Deliverable 1.5: A survey of ontology learning methods and techniques

Asunción Gómez-Pérez, David Manzano-Macho
Universidad Politécnica de Madrid
asun@fi.upm.es dmanzano@delicias.dia.fi.upm.es

Identifier	Deliverable
Class	**Deliverable**
Version	**3243**
Version date	**30/05/2003**
Status	**Final**
Distribution	**Public**
Responsible Partner	**UPM**

OntoWeb Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-2000-29243.

Next Web Generation
Leopold Franzens University of Innsbruck
Insitute of Computer Science
Next Web Generation - Research Group
Technikerstraße 13
6020 Innsbruck
Austria
Contactperson: Dieter Fensel
E-mail: dieter.fensel@uibk.ac.at

This document has been edited by:
Asunción Gómez-Pérez, David Manzano-Macho
Departamento de Inteligencia Artificial
Facultad de Informática (Universidad Politécnica de Madrid)
Campus de Montegancedo s/n. 28660 Boadilla del Monte. Madrid. Spain

The main contributors of this document are:
Asunción Gómez-Pérez, David Manzano-Macho, Enrique Alfonseca, Rafael Núñez, Ian Blacoe, Steffen Staab, Oscar Corcho, Ying Ding, Jan Paralic, Raphael Troncy

Revision Information

Revision date	Version	Changes
24-01-2003	0.1	Table of Content proposal
26-01-2003	0.1	UPM distributes first version of the deliverable
28-01-2003	0.1	UPM provides the description of the Aussenac-Gilles and colleagues approach (3.2.3), Kietz and colleagues' method (3.2.11), Nobecourt approach (3.2.15), and the first version of the ontology learning from text tools (section 3.3)
31-01-2003	0.2	Ying Ding provides the description of Hwang's method (3.2.9), Wagner's approach (3.2.17), WOLFIE tool (3.3.18), and Jannink and Wiederhold's approach (4.2.1)
10-02-2003	0.3	UPM provides the Aguirre and colleagues' method (3.2.1), Faatz and Steinmetz approach (3.2.5), and OntoBuilder tool (6.3.1)
15-02-2003	0.4	UPM provided the criteria to be described for each method and tool
20-02-2003	0.5	Steffen Staab provides the Deliverable 11 of WonderWeb Project (IST Project 2001-33052), in which appears the description of the Volz and colleagues' approach (6.2.4) and the OntoLift prototype (7.2.4)
22-02-2003	0.5	Jan Paralic sends a paper with several descriptions of techniques used for conceptual clustering
25-02-2003	0.6	UPM provides the Gupta and colleagues' (3.2.6) and Xu and colleagues' approaches (3.2.18)
27-02-2003	0.6	Raphael Troncy provides the description of the Bachimont's method (3.2.4) and DOE tool (3.3.4)
15-03-2004	0.7	UPM provides the description of the Khan and Luo's method (3.2.10) and Hahn and colleagues' method (3.2.7)
31-03-2003	0.8	Enrique Alfonseca sends the description of the Alfonseca and Manandhar's method (3.2.2), Rigau and colleagues' method (4.2.2) , Hearst's approach (3.2.8), Welkin (3.3.17) and SEID tool (4.3.1)
15-04-2003	0.8	UPM provides the final version of the section 7 and the description of the DODDLE tool (4.3.2)
5-05-2003	0.9	Update of the section 3.3 provided from Rafael Núñez
15-05-2003	0.9	UPM provides the final version of the final version of the section 6, adding from the section 6.2.1 to 6.2.3
17-05-2003	0.9	UPM provides the final version of the section 3.2, adding from the section 3.2.12 to 3.2.16

30-05-2003	1.0	Delivery of the first version
30-05-2003	1.0	Steffen Staab suggests to add the Hearst approach and to actulize the KAON reference
4-06-2003	1.0	Nathalie Aussenac-Gilles send some comments and addictions about her method (3.2.3), adds the description of Caméléon (3.3.2), provides some comments about hearst's approach (3.2.8) and Prométhée (3.3.9)
6-06-2003	1.0	Delivery of the final version

Contents

Executive summary

This deliverable presents a survey of the most relevant methods, techniques and tools used for building ontologies from text, machine readable dictionaries, knowledge bases, structured-data, semi-structured data and unstructured data.

The deliverable is organized in eight sections.

The first one aims to emphasize the relevance of the research in the ontology learning for the ontology community and Semantic Web.

The second section shows different approaches for learning ontologies in a semi-automatic fashion.

The third section overviews of the ontology learning from text, presenting a summary of the most relevant methods and tools used to perform it.

In the fourth section, the ontology learning from dictionary is presented as well as the most relevant methods and tools that use a machine readable dictionary to achieve the goal of building an ontology are summarized.

The fifth section deals with the methods used for ontology learning from knowledge bases.

In the sixth section, an overview of the methods and tools used for ontology learning from semi-structured data is presented.

The eighth section shows methods and tools for learning ontologies from databases.

And finally, the last section aims to be a conclusion of the state of the art in the ontology learning area.

For each approach inside this deliverable, we have followed the same structure. First, we describe the methods, later the tools, and finally, some conclusions. In order to allow the comparison of different methods and tools, we have tried to describe the following topics for each one.

- For each method, we will present a general description including its goals and scope of the learning process, general steps used to learn, the knowledge source used for learning (if the method needs other type of sources in addiction to text), the main techniques applied in the process, the possibility of reusing other existing ontologies, the main goals looked for the method, the domain in which it has been applied and tested, if there are a tool associated, how the evaluation of the knowledge learnt is performed, a list of the most relevant ontologies built following it, the URL where more information about it can be found, and relevant bibliography.
- For each tool, we will present a general description including its main goals, the main techniques used by the tool in the learning process, the method followed, the user intervention in the process, the types of sources used by the method, the software architecture, the possibility of interoperate with other tools, the import and export facilities that the tool provides, the interface facilities, a URL where you can find more information, and relevant bibliography.

1 Introduction

The Semantic Web has marked another stage in the ontology field. According to Berners-Lee [Berners-Lee, 1999], the Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. This cooperation can be achieved by using shared knowledge-components, and so ontologies and PSMs have become key instruments in developing the Semantic Web idea. Ontologies represent static domain knowledge and PSMs will be used inside Semantic Web Services that model reasoning processes and deal with that domain knowledge. For this reason, it is necessary to develop methods and techniques that allow reducing the effort necessary for the knowledge acquisition process, being this the goal of the ontology learning.

Acquiring domain knowledge for building ontologies requires much time and many resources. In this sense, we can define ontology learning as the set of methods and techniques used for building an ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources. Other terms are also used to refer to the semi-automatic construction of ontologies like ontology generation, ontology mining, ontology extraction, etc. Several approaches exist for the partial automatization of the knowledge acquisition process. To carry out this automatization, natural language analysis and machine learning techniques can be used.

Alexander Maedche and Steffen Staab [Maedche and Staab, 2001] distinguish different ontology learning approaches focus on the type of input used for learning. In this sense, they propose the following classification: ontology learning from text, from dictionary, from knowledge base, from semi-structured schemata and from relational schemata.

Ontology Learning puts a number of research activities, which focus on different types of inputs, but share their target of a common domain conceptualisation It is a complex multi-disciplinary field that use the knowledge of natural language processing, data and web mining, machine learning and knowledge representation.

References

- Maedche A, Staab S. (2001) *Ontology Learning for the Semantic Web.* IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2)
- Berners-Lee T. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. HarperCollins Publishers, New York

2 Approaches for Ontology Learning

Alexander Maedche and Steffen Staab distinguish different ontology learning approaches focus on the type of input: ontology learning from text, from dictionary, from knowledge base, from semi-structured schemata and from relational schemata

Ontology learning methods from texts consist of extracting ontologies by applying natural language analysis techniques to texts. The most well-known approaches from this group are:

- **Pattern-based extraction** [Morin, 1999] [Hearst, 1992]. A relation is recognized when a sequence of words in the text matches a pattern. For instance, a pattern can establish that if a sequence of *n* names is detected, then the *n-1* first names are hyponyms of the *n*th
- **Association rules**. They were initially defined on the database field as follows: "*Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X implies Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y*" [Agrawall et al., 1993]. Association rules are used on the data mining process to discover information stored on databases if we already have a rough idea of what we are looking for [Adriaans and Zantinge, 1996]. The association rules method for ontology learning has been originally described and evaluated in [Maedche and Staab, 2000]. The association rules have been used [Maedche and Staab, 2001] to discover non–taxonomic relations between concepts, using a concept hierarchy as background knowledge.
- **Conceptual clustering** [Faure et al., 2000]. Concepts are grouped according to the semantic distance between each other to make up hierarchies. The formulae to calculate the semantic distance between two concepts may depend on different factors and must be provided in these methods.
- **Ontology pruning** [Kietz et al., 2000]. The objective of ontology pruning is to build a domain-ontology based on different heterogeneous sources. It has the following steps. First, a generic core ontology is used as a top level structure for the domain-specific ontology. Second, a dictionary which contains important domain terms described in natural language is used to acquire domain concepts. These concepts are classified into the generic core ontology. Third, domain-specific and general corpora of texts are used to remove concepts that were not domain specific. Concept removal follows the heuristic that domain-specific concepts should be more frequent in a domain-specific corpus than in generic texts.
- **Concept learning** [Hahn et al., 2000]. A given taxonomy is incrementally updated as new concepts are acquired from real-world texts.

Ontology learning from dictionary bases its performance on the use of a machine readable dictionary to extract relevant concepts and relations among them.

Ontology learning from a knowledge base aims to learn an ontology using as source existing knowledge bases.

Ontology learning from semi-structured data looks for eliciting an ontology from sources which have any predefined structure, such as XML schemas.

Ontology learning from relation schemas aims to learn an ontology extracting relevant concepts and relations from knowledge in databases.

References.

- Morin E (1999) *Automatic acquisition of semantic relations between terms from technical corpora.* Proc. Of the Fifth Int. Congress on Terminology and Knowledge Engineering (TKE-99), TermNet-Verlag, Vienna
- Hearst M.A. (1992) *Automatic acquisition of Hyponyms from large text corpora.* In Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France, July 1992.
- Agrawal R, Imielinski T, Swami A (1993) *Mining association rules between sets of items in large databases.* In Proc. Of the ACM SIGMOD Conference on Management of Data, 207-216

- Adriaans P, Zantinge D. (1996) *Data Mining.* Addisson-Wesley, 1996
- Maedche A, Staab S. (2001) *Ontology Learning for the Semantic Web.* IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2)
- Maedche, A. and Staab, S. (2000) *Discovering Conceptual Relations from Text*. In: W.Horn (ed.): ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, August 21-25, 2000. IOS Press, Amsterdam, 2000. http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf
- Faure D, Poibeau T (2000) *First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX.* In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany
- Kietz JU, Maedche A, Volz R (2000) *A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet*. In: Aussenac-Gilles N, Biébow B, Szulman S (eds) EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France. CEUR Workshop Proceedings 51:4.1–4.14. Amsterdam, The Netherlands (*http://CEUR-WS.org/Vol-51/*)
- Hahn U, Schulz S (2000) *Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine.* In Canadian Conference on AI 2000: 176-186

3 Ontology Learning from texts

3.1 Introduction

This section shows different methods and tools for ontology learning from text. All the methods presented here use selected texts for learning the structure and contents of an ontology. However, they use different approaches in order to manage the texts and extract the ontology.

For each method, we will present a general description including its goals and scope of the learning process, general steps used to learn, the knowledge source used for learning (if the method needs other type of sources in addiction to text), the main techniques applied in the process, the possibility of reusing other existing ontologies, the main goals looked for the method, the domain in which it has been applied and tested, if there are a tool associated, how the evaluation of the knowledge learnt is performed, a list of the most relevant ontologies built following it, the URL where more information about it can be found, and relevant bibliography. The methods and approaches presented in this section are: Aguirre and colleagues' method, Alfonseca and Manandhar's method, Aussenac-Gilles and colleagues' approach, Bachimont's method, Faatz and Steinmetz approach, Gupta and colleagues' approach, Hahn and colleagues' method, Hearst's approach, Hwang's method, Khan and Luo's method, Kietz and colleagues' method, Lonsdale and colleagues' method, Missikoff and colleagues' method, Moldovan and Girju's method, Nobécourt approach, Roux and colleagues' approach, Wagner approach, and Xu and colleagues' approach.

For each tool, we will present a general description including its main goals, the main techniques used by the tool in the learning process, the method followed, the user intervention in the process, the types of sources used by the method, the software architecture, the possibility of interoperate with other tools, the import and export facilities that the tool provides, the interface facilities, a URL where you can find more information, and relevant bibliography. The tools presented in this section are: ASIUM, CORPORUM-Ontobuilder, DOE, KEA, LTG Text Processing Workbench, Mo'K Workbench, Ontolearn, Promethée, SOAT, SubWordNet Engineering Process Tool, SVETLAN', TDIDF[1]-based Term Classification System, TERMINAE, Text-To-Onto, TextStorm and Clouds, Welkin, and WOLFIE.

3.2 Methods for ontology learning from texts

In this section, we will summarize, in alphabetical order, the most relevant methods and approaches used for ontology learning from text. The name of each method is the main reference in which the method or the approach has been described.

3.2.1 Aguirre and colleagues' method

This method by Aguirre *et al.* [Aguirre *et al.*, 2000] aims to enrich the concepts in existing large ontologies using text retrieved from the word wide web. The overall goal of this approach is to overcome two shortcomings of large ontologies like WordNet: the lack of topical links among concepts, and the proliferation of different senses for each concept. To achieve these aims, the method first retrieves documents related to a concept. For each sense of a concept in the ontology, and in order to construct lists of closely related words for each one, the words in the text that are most closely related to the concept are collected. The approach is based on the use of topic signatures, used in text summarization, that have been described in [Hovy and Lin, 1999] and [Lin and Hovy, 2000]. The strategy proposed to build such

[1] *TF/IDF* Term Frequency-Inverse Document Frequency.
Salton G, Buckley C. (1998) *Term-weighting approaches in automatic text retrieval*. Information Processing and Management: an International Journal, v.24 n.5, p.513-523, 1988

lists is as follows: Firstly, the information contained in the ontology is used to build the queries that retrieve relevant documents relative to the given concept sense. Then the texts retrieved are organized in collections, one per word sense. Finally, for each collection, the words and their frequencies are extracted and compared with the data in those other collections that cover other senses of the same concept. An overview of the whole process can be seen in the figure 1.

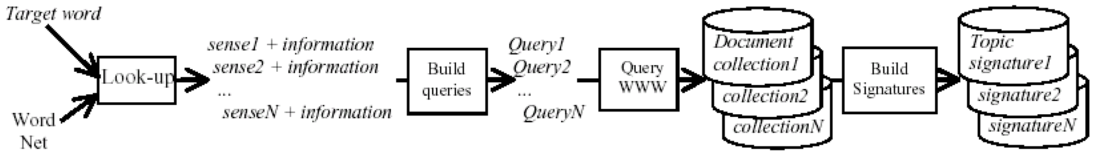


Figure 1. Enriching the concepts in existing ontologies using the Word Wide Web

The method proposes four steps to enrich an existing ontology, and these are now explained in more detail:

1. *Retrieve relevant documents for each concept.* The goal of this step is to retrieve documents related to an ontology concept from the web. Queries are constructed for each concept sense using the information containing in the ontology, such as synonyms of the concept, hyperonyms, attributes, etc. The documents that could belong to more than one sense are discarded, and documents related to the same concept sense are grouped together to form collections, one for each sense.
2. *Build topic signatures.* The documents in each collection, related to a specific concept sense, have to be processed in order to extract the words and their frequencies using a statistical approach. Then, the data from one collection is compared with the data in the other collections. The words that have a distinctive frequency for one of the collections are grouped in a list, which then constitutes the topic signature for each concept sense.
3. *Clustering word senses.* Given a word, the concepts that lexicalise its word sense are hierarchically clustered. To carry out this task different topic signatures are compared to discover shared words, in order to determine overlaps between the signatures. Various semantic distance metrics and clustering methods can be used for this purpose.
4. *Evaluation.* This is performed in the same way as a word sense disambiguation task. The topic signatures and hierarchical clusters are used to tag a given occurrence of a word in another corpus with the intended concept using different disambiguation algorithms.

The approach has been tested with WordNet and the benchmark corpus SemCor [Miller *et al.*, 1993] to perform the evaluation task.

Main techniques used: statistical approach, topic signatures, clustering methods
Reuse of other ontologies: WordNet
Source: text
Main goal: enrich concepts, control over sense proliferation.
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: information not available in papers
Relevant ontologies built following it: Ontology enrichment only
URL: information not available in papers
References

- Agirre, E., Ansa, O., Hovy, E., and Martinez, D. (2000). *Enriching very large ontologies using the WWW*. In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00).

- Lin, C.-Y. and Hovy E.H. (2000). *The Automated Acquisition of Topic Signatures for Text Summarization*. Proc. of the COLING Conference. Strasbourg, France. August 2000.
- Hovy, E.H. and Lin C.-Y. (1999). *Automated Text Summarization in SUMMARIST*. In M. Maybury and I. Mani (Eds), Advances in Automatic Text Summarization. Cambridge: MIT Press.

3.2.2 Alfonseca and Manandhar's method

This method consists of automatically acquiring contextual properties of the words that co-occur with each one of a set of concepts. It can then be used to either cluster concepts inside an ontology, or to refine the ontology by adding new concepts.

The principle is based on the hypothesis of Distributional Semantics: "The meaning of a word is highly correlated to the contexts in which it appears". This hypothesis can be generalised to cover complex phrases (such as whole Noun Phrases) instead of words. The contexts can be encoded as vectors of context words, as in the case of the topic signatures [Lin and Hovy, 2000] Using the topic signatures, each concept would be represented by the set of words that co-occur with it, and the frequencies with which they appear. Several similarity metrics, such as TFIDF or chi-square, can then be used to measure the distance between the different concepts.

Alfonseca and Manandhar (Alfonseca and Manandhar, 2002a) describe a top-down classification algorithm for extending existing ontologies such as WordNet with new concepts.

The quality of the topic signatures can be improved by including only those context words that have some syntactic relation with the concepts in the ontology. For instance, it is possible to only consider the list of verbs for which a concept appears as subject, or as direct object; or to consider only the adjectives that modify the concept. A possible method to combine different kinds of signatures in a single system is described by Alfonseca and Manandhar (Alfonseca and Manandhar, 2002b).

This method was developed as part of the Ensenada CICYT project (2002-2005), funded by the Spanish Ministry, a project which includes knowledge acquisition from free texts for automatic generation of e-learning materials.

As it is based on contextual information, this method requires that we have available several occurrences of the concepts to be classified, so that there is enough contextual information to generate the topic signatures. The method has been used to automatically classify high-frequency concepts from historical texts for generating e-learning web sites (Darwin's *The Voyages of the Beagle*, Osler's *The Evolution of Modern Medicine* and Hegel's *Lectures on the History of Philosophy*), and test data has also been constructed from novels (Tolkien's *The Lord of the Rings* and Homer's *The Iliad*).

Main techniques used: topic signatures and different metrics of semantic distance.
Reuse of other ontologies: WordNet
Source: an existing ontology and free (unannotated) text.
Main goal: to extend an existing ontology with new concepts in an automatic and unsupervised way.
Domain in which it has been applied: the sub-ontology of physical entities inside WordNet.
Tool associated: Welkin
Evaluation of the knowledge learnt: a testing benchmark framework has been constructed, and five different metrics are evaluated, from single accuracy (the percentage of the new concepts that are correctly placed in the existing ontology) to metrics that take into account the distance inside the ontology between the location where the new concepts should have been placed and the location where they have been placed by the algorithm.
Relevant ontologies built following it: extensions of WordNet with new entities found in texts.
URL: http://www.ii.uam.es/~ealfon
References

- Alfonseca E. and Manandhar S. (2002), *An unsupervised method for general named entity recognition and automated concept discovery*. In Proceedings of the 1st International Conference on General WordNet, Mysore, India.

- Alfonseca E. and Manandhar S. (2002). *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*, EKAW-2002, Siguenza, Spain. Published in Lecture Notes in Artificial Intelligence 2473 (Springer Verlag).
- Lin, C.-Y. and Hovy E.H. (2000). *The Automated Acquisition of Topic Signatures for Text Summarization*. Proc. of the COLING Conference. Strasbourg, France. August, 2000.

