extraction: Distilling Structured Data from Unstructured Text. *Queue*, 3(9):48–57, 2005.

(McCallum, 2012) Q. Ethan McCallum, editor. *Bad Data Handbook: Mapping the World of Data Problems*. O'Reilly Media, Sebastopol, CA, USA, 2012.

(Mehdad et al., 2010) Yashar Mehdad, Matteo Negri, and Marcello Federico. Towards Cross-Lingual Textual Entailment. In *Proceedings of the 11th Annual Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 321–324, Los Angeles, 2010.

(Mendes and Jakob, 2013) Pablo N. Mendes and Max Jakob. Semantic Exploration of Open Source Software Project Descriptions. In *Proceedings of the 2nd International Workshop on Intelligent Exploration of Semantic Data (IESD)*, Paris, France, 2013.

(Mendes et al., 2011) Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, Graz, Austria, 2011.

(Merchant et al., 1996) Roberta Merchant, Mary Ellen Okurowski, and Nancy Chinchor. The Multilingual Entity Task (MET) Overview. In *Proceedings of the TIPSTER Text Program: Phase II*, pages 445–447, Vienna, VA, USA, 1996.

(Meunier, 2014) Vincent Meunier. ISO 25964 et formalisation de la distinction entre concept et terme. Présupposés conceptuels et implications opérationnelles. Master's thesis, Université libre de Bruxelles (ULB), 2014.

(Mikolov et al., 2013) Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v2, 2013.

(Miles and Bechhofer, 2009) Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 2009.

(Miliaraki et al., 2015) Iris Miliaraki, Roi Blanco, and Mounia Lalmas. From "Selena Gomez" to "Marlon Brando": Understanding Explorative Entity Search. In *Proceedings of the 24th International World Wide Web Conference*, Florence, Italy, 2015.

(Miller, 1995) George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

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

(Miwa et al., 2010) Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. Event Extraction with Complex Event Classification using Rich Features. *Journal of Bioinformatics and Computational Biology*, 8(1):131–146, 2010.

(Moens, 2006) Marie-Francine Moens. *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

(Moles, 1995) Abraham A. Moles. *Les sciences de l'imprécis*. Éditions du Seuil, Paris, 1995.

(Mollá et al., 2006) Diego Mollá, Menno Van Zaanen, and Daniel Smith. Named Entity Recognition for Question Answering. *Proceedings of the 4th Australasian Language Technology Workshop (ALTW)*, pages 51–58, 2006.

(Mooney and Nahm, 2003) Raymond J. Mooney and Un Yong Nahm. Text Mining with Information Extraction. In *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, pages 141–160, Bloemfontein, South Africa, 2003.

(Moreau et al., 2007) Fabienne Moreau, Vincent Claveau, and Pascale Sébillot. Combining Linguistic Indexes to Improve the Performances of Information Retrieval Systems: A Machine Learning Based Solution. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 369–387. Le centre de hautes études internationales d'informatique documentaire, 2007.

(Moretti, 2005) Franco Moretti. *Graphs, Maps, Trees : Abstract Models for a Literary History*. Verso, London, 2005.

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

(Moura and Davis, 2014) Tiago Henrique V. M. Moura and Clodoveu Augusto Jr. Davis. Integration of Linked Data Sources for Gazetteer Expansion. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, Dallas, TX, USA, 2014. ACM.

(Nadeau and Sekine, 2007) David Nadeau and Satoshi Sekine. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

(Nebhi, 2013) Kamel Nebhi. A Rule-Based Relation Extraction System using DBpedia and Syntactic Parsing. In *Proceedings of the 1st International Workshop on NLP and DBpedia*, Sydney, Australia, 2013.

(Olson, 2003) Jack E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, San Francisco, 2003.

(Oostdijk et al., 2008) Nelleke Oostdijk, Martin Reynaert, Paola Monachesi, Gertjan Van Noord, Roeland Ordelman, Ineke Schuurman, and Vincent Vandeghinste. From D-Coi to SoNaR: A Reference Corpus for Dutch. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

(Otlet, 1934) Paul Otlet. *Traité de documentation: le livre sur le livre, théorie et pratique*. Editions Mundaneum, 1934.

(Palmer and Day, 1997) David D. Palmer and David S. Day. A Statistical Profile of the Named Entity Task. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 190–193, Stroudsburg, PA, USA, 1997.

(Paskin, 1999) Norman Paskin. Toward Unique Identifiers. *Proceedings of the IEEE*, 87(7):1208–1227, 1999.

(Pereira and Warren, 1980) Fernando C. N. Pereira and David H. D. Warren. Definite Clause Grammars for Language Analysis – A Survey of the Formalism and a Comparison with Augmented Transition Networks. *Artificial Intelligence*, 13 (3):231–278, 1980.

