

File ID 182357  
Filename 1: Introduction

---

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation  
Title Ontology enrichment from heterogeneous sources on the web  
Author V. de Boer  
Faculty Faculty of Science  
Year 2010  
Pages viii, 162

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/345217>

---

*Copyright*

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use.*

---

---

## INTRODUCTION

---

The Semantic Web is a proposed web of linked data that co-exists with the current World Wide Web of linked documents. One way of building this Semantic Web is to extract the classes, instances and relations from the text documents on the current Web. This can be done either by hand or (semi-)automatically. This thesis is about methods that automatically extract structured information from text documents on the Web so that they can be used by Semantic Web applications. The methods we propose for this exploit a property of the current Web of documents: the fact that a lot of information is redundantly available. By using simple and therefore generally applicable extraction methods, we can aggregate evidence from multiple sources, contributing to the performance of these methods. In this introduction, we first provide some more background on this task.

### 1.1 BACKGROUND

Since the end of the Twentieth century, the Internet has rapidly grown in size, number of users and in functionality. Where the Internet was first only conceived as a means of communication, with the publication of the World Wide Web project by Tim Berners-Lee and Robert Cailliau in the early 1990s [Berners-Lee and Cailliau, 1990], its main application quickly became the interconnected network of hypertext documents that can be explored using web browsers. Humans with an information need use these web browsers to locate and extract relevant information from natural language fragments, tables and lists that appear in these documents. With the growing popularity of the Web, more functionality emerged as new technologies allowed for interactive applications and user-generated content. To acknowledge this progression of the Web, these combined technologies have been dubbed the Web 2.0. While this new version of the Web opens up many more usage possibilities, the Web still continues to evolve.

One of these currently evolving developments is that of the Semantic Web. The concept of the Semantic Web was introduced by Tim Berners-Lee together with Jim Hendler and Ora Lassila in their 2001 paper [Berners-Lee et al., 2001]. Where the old World Wide Web is in its essence a web of linked documents intended for human readers, the Semantic Web is an initiative that attempts to establish a web of interlinked data, information and knowledge that can be used and interpreted by programs.

For example, to find out which Dutch painters belonged to the "Impressionism" movement from the Web of documents one has to read through relevant documents to find the answer. In the proposed Semantic Web, such a question can be answered directly by a program when the relevant knowledge is available. By traversing and interpreting the interlinked data, the program should be able to produce a correct answer to the question.

For the Semantic Web, a set of new technologies including the Resource Description Framework (RDF) and the Web Ontology Language (OWL), are proposed to facilitate the formalization and sharing of this information. This formalization takes place in the form of ontologies, which make up the backbone of the Semantic Web. These ontologies define for a given domain, the types of concepts and relations that are used in that domain (such as the concept of a painter, an art style or the relation between the two). Knowledge bases that conform to these ontologies hold the actual information instances of the classes and the relations (such as the fact that Vincent van Gogh belongs to the Post-Impressionism movement). By establishing relations between classes in ontologies and instances in knowledge bases, this web of interlinked information expands and increasingly more complex questions can be answered.

## 1.2 SEMI-AUTOMATIC ONTOLOGY ENRICHMENT

### 1.2.1 *Ontology Learning, Population and Enrichment*

Ontology Engineering concerns itself with constructing the ontologies that are used in the Semantic Web and filling the knowledge base with instances of these ontologies. For smaller domains, creating the ontologies and filling them by hand is a feasible task. For larger domains, manual ontology engineering is too time-consuming and (semi-)automatic methods are needed (see for instance work by [Maedche and Staab \[2001\]](#) or [Cimiano et al. \[2004\]](#)). Since the Web of documents already contains a large amount of information in almost every domain imaginable, this is a good source that these automatic methods can use as a corpus from which to extract the target information (cf. [\[Craven et al., 1998\]](#)). By doing this, we are extracting the Semantic Web from the current World Wide Web. In this thesis, we present a number of methods that (semi-)automatically extract semantic knowledge from HTML documents.

We distinguish a number of automatic ontology engineering tasks. For this, we first clarify the notions used throughout this thesis. An *ontology* is a shared formalization of a conceptualization of some domain (cf. [\[Gruber, 1995\]](#), [\[Uchold and Gruninger, 1996\]](#)). It contains classes and relations between those classes. A *knowledge base* that conforms to this ontology contains the instances of the classes and instances of the relations. In other words, it holds the actual data of the domain.