3.2.3 Aussenac-Gilles and colleagues' approach

This is a method for ontology learning based on knowledge elicitation from technical documents [Aussenac-Gilles et al., 2000a] and [Aussenac-Gilles et al., 2000b]. The method allows creating a domain model by means of the analysis of a corpus using natural language processing (NLP) tools and linguistics techniques. The central role in this method is given to the text. The method combines knowledge acquisition tools based on linguistic with modelling techniques that allows keeping links between models and texts. The method uses texts and may use other existing ontologies or terminological resources to build the ontology. The ontologist, helped by the tools, selects and combines the results to build up the ontology. This method, like Bruno Bachimon's one presented in section 3.2.4, has benefited from the cross-disciplinary works carried out in the French GDR-I3 and CNRS Special Interest Group on "Terminology and AI" (TIA[2]) since 1995.

The method proposes to perform ontology learning in three levels: linguistic, normalization and formal level. The *linguistic level* is composed by terms and lexical relations extracted from texts by means of a linguistic analysis. These elements are used to create lexical clusters and convert them into concepts and conceptual relations at the *normalization level*. The process goes from terminological analysis to conceptual analysis, this means from terms to concepts and from lexical relations to semantic ones. Finally, concepts and relations are formalized by means of a *formal language*.

The activities that are proposed in this method are described next (activities 1 and 2 are performed in the linguistic level, activity 3 is performed in the normalization level and activity 4 is performed in the formal level):

1. *Corpus constitution*. Texts are selected among the available technical documentation from the ontology requirements. The authors recommend that the selection of texts be made by an expert in texts of the domain. Also according to the authors, the corpus has to cover the entire domain specified by the application. To perform this activity is very useful to have a glossary of terms of the domain. Thus, the expert selects texts containing the terms of the glossary.

2. *Linguistic study.* This activity consists in selecting adequate linguistic tools and techniques and applying them to the texts. The main difficulty is to select the tools to be used, which strongly depend on the language to be processed. As a result of this activity, domain terms, lexical relations, and groups of synonyms will be obtained.

3. *Normalization.* The result of this activity is a conceptual model expressed by means of a semantic network. This conceptual model is rather informal, however, it can be easily understood by the ontology designer. Normalization includes a linguistic step and a conceptual modelling step.

   During the *linguistic step*, the designer has to choose the terms and the lexical relations (hyperonym, hyponym, etc.) to be modelled. According to the authors, this choice is mainly subjective, the terms and relations are kept when they seem important both for the domain and for the application where the ontology will be used. Also in this linguistic step, the designer adds a natural language definition for these terms taking into account the senses they have in the source texts. If there are terms with several meanings, the most relevant of the domain are kept.

[2] http://www.biomath.jussieu.fr/TIA

During the *conceptual step*, concepts and semantic relations are defined in a normalized form using the labels of the concepts and relations

The result has to be checked according to differentiation rules. The ontologist may require look for additional knowledge in the documents or from the expert. The differentiation rules require that for any given concept, the following information should be made explicit in the model:

- The concept must have at least one common attribute or relation with its father concept (generally an inherited attribute or relation).
- The concept must have at least one specific attribute or relation that make it different from its father concept.
- The concept must have at least one property that makes it different from its brothers (this may be a specific attribute, relation or value or an inherited attribute or relation).

4. *Formalization*. It includes ontology validation and implementation.

The tools that give support for different steps of this method are: LEXTER [Bourigault, 1996] (as NLP tool for terminology extraction), GEDITERM [Aussenac-Gilles, 1999] (to define, model and consult a terminology connected to a semantic network), Caméléon [Aussenac-Gilles and Seguela, 2000] (to extract relations), and TERMINAE [Biébow et al., 1999] (to consult a corpus and as modelling tool).

Main techniques used: term extraction based on distributional analysis, relation extraction based on linguistic patterns, and knowledge extraction with syntactic patterns with a concordancer
Reuse other ontologies: allowed: OWL and RDFs ontologies can be imported as a first kernel to build up a new one.
Source: texts (must be selected according to very precise criteria connected to the target application and users' needs), existing ontologies or terminologies, and human expert knowledge for validation.
Main goal: learn concepts and relations among them
Domain in which it has been applied: knowledge engineering
Tool associated: GEDITERM and TERMINAE
Evaluation of the knowledge learnt: by the user and a domain expert (that are well aware of the end user's needs). Human expertise is used as less as possible during well-prepared validation sessions. This validation is very useful to better meet the users' requirements.
Relevant ontologies built following it: three ontologies have been built following this approach: one for the Th(IC)2 project, about tools in knowledge engineering (about 80 concepts), other for a private company about fiber glass manufacturing [Aussenac-Gilles et al., 2003] [Aussenac-Gilles and Busnel, 2002] (about 100 concepts), and a third one a case study in the tourism domain for the EON experiment.
URL: http://www-lipn.univ-paris13.fr/~szulman/TERMINAE.html

References

- Aussenac-Gilles, N, Biébow B, Szulman S. (2000a) *Corpus Analysis For Conceptual Modelling.* Workshop on Ontologies and Text, Knowledge Engineering and Knowledge Management: Methods, Models and Tools, 12th International Conference EKAW’2000, Juan-les-pins, France, Springer-Verlag.
- Aussenac-Gilles N, Biébow B, Szulman S (2000b) *Revisiting Ontology Design: A Methodology Based on Corpus Analysis*. In: Dieng R, Corby O (eds) 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW’00). Juan-Les-Pins, France. Springer-Verlag, Lecture Notes in Artificial Intelligence (LNAI) 1937, Berlin, Germany, pp 172–188.
- Aussenac-Gilles N. and Seguela P. (2000) *Les relations sémantiques: du linguistique au formel. Cahiers de grammaire.* N° spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : Presse de l'UTM. Pp 175-198.
- Aussenac-Gilles N. (1999). Gediterm, un logiciel de gestion de bases de connaissances terminologiques. Terminologies Nouvelles, 19 :111-123

- Biébow B, Szulman S. (1999) TERMINAE: a linguistic-based tool for the building of a domain ontology. In EKAW'99 – Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, pages 49-66, Berlin, 1999. Springer-Verlag.
- Bourigault D, Gonzalez I, Gros C. (1996). LEXTER, a Natural Language Tool for Terminology Extraction. In Proceedings of the seventh EURALEX International Congress, Goteborg, Sweden.
- Aussenac-Gilles N., Biebow B., Szulman S. (2003) *D'une méthode à un guide pratique de modélisation de connaissances à partir de textes*. 5 e rencontres Terminologie et IA, TIA 2003. Ed. F. Rousselot. Strasbourg (F), ENSSAIS, Avril 2003. pp 41-53.
- Aussenac-Gilles N. and Busnel A. (2002) Méthode de construction à partir de textes d'une ontologie du domaine de l'industrie de la fibre de verre. Rapport final, contrat de recherche entre IRIT et Saint-Gobain Recherche. Rapport Interne IRIT/2002-28-R. Sept. 2002.

3.2.4 Bachimont's method

This is a method proposed by Bruno Bachimont [Bachimont *et al.*, 2002] for building ontologies, taking into account linguistic techniques that have come from Differential Semantics. The overall process is summarized in the figure 2.

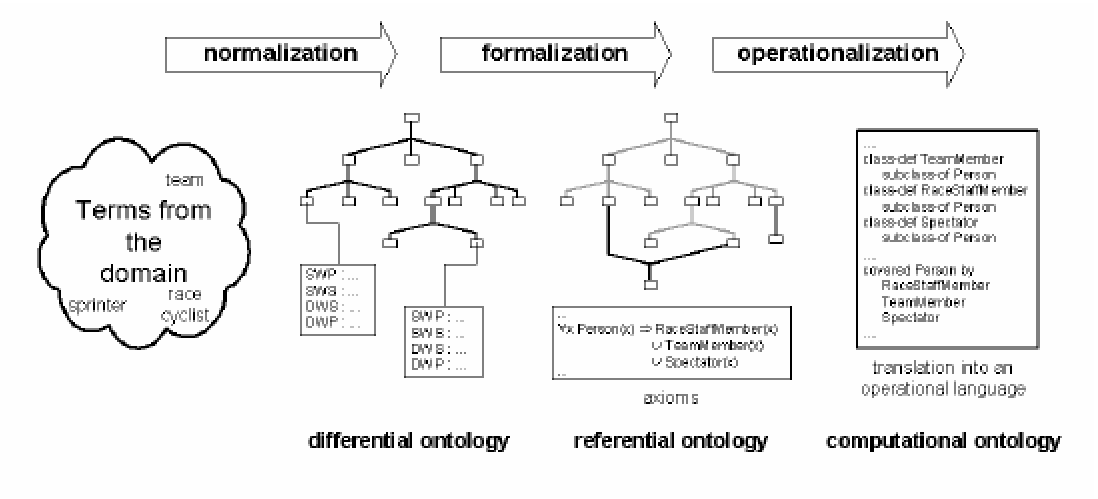


Figure 2. Bachimont's method for building ontologies.

In this method the construction of ontologies follows 3 steps.

1. *Semantic Normalization.* The user has to choose the relevant terms of a domain and normalize their meaning, expressing the similarities and differences of each notion with respect to its neighbours. To achieve this goal, the user has to justify the place of the notion in the hierarchy by determining which concept it is similar to and how it is different from its parent and siblings.
2. *Knowledge Formalization.* Using the taxonomy obtained in the first step, this step needs to disambiguate the notions and to clarify their meanings for a domain-specific expert to carry out the formalisation of the knowledge. Hence, the user can constrain the domains of a relation, define new concepts, add properties to these concepts or add general axioms.
3. *Operationalization.* The third step transcribes the ontology into a specific knowledge representation language.

This methodology is partially supported by the DOE tool.

Main techniques used: NLP techniques
Reuse of other ontologies: not proposed
Source: text, terms proposed by the expert
Main goal: build a taxonomy
Domain in which it has been applied: audio-visual documents.
Tool associated: DOE (Differential Ontology Editor)
Evaluation of the knowledge learnt: expert
Relevant ontologies built following it: Cycling Ontology
URL: http://opales.ina.fr/public/

References:

- Bachimont B., Isaac A., and Troncy R. (2002). *Semantic commitment for designing ontologies: a proposal.* In A. Gomez-Perez and V.R. Benjamins (Eds.): EKAW 2002, LNAI 2473, pp. 114–121, 2002. Springer-Verlag Berlin Heidelberg 2002
- Bachimont B. (2000). *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances*. In Ingénierie des Conniassances : Evolutions récentes et nouveaux défis, Eyrolles, 2000
- Bachimont B. (1996) *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser*. PhD Thesis, Ecole Polytechnique, 1996.

3.2.5 Faatz and Steinmetz approach

Faatz and Steinmetz [Faatz and Steinmetz, 2002] present an approach that aims to enrich an existing ontology by extracting meaning from the world wide web. The enrichment process is based on the comparison between statistical information of word usage in a corpus, and the structure of the ontology itself. Each concept in the ontology should have one or more phrases or words in natural language associated with it. Using this information, the approach proposes a method to calculate the semantic similarity between words in order to enrich the concept definition, and to create clusters of words related to a new concept. The new concepts will be proposed to an domain expert who will decide whether to add them to the ontology.

The approach proposes the following general steps in order to create new concepts from textual documents:

1. *Corpus constitution*. The sources used for learning are a special corpus of text derived from world wide web search results.
2. *Detect a set of candidate concepts from the corpus*. The core idea of this step is to compute enrichment rules which do not contradict the semantic distance information already given by the ontology to be enriched. The corpus is statistically analysed, and a list is generated of co-occurrences for each word in the corpus. New words, related to the descriptors of each concept, are extracted based on a semantic distance function. These words, or their possible clusters, will be candidates to be new concepts.
3. *Select a subset of candidate concepts*. The list created in the previous step is proposed to a domain expert who will decide if they are important or not for the domain.

The approach has been demonstrated in the medical domain using two types of corpus: one of these was created directly from world wide web, and the other consisted of previously selected documents.

Main techniques used: statistical approach, semantic relativeness
Reuse of other ontologies: medical ontologies
Source: corpus from WWW
Main goal: enrich the ontology with new concepts

Domain in which it has been applied: medical
Tool associated: reuse of another ontology workbench with the authors' algorithms added.
Evaluation of the knowledge learnt: domain expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

- Faatz A. and Steinmetz R. (2002). *Ontology enrichment with texts from the WWW.* Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, 20th August 2002, Helsinki, Finland

3.2.6 Gupta and colleagues' approach

Gupta and colleagues [Gupta et al., 2002] present an approach to acquire and maintain sublanguage WordNets[3] from domain specific textual documents. The approach aims to enable rapid development of SubWordnets for NLP applications.

The approach proposes an iterative three-step lexicon engineering cycle for developing SubWordNets as follows:

1. *Discover Concept Elements*: The goal of this step is to discover concept elements, which include words, generated multi-word phrases, and potential relationships among these elements that occur in input sublanguage documents. For example, "Marine Mountain Warfare Training" and "Maritime Interception Operation Training" would be discovered as multi-word phrases in the Navy Lessons domain. An unnamed relation between them could be discovered and suggested to the user. Subsequently, a user could identify the relation as of meronym/holonym type. This step typically uses a combination of shallow language and text processing along with learning, discovery, and extraction techniques.
2. *Identify Concepts*: The objective of this step to identify new concepts and relations from phrases and relations discovered in the previous step. Concept identification is supported by grouping phrases into concept nodes and establishing concordance with synsets in WordNet. The new concept nodes and relationships can be used to update the SubWordNet.
3. *Maintain Concepts (Update SubWordNet)*: This step allows controlled insertion, deletion, and updating of concepts and relations derived from the previous step in a SubWordNet while maintaining its integrity.

Users can iterate through these steps with as many sublanguage documents as needed to develop SubWordNets and to maintain them on an ongoing basis.

Main techniques used: NL techniques, especially extraction techniques
Reuse other ontologies: WordNet
Source: textual documents
Main goal: build sublanguage WordNets
Domain in which it has been applied: Navy lessons
Tool associated: SubWordNet Engineering tool
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

[3] *Sublanguage WordNets* are specific WordNets build with a domain specific lexicon. Sublanguage WordNets has the same structure than WordNet and covers an specific language relative to a concrete domain.

- Gupta, K.M., Aha, D.W., Marsh, E., and Maney, T. (2002). *An architecture for engineering sublanguage WordNets*. In Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages.

3.2.7 Hahn and colleagues' method

Hahn and colleagues present a method for the maintenance [Hahn et al., 1998] and growth [Hahn and Markó, 2001] of domain-specific taxonomies based on natural language text understanding. A given taxonomy is incrementally updated as new concepts are acquired from real-world texts. The acquisition process is focused around the linguistic and conceptual "quality" of various forms of evidence underlying the generation and refinement of concept hypotheses. On the basis of the quality of evidence, concept hypothesis are ranked according to credibility and the most credible ones are selected for assimilation into the domain ontology. In this approach, learning is achieved by the refinement of multiple hypotheses about the concept membership of an instance. New concepts are acquired taken two sources of evidence into account: background knowledge from the domain texts, and linguistic patterns in which unknown lexical items occur.

The model presented for text knowledge elicitation can be summarized in the following general steps.

1. *Language processing.* It aims to determine structural dependency information from the grammatical constructions in which an unknown lexical item occurs in terms of the corresponding parse tree. The conceptual interpretation of parse trees involving unknown lexical items in the terminological knowledge base is used to derive concept hypotheses, which are further enriched by conceptual annotations reflection structural patterns of consistency, analogy, etc. This kind of initial evidence is represented by corresponding sets of linguistic and conceptual quality labels.
2. *Calculation of the quality labels.* As it has been mentioned above, there are two kinds of quality labels to be calculated. The first one is the *linguistic quality label* that reflects structural properties of phrasal patterns or discourse contexts in which unknown lexical items occur, and depending on the type of the syntactic construction, different hypothesis generation rules may fire. The second type is the *conceptual quality labels*, that results from comparing the representation structures of a concept hypothesis with those of alternative concept hypotheses or already existing representation structures in the underlying domain knowledge base from the viewpoint of structural similarity, compatibility, etc.
3. *Quality estimation.* The overall credibility of single concept hypotheses is estimated by taking the available set of quality labels for each hypothesis into account. The final computation of a preference order for the entire set of competing hypotheses. This output is a ranked list of concept hypotheses. Whenever new evidence for or against a concept hypothesis is brought, all concept hypothesis are re-evaluated.
4. *Evaluation.* An empirical evaluation of the text knowledge acquisition process is performanced using different measures that evaluate the learning accuracy and the learning rate. The learning is achieved by the refinement of multiple hypotheses about the concept membership of an instance.

This method has been tested on a medium-sized knowledge base for the information technology domain.

Main techniques used: concept hypothesis based on linguistic and conceptual quality labels
Reuse other ontologies: not proposed
Source: domain text, domain knowledge base
Main goal: learn new concepts
Domain in which it has been applied: information technology
Tool associated: information not available in papers
Evaluation of the knowledge learnt: empirical measures and by an expert

Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

- Hahn U., and Schnattinger K. (1998). *Towards text knowledge engineering.* In: AAAI '98 / IAAI '98 Proceedings of the 15th National Conference on Artificial Intelligence & 10th Conference on Innovative Applications of Artificial Intelligence. Madison, Wisconsin, July 26-30, 1998. Menlo Park, CA; Cambridge, MA: AAAI Press / MIT Press, pp. 524-531.
- Hahn U., and Markó K. (2001). *Joint knowledge capture for grammars and ontologies.* Proceedings of the First International Conference on Knowledge Capture K-CAP 2001: Victoria, BC, Canada
- Hahn U., and Schulz S. (2000). *Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine.*Canadian Conference on AI 2000: 176-186

3.2.8 Hearst's approach

Hearst (Hearst, 1998) describes a procedure, called hyponymy pattern approach, for automatically learning relationships between concepts in an ontology. It consists of looking for concepts that are related in an existing ontology (e.g. WordNet) and determining whether they are associated with each other in a word pattern that expresses that relationship. For instance, *Shakespeare* is a hyponym of *poet* in WordNet. Therefore, if we find in a text the pattern "*poets such as Shakespeare"* we can determine that the pattern "*such as"* usually indicates a hyponymy relationship.

The hyponymy pattern approach has been applied for Ontology Refinement in the On-To-Knowledge system, described by Kietz *et al.* (Kietz *et al*., 2000). However, the degree of errors produced by this method is very high and it is necessary to have the results validated by an expert. This approach is also the one used in Prométhée tool and Caméléon to identify new lexical-syntactic patterns in a corpus. In both of these tools, Hearst's approach is applied to a specific domain using couples of related terms (that may be noun phrases) known in this particular domain to learn specific patterns by analysing the contexts in which these couples of terms occur.

Alfonseca and Manandhar (Alfonseca and Manandhar, 2002) propose a way in which this method can be combined with the contextual signatures method in order to improve the classification of new concepts inside an existing ontology. This approach was developed as part of the Ensenada CICYT project (2002-2005), funded by the Spanish Ministry, a project which includes knowledge acquisition from free texts for automatic generation of e-learning materials.

The method has been used to extend lexical ontologies with new concepts, and to add new relationships between the existing concepts. Hearst (Hearst, 1998) and Alfonseca and Manandhar (Alfonseca and Manandhar, 2002) worked with WordNet, and Kietz (Kietz, 2000) worked with GermaNet, the German version of WordNet.

Main techniques used: use of word patterns that express lexical relationships in order to learn new relationships between the concepts in an ontology.
Reuse of other ontologies: WordNet-like ontologies (e.g. the English WordNet or the German GermaNet)
Source: the original ontology and a corpus of texts.
Main goal: to extend existing ontologies with new concepts, and with new relationships among the existing concepts.
Domain in which it has been applied: general-purpose terms.
Tool associated: Welkin, Ontology Learning Tool in On-To-Knowledge. It is also used by Prométhée and Caméléon
Evaluation of the knowledge learnt: a human expert evaluates the results of the algorithm.
Relevant ontologies built following it: information not available in papers
URL: http://www.ii.uam.es/~ealfon
References

- Alfonseca E. and Manandhar S. (2002), Improving an Ontology Refinement Method with Hyponymy Patterns. Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.
- Hearst M. A. (1998), Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database, MIT Press, pp. 132--152.
- Kietz J., Maedche A., and Volz R. (2000), A Method for Semi-automatic Ontology Acquisition from a Corporate Intranet, Workshop ``Ontologies and text'', co-located with EKAW'2000.

3.2.9 Hwang's method

Within the InfoSleuth[4] research project at MCC (Microelectronics and Computer Technology Corporation) different technologies have been developed for finding information available in both corporate networks and external networks. It focuses on the problems of locating, evaluating, retrieving, and merging information in an environment in which new information sources are constantly being added.