(Pinker, 1994) Steven Pinker. *The Language Instinct. How the Minds Creates Language*. William Morrow and Company, New York, 1994.

(Pomian, 1984) Krzysztof Pomian. *L'ordre du temps*. Gallimard, Paris, 1984.

(Prost, 1996) Antoine Prost. *Douze leçons sur l'histoire*. Éditions du Seuil, Paris, 1996.

(Pustejovsky, 1991) James Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441, 1991.

(Pustejovsky and Boguraev, 1993) James Pustejovsky and Branimir Boguraev. Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence*, 63(1):193–223, 1993.

(Pustejovsky et al., 2003) James Pustejovsky, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New Directions in Question Answering*, 3:28–34, 2003.

(Raimond et al., 2013) Yves Raimond, Michael Smethurst, Andrew McParland, and Christopher Lowis. 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, 2013.

(Rajaraman and Ullman, 2011) Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, MA, USA, 2011.

(Ramakrishnan et al., 2006) Cartic Ramakrishnan, Krys J. Kochut, and Amit P. Sheth. A Framework for Schema-Driven Relationship Discovery from Unstructured Text. In *The Semantic Web-ISWC 2006*, pages 583–596. Springer, 2006.

(Ramsay and Rockwell, 2012) Stephen Ramsay and Geoffrey Rockwell. Developing Things: Notes towards an Epistemology of Building in the Digital Humanities. In Matthew K. Gold, editor, *Debates in the Digital Humanities*, pages 75–84. University of Minnesota Press, 2012.

(Ratinov et al., 2011) Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, volume 1, pages 1375–1384, Portland, OR, USA, 2011.

(Rau, 1991) Lisa F. Rau. Extracting Company Names from Text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*, volume 1, pages 29–32, Miami Beach, FL, USA, 1991.

(Rayward, 1994) W. Boyd Rayward. Visions of Xanadu: Paul Otlet (1868–1944) and Hypertext. *JASIS*, 45(4):235–250, 1994.

(Redman, 1997) Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Boston, 1997.

(Richards, 1929) Ivor Armstrong Richards. *Practical Criticism: A Study of Literary Judgment*. Harcourt Brace Jovanovich, New York, 1929.

(Richman and Schone, 2008) Alexander E. Richman and Patrick Schone. Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies*, pages 1–9, Columbus, Ohio, 2008.

(Rickert, 1986) Heinrich Rickert. *The Limits of Concept Formation in Natural Science: A Logical Introduction to the Historical Sciences*. Cambridge University Press, 1986.

(Riloff and Lorenzen, 1998) Ellen Riloff and Jeffrey Lorenzen. Extraction-Based Text Categorization: Generating Domain-Specific Role Relationships Automatically. In *Natural Language Information Retrieval*, pages 167–196. Kluwer Academic Publishers, 1998.

(Rizzo and Troncy, 2011) Giuseppe Rizzo and Raphaël Troncy. 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, 2011.

(Rizzo and Troncy, 2012) Giuseppe Rizzo and Raphaël Troncy. 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, Avignon, France, 2012. ACL.

(Rodriquez et al., 2012) Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. Comparison of Named Entity Recognition Tools for Raw OCR Text. In *Proceedings of KONVENS*, pages 410–414, Vienna, 2012.

(Romano et al., 2006) Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. Investigating a Generic Paraphrase-Based Approach for Relation Extraction. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, Trento, Italy, 2006.

(Ruiz and Poibeau, 2015) Pablo Ruiz and Thierry Poibeau. 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, 2015.

(Russell and Norvig, 2009) Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3rd edition, 2009.

(Saggion et al., 2007) Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. Ontology-Based Information Extraction for Business Intelligence. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, pages 843–856, Busan, Korea, 2007.

(Saygin et al., 2000) Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. Turing Test: 50 Years Later. *Minds and Machines*, 10:463–518, 2000.

(Schmidt, 2010) Desmond Schmidt. The Inadequacy of Embedded Markup for Cultural Heritage Texts. *Literary and Lingusitic Computing*, 25(3):337–356, 2010.

(Scholz, 2010) Ronny Scholz. Traiter le multilinguisme dans les discours politiques. Possibilités et limites d'une analyse lexicométrique dans un corpus composé de textes de différentes langues. In *Intervention au séminaire du 19 mars 2010*. Ceditec : Université Paris Est Créteil Val de Marne (UPEC), 2010.

(Schreibman et al., 2008) Susan Schreibman, Ray Siemens, and John Unsworth. *A Companion to Digital Humanities*. John Wiley & Sons, Hoboken, NJ, USA, 2008.

(Searle, 1980) John R. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

(Segers et al., 2011) Roxane Segers, Marieke van Erp, Lourens van der Meij, Lora Aroyo, Guus Schreiber, Bob Wielinga, Jacco van Ossenbruggen, Johan Oomen, and Geertje Jacobs. 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, 2011.