*Ontology Learning* is the automatic identification of the classes and relations in the ontology in some corpus [\[Maedche and Staab, 2001\]](#). For *Ontology Population*, the goal is to extract instances of the classes or the instances of the relations and store them in the knowledge base [\[Buitelaar et al., 2005\]](#). *Ontology Enrichment*

is a term denoting cases of the Ontology Learning or Population task where a partial ontology and knowledge base is already available. In this thesis, we are concerned with methods for automatic Ontology Population, more specifically the extraction of relation instances. We do this for previously defined ontologies that are already partly populated and therefore this is an Ontology Enrichment task. A task related to the learning of relations is that of Ontology Mapping. Here, relations (usually part-of, same-as or hyperonymy/hyponymy relations) are learned between terms in different ontologies or thesauri in order to achieve a mapping between the two or more ontologies/thesauri. In recent years, the Web is more and more used as a source of information from which ontology elements and knowledge base instances and mappings between ontologies can be learned (cf. [Gligorov et al., 2007], [van Hage et al., 2006], [Malaisé et al., 2007]).

### 1.2.2 The Information Extraction Task

Ontology Population is closely related to the notion of *Information Extraction*. Information Extraction is a subtask of Information Retrieval, where the goal is to automatically extract structured data from unstructured text in a corpus [Mooney and Bunescu, 2005, McCallum, 2005]. A popular Information Extraction task is that of *Named Entity Recognition* (NER). Here the goal is to identify in a (set of) text document(s) the occurrences of instances of a set of predefined target classes such as person names, locations or time expressions. Every occurrence of an instance of a target class in the text is to be recognized by the NER system and all hits and misses count toward the final performance. Another subtask of Information Extraction is *Question Answering* (QA), where the goal is to provide a single answer to a natural language question of a user. This answer is to be extracted from a corpus. Here, the performance of the method is only determined by whether or not an answer is correct. In other words, the target information needs to be found only once. The Information Extraction task that this thesis deals with, Ontology Population, shares its goal with QA to extract a target piece of information from a corpus. Although information about multiple occurrences can be aggregated to determine the likelihood that a target instance is correct, for measuring performance, we do not care whether or not we have found every occurrence of an instance in the corpus. Here the input is not a natural language question but a class or relation as defined in the ontology for which instances are to be extracted. The resulting information is stored in the knowledge base.

The general approach to Information Extraction is to extract the desired information from the text using patterns. The type of patterns that can be used varies per method. These patterns can be either constructed by hand or can be derived (semi-)automatically using machine learning techniques. For the machine learning task, there exist effective supervised methods such as that described by Bunescu and Mooney [2006] or Zelenko et al. [2003]. These methods require significant amounts of examples to learn from and manual annotation of large corpora is recognized to be expensive and tedious. Unsupervised learning methods have also been developed, for example by [Grenager et al., 2005] where prior knowledge is used to aid the learning of field segmentation models. However, in

general unsupervised methods unfortunately do not attain performance close to state-of-the-art supervised methods (cf. [Bellare and McCallum, 2009] or, [Ireson et al., 2005]). Semi-supervised learning methods that learn from a small set of examples are increasingly popular for the Information Extraction task. Examples of semi-supervised approaches include work by Michelson and Knoblock [2007], and Bunescu and Mooney [2007]. In Chapter 2, we describe a semi-supervised approach to learning relations.

We can further divide the automatic IE methods by the type of information they use. Broadly speaking, we identify two categories: those that make heavy use of natural language processing (*NLP-based methods*) and those that use patterns that exploit structural regularities in (HTML) documents: *wrapper-based methods*. Modern Information Extraction systems employ a multitude of extraction methods and heuristics and are not necessarily restricted to either NLP-or wrapper-based methods. However, the following issues remain for the individual methods.

NLP-based methods make use of a range of natural language processing techniques including part-of-speech-tagging, stemming, phrase chunking, dependency tree construction etc.. This information is then used in the extraction patterns with which new instances of a target class or relation are extracted [Mooney and Bunescu, 2005]. One problem with these methods is that they are very much dependent on the language -a model trained in one language will not be able to extract new information from documents in a second language. Even within one language, the documents in the corpus are required to be more or less natural language (i.e. complete grammatically correct sentences). Although the latter assumption holds for often-used static corpora such as newspaper corpora, a lot of information in Web documents does not appear in these natural language sentences but in lists, information boxes or short phrases.