As part of the InfoSleuth project, an approach has been developed [Hwang, 1999] to represent and retrieve information from large textual databases. It is based on the use of dynamic ontologies that capture the semantics of information present inside the documents. The ontology is organized in simple taxonomies. Concepts from the taxonomy are then identified within the documents to enable the retrieval process. To carry out the process, NLP and machine learning techniques have been used.

The procedure for generating the ontology has the following steps:

1. *Human experts provide* the system with a small number of *seed-words* that represent high-level concepts. Relevant documents will be collected from the web automatically (with POS-tagged or otherwise unmarked text).

2. The system *processes the incoming documents*, extracts only those phrases that contain seed-words, generates corresponding concept terms and places them in the "right" place in the ontology, and alerts the human experts of the changes. This feature is named "discover-and-alert". At the same time, it also collects candidates for seed-words for the next round of processing. The iteration continues a predefined number of times. The method indexes documents according to the concepts identified within them for future retrieval and also the "context lines" in which the concept has been discovered to show how the concept was used in the text as well as frequency of co-occurrence inside each document.

3. Several kinds of *relations* are extracted. Examples of relations are: "is-a", "part-of", "manufactured-by", "owned-by", etc, which are extracted based on linguistic features. The "assoc-with" relation is used to define all relations that are not an "is-a" relation. The distinction between "is-a" and "assoc-with" relations is based on a linguistic property of noun compounds. The method only can discover some of the attributes associated with certain concepts based on linguistic characters.

4. In each iteration, a *human expert is consulted* to ascertain the correctness of the concepts. If necessary, the expert has the right to make the correction and reconstruct the ontology. While constructing the ontology, the method also allows the indexing of documents for future retrieval

There are some *problems* for automatically generating ontologies with this approach such as: *syntactic structural ambiguity, recognising different phrases that refer to the same concept, word sense problems, etc*.

Main techniques used: NLP, machine learning techniques, and statistical approach
Reuse other ontologies: not proposed
Source: text
Main goal: elicit a taxonomy

[4] http://www.argreenhouse.com/InfoSleuth/index.shtml

Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers.
Evaluation of the knowledge learnt: it is carried out by a domain expert.
Relevant ontologies built following it: information not available in papers
URL: http://www.argreenhouse.com/InfoSleuth/index.shtml.

References

- Hwang, C. H. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In. Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linköping, Sweden, July 29-30, 1999.

3.2.10 Khan and Luo's method

This method [Khan and Luo, 2002] aims to build a domain ontology from text documents using clustering techniques and WordNet. The method constructs the ontology in a bottom-up fashion. Firstly construct a hierarchy using some clustering techniques. Similar documents in content are associated with the same concept in the ontology. Next, a concept for each cluster of documents relative to the same topic in the hierarchy is assigned using a bottom up concept assignment mechanism. To achieve this goal, a topic tracking algorithm [Joachims, 1998] and WordNet are used.

The method proposes the following steps:

1. *Selection of the corpus to be used*. The user provides a selection of documents regarding to same domain.
2. *Hierarchy construction*. Using the set of documents provided in the previous sets, the method aims to create a set of clusters where each cluster may contain more than one document, and put then into the correct place in a hierarchy. Each node in this hierarchy is a cluster of documents. For this purpose, the method proposes to use a modify algorithm, called SOTA algorithm, specifically designed for molecular bio-sequence classification with the main purpose of classification clusters of documents into a hierarchy.
3. *Concept assignment*. After building a hierarchy of clusters, a concept is assigned for each cluster in the hierarchy using a bottom-up fashion. Firstly, concepts associated with documents will be assigned to leaf nodes in the hierarchy. For each cluster of documents, it will be assigned a keyword, called topic that represents its content using a predefined topic categories. Then, this topic will be associated with an appropriate concept in WordNet. And finally, the interior node concepts will be assigned based on the concepts in the descendent nodes and their hyperyms in WordNet. The type of relation between concepts in the hierarchy is ignored; it is only possible to know that there is a relation between them.

Main techniques used: clustering techniques
Reuse other ontologies: WordNet ontology
Source: text documents
Main goal: Learn concepts
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers

References

- Khan L., and Luo F. (2002) *Ontology Construction for Information Selection* In Proc. of 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 122-127, Washington DC, November 2002.

- Joachims T. (1998) *A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization.* Logic J. of the IGPL, 1998.

3.2.11 Kietz and colleagues' method

This method [Kietz et al., 2000] is a generic method used to discover a domain ontology from given heterogeneous resources by the use of natural language analysis techniques. It is a semi-automatic process in the sense of the user takes part in the process. In their approach, they have adopted the balanced cooperative modelling [Morik, 1993], where the work of building the ontology is distributed between several learning algorithms and the user. The method is based on the assumption that most concepts and conceptual structures of the domain to be included in an ontology as well as the terminology of a given domain are described in documents. The authors propose to learn the ontology using as a base a core ontology (it could be: SENSUS, WordNet, etc.) that is enriched with new specific domain concepts. New concepts are identified using NL analysis techniques over the resources previously identified by the user. The resulting ontology is pruned and focused to a specific domain by the use of several approaches based on statistics. Finally, relations between concepts are learnt applying learning methods. Such relations are added to the resulting ontology. A summary of whole process can be seen in the figure 3.

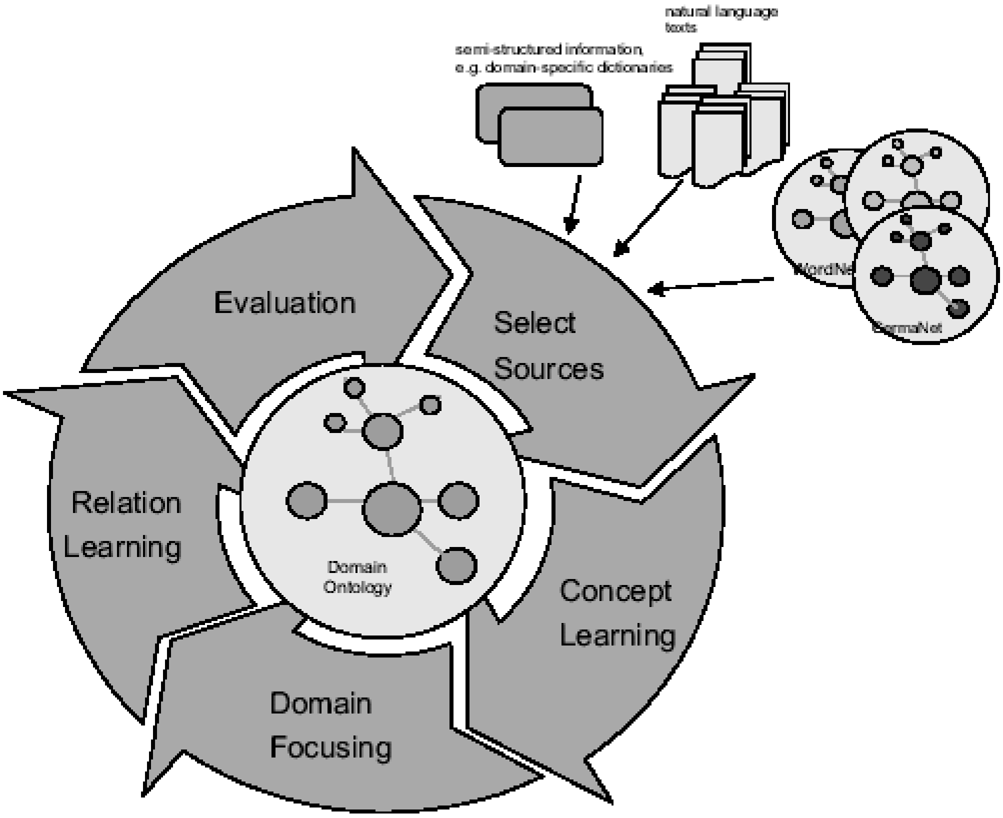


Figure 3. An overview of Kietz and colleagues' method

This method consists of the following steps: select sources, concept learning, domain focusing, relation learning, and evaluation of the resulting ontology. The process is cyclic in the sense that the resulting ontology can be refined applying the method iteratively.

The acquisition process proposed by this method is constituted by the following steps.

1. *Select sources.* The process starts with the selection of a generic (top-level) ontology, which is used as a base in the learning process. This ontology should contain generic and domain concepts. The user must specify which documents should be used in the following steps to refine and extend the previous ontology. By its own nature, sources are heterogeneous in their formats and contents. Sources can be free text documents, semi-structured text, domain text, and generic text. Documents can be general o domain specific.
2. *Concept learning.* Its goal is to acquire new generic and specific concepts to decide if the discovered concepts are specific enough to be included in the ontology. The method proposes to analyse the frequency of the terms. Those terms that are more frequent in a domain-specific corpus than in a generic corpora (and they are not contained in the given ontology) should be proposed to the user to decide whether they should be incorporated to the ontology. The selection of the tools depends on the language to be processed (Spanish, English, German, etc.).
3. *Domain focusing.* Its purpose is to prune the enriched core ontology by removing general concepts.
4. *Relation learning.* Frequency analysis can be used to learn ad hoc relations of the domain. This is founded in the underlying idea that frequent couplings of concepts in sentences can be consider as relevant relations between concepts in the ontology. This approach is used to find frequent correlations between concepts and it is based on the association rule's algorithm proposed in [Skrikant and Agrawal, 1995].
5. *Evaluation.* Its goal is to evaluate the resulting ontology and to decide whether it is necessary to repeat the process again.

This method is supported by the tool Text–To–Onto [Maedche and Volz, 2001].

Main techniques used: statistical approach, NLP techniques
Reuse other ontologies: GermaNet, WordNet
Source: text and existing ontologies
Main goal: prune an existing ontology and to enrich it with new domain concepts and relations among them
Domain in which it has been applied: On-To-Knowledge project
Tool associated: Text-To-Onto
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: http://ontoserver.aifb.uni-karlsruhe.de/texttoonto/

References

- Kietz JU, Maedche A, Volz R (2000) A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In: Aussenac-Gilles N, Biébow B, Szulman S (eds) EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France. CEUR Workshop Proceedings 51:4.1–4.14. Amsterdam, The Netherlands (http://CEUR-WS.org/Vol-51/).
- Maedche, A. and Volz, R. (2001) The Text-To-Onto Ontology Extraction and Maintenance Environment. To appear in Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.
- Morik K. (1993) Balanced Cooperative Modelling. Machine Learning, 11(1), 1993, pages 217-235.

3.2.12 Lonsdale and colleagues' method

This method [Lonsdale et al., 2002] aims to build a new domain ontology reusing an existing big one. It uses as input different lexical resources and domain documents. The method has been developed as a part of the SALT[5] and TIDIE[6] projects. The main goal of this work is to improve the generation of extraction ontologies through use of a terminological database that has been represented in a standardized format. Two stages have been proposed. First, the conversion of a large-scale terminological resource into a format that can be used by a data-extraction ontology system. And the second is how an ontology generation system integrates this terminological information with similar resources, and then uses the composite knowledge base to analyse the content of input documents and generate novel ontological relationships based on the observed concepts and their relationships. Using these variety of information sources to be treated and the range of formats currently in use, the field of lexicon data integration and exchange requires a principled approach to the modelling of data.

The work reported here, integrates information of a novel type: a large-scale terminology database which has some ontological structure and which has been reformatted according to a standard format to be used in the generation process, mainly to XML.

The knowledge sources used in the process has been: the Mikrokosmos ontology, a data frame library, which is a repository of regular expression templates designed to match structured low-level lexical items and which can provide information for a conceptual matching via inheritance; lexicons, in this case has been WordNet; and finally, training documents, which contain domain-specific textual content of interest for a user. The main steps of this method can be seen in the figure 4

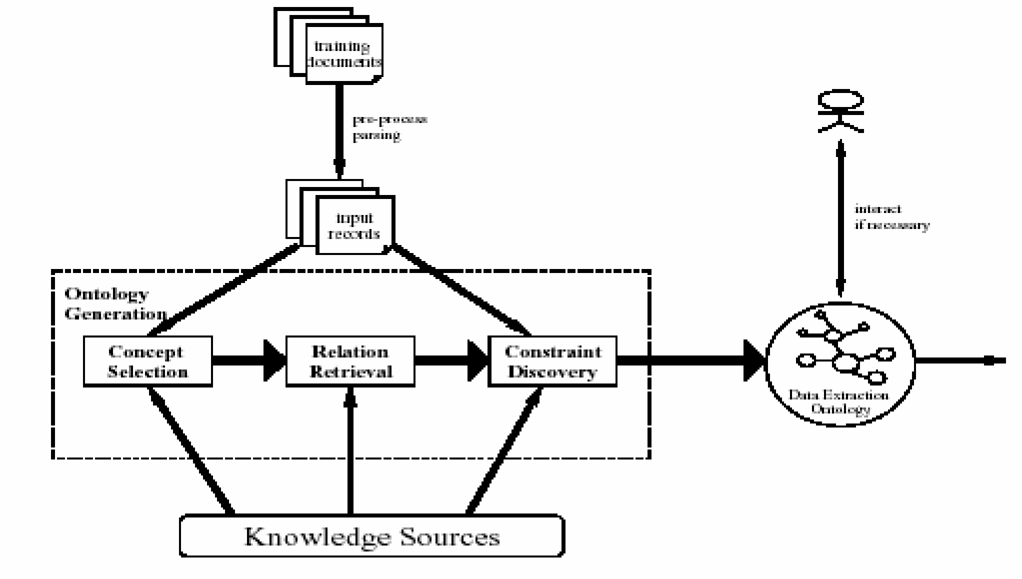


Figure 4 Ontology generation process

The steps proposed are:

1. *Pre-processing of the knowledge sources.* Given the abovementioned knowledge sources, it is necessary to do a previous preprocessing over them. First, an integrated repository of conceptual information is created by mapping lexicon content and data frame templates to nodes in the merged ontology. Secondly, the collection of training documents must be processed to extract its pertinent information. It is assumed that the documents are encoded in HTML. These documents are first parsed to isolate linguistic content to be later tokenised and regularized.

2. *Concept selection.* This step involves finding which subset of the ontology's concept is of interest to a user. Concepts are selected via string matches between textual content and ontological data. Three different selection heuristics have been proposed to select concepts: concept-name matching which selects concepts according to matches from conceptual names in the ontology to parsed sentences in

[5] Standards-based Access service to multilingual Lexicons and Terminologies, http://www.ttt.org/salt/index.html

[6] Target-based Independent-of-Document Information Extraction, http://www.deg.byu.edu

the documents; concept-value matching aims to match a instance in the ontology with a value found in a sentence; data-frame pattern matching. String matches process proposed are calculated straightforwardly, with two assumptions: word synonyms are considered following WordNet synsets, and compose words are consider synonyms of both involved words. In this step, it is performed a concept conflict resolution to arrive at an internally consistent set of selected concepts. There are proposed two levels: document-level resolution, and knowledge level resolution.

3. *Relationships retrieval.* Once concepts have been matched, schemas representing the relationships between these concepts must be generated. The ontology is structured as a directed graph whose nodes are concepts. All the concepts generated constitute a directed subgraph, either connected or unconnected, and the relationships among these concepts can be represented by paths among them. The technique followed to find these relations is based on the graph theory, and the same algorithm than in the MikroKosmos project has been used. With these new relations, the schemas can be built.
4. *Constraint discovery.* This step aims to determine the constraints on the relationships discovered in the previous step.
5. *Refining the results.* The final ontology must be checked and refined by the user to avoid mistakes, and inconsistencies. The amount of work to be done depends of the quality of the sources used in the process.

The method has been proved on various of U.S. Department used together with Eurodicautom terminology bank

Main techniques used: NLP techniques, mappings, several linguistic heuristic, graph theory
Reuse other ontologies: allowed
Source: terminological databases, domain ontologies, WordNet, text documents
Main goal: discover new relations
Domain in which it has been applied: financial
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by the user
Relevant ontologies built following it: information not available in papers
URL: http://www.ttt.org/salt/index.html
References

- Lonsdale D, Ding Y, Embley D.W, and Melby A. (2002) *Peppering Knowledge Sources with SALT; Boosting Conceptual Content for Ontology Generation*. Proceedings of the AAAI Workshop on Semantic Web Meets Language Resources, Edmonton, Alberta, Canada, July 2002.

3.2.13 Missikoff and colleagues' method

OntoLearn [Missikoff *et al*., 2002] is a method for ontology construction and enrichment using NL and machine learning techniques. The method proposes using WordNet as a source of prior knowledge to build a core domain ontology, after pruning all of the unspecific domain concepts. The approaches followed by the method are: statistical, to determine the relevance of one term for the domain; and semantic interpretation, based on machine learning techniques, to identify the right sense of terms and the semantic relations among them. A summary of whole process can be seen in the figure 5

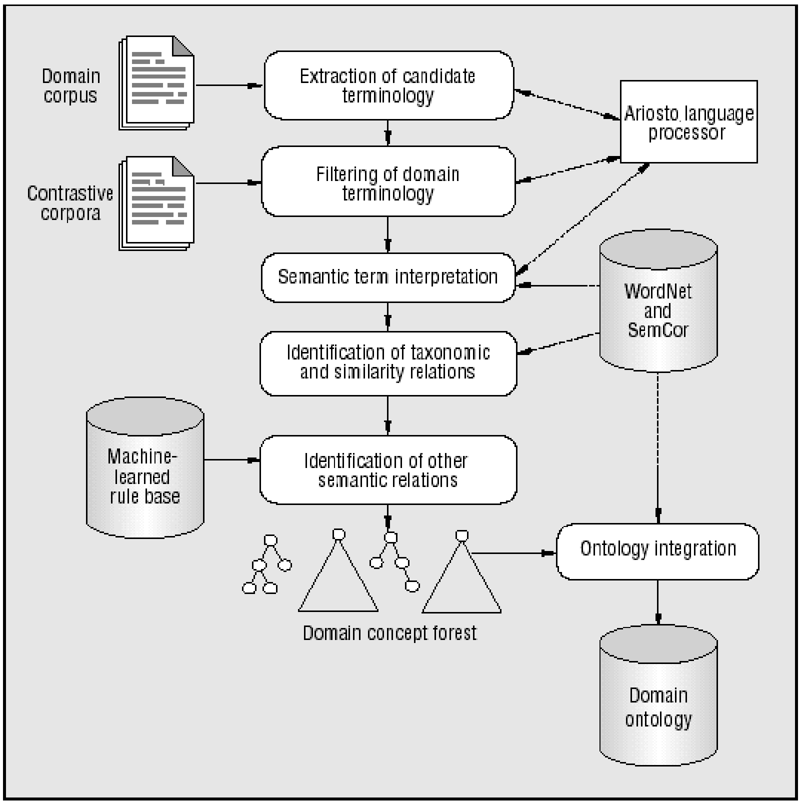

Domain corpus
Contrastive corpora
Extraction of candidate terminology
Filtering of domain terminology
Semantic term interpretation
Identification of taxonomic and similarity relations
Identification of other semantic relations
Ariosto language processor
WordNet and SemCor
Machine-learned rule base
Ontology integration
Domain concept forest
Domain ontology

Figure 5 OntoLearn Method: semantic interpretation.

The method proposes three main steps to achieve its goals [Velardi et al., 2003]: terminology extraction, semantic interpretation, and creation of a specialized view of WordNet

1. *Terminology extraction*. Terms and combinations of terms, such as "*last week*", are extracted from a parsed corpus using NL techniques. Terms are considered as the surface appearance of relevant domain concepts. High frequency in a corpus is a property observable for terminological as well as non-terminological expressions. The method proposes to use a measure of the specificity of a terminology candidate with respect to the target domain via comparative analysis across different corpora. For this purpose, two different elements are defined to determine a threshold for the relevance of one terminology expression for the domain. The first element is the *domain relevance score,* which is a measure of the amount of information captured in the target corpus relative to the entire collection of corpora used for the learning process. The second element is the *domain consensus* which captures those terms that appear frequently across a given domain's documents.

2. *Semantic interpretation*. The main goals of this step are to determine the right concept sense for each component of a complex term, like a semantic disambiguation process, and then to identify the semantic relations holding among the concepts to build a complex concept. At the end of this step, a domain concept forest will be obtained, showing the taxonomic and other relationships among complex domain concepts represented by expressions. To carry out this step, it is necessary to use semantic and linguistic resources (the method has been tested with WordNet) to assist in the semantic interpretation of terms. This step consists of two main processes, the first of which is a *semantic disambiguation process*. The sense of each word is defined as a synset of synonyms (or the right synset in WordNet in which the word can be placed). The second process is extracting semantic relations that hold between the components of complex terms extracted in the previous step.