(Sekine et al., 2002) Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, 2002.

(Shen et al., 2015) Wei Shen, Jianyong Wang, and Jiawei Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

(Shirky, 2003) Clay Shirky. The Semantic Web, Syllogism, and Worldview. `http://www.shirky.com/writings/semantic_syllogism.html`, 2003.

(Singh et al., 2011) Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Large-Scale Cross-Document Coreference using Distributed Inference and Hierarchical Models. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, volume 1, pages 793–803, Portland, OR, USA, 2011.

(Singh et al., 2012) Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Wikilinks: A Large-Scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015*, 2012.

(Snow et al., 2008) Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, 2008. ACL.

(Spaeth, 2004) Donald Spaeth. Representing Text as Data. *Historical Methods*, 37(2): 218–239, 2004.

(Speck and Ngomo, 2014) René Speck and Axel-Cyrille Ngonga Ngomo. Named Entity Recognition using FOX. In *Proceedings of the Posters & Demonstrations Track within the 13th International Semantic Web Conference (ISWC)*, Riva del Garda, Italy, 2014.

(Speer and Havasi, 2012) Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3679–3686, Istanbul, Turkey, 2012.

(Steiner, 2014) Thomas Steiner. Bots vs. Wikipedians, Anons vs. Logged-Ins. In *Proceedings of the 23rd International World Wide Web Conference*, pages 547–548, Seoul, Korea, 2014.

(Stern, 2013) Rosa Stern. *Identification automatique d'entités pour l'enrichissement de contenus textuels*. PhD thesis, Université Paris-Diderot – Paris VII, 2013.

(Stone, 1979) Lawrence Stone. The Revival of Narrative: Reflections on a New Old History. *Past and Present*, 85(1):3–24, 1979.

(Sundheim, 1995) Beth M. Sundheim. Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Conference on Message Understanding (MUC)*, pages 13–31, Stroudsburg, PA, USA, 1995.

(Svensson, 2010) Patrik Svensson. The Landscape of Digital Humanities. *Digital Humanities Quarterly*, 4(1), 2010.

(Taine, 1875) Hippolyte Taine. *Les origines de la France contemporaine*, volume 1. Laffont, Paris, 1875.

(Tamilin et al., 2010) Andrei Tamilin, Bernardo Magnini, Luciano Serafini, Christian Girardi, Mathew Joseph, and Roberto Zanoli. Context-Driven Semantic Enrichment of Italian News Archive. In *Proceedings of the 7th international conference on The Semantic Web: research and Applications*, volume 1, pages 364–378, Heraklion, Greece, 2010.

(Tang et al., 2015) Jie Tang, Zhanpeng Fang, and Jimeng Sun. Incorporating Social Context and Domain Knowledge for Entity Recognition. In *Proceedings of the 24th International World Wide Web Conference*, Florence, Italy, 2015.

(Tanner et al., 2009) Simon Tanner, Trevor Muñoz, and Pich Hemy Ros. Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. *D-Lib Magazine*, 15(7/8), 2009.

(Tjong Kim Sang, 2002) Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 6th Conference on Computational Natural Language Learning (CoNLL)*, pages 155–158, Taipei, Taiwan, 2002.

(Tjong Kim Sang and De Meulder, 2003) Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL)*, pages 142–147, Edmonton, Canada, 2003.

(Tsymbal, 2004) Alexey Tsymbal. The Problem of Concept Drift: Definitions and Related Work. Technical report, Computer Science Department, Trinity College Dublin, 2004.

(Turing, 1950) Alan M. Turing. Computing Machinery and Intelligence. *Mind*, LIX: 433–460, 1950.

(Tylenda et al., 2014) Tomasz Tylenda, Sarath Kumar Kondreddi, and Gerhard Weikum. Spotting Knowledge Base Facts in Web Texts. In *Proceedings of the 4th Workshop on Automated Knowledge Base Construction*, pages 1–6, Montreal, Canada, 2014.

(Valsecchi et al., 2015) Fabio Valsecchi, Matteo Abrate, Clara Bacciu, Maurizio Tesconi, and Andrea Marchetti. DBpedia Atlas: Mapping the Uncharted Lands of Linked Data. In *Proceedings of the 8th Workshop on Linked Data on the Web (LDOW)*, Florence, Italy, 2015.

(van Hooland, 2009) Seth van Hooland. *Metadata Quality in the Cultural Heritage Sector: Stakes, Problems and Solutions*. PhD thesis, Université libre de Bruxelles (ULB), 2009.

(van Hooland and Verborgh, 2014) Seth van Hooland and Ruben Verborgh. *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*. Facet Publishing, London, 2014.