Wrapper-based methods on the other hand are explicitly designed to exploit these structures in corpus documents [Kushmerick et al., 1997]. Wrappers are essentially patterns that extract target information from a text. These patterns can consist of textual and structural information and are very useful to extract terms that occur in the same context as similar terms. These wrappers can be made by hand or, in wrapper induction, learned from a set of labeled instances. Wrappers perform well on semi-structured documents but have a hard time extracting information from natural language, as the number of possible variations of denoting the same information is very large. When the goal is to extract information from a static, offline corpus, a method that is suited for that specific corpus can be used as these methods can efficiently exploit the particularities of the documents in that corpus. For a newspaper corpus consisting of natural language documents in one language, NLP-based methods will perform well. For a corpus of (semi-)structured documents, wrapper-based techniques are most suitable.

### 1.2.3 *Heterogeneity of the Web*

The World Wide Web is highly heterogeneous in multiple ways. First, it consists of documents in many different languages. Second, the document types are

heterogeneous. The Web consists of 'free text' documents containing more or less grammatically correct sentences, structured documents containing tables and lists and semi-structured documents (a combination of both). Thirdly, the 'genre' of web documents differs: the web consists of commercial, encyclopedic or personal documents, blog posts, news articles, homepages etc. In documents from different genres, the same information will be listed in different forms, introducing heterogeneity. Methods that can cope with more different types of documents are applicable to a larger part of the Web corpus and will therefore be able to extract more heterogeneous information. This makes it possible to exploit the redundancy of information on the Web.

The 'redundancy' from this thesis' title refers to the property of the Web that a lot of information occurs in multiple documents. This is partly because of the size of the Web and because it has a lot of different authors copying, reusing and rephrasing information. This means that a lot of information is available in different formats in different document types. Methods that are able to extract information from many different types of documents and combine this information can use frequencies of occurrences to determine the likelihood that an extracted instance or relation is correct. Moreover, methods that can exploit redundancy are more robust: if the target information cannot be extracted from one source, chances are it can be extracted from other documents on the Web. The more generally applicable a method is, the more different types of documents it can handle and the more the redundancy of information on the Web can be exploited.

A strategy for generally applicable methods is to use simple co-occurrence of terms. Generally speaking, this strategy assumes that when terms often co-occur in the same context, there is a relation between them. Extraction by co-occurrence can be seen as a generalization of the wrapper- and NLP-based methods where the patterns used are extremely generalized and no document-specific information is used other than that two terms must co-occur within a sentence, paragraph or document. Co-occurrence-based methods are simple by design and thus rely less on language or document structure. They are more generally applicable than NLP- or wrapper-based techniques. An intrinsic drawback of a more general method such as a co-occurrence-based method is that they are very coarse and that precision can be relatively low. They are however able to extract information from a larger part of the corpus. By exploiting the redundancy of information in this larger corpus, we hypothesize that we can achieve the same or better accuracy than more specific methods can achieve.

In Chapter 2, we investigate this hypothesis by presenting a semi-supervised relation instantiation method that works on large web corpora with documents of heterogeneous structure. For this we use a simple co-occurrence-based extraction strategy. In Chapter 3, we also use this strategy for a more specific Ontology Enrichment task: to find time periods for concepts. In Chapter 4, we investigate how exactly the generality of patterns influences the performance of an extraction method. In that chapter, we use hand-made patterns to extract the target information. In Chapter 5, we combine the different methods and ontological background knowledge to see how the use of multiple heterogeneous methods increases the Ontology Enrichment performance.

### 1.3 MULTIMEDIAN E-CULTURE REPOSITORIES AND VOCABULARIES

The research in this thesis has been carried out within the context of the MultimediaN E-culture project. The main goal of this project is to construct an application that demonstrates Semantic Web and novel information presentation technologies within a cultural heritage context [Schreiber et al., 2006]. The tool shows how the use of (Semantic) Web standards facilitates interoperability between different vocabularies and different cultural heritage collections. For this purpose, object descriptions of different Dutch cultural heritage institutions are mapped to common vocabularies. Relations between different instances describing art objects and other art terms such as art styles or people result in a large network of knowledge through which a user can browse, search and interrelate different types of information.

The total database of the current version of the MultimediaN E-culture demonstrator holds more than 100.000 descriptions of art objects from four Dutch cultural heritage institutions and one online art archive. The demonstrator uses a number of thesauri and vocabularies to describe the objects descriptions including the three Getty vocabularies<sup>1</sup> (i.e. the Getty Art & Architecture Thesaurus (AAT, >31,000 terms), Union List of Artists Names (ULAN, >130,000 terms) and the Thesaurus of Geographical Names (TGN, >890,000 terms), the SVCN dutch cultural heritage thesaurus (>11,000 terms)<sup>2</sup> and the lexical resource WordNet [Fellbaum, 1998] (>115,000 terms). In total, more than one million classes and instances are stored in these vocabularies, each having one or more labels describing them. Throughout this thesis, when we present experiments to empirically evaluate the proposed Ontology Enrichment methods, we will return to extracting instances of relations defined in the vocabularies and enrich the knowledge base with them. One recurring relation on which we specifically focus is a relation between art styles in the AAT and artists in the ULAN. This relation is not present in the original vocabularies. The interlinkedness of the knowledge base will benefit greatly from adding these links, relating not only artists to art styles, but also artists to other artists (e.g. artists with the same style), artists to works, works to styles etc.. The relation instances that we have discovered in this thesis concerning these vocabularies were added to the demonstrator's knowledge base where they provide additional browsing and search results.