3. *Creating the domain ontology*. This step aims to integrate the taxonomy obtained in the previous step with a core domain ontology. In the case that an existing domain ontology is not available, the method proposes to create a new one from WordNet, pruning concepts that are not related to the domain, and extending it with the new domain concept trees under the appropriate nodes.

This method has been developed and tested inside the *Harmonise*[7] and *Fetish*[8] European projects, both in the tourism domain.

Main techniques used: NL, statistical approach and machine learning.
Reuse other ontologies: WordNet (after removing all unspecific domain concepts)
Source: text
Main goal: to build trees of domain concepts and to fuse then with an existing core domain ontology
Domain in which it has been applied: tourism
Tool associated: OntoLearn
Evaluation of the knowledge learnt: by an expert.
Relevant ontologies built following it: Tourism ontology, called OntoTour
URL: information not available in papers.
References

- Navigli R., Velardi P, and Gangemi A. (2003). *Ontology Learning and its application to automated terminology translation.* IEEE Intelligent Systems, vol. 18, n.1, January February 2003(2003)
- Missikoff M., Navigli R., and Velardi P. (2002). *The Usable Ontology: An Environment for Building and Assessing a Domain Ontology* Research paper at International Semantic Web Conference (ISWC) 2002, June 9-12th, 2002 Sardinia, Italia

3.2.14 Moldovan and Girju's method

This is a method for discovering domain-specific concepts and relationships in an attempt to extend an existing ontology, like WordNet, with new knowledge acquired from parsed text. The source for discovering new knowledge is a non-specific domain corpus, and is augmented by using other lexical resources like domain specific and general dictionaries. The user provides a number of domain-specific concepts that are used as seed concepts to discover new concepts and relations from the source. The user performs the validation of the process and confirms the correctness of the new concepts and relations learnt.

To enrich an existing ontology with new concepts and relations, the following five steps are proposed by the method [Moldovan and Girju, 2001]:

1. *Select seed concepts.* Some seed-concepts, that a user considers important for the target domain ontology, are selected. This set of seed-concepts is extended with each concept's corresponding synonyms to form a synset. The knowledge that is to be acquired has to be related to one or more of these seed-concepts, and consists of new concepts not defined in the existing ontology as well as new relations. The new relations link the new concepts with other concepts, some of which may already be present in the existing ontology.
2. *Discover new concepts.* To discover new concepts from a general corpus, the method proposes the following phases [Moldoban and Girju, 2000]. Firstly, *documents that contain the seed-concepts are retrieved* and stored before they are processed. Only the nouns are considered as candidate concepts by the method. Secondly, for each document, *sentences that contain the seed concepts are extracted.* Only the noun phrases are considered. Thirdly, *each of the previous sentences are POS-tagged and parsed.* There are two possible types of sentences to be selected. One of these is when the seed is the head noun of the phrase (the phrase would take the form [word, word, word, ..., seed]). The other possibility is when the seed is not the head noun of the phrase (the sentence would take the form [word, word, ..., seed, ..., word, word]). Finally, after parsing all sentences, *new concepts are extracted*. To carry out this process, three main points have been proposed by the authors [Moldoban and Girju, 2001]:

[7] Harmonise EC project IST-2000-29329
[8] Fetish EC project IST-13015

a. Search the identified noun-sentences in the corpus for concepts present in the existing ontology. The purpose is to find words that commonly co-occur, one of which must be a existing concept of the ontology. A dictionary may be used to find more of these compound concepts.

b. For each co-location the process takes the words that modify a noun (that is the adjectives). Three types of adjectives are considered: descriptive (express an attribute of the modified noun), participial (derived from the participle form of verbs), and relational (related semantically or morphologically with the modifying noun). The method proposes to consider as *new concepts* only those that are formed with relational and participial adjectives and discard the descriptives. Based on this, the method proposes to take out all the adjectives from the previous step, with the following exceptions: when the adjective is part of a concept determined from the existing ontology or from a dictionary; or when the adjective is a relational or participial adjective.

c. User validation. The user inspects the list of the remaining noun phrases and decides whether to accept or decline each concept proposed.

3. *Discover lexical-syntactic patterns*. The main aim of this step is to discover semantic relations between concepts (between two new concepts or between a new one and one present in the existing ontology). The method then proposes to create a new corpus, different than the corpus used in the previous step. New noun-sentences are extracted from this corpus. The objective is to search for lexico-syntactical patterns comprising the concepts of interest, extracted in the previous step, inside the new group of sentences.

4. *Discover new relations between concepts.* To carry out this process three elements are used: the new concepts discovered in Step 2, the group of noun-sentences extracted in that step, and the lexical-syntactic patterns resulting from the Step 3. For each new concept the process tries to find all of the syntactic relations established in Step 3 in which the concept is involved. The relation is created between the two concepts linked by the syntactic relation. The validation of the process is performed by the user.

5. *Classification and integration*. In this step a new taxonomy is created for the newly acquired concepts. This new taxonomy will be integrated with the existing ontology using the relations discovered in the previous step between a new concept and other concepts in the existing ontology [Harabagiu and Moldoban, 2000].

To carry out the process and for evaluating the learning process, WordNet has been used. The method can be applied to the learning of an ontology from machine readable dictionaries.

Main techniques used: NLP techniques
Reuse other ontologies: WordNet
Source: non-specific domain corpus, lexical resources, and dictionaries
Main goal: to enrich an existing ontology
Domain in which it has been applied: financial
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers

References

- Harabagiu, S. M.; Moldovan D. I. (2000). Enriching the WordNet Taxonomy with Contextual Knowledge acquired from text. In Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language, (Eds) S. Shapiro and L. Iwanska, AAAI/MIT Press, 2000, pages 301-334.
- Moldovan, D. I.; Girju, R. C. (2001). An interactive tool for the rapid development of knowledge Bases. In International Journal on Artificial Intelligence Tools (IJAIT), vol 10., no. 1-2, March 2001.

- Moldovan, D. I.; Girju, R. C. (2000). Domain-Specific Knowledge Acquisition and Classification using WordNet. In Proceedings of International Florida Artificial Intelligence Research Society (FLAIRS-2000) conference, Orlando, Fl., May 2000.
- Moldovan, D. I.; Girju, R. C.; Rus, V. (2000). Domain-Specific Knowledge Acquisition from Text. In Proceedings of the Applied Natural Language Processing (ANLP-2000) conference, Seattle, WA., April-May 2000.

3.2.15 Nobécourt approach

This work [Nobécourt, 2000] presents an approach to build domain ontologies from texts using NLP techniques and a corpus[9]. The method proposes two activities: modelling and representation.

1. The *modelling* activity includes a linguistic and a conceptual activity.
    - The goal of the linguistic activity is to extract the main domain terms (called "conceptual primitives") from a corpus.
    - Once the domain terms are identified, the conceptual activity starts. Domain experts look for relevant terms of the domain in the previous list to identify the main sub-domains of the ontology. Such terms are modelled as concepts or properties and they constitute the first skeleton of the ontology. Concepts of the ontology are described in natural language and they also constitute a new source document, which is used as a new input for the method. Again, from this new document, a new list of conceptual primitives are created. The ontologist compares this new list with the old one to find new conceptual primitives or new primitives that express relationships between concepts. The proposed process refines iteratively the skeleton.
2. The *representation activity* consists in the translation of the modelling schemata into an implementation language.

This method is technologically supported by TERMINAE [Biébow et al., 1999].

Main techniques used: linguistic analysis
Reuse other ontologies: not proposed
Source: text
Main goal: learn concepts and relations
Domain in which it has been applied: information not available in papers
Tool associated: TERMINAE
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers

References

- Biébow B, Szulman S. (1999) TERMINAE: a linguistic-based tool for the building of a domain ontology. In EKAW'99 – Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, pages 49-66, Berlin, 1999. Springer-Verlag.
- Nobécourt J (2000) A method to build formal ontologies from text. In: EKAW-2000 Workshop on ontologies and text, Juan-Les-Pins, France.

[9] A *corpus of texts* is a set of texts that should be representative of the domain (complete), prepared to be processed by a computer, and accepted by the domain experts. Definition extracted from Enery TMC, Wilson A (2001) *Corpus linguistics: an introduction*. Edinburgh University Press, Edinburgh, United Kingdom

3.2.16 Roux and colleagues' approach

This work by Roux et al. [Roux et al., 2000] aims to enrich an existing ontology with new concepts extracted from a parsed domain corpus using NL techniques. The approach is based on conceptual graphs [Sowa J.F., 1984] and the idea behind it is to use the syntactic dependencies, focused on verb patterns extracted at the linguistic level, to build a semantic representation. According to these verb patterns concepts are added into the ontology. There are two restrictions on the application of the approach. Firstly, the concept can only be an expression or a proper noun. Secondly, the data in the text should be easily identified by their immediate context.

The approach proposes a general step to perform its goals. When a new word appears in the text that has not yet been referenced as a concept in the existing ontology, it is necessary to add this new word as a new concept. As the ontology comprises concepts that are connected to each other along semantic paths, this necessitates classification of this new concept in order to find its correct place in the ontology. The approach is focused on managing new concepts with certain configurations of verbs (verb patterns) that will assign their position in the ontology. The verb patterns, used in this approach, are graphs in which one of the nodes is a verb that expects certain semantic attributes. These graphs will serve to connect the new term that matches in a specific semantic context, under corresponding nodes inside the ontology.

Main techniques used: verb-patterns
Reuse other ontologies: an existing ontology of the domain
Source: text
Main goal: enrich a taxonomy with new concepts
Domain in which it has been applied: Genetic
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

- Roux C., Proux D., Rechermann F., and Julliard L. (2000). *An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions*. Position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000), Berlin, Germany. August 2000.
- Sowa J.F. (1984). *Conceptual Structures.* Information Processing in Mind and Machine, Reading, Mass. : Addison-Wesley, 1984

3.2.17 Wagner approach

Wagner [Wagner, 2000] presents an approach for the automatic acquisition of selectional preferences of verbs by means of statistical corpus analysis for automatic ontology enrichment. He also introduced a modification of the approach by Ade and Li [Abe and Li, 1996] that is based on the well-founded principle of Minimum Description Length.

There are several relations encoded into lexical semantic ontologies. This approach is focused upon enriching the hierarchical relations, in particular the thematic role relations that connect a verbal concept with those nominal concepts that typically occur as their complements. For example, the verbal concept <eat> should have AGENT pointers to the nominal concepts <human> and <animal>, and a PATIENT pointer to <food>.

The approach uses statistical methods for learning those thematic relations and for encoding them into the semantic ontology at the right level of generalization. However, one of the drawbacks of this approach is that the learning algorithms are fed with word forms rather than word senses.

Main steps proposed: not proposed

Main techniques used: statistical approaches
Reuse of other ontologies: EuroWordNet
Source: domain corpus
Main goal: to enrich the ontology with new semantic relations
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: performed by an expert
Relevant ontologies built following it: information not available in papers
URL:information not available in papers
References

- Wagner, A. (2000). *Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis*. In Proceedings of the ECAI-2000 Workshop on Ontology Learning*,* Berlin, August 2000, 37-42.
- Abe, N. and Li, H. (1996). *Learning word association norms using tree cut pair models*. In. *Proc. Of 13th Int. Conf. On Machine Learning*, 1996.

3.2.18 Xu and colleagues' approach

The approach presented in [Xu *et al.*, 2002] aims to acquire domain-relevant terms and their relations using unsupervised hybrid text-mining techniques. The approach is based on using two different text-mining techniques to learn lexico-sysntactic patterns, such as near synonymy relations, which indicate domain-relevant syntactic relations between the extracted terms. The first one uses an existing ontology as initial knowledge for learning lexico-syntactic patterns, while the second is based on different co-location acquisition methods to deal with the free word-order language. The input for the process consists of a collection of pre-classified and linguistically annotated documents. In summary, the approach is based on using language parsing, an existing general ontology and statistical measures.

The steps proposed by this approach are:

1. *Extract single-word terms* using a word-term classification.
2. *Learn multi-word terms and identify the lexico-syntactic patterns* using term co-location methods.
3. *Learn patterns* using a set of known relations (initialised with GermaNet or WordNet). The linguistic patterns are determined using the known relations coded in GermaNet or WordNet. Similar patterns are grouped to build clusters of patterns, and are finally assigned to the correct relation type.
4. *Extract related terms via the application of learned lexico-syntactic patterns to the corpus* using a relation extractor which looks for to lexico-syntactic patterns.

The overall process can be explained as the following. Firstly, it is necessary to *mine relevant terms* for the domain, for which several measures based on categorized documents are applied. Next, the process aims to *learn relations with lexico-syntactic patterns* between the terms extracted from the corpus, using the relations contained in GermaNet or WordNet to assign synonymy, hyponymy and meronymy relations. To perform this activity, text fragments containing these semantic relations are extracted and similar relations are grouped to build clusters of patterns. At the end of this process, two types of patterns can be identified: domain-specific patterns, that define reliable domain-specific relations; and domain independent patterns. With these grouped relations, together with the extracted terms, *clusters of terms* can be created. Finally, with a learning term co-location activity, terms are put into the correct place *in the taxonomy* using the patterns mentioned before, and with statistical measures calculated for each pattern.

Main techniques used: NL techniques, statistical and text-mining approaches
Reuse other ontologies: existing taxonomies, GermaNet and WordNet
Source: linguistically annotated and pre-classified text.
Main goal: learn concepts and relations between them.

Domain in which it has been applied: management succession, stock market and drugs in German language.
Tool associated: TFIDF
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

- Xu F., Kurz D., Piskorski J., and Schmeier S. (2002). *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping*. In Proceedings of LREC 2002, the third international conference on language resources and evaluation, Las Palmas, Canary island, Spain, May 2002.

3.3 Ontology Learning Tools from text

In this section, we will summarize the most relevant tools used for ontology learning from text.

3.3.1 ASIUM

ASIUM [Faure and Nedellec, 1999] is an acronym for "Acquisition of SemantIc knowledge Using Machine learning methods". It has been developed at the Computer Science Research Laboratory (LRI) in the University of Paris-Sud[10]. The main aim of ASIUM is to help the expert in the acquisition of semantic knowledge from technical texts using syntactic analysis. ASIUM takes as input French texts in natural language and associates a rate of appearance in the text. The learning method is based on conceptual and hierarchical clustering. Basic clusters are formed by words that occur with the same verb after the same preposition [Faure and Nedellec, 1998]. The tool uses a metric to compute the semantic similarity between clusters, which is used by the ontologist to decide if a new concept is created. Clusters are successively aggregated by the conceptual clustering method to form the concepts of the ontology. The ontologist defines a minimum threshold for gathering clusters into concepts. The learning is intertwined and validated by the ontologist.

The tool follows two steps to achieve its performances. The first one, the *Factorisation* (conceptualisation): The head words are associated with its frequency of appearance in the text in order to calculate the distance among concepts. Those who appear in similar contexts are added, by means of an algorithm of conceptual clustering to form the concepts of the ontology. For this purpose, a technique to estimate the semantic similarity among concepts has been used [Liu 1996], [Bisson 1994] and [Bisson 1992]. The second step is *Clustering* (ontology building). Due to the fact that a hierarchy would be too much restricted to represent the complexity of the ontology in many domains, the authors have adopted the skill of pyramidal clustering. The ontology is constructed level-by-level.

Goal and scope of the tool: to find taxonomic relations among terms in natural language texts in French without annotating them
Learning technique used by the tool: conceptual clustering
Method followed for ontology learning: two steps: (1) Factorisation (conceptualisation) and (2) Clustering (ontology building)
User/Expert intervention in the process: the user participation is needed not only to tag the new concepts but also to control the generality level of verbal frames, in order to refine the learned clusters and to handle noise. Each of the conceptualisation and clustering steps are validated by the expert. If a cluster or a pyramid level is updated each process starts again
Types of sources used by the method: ASIUM uses texts syntactically analysed by SILEX [Constant, 95]

[10] http://www.lri.fr/

SW Architecture: information not available in papers
Interoperability with other tools: it is possible to use ASIUM as knowledge acquisition tool in any other ontology development one
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.lri.fr/~faure/Demonstration/Presentation_Demo.html
References.

- Faure D, Poibeau T (2000) First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany.
- Faure D, Nédellec C. (1999) Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In D. Fensel and R. Studer editors, Proc. Of the 11th European Workshop (EKAW'99), LNAI 1621, pages 329-334. Springer-Verlag.
- Faure D, Nédellec C. (1998) A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In LREC workshop on adapting lexical and corpus resources to sublanguages and applications, Granada, Spain.
- Liu W. Z.(1996) *An Integrated Approach for Different Attribute Types in Nearest Neighbour Classification*. The Knowledge Engineering Review, 1996.
- Bisson G. (1994) *Conceptual Clustering.* In Mlnet Summer School on Machine Learning and Knowledge acquisition. Dourdan, France, September 1994. 5-10
- Bisson G. (1992) *Learning in FOL with a similarity measure.* In Tenth National Conference of Artificial Intelligence. San José, California. July 1992. 12-16

3.3.2 Caméléon

Caméléon [Aussenac-Gilles and Seguela, 2000] has been developed in the frame of a joint convention between the IRIT laboratory (Institut de Recherches en Informatique de Toulouse) and CEA (Commisariat à l'Energie Atomatique) - CEN de Cadarache. Caméléon assists the learning of conceptual relations to enrich conceptual models for the REX Knowledge management System. Caméléon relies on linguistic principals for relation identification: lexico-syntactic patterns are good indicators of semantic relations. Some patterns may be regular enough to indicate the same kind of relation from one domain to another. Other patterns are domain specific and may reveal domain specific relations.

Learning relations conceptual with Caméléon is a two-fold process. The first part is dedicated to the identification of the relevant patterns and relations for the current corpus. Generic patterns from a generic base are available in the tool and must evaluated for the current corpus. They may be modified or rejected. New specific patterns may be identified either by manually reading some term contexts, or by using Hearst's principle with couples of domain specific related terms. The second part is dedicated to using these patterns to identify lexical relations and manually enrich a conceptual model from them. Patterns are used to list all possible lexical relations in the texts. Their evaluation provides suggestions of conceptual relations. For each concept in the model, lexical relations are presented and must be validated to enrich the model.

Goal and scope of the tool: to tune generic lecixo-syntactic patterns or build new ones for a specific domain and corpus, to find taxonomic and non taxonomic lexical relations in texts, and to enrich a conceptual model from these lexical relations

Learning technique used by the tool: reuse and tuning of generic patterns (mainly for taxonomic relations), hearst's proposal: couples domain specific related terms are localised on texts and concepts are used to define new patterns, and pattern identification in text help to learn lexical relations and their validation leads to conceptual relations

Method followed for ontology learning: is a cyclic process with four steps: (1) evaluate and adjust generic patterns on the corpus, (2) identify new domain specific patterns and relations, (3) evaluate the lexical relations identified thanks to these patterns, (4) evaluate the opportunity to add conceptual relations to the model. Back to step 2 until no new pattern or relation is identified

User/Expert intervention in the process: validates or adapts patterns, defines new domain specific patterns and relations, enriches the model with concepts and relations. Domain expert just validates the model

Types of sources used by the method: works with texts processed by taggers like Tree Tagger (for English) or Cordial Unievrsité (for French). Caméléon has its own base of generic patterns that is continuously enriched after each new experiment.

SW Architecture: client-server. The server is used to store in a MySQL database pattern bases, tagged texts, extracted terms and lexical relations found in these texts; the client is a JAVA application that provides editors and result browsers.

Interoperability with other tools: imports lists of terms from any term extractor (Nomino or Syntex have been tested).

Import and export facilities to ontology languages: exports ontologies in OWL
User interface facilities: pattern evaluation module and model enrichment module with relation validator and editor
URL where you can find information about the method or tool: information not available in papers

References.

- Aussenac-Gilles N. and Seguela P. (2000) Les relations sémantiques: du linguistique au formel. Cahiers de grammaire, N° spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : Presse de l'UTM. Pp 175-198.
- Seguela P. (1999) Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés, in Actes de TIA'99 (Terminologie et Intelligence Artificielle), Nantes, Terminologies Nouvelles n°19, 52-60.