(van Hooland et al., 2013) Seth van Hooland, Ruben Verborgh, Max De Wilde, Johannes Hercher, Erik Mannens, and Rik Van de Walle. Evaluating the Success of Vocabulary Reconciliation for Cultural Heritage Collections. *Journal of the American Society for Information Science and Technology*, 64(3):464–479, 2013.

(van Hooland et al., 2015) Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. *Digital Scholarship in the Humanities*, 30(2):262–279, 2015.

(van Rijsbergen, 1975) Cornelis Joost van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1975.

(van Zaanen and Freeman, 2004) Menno van Zaanen and Robert John Freeman. Reducing Subjectivity of Natural Language Processing System Evaluation. Unpublished paper, 2004.

(Verborgh and De Wilde, 2013) Ruben Verborgh and Max De Wilde. *Using OpenRefine*. Packt Publishing, Birmingham, 2013.

(Verborgh et al., 2015) Ruben Verborgh, Seth van Hooland, Aaron Straup Cope, Sebastian Chan, Erik Mannens, and Rik Van de Walle. The Fallacy of the Multi-API Culture: Conceptual and Practical Benefits of Representational State Transfer (REST). *Journal of Documentation*, 71(2):233–252, 2015.

(Vossen et al., 2012) Piek Vossen, Aitor Soroa, Benat Zapirain, and German Rigau. Cross-Lingual Event-Mining using Wordnet as a Shared Knowledge Interface. In *Proceedings of the 6th Global Wordnet Conference*, pages 382–390, Matsue, Japan, 2012.

(Vrandečić, 2012) Denny Vrandečić. Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of the 21st International Conference Companion on World Wide Web*, pages 1063–1064, Lyon, France, 2012.

(Vrandečić and Krötzsch, 2014) Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

(Wang, 1998) Richard Y. Wang. A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2):58–65, 1998.

(Watrin, 2006) Patrick Watrin. *Une approche hybride de l'extraction d'information: sous-langages et lexique-grammaire*. PhD thesis, Université Catholique de Louvain, 2006.

(Widmer and Kubat, 1996) Gerhard Widmer and Miroslav Kubat. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23(1):69–101, 1996.

(Wiener, 1994) Lauren Ruth Wiener. *Les avatars du logiciel*. Éditions Addison-Wesley France, Paris, 1994.

(Wilks, 2008) Yorick Wilks. The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics? *Intelligent Systems, IEEE*, 23(3):41–49, 2008.

(Wilks and Brewster, 2009) Yorick Wilks and Christopher Brewster. Natural Language Processing as a Foundation of the Semantic Web. *Foundations and Trends in Web Science*, 1(3–4):199–327, 2009.

(Wismann, 2012) Heinz Wismann. *Penser entre les langues*. Albin Michel, Paris, 2012.

(Wong et al., 2009) Wilson Wong, Wei Liu, and Mohammed Bennamoun. Acquiring Semantic Relations using the Web for Constructing Lightweight Ontologies. In *Advances in Knowledge Discovery and Data Mining*, pages 266–277. Springer, 2009.

(Wubben and van den Bosch, 2009) Sander Wubben and Antal van den Bosch. A Semantic Relatedness Metric based on Free Link Structure. In *Proceedings of the 8th International Conference on Computational Semantics*, pages 355–358, Tilburg, The Netherlands, 2009. ACL.

(Yakushiji et al., 2001) Akane Yakushiji, Yuka Tateisi, Yusuke Miyao, and Jun'ichi Tsujii. Event Extraction from Biomedical Papers using a Full Parser. In *Proceedings of the 6th Pacific Symposium on Biocomputing*, pages 408–419, Hawaii, USA, 2001.

(Yosef et al., 2011) Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. *Proceedings of the 37th International Conference on Very Large Databases (VLDB)*, 4(12):1450–1453, 2011.

(Zampieri et al., 2014) Marcos Zampieri, Liling Tan, Nikola Ljubešic, and Jörg Tiedemann. A Report on the DSL Shared Task 2014. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland, 2014.

(Zaveri et al., 2013) Amrapali Zaveri, Dimitris Kontokostas, Mohamed A. Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. User-Driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104, Graz, Austria, 2013.

(Zhang et al., 2015) Junsheng Zhang, Yunchuan Sun, and Antonio J. Jara. Towards Semantically Linked Multilingual Corpus. *International Journal of Information Management*, 35(3):387–395, 2015.

(Zheng et al., 2014) Jin Guang Zheng, Daniel Howsmon, Boliang Zhang1 Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. Entity Linking for Biomedical Literature. In *Proceedings of the 8th International Workshop on Data and Text Mining in Bioinformatics*, Shanghai, China, 2014.

(Zhou and He, 2011) Deyu Zhou and Yulan He. Biomedical Events Extraction using the Hidden Vector State Model. *Artificial Intelligence in Medicine*, 53(3):205–213, 2011.