### 1.4 SCALABILITY ISSUES

An important issue in the design of Ontology Enrichment methods is the large size of the repositories and vocabularies used. This size poses some restrictions on the methods that can be used to enrich the ontologies as the methods must be scalable, to be able to cope with the large numbers of instances. For example, in Chapter 4 we will see that loading the entire ULAN, containing more than 130,000 artists, with different name spellings into a text analysis tool causes a number of problems for discovering these artists in text documents. Some preprocessing

<sup>1</sup> [http://www.getty.edu/research/conducting\\_research/vocabularies/](http://www.getty.edu/research/conducting_research/vocabularies/)

<sup>2</sup> [www.svcn.nl](http://www.svcn.nl)

has to be performed to reduce the number of objects loaded in the tool's memory. Another example is given in Chapter 2. Here, we show that for methods that use a web search engine to check the likelihood of a relation, this requires a number of lookups equal to or greater than the sizes of the two sets multiplied. For the discovery of relation instances between two large sets of instances, methods that consider all possible combinations of instances do not scale.

At the same time, the size of the vocabularies used introduces an ambiguity in the terms itself: the more terms there are in a vocabulary, the more difficult it will be to unambiguously map a term in the text to a single instance in the vocabulary. For example, the term "van Gogh" matches three instances in the ULAN, one of which is the famous Dutch painter Vincent van Gogh and in the TGN, the string "Paris" occurs in labels of 102 separate vocabulary instances. Examples of disambiguation measures that are used in this thesis are extracting more descriptive terms that are to be matched ("Vincent van Gogh" instead of just "van Gogh") or to do preprocessing of the vocabulary (such as only considering larger cities in Europe for Paris in the TGN).

We design our methods in such a way that they are *efficient* in the sense that they can deal with large vocabularies and can extract new information for large numbers of instances. The methods must also deal with the problem of the inherent ambiguity of terms in large vocabularies.

## 1.5 APPROACH

In each of the Chapters 2, 3 and 4 we present a method that (semi-)automatically extracts ontology constructs from the Web. These methods are based on the observation that a great deal of information is redundantly available on the Web. We explore the behavior of these methods in real-world Ontology Enrichment tasks in the cultural heritage domain and evaluate their reusability with experiments in other domains. Each of these methods makes use of a web search engine to construct a working corpus. In the working corpus, simple methods are used to extract entities that can be used to enrich the ontology. These extraction methods do not use 'deep' natural language processing techniques and in general are designed to be as domain-, language- and corpus-independent as possible. The general methods have the ability to extract and combine information from multiple sources and can therefore benefit from the redundancy of information on the Web. We will show that by exploiting this redundancy, we can get satisfactory performance on the Ontology Enrichment tasks.

A second way of using redundancy to our benefit is to use multiple Information Extraction/Ontology Enrichment methods together with available background knowledge. In Chapter 5, we present a method to combine the results of these different information sources. When the different extraction methods vary in the way they extract new instances, we can expect that they will have different uncorrelated errors. By combining the results of multiple methods, mistakes made by one method can be corrected by others. In the same fashion, we consider background knowledge in the ontology and knowledge base to filter candidate knowledge base instances that are the result of Information Extraction methods.

By combining evidence from these possibly redundant methods and background knowledge sources, we can expect increased performance.