3.3.3 CORPORUM-Ontobuilder

CORPORUM-Ontobuilder has been developed by CognIT. Its main aim is to be able to extract ontologies (mainly taxonomies) from natural language texts. The tool uses several linguistic techniques that drive the analysis and information extraction functionalities. CORPORUM-Ontobuilder extracts information from structured and unstructured documents using the tools named OntoWrappper [Engels, 2002] and OntoExtract [Engels, 2001]. Ontowrapper extracts information from on-line resources (e.g. names, email addresses, telephone numbers, etc.) and OntoExtract obtains taxonomies from natural language texts. Ontoextract is also able (through semantic analysis of the content of web pages) to provide initial concept taxonomies, to refine existing concept taxonomies (include more concepts), to find relations between key terms in documents and to find concept instances within documents. Concept taxonomies are created in RDF(S).

Goal and scope of the tool: to extract an initial ontology and refine it
Learning technique used by the tool: linguistic and semantic techniques
Method followed for ontology learning: own method
User/Expert intervention in the process: not necessary
Types of sources used by the method: text
SW Architecture: information not available in papers
Interoperability with other tools: OntoWrapper, OntoExtract
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://ontoserver.cognit.no
References.

- Engels R (2001) CORPORUM-OntoExtract. Ontology Extraction Tool. Deliverable 6 Ontoknowledge. http://www.ontonowledge.org/del.shtml
- Engels R (2001) CORPORUM-OntoWrapper. Extraction of structured information from web based resources. Deliverable 7 – Ontoknowledge. http://www.ontonowledge.org/del.shtml

3.3.4 DOE: Differential Ontology Editor

DOE is a simple ontology editor that allows the user to build ontologies in three steps according to the method proposed by Bruno Bachimont. In the first step, the user develops taxonomies of concept and relation by justifying explicitly their position in the hierarchy. For each notion, the editor proposes mainly four principles to fill, which come from the Differential Semantic according to François Rastier. Hence, the user has to explicit in what a notion is similar but more specific than its parent and in what this notion is similar but different from its siblings. In a second step, the two taxonomies are imported and the user can add constraints onto the domains of the relations. Finally, in a third step, the ontology can be translated into a KR language via XSLT style sheet, based on conceptual graph or DAML+OIL, OIL and RDF(S). DOE is not intended as a full ontology development environment but is rather a complement of others editors.

DOE is not intended as a full ontology development environment. It will not actively support many activities that are involved traditionally in ontology construction, but is rather a complement of others editors, offering linguistic techniques which attach a lexical definition to the notions used and consequently justify the taxonomy.

Goal and scope of the tool: to help to the ontologist in the process of building an ontology
Learning technique used by the tool: differential semantic
Method followed for ontology learning: Bachimont's method
User/Expert intervention in the process: whole process
Types of sources used by the method: text
SW Architecture: standalone
Interoperability with other tools: none
Import and export facilities to ontology languages: import from DOE Format and RDFS; export to CGXML, DOE Format, KML, RDFS, OIL, and DAML+OIL
User interface facilities: visual editor**]**
URL where you can find information about the method or tool: http://opales.ina.fr/public/
References.

- Bachimont B. (1996) *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser*. PhD Thesis, Ecole Polytechnique, 1996.
- Bachimont B. (2000). *Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances*. In Ingénierie des Conniassances : Evolutions récentes et nouveaux défis, Eyrolles, 2000.

3.3.5 KEA: Keyphrases Extraction Algorithm

Kea [Jones and Paynter, 2002] automatically extracts keyphrases from the full text of documents. The set of all *candidate phrases* in a document are identified using lexical processing, *features* are computed for each candidate, and machine learning is used to generate a classifier that determines which candidates should be assigned as keyphrases. The machine learning scheme first builds a prediction model using training documents with known keyphrases, and then uses the model to find keyphrases in new documents. Two features are used in the standard algorithm: TF/IDF[11] and position of first occurrence.

[11] *Term Frequency-Inverse Document Frequency*

The TF/IDF requires a corpus of text from which document frequencies can be calculated; the machine learning phase requires a set of training documents with keyphrases assigned.

Kea's extraction algorithm has two stages:

1. Training stage that uses a set of training documents for which the author's keyphrases are known. For each training document, candidate phrases are identified and different features values are calculated. To reduce the size of the training set, any phrase that occurs only once in the document is discarded. Each phrase is then marked as a keyphrase or a non-keyphrase, using the actual keyphrases for that document.
2. Extraction stage. To select keyphrases from a new document, Kea determines candidate phrases and feature values, and then applies the model built during the training stage. The model determines the overall probability that each candidate is a keyphrase, and then a post-processing operation selects the best set of keyphrases.

The success of the procedure can be evaluated on a large test corpus, in terms of how many authors assigned keyphrases are correctly identified. We are also conducting evaluations using human assessors to rate keyphrases.

Goal and scope of the tool: to extract keyphrases that represent the content of a document
Learning technique used by the tool: statistical approach, machine learning and lexical processing
Method followed for ontology learning: not propose
User/Expert intervention in the process: evaluate the process
Types of sources used by the method: text documents
SW Architecture: standalone and it is implemented in Java
Interoperability with other tools: WEKA machine learning workbench
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.nzdl.org/Kea/
References.

- Jones, S. and Paynter, G.W. (2002) *Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications*. Journal of the American Society for Information Science and Technology (JASIST)

3.3.6 LTG (Language Technology Group) Text Processing Workbench

LTG (Language Technology Group) Text Processing Workbench[12] [Mikheev and Finch, 1997] has been developed by Language Technology Group (LTG) in the University of Edinburgh. It is a set of computational tools for uncovering internal structure in natural language texts written in English. The main idea behind the workbench is the independence of the text representation and text analysis.

In LTG, ontology learning is performed in two sequential steps: representation and analysis. At the representation step, the text is converted from a sequence of characters to features of interest by means of annotation tools. At the analysis step, those features are used by statistics-gathering tools and inference tools for finding significant correlations in the texts. The analysis tools are independent from a particular assumption on the nature of the feature-set and work on the abstract level of feature-elements which are represented as SGML items. The workbench is being used both for lexicographic purposes and for statistical language modelling.

[12] http://www.ltg.ed.ac.uk/%7Emikheev/workbench.html

It supports an incremental process of corpus analysis starting from a rough automatic extraction and organization of lexical-semantic regularities and ending with a computer-supported analysis of extracted data and a semi-automatic refinement of obtained hypotheses. To do this the workbench uses methods from computational linguistics, information retrieval and knowledge engineering.

Goal and scope of the tool: to discover internal structure of texts in natural language
Learning technique used by the tool: statistic Inference
Method followed for ontology learning: two steps: representation (annotation of text in SGML) and analysis (the annotated text is analysed by the statistical inference tools)
User/Expert intervention in the process: N/A
Types of sources used by the method: texts in natural language
SW Architecture information not available in papers
Interoperability with other tools: it is possible to use LTG Text Processing Workbench as knowledge acquisition tool in any other ontology development tool
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.ltg.ed.ac.uk/%7Emikheev/workbench.html
References.

- Mikheev, A. Finch, S.(1997) *A Workbench for Finding Structure in Texts*. Proceedings of ANLP-97 (Washington D.C.). ACL March 1997. pp 8.

3.3.7 Mo'K Workbench

Mo'K Workbench [Bisson et al., 2000] is a configurable workbench that supports the semiautomatic construction of ontologies from a corpus using different conceptual clustering methods. it has been developed by INRIA, LRI Univ. Paris-South[13]. Mo'k assists ontologists in the exploratory process of defining the most suitable learning method. In this sense, Mo'K supports the elaboration, comparison, characterization and evaluation of different conceptual clustering methods. It also permits fine-grained definitions of similarity measures and class construction operators, easing the task of method instantiation and configuration.

The learning process proposed by this workbench takes a corpus as input. No additional knowledge is used to label the input, to guide the learning, or to validate the learned results. Trough NLP techniques, the tool extracts from the corpus a list of triplets. A triplet is composed by a verb, a word and a syntactic role of this word in a sentence. Using the triplets, Mo'K calculates the number of occurrences of each one. Triplets with low number of occurrences or too many occurrences are removed from that list. Finally, Mo'K calculates the semantic distance between the triplets in the previous list to form conceptual clusters.

Goal and scope of the tool: to obtain concept taxonomy from domain tagged text
Learning technique used by the tool: conceptual clustering
Method followed for ontology learning: following Harris hypotheses [Harris, 1989], syntactic relations are used among words to derive semantic relations
User/Expert intervention in the process: information not available in papers
Types of sources used by the method: domain tagged texts. In addition, there can be used other ontologies, dictionaries and other similar resources
SW Architecture information not available in papers
Interoperability with other tools: it is possible to use Mo'K Workbench as knowledge acquisition tool in any other ontology development tool

[13] http://www.inria.fr/index.en.html

Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: information not available in papers
References.

- Bisson G, Nedellec C, Cañamero D. (2000) *Designing Clustering Methods for Ontology Building. The Mo'K Workbench*. In S. Staab, A. Maedche, C. Nedellec, P. WiemerHasting (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00, Berlin, Germany, August 20-25.

3.3.8 OntoLearn Tool

The OntoLearn tool [Velardi et al., 2002] aims to extract relevant domain terms from a corpus of text, relate them to appropriate concepts in a general-purpose ontology, and to detect relations among the concepts. To carry out these tasks, natural language analysis and machine learning techniques are used. The tool has been tested inside the Harmonise[14] European project.

OntoLearn extracts terminology from a corpus of domain text, such as specialized Web sites. The system then filters the terms using natural language processing and statistical techniques that perform comparative analysis across different domains, or contrasting corpora. This analysis identifies terminology that is used in the target domain but is not seen in other domains. Next, it uses the WordNet and SemCor lexical knowledge bases to perform semantic interpretation of the terms. The tool then relates concepts according to taxonomic (kind-of) and other semantic relations, generating a domain concept forest. For this purpose, WordNet and a rule-based inductive-learning method have been used to extract such relations. Finally, OntoLearn integrates the domain concept forest with WordNet to create a pruned and specialized view of the domain ontology. The validation of the process is performed by an expert.

Goal and scope of the tool: to enrich a domain ontology with concepts and relations
Learning technique used by the tool: NLP and machine learning (semantic interpretation)
Method followed for ontology learning: OntoLearn method: semantic interpretation
User/Expert intervention in the process: evaluation
Types of sources used by the method: text
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: information not available in papers
References.

- Velardi P., Navigli R., and Missikoff M. (2002). *Integrated approach for Web ontology learning and engineering.* IEEE Computer - November 2002.
- Velardi P., Missikoff M., and Fabriani P. (2001). *Using Text Processing Techniques to Automatically enrich a Domain Ontology*. ACM conference on Formal Ontologies in Information Systems (FOIS 2001), Maine, USA (2001)

[14] http://www.harmonise.org/

3.3.9 Prométhée

Prométhée [Morin, 1998 and 1999] has been developed at the Institut de Reserche en Informatique de Nantes (IRIN)[15]. It is a machine learning based tool for the extraction and refinement of lexical-syntactic patterns relative to conceptual specific relations from technical corpora [Martienne&Morin, 1999]. It uses patterns bases, which enrich with the ones extracted in the learning.

To refine patterns, the authors propose the Eagle [Martienne and Quafafou, 1998] learning system. The Eagle system is based on the inductive paradigm *learning from examples* [Muggleton, 1991], which consists of the extraction of intensional descriptions of target concepts from their extensional descriptions, as well as previous knowledge on the given domain [Mitchell, 1997]. This specifies general information, like the objects characteristics and their relations. Eagle extracts *intensional* descriptions of concepts from their *extensional* descriptions. The learned definitions are later used in recognition and classification tasks.

The interface between the two systems is as follows: (1) Prométhée extracts lexical-syntactic patterns, (2) some instances of these patterns are then produced from the corpus and classified among the examples of the patterns, and (3) from these labelled patterns, Eagle produces descriptions that are interpreted as restrictions refining the patterns.

Goal and scope of the tool: extraction and refinement of lexical-syntactic patterns relative to conceptual specific relations.
Learning technique used by the tool: learning from examples.
Method followed for ontology learning: Eagle extracts *intensional* descriptions of concepts from their *extensional* descriptions
User/Expert intervention in the process: whole process
Types of sources used by the method: pattern bases
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/promethee/promethee.html
References:

- Morin E. (1999) *Acquisition de patrons lexico-syntaxiques caractéristiques dúne relation sémantique.* TAL (Traitement Automatique des Langues). 40/1: 143-166, 1999 (Prométhée).
- Morin E. (1998) *Prométhée un outil d'aide a l'acquisition de relations semantiques entre temes.* In Actes, 5th National Conference on Traitement Automatique des Langues Naturelles (TALN'98), pages 172-181, Paris, France, June 1998

3.3.10 SOAT: a Semi-Automatic Domain Ontology Acquisition Tool

SOAT tool [Wu and Hsu, 2002] allows a semi-automatic domain ontology acquisition from a domain corpus. The main objective of the tool is to extract relationships from parsed sentences based on applying phrase-rules to identify keywords with strong semantic links like hyperonym or synonym. The acquisition process is based on using InfoMap [Hsu et al., 2001], a knowledge representation framework, that integrates linguistic, commonsense, and domain knowledge. InfoMap has been developed to perform natural language understanding, and to capture the topic words, usually pairs of noun and verb, or noun and noun in a sentence. InfoMap has two major relations between concepts: taxonomic relations (category and synonym) and non-taxonomic (attribute and event).

[15] http://www.sciences.univ-nantes.fr/irin/

The acquisition process carry out by SOAT includes to collect domain keywords and find the relationships among them. To perform this activity, a set of rules has been defined for extracting keywords from a sentence related to concepts in InfoMap with a strong semantic relation between them. The tool receives as input a domain corpus with the POS-tag. A keyword, usually the name of the domain, is selected in the corpus as root. Then, with this keywords, the process aims to find a new related keyword with the previous by means of applying the extraction rules and add the new keyword into the ontology according to the rules and the structure fixed in InfoMap. This new keyword is now taken as root to repeat the process during a determined number of times or until being impossible to find a new related keyword. The user intervention is necessary to verify the results of the acquisition and to refine and update the extraction rules. The restrictions of SOAT is that the quality of the corpus must be very high in the sense that the sentences must be accurate and enough to include most of the important relationships to be extracted.

Goal and scope of the tool: acquisition of relationships using a predefined knowledge representation framework
Learning technique used by the tool: phrase-patterns, defined as set of linguistic templates
Method followed for ontology learning: not specified, the tool follows its own method
User/Expert intervention in the process: information not available in papers
Types of sources used by the method: text
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: none
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.iis.sinica.edu.tw/IASL/en/index.htm
References.

- Wu S.H, Hsu W.L. (2002). *SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus.* In the 19th International Conference on Computational Linguistics, Howard International House and Academia Sinica, Taipei, Taiwan
- Hsu W.L., Wu S.H., and Chen, Y.S. (2001). *Event identification based on the Information Map – InfoMap.* In symposium NLPKE of the IEEE SMC Conference, Tuckson, Arizona, USA.

3.3.11 SubWordNet Engineering Process tool

In [Gupta et al., 2002] is presented an architecture to interactively acquire and maintain sublanguage WordNets follows the Iterative SubWordNet Engineering process approach explained in this deliverable. The architecture builds upon WordNet semantic structure and includes integrated capabilities for concept element discovery, concept identification, and concept maintenance. The architecture to perform each of these capabilities has been modularised into three layers: the graphical user interface (GUI) layer, the process layer, and the data layer.

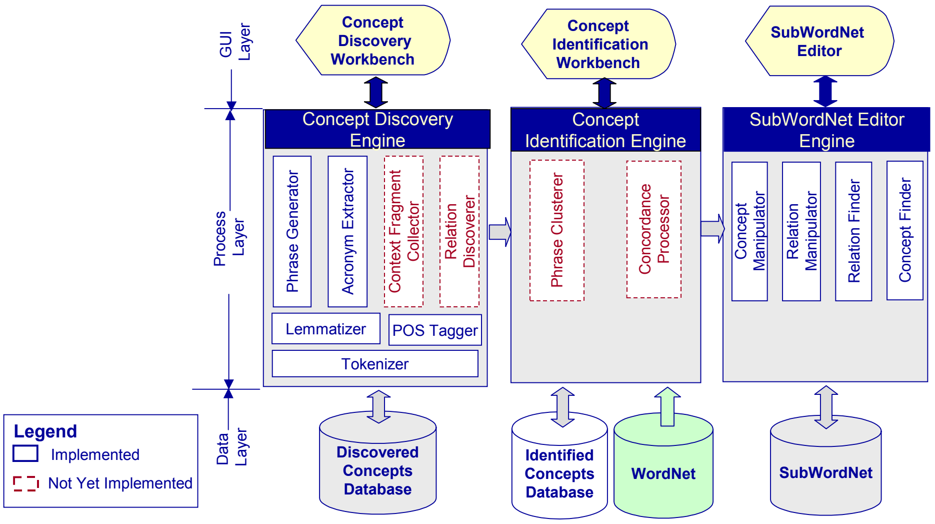


Figure 6 Architecture for Engineering SubWordNets

- The *concept discovery capability* includes a Concept Discovery Workbench, a Concept Discovery Engine, and a Discovered Concepts Database modules. This module provides a GUI that allows users to select the documents, manipulate discovered concept elements, and also provides summary distributional information of words and phrase for assisting users. It also includes several NLP components to discover relations using collocation statistics, lexical patterns, etc.
- The *concept identification capability* includes Concept Identification Workbench, Concept Identification Engine, and the Identified Concepts Database modules. This module has been designed to support concept identification, phrase clustering, and to establish concordance between concept nodes and WordNet synsets.
- For the *concept maintenance capabilities* includes the SubWordNet Edito*r* as the GUI layer, the SubWordNet Editor Engine as the process layer, and the SubWordNet Database as the data layer.

Goal and scope of the tool: to build Sublanguage WordNets
Learning technique used by the tool: different NLP techniques and several statistical approaches
Method followed for ontology learning: approach for iterative SubWordNet engineering process
User/Expert intervention in the process: whole process
Types of sources used by the method: textual documents
SW Architecture information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: visual editor
URL where you can find information about the method or tool:
http://www.aic.nrl.navy.mil/~aha/cbr/luikm.html
References.

- Gupta, K.M., Aha, D.W., Marsh, E., and Maney, T. (2002). *An architecture for engineering sublanguage WordNets*. In Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages

3.3.12 SVETLAN'

SVETLAN' [Chaelandar and Grau, 2000] has been developed by the LIR research group of the HMC Department of the University of Paris-Sud[16]. SVETLAN' is a domain-independent tool that creates clusters from words appearing on texts. Its learning method is based on a distributional approach: nouns playing the same syntactic role in sentences with the same verb are aggregated in the same class.

The learning process has the following steps: syntactic analysis, aggregation and filtering. In the syntactic analysis step, the tool retrieves sentences of the original texts in order to find the verb inside the sentence. This is based on the assumption that verbs allow categorizing nouns. The output of this step is a list of triplets that contain the verb, the noun and the syntactic relation between them. The aggregation step constructs groups of nouns with similar meanings. The filtering step is based on the weight of the nouns inside their classes. It removes nouns from these groups if they are not very relevant for the class. The threshold is established by the ontology developer. The process doesn't require validation and it is completely independent from the ontology developer.

Goal and scope of the tool: to build a hierarchy of concepts
Learning technique used by the tool: conceptual clustering
Method followed for ontology learning: syntactic analysis, clustering and filtering
User/Expert intervention in the process: validation
Types of sources used by the method: French texts in natural language
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: http://www.limsi.fr/Individu/gael/ManuscritThese/
References.

- Chaelandar G, Grau B. (2000) SVETLAN'- A System to Classify Words in Context. In S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.) Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25.

3.3.13 TFIDF-based term classification system

The system presented here [Xu *et al.*, 2002] aims to detect relevant domain terms and to learn relations that hold among them. The tool gives support for the approach described in the section 3.2.18 of this deliverable.

The tool has the following three main components. A *based single-word term classifier* that is used to extract single words from a corpus. A *lexico-syntactical pattern finder*, that has two sub-modules: the first one is for learning patterns based on the set of known relations (using GermaNet or WordNet) and implements the interfaces needed to interoperate with these two systems; the second is for learning patterns based on term co-location methods. The final component is the *relation extractor*.

The system receives as input a domain corpus, that is annotated and parsed using a shallow NLP tool. In this case the tool used is SPPC [Xu *et al.*, 2002] (Shallow Processing Production Center). This NL tool provides a domain independent extraction for processing German free-text documents and consists of a tokenizer, lexical processor, part-of-speech filtering, named-entity finder and chunk recogniser, in addition to other components.