## 1.6 RESEARCH QUESTIONS

In this thesis, we present a number of methods for the (semi-)automatic Ontology Enrichment, more specifically the extraction of relation instances and class instances. Through this, we investigate the main research question of this thesis:

*How can we efficiently automatically enrich ontologies with large knowledge bases from heterogeneous textual sources on the World Wide Web?*

We refine this general research question into a number of more specific questions that we answer in each of the following chapters. Each question concerns a different specific Ontology Enrichment task for which we present a number of methods in the following chapters. The first question focuses on relation instantiation from heterogeneous sources:

*1. Given a small set of instances of a relation, how can we use a semi-supervised learning method to iteratively extract more instances of this relation from Web documents using simple term matching methods?*

This sub-question considers the task of enriching an ontology by adding more instances of a relation to the knowledge base. More specifically, we assume that a small set of examples of the relation are already present in the knowledge base and that these can be used as labeled examples for a semi-supervised Information Extraction task. The performance of relation extraction techniques that use Natural Language Processing or Wrapper-induction is heavily dependent on the language and the structure of the text. By using simple term matching techniques, we aim for a method that is independent of the language and structure of the documents. In Chapter 2, we propose a method for the extraction of relation instances that uses this bootstrapping approach. By combining information from different documents, we can exploit the redundancy of information on the web. We investigate whether a metric based on co-occurrence with the seed set can correctly rank the candidate relation instances.

*2a. How can we extract time periods from the Web.*

This research question concerns a very specific type of Ontology Enrichment task, that of temporal information. Again, NLP- or wrapper-based are limited to a certain type of documents. Our approach attempts to use simple term matching to extract temporal information from as many documents as possible and combine that using simple statistics. This way, we can exploit the redundancy of temporal information on the web. The extraction method we propose results in a statistical distribution of temporal terms and we argue that many (historical) time periods have a fuzzy rather than a discrete nature. A second question regarding Ontology Enrichment with extracted time periods is:

*2b. How can extracted probabilistic models of time periods be converted to vocabulary instances, while still retaining their fuzzy nature.*

In Chapter 3 we investigate a structured vocabulary that can be used to denote time and time periods. Through this structured vocabulary, complex time indications such as "Beginning of the Twentieth Century" can be stored in the knowledge base. We investigate a method to transform the extracted frequency distributions of dates into terms denoting periods from this structured vocabulary.

*3. For pattern-based web Information Extraction methods whose patterns contain references to semantic classes, how does the generality of patterns affect their performance.*

Pattern-based Information Extraction is often used for Ontology Enrichment or other Information Extraction tasks. New text analysis tools such as the tOKo text analysis tool [Anjewierden, 2006] offer the option to include semantic classes in the pattern, allowing for a wider variety of generality of the patterns (cf. [Califf and Mooney, 2003]). For Information Extraction, specific patterns result in a high precision but often have a low recall. More general patterns have higher recall, at the cost of precision. By using more general patterns in combination with a threshold on the frequency of the occurrences of the results, we also exploit the redundancy of information in the working corpus. In Chapter 4, we investigate how the performance of patterns is influenced by their generality.

*4. How can multiple Information Extraction methods and background knowledge be combined to improve the Ontology Enrichment performance?*

In Ontology Enrichment, there is often a lot of background knowledge available that can be used to improve the performance. The use of background knowledge to check the likelihood of a candidate relation instance can reduce the uncertainty of that relation instance. At the same time, combining information from multiple Information Extraction methods can also reduce this uncertainty. In Chapter 4, we investigate how we can introduce background knowledge to the relation instantiation task and combine information from multiple sources.

## 1.7 OUTLINE OF THE THESIS

This thesis is outlined as follows. In Chapter 2, we present a semi-supervised method for the extraction of relation instances from heterogeneous sources on the web. The method starts with a small seed set of relation instances and iteratively adds new instances extracted from a working corpus retrieved from the web. The method is able to exploit the redundancy of information on the web. We evaluate it using a number of experiments and compare it to other methods.

In Chapter 3, we discuss a method to extract time periods for historical concepts. Furthermore, we present a number of rewriting rules that transform these extracted probabilistic descriptions of the time periods into instances of a structured vocabulary. We evaluate both the performance of the time Information Extraction method and the rewritten descriptions for a number of different concepts in different domains.

Chapter 4 describes a relation instantiation method that uses a text analysis tool that incorporates semantic classes in its pattern search functionality. Using this,

we can efficiently extract candidate instances of relations. With this method, we are also able to extract class instances in addition to the relation instances. We do a number of experiments in different domains using patterns of varying generality and report on the performance of these patterns. Using more general patterns has the result that more information from different source can be considered, therefore counterbalancing the loss in precision caused by not using more specific patterns. We again evaluate the method on a number of relation instantiation tasks in a number of domains, including the cultural heritage domain.

In *Chapter 5*, we present a method for combining the results from different Information Extraction methods and background knowledge in the ontology and knowledge base. We present a number of rules through which background knowledge related to a target relation instance can be used to yield additional positive or negative evidence for a candidate relation instance. Combining all this information, we can assign a higher or lower likelihood to candidate relation instances, thereby increasing the total performance of the Ontology Enrichment process.

In *Chapter 6*, we present the overall conclusions from this thesis