Goal and scope of the tool: to learn concepts and relations between them
Learning technique used by the tool: text-mining and statistical approach

[16] http://www.limsi.fr/

Method followed for ontology learning: unsupervised hybrid text-mining approach to acquire domain terms and their relations
User/Expert intervention in the process: evaluation
Types of sources used by the method: text
SW Architecture: information not available in papers
Interoperability with other tools: GermaNet and SPPC NLP tool
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: information not available in papers
References.

- Xu F., Kurz D., Piskorski J., and Schmeier S. (2002). *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping*. In Proceedings of LREC 2002, the third international conference on language resources and evaluation, Las Palmas, Canary island, Spain, May 2002.

3.3.14 TERMINAE

TERMINAE[17] [Biébow et al., 1999]] and [Szulman et al., 2002] has been developed in the Laboratoire d'Informatique of Paris-Nord at the University of Paris-Nord (LIPN)[18]. It integrates linguistic tools and knowledge engineering tools. The linguistic tool allows defining terminological forms from the analysis of term occurrences in a corpus. The ontologists analyse the uses of the term in the corpus to define the meanings of the terms. The knowledge engineering tool involves an editor and a browser for the ontology. The tool helps to represent terminological forms as a concept (called terminological concept). The tool helps to represent a terminological form as a concept (called terminological concept).

TERMINAE uses a method to build concepts from the study of the corresponding term in a corpus. First, the tool establishes the list of terms, which requires the constitution of a relevant corpus on the domain. Using a term extractor tool, a set of candidate terms are proposed to the ontologist, which selects a set of terms. After that, the ontologist conceptualises the terms and analyses the uses of the term in the corpus to define all the meanings of the term. The ontologist gives a definition in NL for each meaning, and then translates the definition into an implementation language.

Goal and scope of the tool: to build an ontology
Learning technique used by the tool: conceptual clustering
Method followed for ontology learning: an expert extracts "terms" from the "term candidates" list and defines "notions" for the "term" meanings
User/Expert intervention in the process: validation
Types of sources used by the method: French or English texts
SW Architecture: information not available in papers
Interoperability with other tools: TERMINAE imports lists of extracted terms from the LEXTER and SYNTER term extractors
Import and export facilities to ontology languages: is able to read RDFs and OIL ontologies, and to generate XML, RDFs or OIL files from an ontology built with this tool
User interface facilities: hierarchy editor and graphic visualisation tool, traceability functions to go from a concept editor to the correspond term editors or to its formal representation in description logic
URL where you can find information about the method or tool:
http://www-lipn.univ-paris13.fr/~szulman/TERMINAE.html
References.

-

[17] www-lipn.univ-paris13.fr/
[18] http://www-lipn.univ-paris13.fr/index-english.html

- Szulman, S., Biebow B., Aussenac-Gilles N. (2002) *Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE. Traitement Automatique de la Langue (TAL)*. Numéro spécial sur le Structuration de Terminologie. Eds A. Nazarenko, T. Hammon. Vol43, N°1; pp 103-128. 2002
- Biébow B, Szulman S. (1999) *TERMINAE: a linguistic-based tool for the building of a domain ontology*. In EKAW'99 Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, pages 49-66, Berlin, 1999. Springer-Verlag.

3.3.15 Text-To-Onto

Text-To-Onto [Maedche and Volz, 2000] [Maedche and Staab, 2003] has been developed at the AIFB Institute in the University of Karlsruhe[19]. The tool integrates an environment for building domain ontologies from an initial core ontology. It also discovers conceptual structures from different German sources using knowledge acquisition and machine learning techniques. Text-To-Onto has implemented some techniques for ontology learning from free text and semi-structured text, dictionaries, legacy ontologies and databases. The result of the learning process is a domain ontology that contains domain-specific and domain-independent concepts. Domain-independent concepts are withdrawn to better adjust the vocabulary of the domain ontology. The result of the process is a domain ontology that only contains domain concepts learnt from the input sources related before. The whole process is supervised by the ontologists. This is a cyclic process, in the sense that it is possible to refine and complete the ontology if we repeat the process.

Goal and scope of the tool: to find taxonomic and non-taxonomic relations
Learning technique used by the tool: statistical approach, pruning techniques and association rules
Method followed for ontology learning: method based on Srikant's and Agrawal [Srikant&Agrawal, 1995] algorithm
User/Expert intervention in the process: validation
Types of sources used by the method: machine readable dictionaries and other ontologies
SW Architecture is part of KAON tool suite
Interoperability with other tools: Text-To-Onto is a component of integrated environment for ontological manual and semiautomatic engineering, KAON[20]
Import and export facilities to ontology languages: imports from DAML+OIL, RDF(S), and exports to DAML+OIL and RDF(S)
User interface facilities: visual editor
URL where you can find information about the method or tool: http://ontoserver.aifb.uni-karlsruhe.de/texttoonto/
References.

- Maedche A. and Staab S. (2003) *Ontology Learning*. In S. Staab & R. Studer (eds.) Handbook on Ontologies in Information Systems. Springer 2003. http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf
- Maedche, A. and Volz, R. (2001) *The Text-To-Onto Ontology Extraction and Maintenance Environment.* To appear in Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.
- Srikant R, Agrawal R.(1995) *Mining generalized association rules*. In Proc. Of VLDB'95, pages 407-419.

[19] http://www.aifb.uni-karlsruhe.de/

[20] http://kaon.semanticweb.org/

3.3.16 TextStorm and Clouds

The framework presented here has been developed within the Dr. Divago project [Pereira, 98] for semi-automatically constructing a semantic network which contains only concepts and their relations, using a relevant text for the target domain.

It is composed of two main modules: TextStorm [Oliveira et al., 2001], and Clouds [Pereira et al., 2000], that perform complementary activities to construct a semantic network, as can be seen in the figure 7. TextStorm deals with the task of extracting relations between concepts from a text file using NL techniques, while Clouds is concentrated on completing these relations and extrapolating rules about the knowledge previously extracted using NL techniques.

TextStorm is a NL tool that extracts binary predicates from a text using syntactic and discourse knowledge. The process starts with providing the system with texts that contain relevant features of the target domain. After this, the text is tagged using a WordNet database to find all parts of speech to which a word may belong and to classify words in the parsing process. Then, the text is parsed using an augmented grammar to obtain a lexical classification of the words in the parsing process. The predicates on which the tool is focused are all of those that relate two concepts in sentences. These predicates are only verbal phrases that contain two nouns (subject and direct object) connected with a verb that specifies an existent relation between them. TextStorm creates a list with this information that will be the input for the Clouds tool. To perform the process, the system needs to interact with the user, who has the responsibility to resolve inconsistencies and to decide the relevance of a sentence for the domain.

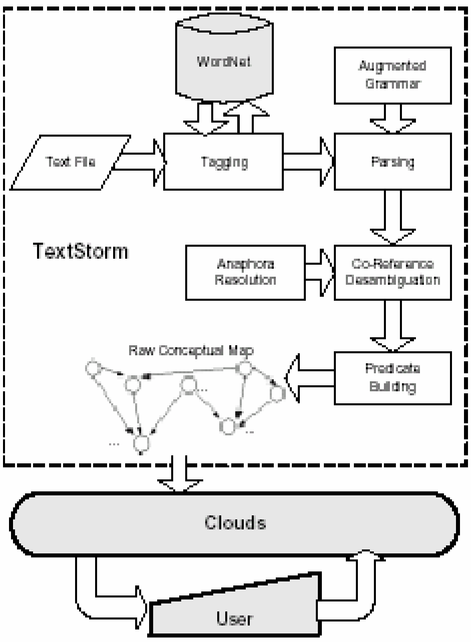


Figure 7. TextStorm and Clouds framework

Clouds is responsible for the construction of a semantic network in an interactive way. Using the previous list, with all binary predicates extracted from the text, Clouds builds a hierarchical tree of concepts, and learns some particulars of the domain using two different techniques. The first one is called a *best current hypothesis based algorithm* to learn the categories of the arguments of each of the relations. The other is an *Inductive Logic Programming based algorithm* to learn the contexts that are recurrent in each relation. To perform the process, Clouds will ask the user about new concepts and new relations that it suspects the existence of.

Goal and scope of the tool: to build a taxonomy focusing only on subclass relations
Learning technique used by the tool: NLP, current hypothesis algorithm, Inductive Logic Programming

algorithm
Method followed for ontology learning: own method
User/Expert intervention in the process: present in whole process, but the user has the particular responsibility of resolving inconsistencies and determining the relevance of the new concepts
Types of sources used by the method: text
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: not available
References.

- Novak J. D. and Gowin D. B. (1984). Learning How To Learn. New York: Cambridge University Press. 1984.
- Oliveira A., Pereira F.C., and Cardoso A. (2001). Automatic Reading and Learning from Text. In Proceedings of International Symposium on Artificial Intelligence, ISAI'2001. December, 2001
- Pereira, F. C. (1998). Modelling Divergent Production: a multi domain Approach. European Conference of Artificial Intelligence, ECAI'98, Brighton, UK, 1998.
- Pereira, F. C.; Oliveira, A. and Cardoso, A. (2000). Extracting Concept Maps with Clouds. Argentine Symposium of Artificial Intelligence (ASAI 2000), Buenos Aires, Argentina, 2000.

3.3.17 Welkin

Welkin (Alfonseca and Rodríguez, 2002) is a tool for automatically generating e-learning materials from unrestricted texts. In particular, one of the modules of the architecture tries to create an ontological representation of the terms of interest that appear in the text, to internally represent the different sections in the e-learning web sites. The aim of this module is the analysis of the texts to identify relevant terminology, and the classification of those terms inside lexical ontologies such as WordNet.

Contextual information is used for the classification. Each of the concepts in the original ontology is extended with information about which words can appear in their contexts, and which of those have syntactic relationships to it. A distance metric, based upon the representations of the contexts, is then used to classify the new terms inside the ontology. The final part of the procedure is performed by a module that looks for word patterns that express relationships between the concepts.

This method was developed as part of the Ensenada CICYT project (2002-2005), funded by the Spanish Ministry, a project which includes knowledge acquisition from free texts for automatic generation of e-learning materials.

This procedure has been used to automatically classify high-frequency concepts from historical texts for generating e-learning web sites (Darwin's The Voyages of the Beagle, Osler's The Evolution of Modern Medicine and Hegel's Lectures on the History of Philosophy), and test data has also been constructed from novels (Tolkien's The Lord of the Rings and Homer's The Iliad).

Goal and scope of the tool: automatic extension of existing general-purpose ontologies with new terms identified in unrestricted text
Learning technique used by the tool: the metrics to measure the similarities between contexts (which are represented as lists of words and frequencies) are those that are standard in the field of Information Retrieval: TF/IDF, chi-square, etc
Method followed for ontology learning: contextual signatures, hyponymy patterns
User/Expert intervention in the process: this is not required, but it is advised if a high accuracy is needed
Types of sources used by the method: a general-purpose ontology (such as WordNet) and a large corpus for collecting the contextual information (such as the Internet with a search engine)

SW Architecture: the architecture is implemented as a pipeline, where the text is processed with different modules for linguistic processing (stemmer, PoS taggers, and a cascade of shallow parsers). The relevant terms are then identified, their contexts are collected, and the terms are finally classified inside WordNet
Interoperability with other tools: none
Import and export facilities to ontology languages: none
User interface facilities: none, all the modules are unix console commands
URL where you can find information about the method or tool. http://www.ii.uam.es/~ealfon
References.

- Alfonseca E., and Rodríguez P. (2002), *Automatically Generating Hypermedia Documents depending on User Goals*, Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems, AH-2002, Málaga, Spain.

3.3.18 WOLFIE (WOrd Learning From Interpreted Examples)

The system WOLFIE [Thompson and Mooney, 1997] learns a semantic lexicon from a corpus of sentences paired with representations of their meaning. The lexicon learned consists of words paired with representations of their meaning, and allows for both synonymy and polysemy. WOLFIE is part of an integrated system that learns to parse novel sentences into their meaning representations.

The system combines the following features. First, arbitrary amounts of both polysemy and synonymy can be handled. Second, WOLFIE interacts with the system CHILL [Zelle, 1995] that learns to parse database queries directly into logical form. And finally, the algorithm used for learning is fast and accurate, and deals with the best selection for phrase meanings based on several heuristic. The idea behind its algorithm is that each choice of a lexical item may constrain the possible meanings of phrases not yet learned. To perform its goal, the system makes a few assumptions about the problem.

- The meaning of sentence is composed from possible meanings of words and phrases in that sentence.
- The sentence representation contains no noise.
- The meaning for each occurrence of a word in a sentence appears only once in the sentence's representation.

The system has been tested on acquiring a semantic lexicon for task of answering geographical database queries using a corpus of queries collected from human subjects and annotated with their executable logical form. The output produced by WOLFIE has been used to assist a larger language acquisition system like CHILL that uses WOLFIE as a parser acquisition system.

Goal and scope of the tool: to learn a semantic lexicon
Learning technique used by the tool: Natural language processing and statistical approaches
Method followed for ontology learning: information not available in papers
User/Expert intervention in the process: validation
Types of sources used by the method: pre-processed corpus (sentences paired with their meaning) and with examples provided by an expert
SW Architecture: information not available in papers
Interoperability with other tools: CHILL
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: information not available in papers
References.

- Thompson, C.A. & Mooney, R. J. (1997). Semantic Lexicon Acquisition for Learning Parsers. Technical Note. January 1997.

- Zelle, J. M. (1995). Using Inductive Logic Programming to automate the construction of natural language parsers. PhD Dissertation, University of Texas, Austin, TX. Also appears as Artificial Intelligence Laboratory Technical Report AI 96-249.

3.4 Conclusions

In this section, we have presented an overview of the most important methods and tools used to achieve the ontology learning process from text. In the next tables, we show a summary of all the methods and tools that have been described in this section. Table 1 summarizes the ontology learning methods from text according to the same description criteria presented before. In table 2, a summary of the topics described in the tools section is given.

Table 1. Summary of ontology learning methods from text. (1/3)

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Aguirre and colleagues' method	To enrich concepts in existing ontologies	Statistical approach Clustering Topic signatures	Yes	Domain Text WordNet	Information not available in papers	User	Agirre et al., 2000
Alfonseca and Manandhar's method	To enrich an existing ontology with new concepts	Topic signatures Semantic distance	Yes	Domain text WordNet	Welkin	Expert	Alfonseca et al., 2002
Aussenac-Gilles and colleagues' approach	To learn concepts and relations among them	Linguistic analysis Clustering techniques	Yes	Domain Text Domain ontologies	GEDITERM TERMINAE	User	Aussenac-Gilles and colleagues, 2000a and 2000b
Bachimont's method	To build a taxonomy	NLP techniques	No	Domain text	DOE	Expert	Bachimont et al., 2002
Faatz and Steinmetz approach	To enrich an existing ontology with new concepts	Statistical approach Semantic distance	Yes	Domain corpus Domain ontology	Any ontology workbench	Expert	Faatz et al 2002
Gupta and colleagues' approach	To build sub-languages in WordNet	NLP techniques Term-extraction techniques	Yes	Domain text WordNet	SubWordNet Engineering tool	Expert	Gupta et al. 2002
Hahn and colleagues' method	To learn new concepts	Concept hypothesis based on linguistic and conceptual quality labels Statistical approach	No	Domain text	Information not available in papers	Empirical measures and by an expert	Hahn et al 1998
Hearst's approach	To enrich an existing ontology	NLP techniques Linguistic patterns	Yes	Domain Text WordNet	Welkin	Expert	Hearst 1998 and Alfonseca et al. 2002

Table 1. Summary of ontology learning methods from text (2/3)

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Hwang's method	To elicit a taxonomy	NLP techniques ML techniques Statistical approach	No	Domain Text	Information not available in papers	Expert	Hwang 1999
Khan and Luo's method	To learn concepts	Clustering techniques Statistical approach	Yes	Domain text WordNet	Information not available in papers	Expert	Khan et al. 2002
Kietz and colleagues' method	To learn concepts and relations among them to enrich an existing ontology	NLP Statistical approach	Yes	Domain and non-specific domain Text Domain ontologies WordNet	Text-To-Onto	User	Kietz et al., 2000 Maedche el al. 2001
Lonsdale and colleagues' method	To discover new relationships in an existing ontology	NLP Mappings Linguistic technique	Yes	Terminological databases Domain ontology WordNet Domain text	Information not available in papers	User/Expert	Lonsdale et al., 2002
Missikoff and colleagues' method	To build taxonomies and to fuse with an existing ontology with	NLP Statistical approach ML techniques	Yes	Domain text WordNet	OntoLearn	Expert	Navigli et al., 2003 and Missikoff et al. 2002

Table 1. Summary of ontology learning methods from text. (3/3)

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Moldovan and Girju's method	To enrich an existing ontology	NLP techniques	Yes	Domain Text Lexical resources WordNet	Information not available in papers	Expert	Moldoban and Girju 2000 and 2001, and Harabagiu et al. 2000
Nobécourt approach	To learn concept and relations among them	Linguistic analysis	No	Domain Text	TERMINAE	User/expert	Nobécourt 2000
Roux and colleagues' approach	To enrich a taxonomy with new concepts	Verb-patterns	Yes	Domain text Domain ontology	Information not available in papers^{o}	Expert	Roux et al., 2000
Wagner approach	To enrich an existing ontology with new relationships	Statistical approach	Yes	WordNet	Information not available in papers	Expert	Wagner 2000
Xu and colleagues' approach.	To learn concepts and relations between them	NLP techniques Statistical approach Text-mining techniques	Yes	Annotated text corpus WordNet	TFIDF	Expert	Xu et al., 2002

Table 2. Summary of ontology learning tools from text (1/4)

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
ASIUM	To learn taxonomic relations	Conceptual clustering techniques	Own method	Text syntactically analysed	Whole process	Can be used to perform the knowledge acquisition to any other ontology development tool	Faure et al 2000, 1999, and 1998
Caméléon	To tune generic lecixo-syntactic patterns or build new ones. To find taxonomic and non taxonomic lexical relations in texts and to enrich a conceptual model with these lexical relations	To reuse and tuning of generic patterns (mainly for taxonomic relations), hearst's proposal, and pattern identification in text help to learn lexical relations and their validation leads to conceptual relations	Own method	Texts processed by taggers Its own base of generic patterns	Validates, adapts, or defines new domain specific patterns and relations Domain expert just validates the model	Imports lists of terms from any term extractor	Aussenac-Gilles and Seguela 2000
Corporum-Ontobuilder	To extract initial taxonomy	Linguistic and semantic techniques	Own method	Text	Not necessary	OntoWrapper and OntoExtract	Engels 2001 and 2000
DOE	To help to the ontologist in the process of building an ontology	Differential Semantic	Bachimont's method	NL text	Whole process	None	Bachimont 2000

Table 2. Summary of ontology learning tools from text (2/4)

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
KEA	To Summarize documents extracting keywords	Statistical approach ML techniques Lexical processing	Own method	NL text	Evaluation	WEKA ML Workbench	Jones and Paynter, 2002
LTG	To discover internal relations of texts in NL	Statistic Inference Linguistic technique	Own method	NL text	Whole process	Can be used to perform the knowledge acquisition to any other ontology development tool	Mikheev and Finch, 1997
MO'K Workbench	To learn concept taxonomy	Conceptual clustering	Own method	Tagged text	Whole process	Can be used to perform the knowledge acquisition to any other ontology development tool	Bisson et al. 2000
OntoLearn	To enrich a domain ontology	NLP techniques ML techniques	Missikoff and colleagues' method	NL text	Evaluation	None	Velardi et al., 2002 and 2001
Prométhée	Extraction and refinement of lexical-syntactic patterns	Learning from examples	Own method	Pattern bases	Whole process	Information not available in papers	Morin 1999, 1998
SOAT	Acquisition of relationships	Phrase-patterns	Own method	NL text	Information not available in papers	Information not available in papers	Wu et al 2002

Table 2. Summary of ontology learning tools from text (3/4)

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
SubWordNet Engineering Process	Build a Sub WordNet	NLP techniques Statistical approaches	Gupta and colleagues' approach	NL text	Whole process	Information not available in papers	Gupta et al., 2002
SVETLAN'	Build a concept hierarchy	Conceptual clustering	Own method	NL text	Validation	Information not available in papers	Chaelandar and Grau, 2000
TFIDF	To learn concepts and relation between them	Text-mining Statistical approach	Hybrid text-mining approach to acquire domain terms	NL text	Evaluation	SPPC NLP tool	Xu et al., 2002
TERMINAE	To build an initial ontology	Conceptual clustering	Own method	NL text	Validation	Information not available in papers	Biébow and Szulman 1999
Text-To-Onto	To find taxonomic and non-taxonomic relations	Statistical approach Pruning techniques Association rules	Kietz and colleagues' method	NL text Dictionaries Ontologies	Validations	KAON tool suite	Maedche and Volz, 2001
TextStorm and Clouds	To build a taxonomy	NLP techniques Linguistic hypothesis	Own method	NL text	Whole process	Information not available in papers	Oliveira et al., 2001

Table 2. Summary of ontology learning tools from text (4/4)

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
Welkin	To enrich automatically existing general purpose ontologies with new terms	Semantic Similarity measures	Alfonseca and Manandhar's method Hearst's approach	Domain corpus WordNet	Not necessary	None	Alfonseca and Rodríguez, 2002
WOLFIE	To learn a semantic lexicon	NLP techniques Statistical approach	Own method	Pre-processed corpus Examples	Validation	CHILL	Thompson and Mooney, 1997

From the *methodological perspective*, we conclude:

- It does not exist a detailed methodology or method that guides the ontology learning process from text. There are methods that provide general guidelines.
- The methods presented in this section are mainly based on natural language analysis techniques, and use a corpus that guide the overall process. Only Maedche and colleagues' work uses domain and general corpora to remove unspecific domain concepts from an existing ontology. The other ones only use domain documents to learn new concepts and relations.
- The most common ontology used by many methods is WordNet, which is used as initial ontology enriched with new concepts or relations.
- All these methods require the participation of an ontologist to evaluate the final ontology and the accuracy of the learning process.

From a *technological perspective*, we conclude that:

- Most of these tools perform NLP to extract linguistic and semantic knowledge from the corpus used for learning.
- The tools can be grouped in three main groups according with the technique followed to learn: the tools mainly based on conceptual clustering (ASIUM, MO'K, SVETLAN', TERMINAE), the tools based on statistical approaches (LTG, Text-To-Onto, TFIDF, WOLFIE, SubWordNet Engineering, KEA), and the tools based on linguistics and/or semantic approaches (Prométhée, Corporum-Ontobuilder, TextStorm, Welkin, OntoLearn, DOE, SOAT).
- It does not exist a fully automatic tool that carries out the learning process. Some tools are focused to help in the acquisition of lexical-semantic knowledge, others help to elicit concepts or relations from a pre-processed corpus with the help of the user, etc.
- There are neither tools that evaluate the accuracy of the learning process nor to compare different results obtained using different learning techniques.

4 Ontology Learning from dictionary

4.1 Introduction

This section shows different methods and tools for ontology learning from dictionary. Different methods and tools have been developed to achieve the goal of building an ontology based their performance on the use of a machine readable dictionary (MRD). However, some methods and tools can be used for other ontology learning approach, such as ontology learning from text. The reason to have included them here is that they base their functionality in using mainly a MRD.

For each method, we will present a general description including its goals and scope of the learning process, general steps used to learn, the source used for learning (if the method needs other type of sources in addiction to MRD), the main techniques applied in the process, the possibility of reusing other existing ontologies, the main goals looked for the method, the domain in which it has been applied and tested, if there are a tool associated, how the evaluation of the knowledge learnt is performed, a list of the most relevant ontologies built following it, the URL where more information about it can be found, and relevant bibliography. The methods presented in this section are: Jannink and Wiederhold's approach and Rigau and colleagues' method.

For each tool, we will present a general description including its main goals, the main techniques used by the tool in the learning process, the method followed, the user intervention in the process, the types of sources used by the method, the software architecture, the possibility of interoperate with other tools, the import and export facilities that the tool provides, the interface facilities, a URL where you can find more information, and relevant bibliography. The tools summarized in this section are: DODDLE and SEISD.

4.2 Methods for ontology learning from dictionary

In this section, we will summarize, in alphabetical order, the most relevant methods and approaches used for ontology learning from dictionary. The name of each method is the main reference in which the method or the approach has been described

4.2.1 Hearst's method

This method [Hearst, 1992] aims to acquire automatically hyponymy lexical relations from a corpus to build up a general domain thesaurus, using WordNet to verify and augment its performance. The process uses a set of predefined lexico-syntactic patterns that are easily recognizable. These patterns occur frequently and across text genre boundaries, and that indicate the lexical relation of interest. The method aims to discover these patterns and suggest that other lexical relations will also be acquirable in this way, apart from the initial set. All of them will be used to build up the thesaurus, and also can be useful for different purposes, such as lexicon augmentation or semantic relatedness information.

The method has proposed the following procedure to discover automatically new patterns:

1. *Decide on a lexical relation that is of interest.* In this case, this is a subset of the hyponymy relation.
2. *Gather a list of terms for which this relation is known to hold.* This list can be found automatically using this method, bootstrapping from patterns found by hand, or by bootstrapping from an existing lexicon or knowledge base.
3. *Find place in the corpus* where these expressions occur syntactically near one another and record the environment.
4. *Find the commonalities* among these environments and hypothesis, that common ones yield patterns that indicate the relation of interest.
5. Once a new pattern has been positively identified, use it to *gather more instances of the target relation* and go to the step 2.

To validate this acquisition method, the author has proposed to compare it with the information found in WordNet. In this comparison, three kinds of outcomes are possible. (1) To verify. If the two terms presented in the hyponymy relation are in WordNet, and if the relation between them is in the hierarchy, the thesaurus is verified. (2) To critique. If the two terms presented in the new relation are in WordNet, but the relation is not in the hierarchy, the thesaurus is critiqued, and a new set of hyponym connections is suggested to be added to WordNet. (3) To augment. If one or both terms presented in the new relation are not present, these nouns phrases and their relation are suggested as new entries to WordNet.

Main techniques used: patterns: linguistic patterns
Reuse other ontologies: WordNet
Source: text
Main goal: to create a thesaurus, and also to enrich WordNet with the same procedure with new lexical-syntactic relations
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: using WordNet to compare its results,
Relevant ontologies built following it: information not available in papers
URL: information not available in papers

References.

- Hearst M.A. (1992) *Automatic acquisition of Hyponyms from large text corpora.* In Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France, July 1992.

4.2.2 Jannink and Wiederhold's approach

The work presented here [Jannink and Wiederhold, 1999] aims to convert dictionary data to a graph structure to support the generation of domain or task ontology. The resulting text is tagged to mark the parts of the definitions, similar to the XML structure. According to the research purpose, only headwords and definitions having many-to-many relationships are considered. This results in a directed graph that has two properties: that each headword and definition grouping is a node; and each word in a definition node is an arc to the node having that headword. This research has been part of the Scalable Knowledge Composition (SKC) project that aims to develop a novel approach to resolve semantic heterogeneity in information systems. An ontology algebra has therefore been developed to represent the terminologies from distinct and typically autonomous domains. This research effort is funded by United States Air Force Office of Scientific Research (AFOSR).

The approach uses an algebraic extraction technique, described in detail in [Jannink, 1999], to generate the graph structure and to create the thesaurus entries for all words defined in the structure, including some stop words. The reason for using the dictionary as a structuring tool is that head words are distinguished terms from the definition text, which provides the extra information allowing for types of analysis that are not currently performed in traditional data mining and information retrieval.

The basic hypothesis for this work is that structural relationships between terms are relevant to their meaning. The general steps to extract these relationships are:

1. *Graph extraction from the dictionary*. Substantial manipulation is required to bring the dictionary data into a format ready for generating a graph, and only headwords and definitions having many-to-many relationships are considered.
2. *Application of PageRank Algorithm*. The dictionary graph contains both source and sink nodes: sources are words which are never used in other words' definitions; sinks are words whose definitions are not found in the dictionary.
3. *Relative the importance of the relation*. The idea is to identify the most important relations for a given individual node in the graph. A statistical approach is used for this purpose.

The output of the approach is a set of terms that are related by the strength of the associations in the arcs that they contain. These associations were computed according to local hierarchies of subsuming and specializing relationships, and the kinship relation relates sets of terms.

Main techniques used: statistical approach, PageRank Algorithm
Reuse other ontologies: no
Source: MRD
Main goal: build a taxonomy
Domain in which it has been applied: transport
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert who compares the output with WordNet
Relevant ontologies built following it: information not available in papers
URL: information not available in papers.
References

- Jannink, J. (1999). *Thesaurus Entry Extraction from an On-line Dictionary*. In Proceedings of Fusion '99, Sunnyvale CA, July 1999.
- Jannink, J. & Wiederhold, G. (1999). *Ontology maintenance with an algebraic methodology: A case study.* In. Proceedings of AAAI workshop on Ontology Management, July 1999.

4.2.3 Rigau and colleagues' method

This method consists of learning lexical ontologies from dictionaries. When using monolingual dictionaries, each definition is analysed in order to find a hyperonym of the word being defined (the so-called genus word). Afterwards, a Word-Sense Disambiguation (WSD) procedure is used on the genus word to discover which meaning it is used with. For instance, if the word "lily" is defined as "any liliaceous plant of the genus Lilium having showy pendulous flowers", the word "plant" is identified as the genus word. This word is then disambiguated, in order to choose the word sense from biology rather than the industrial meaning.

For this method, a new WSD procedure was developed based on conceptual distance. Furthermore, bilingual dictionaries and WordNet have been combined with several WSD techniques to infer bi-lingual ontologies.

This method was developed as part of the EuroWordNet project (1996-1999), [21], which aimed at developing lexical ontologies for several European languages.

The approach was applied to the generation of Spanish and Catalan versions of WordNet. The accuracy of the WSD procedures was measured independently, and the generated ontologies were also evaluated. The method has been applied to the automatic mapping of different versions of these ontologies (such as successive versions of WordNet between 1.5 and 1.7.1), and it has also been successfully applied to other languages such as Korean (Lee et al., 2000).

Main techniques used: analysis of dictionary definitions and word-sense disambiguation of a genus word.
Reuse of other ontologies: WordNet 1.5 (the original English version)
Source: Machine-readable monolingual dictionaries (The Spanish *Diccionario General Ilustrado de la Lengua Española*), bilingual Spanish-English Dictionaries, and the English version of WordNet
Main goal: to semi-automatically develop versions of WordNet for the Spanish language, and has also been applied to producing the Catalan version of WordNet
Domain in which it has been applied: monolingual and multi-lingual general-purpose ontologies (WordNet / EuroWordNet)
Tool associated: SEISD

[21] http://www.illc.uva.nl/EuroWordNet/

Evaluation of the knowledge learnt: accuracy of the WSD procedures; the accuracy of the links obtained from bilingual dictionaries was also measured
Relevant ontologies built following it: Catalan and Spanish versions of EuroWordNet
URL: http://www.lsi.upc.es/~rigau/
References

- Rigau G. (1998). Automatic Acquisition of Lexical Knowledge from MRDs, Ph.D. Thesis, Departament de Llenguatges i Sistemes Informàtics.-- Universitat Politècnica de Catalunya.
- Rigau G., Rodríguez H. and Agirre E (1998). *Building Accurate Semantic Taxonomies from Monolingual MRDs*. Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL'98. Montreal, Canada. 1998. Online paper in: http://www.lsi.upc.es/~rigau/
- Lee C., Lee G., and Yun S. J. (2000). Automatic WordNet Mapping Using Word Sense Disambiguation, 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000), Hong Kong.

4.3 Tools for ontology learning from dictionary

4.3.1 SEISD (Sistema d'Extracció dÌnformació Semàntica de Diccionaris / System for Extraction of Semantic Information from Dictionaries)

SEISD (Rigau, 1998) automatically learns monolingual and bilingual lexical ontologies from either semi-structured sources (dictionaries) or from a combination of dictionaries and existing ontologies.

The idea is to analyse the dictionary definitions in order to learn relationships between the words defined, such as hyperonym or part-of relationships. It involves two steps:

1. Identification of the genus term in the definition (performed using a specialised grammar)
2. Word-Sense Disambiguation of that term.

This tool was developed as part of the EuroWordNet[22] project (1996-1999), which was aimed at the development of lexical ontologies for several European languages.

This method has been applied in the construction of the Spanish and Catalan versions of WordNet. The accuracy of the WSD procedures was measured independently, and the generated ontologies were also evaluated. The method has been applied to the automatic mapping of different versions of these ontologies (such as successive versions of WordNet between 1.5 and 1.7.1).

Goal and scope of the tool: automatic acquisition of massive lexical information from monolingual and bi-lingual dictionaries
Learning technique used by the tool: uses a specialised grammar for the identification of the genus word in the dictionary definitions, and a combination of heuristics and the conceptual distance technique for Word-Sense disambiguation of the genus term. The conceptual distance procedure roughly consists of choosing the meaning of the word that is most closely related to those words that appear in its context, within WordNet
Method followed for ontology learning: Rigau and colleagues' method
User/Expert intervention in the process: apparently, it can be executed with no human intervention, as the analysis of the definitions and the WSD procedures can be fully automatic
Types of sources used by the method: Machine-Readable Monolingual Dictionaries (The Spanish *Diccionario General Ilustrado de la Lengua Española*), Bi-lingual Spanish-English Dictionaries, and the English version of WordNet

[22] http://www.illc.uva.nl/EuroWordNet/

SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: as seen in the diagrams in (Rigau, 1998), there seem to be internal-use tools for visualisation and validation of results
URL where you can find information about the method or tool: http://www.lsi.upc.es/~rigau/
References.

- Rigau G. (1998). *Automatic Acquisition of Lexical Knowledge from MRDs*, Ph.D. Thesis, Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.

4.3.2 DODDLE: Domain ontology rapid development environment

DODDLE [Yamaguchi, 1999] aims to construct domain ontologies, in particular, a hierarchically structured set of domain terms without concept definitions, reusing a machine readable dictionary (MDR) and making it adjusted to specific domains. The tool deals with concept drift, which means that the senses of concepts change depending on application domains. For this purpose, two strategies have been following: match result analysis, and trimmed result analysis, and both try to identify which part may stay or should be moved from the initial ontology, analysing spell match results between given input domain terms and a MDR. The tool has been proved in the legal domain. An overview of the overall architecture is presented in the figure 8. Domain terms are supposed to be already identified and given to the tool as input. Since DODDLE just generates a hierarchically structured set of domain terms, it supports a user in structuring terms into categories and giving names to the categories.

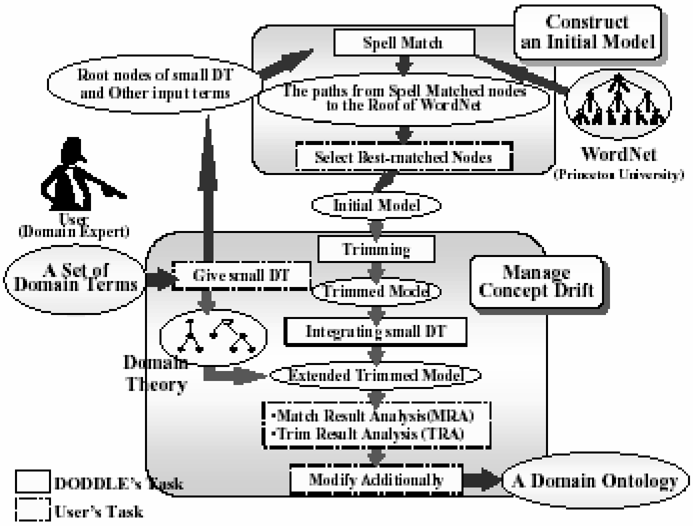


Figure 8. DODDLE overview.

In order to analyse the concept drift between a MRD and a domain ontology, there are involved two main activities. The first one is to build an initial model from a MRD, extracting information relevant to given domain terms from a MRD. And the second one is to manage concept drift, making an initial model adjust to the domain. The process has the following main steps using WordNet as a MRD:

1. *The user gives a non-structured set of domain terms as input*. The user can also give small trees including domain concepts.

2. *Integrate these trees into the initial model*. The spell match has been done between just root nodes and a MRD. So DODDLE does spell match between other all input terms except inner and leaf nodes of the provided trees and a MRD. The spell match links these terms to A MRD. The result of this process is a hierarchically structured set of all the nodes on the path from these terms to the root of a MRD.
3. *Select the correct sense for each concept.* Because a matched node from a MRD sometimes has one or more senses, DODDLE supports the user in doing the selection of the sense by showing the user a detailed descriptions on each sense and where each sense is put in the concept hierarchy structure from a MRD. The output of the process is an initial model.
4. *Manage the concept drift*. Using the previous initial model, DODDLE tries to manage the concept drift while removing unnecessary internal terms in the initial model producing a trimmed model, integrating small trees into the previous trimmed model, and finding out which part should be drifted in the trimmed model.
5. *User modifications*. After moving the unnecessary parts from the ontology and doing additional modifications by the user, the final ontology is shown as a hierarchical structure of domain concepts.

Goal and scope of the tool: build an ontology using a MRD and manage concept drift
Learning technique used by the tool: spell match, trimmed
Method followed for ontology learning: own method
User/Expert intervention in the process: information not available in papers
Types of sources used by the method: MRD, trees of concept
SW Architecture: information not available in papers
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: information not available in papers
User interface facilities: information not available in papers
URL where you can find information about the method or tool: information not available in papers
References.

- Yamaguchi, T. (1999). *Constructing domain ontologies based on concept drift analysis.* In., Proceedings of IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, in conjunction with the Sixteenth International Joint Conference on Artificial Intelligence, August, Stockholm, Sweden.

4.4 Conclusions

In this section, we have presented an overview of the most important methods and tools used to achieve the ontology learning process from text. Tables 3 and 4 summarize the ontology learning methods and tools from dictionary according to the description criteria identified before. We conclude that:

- The performance of the methods and tools are based on the use of semantic and linguistic analysis, such techniques are used to elicit new concepts or relations from a dictionary.
- The tools analysed perform mainly syntactic analysis and need a user to validate their results.
- WordNet is commonly used as dictionary for all these approaches.

Table 3. Summary of ontology learning methods from dictionary

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Hearst's method	To create a thesaurus, and also to enrich WordNet with new lexical-syntactic relations	Linguistic patterns	WordNet	Text WordNet	Information not available in papers	Comparing the relations discovered with WordNet	Hearst 1992
Rigau and colleagues' method	To develop semi-automatically WordNet versions	Word-sense disambiguation Analysis dictionary definitions	WordNet	MRD: monolingual and bilingual	SEISD	WSD procedures	Rigau 1998
Jannink and Wiederhold's approach	To build a taxonomy	Statistical approach PageRank Algorithm	No	MRD	Information not available in papers	Expert comparing the output with WordNet	Jannink 1999

Table 4. Summary of ontology learning tools from dictionary

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
SEID	To acquire automatically lexical information from monolingual dictionary	Specialized grammars Semantic Distance	Rigau and colleagues' method	MRD: monolingual and bilingual	Validation	Information not available in papers	Rigau, 1998
DOODLE	To build an ontology	Spell match Trimmed	Own	MRD Initial taxonomies	Validation	Information not available in papers	Yamaguchi, 1999

5 Ontology learning from knowledge base

5.1 Introduction

This section presents a method for ontology learning from knowledge base. We did not find on the literature tools that give technological support to that process.

As in the previous sections, for the method we provide a general description (including its goals and scope), general steps used to learn, the source used for learning, the main techniques applied in the process, the possibility of reusing other existing ontologies, the main goals looked for the method, the domain in which it has been applied and tested, if there are a tool associated, how the evaluation of the knowledge learnt is performed, a list of the most relevant ontologies built following it, the URL where more information about it can be found, and relevant bibliography.

5.2 Methods for ontology learning from knowledge base

This section summarizes a method for ontology learning from knowledge base.

5.2.1 Suryanto and Compton's approach

This approach [Suryanto and Compton, 2001 and 2000] aims to elicit an ontology from a knowledge base of rules. Given a knowledge based system built with ripple down rules [Compton and Jansen, 1990] (a tree structure where the nodes are rules), the authors proposes an algorithm to extract the class taxonomy where a class is a set of different rule paths giving the same conclusion, and a rule path for node *n* consists of all conditions from all predecessors rules plus conditions of the particular rule of node *n*. The experimental results are based on a large real-world medical RDR knowledge based system.

The approach takes the initial tree and creates a set of classes. From this group of classes, the approach aims to discover class relations between them. Three types of relations have been considered: *subsumption, mutual-exclusivity* and *similarity.* The central idea of the technique is to group all rules for each class and compute a quantitative measurement for each relation between every pair of classes. This quantitative measure provides the confidence to whether these relations exist. In this way, one class (A) subsumes other class (B) with certain confidence measure, if the class (A) only exists when the class (B) exists, but not the other way around. The mutual-exclusivity relation appears if the class (A) and the class (B) never occur together. Finally, the class (A) and the class (B) are similar when both classes have the same conditions (the same path but different conclusions). With the set of classes created in this process, and these types of relations, the class taxonomy is built up. An expert evaluates the whole process.

Main techniques used: statistical measure, to estimate the existence of a relationship between a pair of classes
Reuse other ontologies: not proposed
Source: Rule knowledge bases.
Main goal: elicit a taxonomy
Domain in which it has been applied: medical pathology
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: http://www.pks.com.au/
References

- Suryanto H, Compton P. (2001) *Discovery of Ontologies from Knowledge Bases.* Proceedings of the First International Conference on Knowledge Capture, Eds. *Yolanda Gil; Mark Musen; Jude Shavlik*, Victoria, British Columbia Canada, 21-23 Oct. 2001, The Association for Computing Machinery, New York, USA, pp171-178

- Suryanto H, Compton P. (2000) *Learning classification taxonomies from a classification knowledge based system.* Proceeding of the First Workshop on Ontology Learning in conjunction with ECAI-2000, Eds. Steffen Staab, Alexander Maedche, Claire Neddellec, Peter Wierner-Hastings, Berlin Germany, 22 Aug. 2000, Berlin, pp1-6
- Compton P. and Jansen A. (1990) *A philosophical basis for knowledge acquisition.* Knowledge acquisition, 1990. p. 241-257.

5.3 Tools for ontology learning from knowledge base

We did not find any relevant tool to perform the ontology learning process using a knowledge base at the time when this deliverable was written.

5.4 Conclusions

We conclude this section saying that this approach for ontology learning is not enough explored by the ontology community.

6 Ontology learning from semi-structured schemata

6.1 Introduction

This section shows different methods and tools for ontology learning from semi-structured schemata. The methods described in this section (following the same criteria than in the previous sections), are: Deitel and colleagues' approach, Doan and colleagues' approach, Papatheodorou and colleagues' method, and Volz and colleagues' approach. In the tools' section, we will describe OntoBuilder.

6.2 Methods for ontology learning from semi-structured schemata

In this section, we present a summary of the most relevant methods and approaches for building an ontology from a semi-structured schema. The name of each method is the main reference in which the method or the approach has been described.

6.2.1 Deitel and colleagues' approach

Deitel and colleagues [Deitel et al, 2001] present an approach for learning ontologies from RDF annotations of Web resources. The general aim of the approach is to learn from the whole RDF graph new domain specific concepts to enrich the ontology from which the RDF annotations participate. To perform the learning process, a particular approach of concept formation is adopted, considering an ontology as a concept hierarchy, where each concept is defined in extension by a cluster of resources and in intension by the most specific common description of these resources. A resource description is a RDF subgraph containing all resources reachable from the considered resource through properties. This approach leads to the systematic generation of all possible clusters of descriptions from the whole RDF graph incrementing the length of the description associated to each particular concept in the source graph.

The approach systematically considers all concepts covering a set of nodes (resources) of the whole RDF graph, being defined then in extension as a set of resources. To extract the description of a resource from the graph, this approach follows a resource extraction criterion, called *description of length n* of a resource, that is the largest connected subgraph in the whole RDF graph containing all possible paths of length smaller or equal to *n,* starting from or ending to the considered resource. Given the whole graph and the extraction criteria, the approach aims to associate to each set of resource descriptions the hierarchy of the concepts whose extensions correspond to all possible resource clusters.

The proposed steps used for building a hierarchy based on resource description are the following:

1. *Start the process with the extraction resource description of length one*, and repeat the process incrementing the length until covered the maximum path in the graph.
2. *Extraction of resource descriptions of length one from the whole RDF graph*. These descriptions form a group of RDF triple, composed by the resource, its properties and their values, beginning or ending by a particular resource.
3. *Iterative generalization of all the possible pairs of triples*. The generalization of two triples is the most specific triples subsuming them. This is based on the ontological knowledge represented in the RDF Schema relative to the considered RDF annotations.
4. *Construction of the intensions of length one and the triples sharing a same extension are grouped together*. An intension may include redundant triples, one being more general than another. It is cleaned up by deleting triples subsuming another one.

5. *Build the generalization hierarchy based on the inclusion relations between the node extensions.* Several nodes may share the same intension. In this case, the node to be preserved corresponds to the largest extension.

6. *Repeat the process incrementing the length of the resource description to be extracted.*

Main techniques used: techniques based on the Graph Theory
Reuse other ontologies: the ontology used to annotate the web page
Source: RDF graph generated from the ontology used to annotate
Main goal: enrich the ontology with new concepts and relations
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: http://mondeca-publishing.com/s/anonymous/title11884.html, web page of the European IST Comma project where this approach has been partially developed
References.

- Deitel A., Faron C., and Dieng R. (2001) *Learning ontologies from RDF annotations.* In Proceedings of the IJCAI Workshop in Ontology Learning, Seattle, 2001.

6.2.2 Doan and colleagues' approach

The approach presented here [Doan et al., 2000] aims to semi-automatically learn mappings between source schemas and a mediated schema using machine learning. This approach can be applied to elicit knowledge from semi-structured sources. A *Learning Source Descriptions (LSD)* system gives support for the overall process. The underlying idea is that, after a set of data sources have been manually mapped to a mediated schema, the system should be able to glean significant information from these mappings and to successfully propose mappings for subsequent data sources.

The approach has two main phases:

1. *Learning phase.* The system is trained by manually matching schema elements from the source with the elements of the mediated schema. The system can learn from the given examples, from the name of the elements provided as examples, from their properties, and from the proximity of elements. In other words, it can learn from the types of information that a learner can exploit, such as names, formats, word frequencies, positions, and characteristics of the value distributions for the elements given as examples. The system creates recognisers for each of these elements, to be used as a pattern for the subsequent process. There are different learner modules that recognize certain kinds of patterns.

2. *Classification phase.* Once the system has been trained, and after combining the information produced in the previous phase, the system is ready to perform the schema matching process on new sources.

Main techniques used: machine learning: pattern recognition.
Reuse other ontologies: not proposed
Source: text
Main goal: learn mappings
Domain in which it has been applied: housing market
Tool associated: Learning Source Description (LSD)
Evaluation of the knowledge learnt: by an expert.
Relevant ontologies built following it: information not available in papers
URL: information not available in papers

References

- Doan A., Domingos P. and Levy A. (2000). *Learning Source Descriptions for Data Integration.* Proceedings of the Third International Workshop on the Web and Databases (pp. 81-86), 2000. Dallas, TX: ACM SIGMOD.

6.2.3 Papatheodorou and colleagues' method

The method described here [Papatheodorou et al., 2002] aims to build taxonomies using a data mining approach, called cluster mining [Perkowitz and Etzioni, 1998], from domain repositories written in XML or RDF. The cluster mining approach firstly tries to group similar metadata files into a cluster and then, processing this cluster, extracts a controlled vocabulary to be used for building the taxonomy. This work has been developed within the UNIVERSAL[23] European Union funded project.

The steps proposed by the method are the following:

1. *Data collection and pre-processing*. The main objective is the selection of appropriate keywords from the metadata files that will enable the discovery of similarities between the documents. For this purpose, words such as articles, prepositions and conjunctions are dropped. Then the word roots are extracted using WordNet.
2. *Pattern discovery*. This step aims to build clusters of documents with similar content. For this reason, the cluster mining approach has been used to discover similar documents, to build clusters, and to extract the keywords that represent the content of the document. This activity could be performed using other existing taxonomies as a pre-classification structure.
3. *Pattern post-processing and evaluation*. In this step, the keywords extracted in the previous step are examined and measured using a statistical approach. Then those keywords that best represent the content of the cluster are selected. These keywords constitute a proto-typical model for each cluster and provide the vocabulary necessary to form concepts and to find the relations among them using different measures calculated before.

Main techniques used: cluster mining approach
Reuse other ontologies: not proposed
Source: semi-structured text
Main goal: build a domain taxonomy
Domain in which it has been applied: education
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by an expert
Relevant ontologies built following it: information not available in papers
URL: http://www.educanext.org/
References

- Papatheodorou, C., Vassiliou, A., Simon, B. (2002): *Discovery of Ontologies for Learning Resources Using Word-based Clustering*. ED-MEDIA 2002. Copyright by AACE. Reprinted from the ED-MEDIA 2002 Proceedings, August 2002 with permission of AACE, Denver, USA, August. http://www.wu-wien.ac.at/usr/wi/bsimon/publikationen/EDMEDIA2002.pdf
- Perkowitz M., and Etzioni O, (1998). *Adaptive Web Sites: Automatically synthesizing Web pages.* Proceedings of the Fifteen National Conference in Artificial Intelligence (AAAI 98), AAAI Press.

[23] http://www.ist-universal.org/

6.2.4 Volz and colleagues' approach

This approach [Volz et al., 2003], also called lifting process, tries to capture the semantics of an XML schema by translating non-terminal and terminal symbols into concepts and roles of an ontology. The translation process is performed by means of applying sequentially a set of rules. The sources that can be used are XML, XML Schema or DTD (which firstly needs to be translated to XML Schema).

The approach has two main steps to carry out the process:

1. The schemas are transformed to a regular tree grammar. A description of this type of regular grammars can be found in [Murata et al., 2001]. A regular tree grammar is a 4-tuple compound by: a finite set of non-terminals, a finite set of terminals, a set of start symbols, and a finite set of production rules. Regular grammar allows to the mapping process being independent of the language using to codify the original schema and deleting all unnecessary information of the schema.
2. To capture the semantics of the XML Schema. Then the lifting process is carried out. This process translates non-terminals and terminal to concepts and roles in the ontology. It is a mapping process based on using rules sequentially that transform elements of the tree grammar in concepts or roles in the ontology.

This approach does not consider the integrity constraints of the XML Schema when translating into regular tree grammars.

Main techniques used: mapping techniques
Reuse other ontologies: no proposed
Source: XML schemas, XML, DTD
Main goal: elicit a light ontology
Domain in which it has been applied: information not available in papers
Tool associated: OntoLiFT (into KAON Workbench)
Evaluation of the knowledge learnt: no proposed
Relevant ontologies built following it: information not available in papers
URL: http://www.aifb.uni-karlsruhe.de/WBS/rvo/raphael-bib.html#wonderweb-D11
References

- Volz, R.; Oberle, D.; Staab, S.; Studer, R. (2003) OntoLiFT Prototype. IST Project 2001-33052 WonderWeb Deliverable 11.
- Murata, M.; Lee, D; Mani. M. (2001) Taxonomy of XML Schema Languages using Formal Language Theory. In Extreme Markup Languages, Montreal, Canada, Aug. 2001. http://www.cs.ucla.edu/dongwon/paper/

6.3 Tools for ontology learning from semi-structured schemata

We present now the most relevant tools used for building an ontology from a semi-structured schema.

6.3.1 OntoBuilder tool

OntoBuilder [Modica et al., 2001] tool helps users in the ontology creation process using as source semi-structured data coded in XML or in HTML. The modular architecture of the system is shown in the figure 9. There are three main modules in this system: *the user interaction module, the observer module, and the Ontology Modelling.*

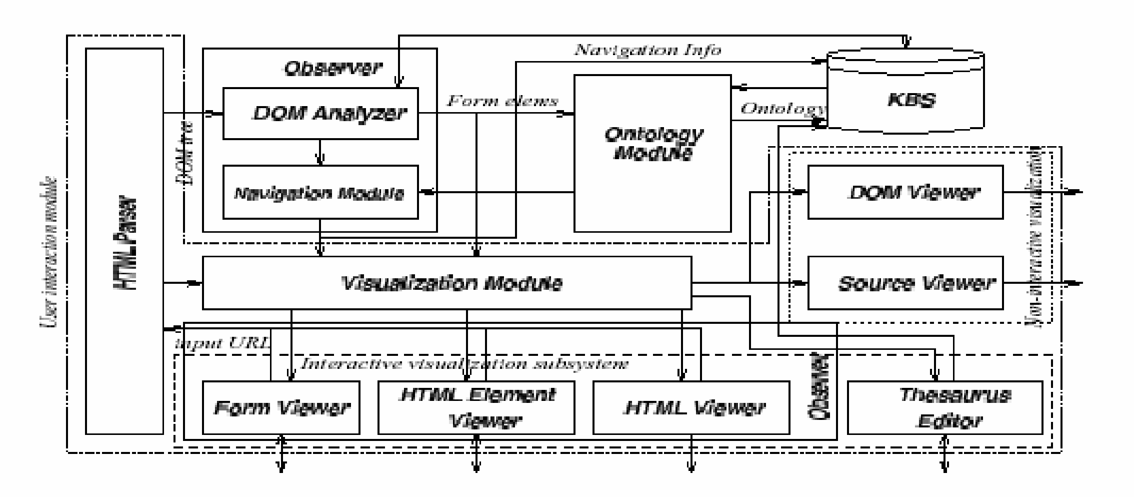


Figure 9. Architecture of OntoBuilder

The process followed to build an ontology has two phases: training and adaptation phases. In the training phase, an initial domain ontology is built using the data provided by the user. The adaptation phase aims to refine and generalize the initial ontology. The user suggests browsing other web sites that contain relevant information for the domain. From each site, a candidate ontology is extracted and merged with the existing ontology. To perform this activity, a thesaurus can be used.

The user interacts with the system through the user interaction module, accesses a set of Web sites of interest through this module, and provides feedback to the system. The system also maintains a user editable thesaurus that it can be used for the ontology creation process.

The performance of the tool has been tested in the car rental domain, using web pages of the most important companies of this domain.

Goal and scope of the tool: to refine an existing domain ontology, adding new concepts by a match process
Learning technique used by the tool: statistical approach
Method followed for ontology learning: own method
User/Expert intervention in the process: whole process
Types of sources used by the method: semi-structure data
SW Architecture: client-server
Interoperability with other tools: information not available in papers
Import and export facilities to ontology languages: import XML and HTML, and the ontology is exported to XML
User interface facilities: Visual Editor, Thesaurus maintenance editor
URL where you can find information about the method or tool: http://www.cs.msstate.edu/~gmodica/Education/OntoBuilder/
References.

- Modica, G., Gal, A. and Jamil, H. M. (2001); *The Use of Machine-Generated Ontologies in Dynamic Information Seeking.* In the Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS 2001), Springer-Verlag LNCS series, September 5-7, 2001, Trento, Italy.

6.4 Conclusions

In this section, we have presented an overview of the most important methods and tools used to achieve the ontology learning process from semi-structured data. In the next tables, we show a summary of all the methods and tools that have been described in this section.

We conclude:

- The technique used for extracting ontological knowledge from semi-structured sources comes from learning mappings to use clustering approaches or patterns recognitions. Otherwise, the evaluation of the final results is performed by the user in all of them.
- The approach for learning ontologies from RDF annotations [Deitel et al., 2001] is the unique one that uses an existing ontology to enrich it with new concepts by means of using all the instances annotated with the ontology. The others do not propose reusing or enriching an existing ontology.
- The OntoBuilder tool [Modica et al., 2001] proposes to use an existing ontology to be refined and enriched with new concepts extracted from a semi-structured source and needs to be helped by a user during the whole learning process..

Table 5. Summary of ontology learning methods from semi-structured data.

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Deitel and colleagues' approach	To enrich an ontology with new concepts and relations	Graph Theory	Yes	RDF graph generated from the ontology	Information not available in papers	Expert	Deitel et al., 2001
Doan and colleagues approach	To learn mappings between the source and the target schema	ML techniques Pattern recognitions	No	Text	Information not available in papers	Expert	Doan et al., 2000
Papatheodorou and colleagues' method	To build a domain taxonomy	Clustering mining approach	No	Semi-structured text	Information not available in papers	Expert	Papatheodorou et al., 2002
Volz and colleagues' approach	To elicit a taxonomy	Mapping approach	No	XML Schemas DTD XML	OntoLiFT (KAON)	No	Volz et al., 2003

Table 6. Summary of ontology learning tools from semi-structured data.

Name	Goal and scope	Learning technique	Method followed to learn	Sources	User intervention	Interoperability	Bibliography
OntoBuilder	To refine an existing ontology, adding new concepts	Statistical approach Matching process	Own method, based on matching process	Semi-structured text	Whole process	Information not available in papers	Modica et al., 2001

7 Ontology learning from relational schemata

7.1 Introduction

In this section we describe different methods and approaches that allow the extraction of ontologies from database schemas.

7.2 Methods for ontology learning from relational schemata

This section summarizes the most relevant methods used for ontology learning from relational schemata in alphabetical order. The name of each method is the main reference, in which the method or the approach has been described.

7.2.1 Johannesson's method

This method [Johannesson, 1994] aims to translate a relational model into a conceptual model with the objective that the schema produced has the same information capacity as the original schema. The conceptual model used is a formalization of an extended Entity-Relation model, which includes subtypes. The method starts transforming the relational schemas into a form appropriate for identifying object structures. Certain cycles of inclusion dependencies are removed, and certain relation schemas are split. After the initial transformations, the relational model is mapped into a conceptual schema. Each relation model gives rise to an object type, and the inclusion dependencies give rise to either attributes or generalization constraints, depending on the structure of the keys in each relation schema. The iterations with the user are needed during the translation process. For each candidate key, a user must decide whether it corresponds to an object type of its own, and for each inclusion dependency where both sides are keys, a user must decide whether it corresponds to an attribute or a generalization constraint.

To understand correctly how the method works, it is necessary to explain the proposed transformation in which the method bases its functionality. Four different transformations have been proposed: *candidate key splitting (*occurs when a relation scheme in third normal form corresponds to several object types), *inclusion dependency splitting (*when a single relation corresponds to several objects types), *folding (*when several relation schemes correspond to a single object type), and *schema mapping (*to map a relational scheme into an object type)

The method proposes the following steps:

1. *Apply the candidate key splitting transformation* repeatedly to selected candidate keys in the relational schema. This will result in a sequence of relational schemas, where each one is obtained by applying the candidate key splitting transformation to the previous one in the sequence. The last one will be the new schema to use in the next step.
2. *Apply the inclusion dependency splitting transformation* repeatedly to the inclusion dependencies in the schema obtained in the previous step which are neither key based nor subsumed until no such inclusion dependencies remain.
3. *Apply the folding transformation* repeatedly to all cycles of generalization indicating inclusion dependencies of the relational schema obtained at the end of the previous step until all such cycles are removed. This will result in a sequence of relational schemas where each one is obtained by applying the folding transformation.
4. *Apply the schema mapping* to the relational schema obtained at the end of the previous step, and the conceptual schema is established. The final revision is performed by a user who decides the correction of the final conceptual schema.

Main techniques used: mapping techniques
Reuse other ontologies: not proposed

Source: relational schemas
Main goal: maps the relational schema with a conceptual schema
Domain in which it has been applied: information not available in papers
Tool associated: information not available in papers
Evaluation of the knowledge learnt: by the user
Relevant ontologies built following it: information not available in papers
URL: information not available in papers
References

- Johannesson P. (1994) *A Method for Transforming Relational Schemas into Conceptual Schemas.* In *10th International Conference on Data Engineering*, Ed. M. Rusinkiewicz, pp. 115 - 122, Houston, IEEE Press, 1994.

7.2.2 Kashyap's method

The method presented here [Kashyap, 1999] has been developed within the InfoSleuth[24] research project at MCC (Microelectronics and Computer Technology Corporation). One of the most important goals for this project is to develop technologies that operate on heterogeneous information sources in a dynamic environment.

The fundamental premise of building a domain ontology from database schemas is that the knowledge specific to the domain is embedded in the data and the schemas of the selected databases. The method uses the database schemas to build an ontology that will then be refined using a collection of queries that are of interest to the database users. The process is interactive, in the sense that the expert is involved in the process of deciding which entities, attributes and relationships are important for the domain ontology. It is iterative in the sense that the process will be repeated as many times as necessary.

The process has *two stages*. In the *first one*, the database schemas are analysed in detail to determine keys, foreign keys, and inclusion dependences. As a result of this process a new database schema is created, and by means of reverse engineering techniques, it is content is mapped into the new ontology. In the *second stage*, the ontology constructed from the database schemas has to be refined to better reflect the information needs of the user and can be used to refine the ontology. The overall process is summarized in the figure 10.

[24] http://www.argreenhouse.com/InfoSleuth/index.shtml

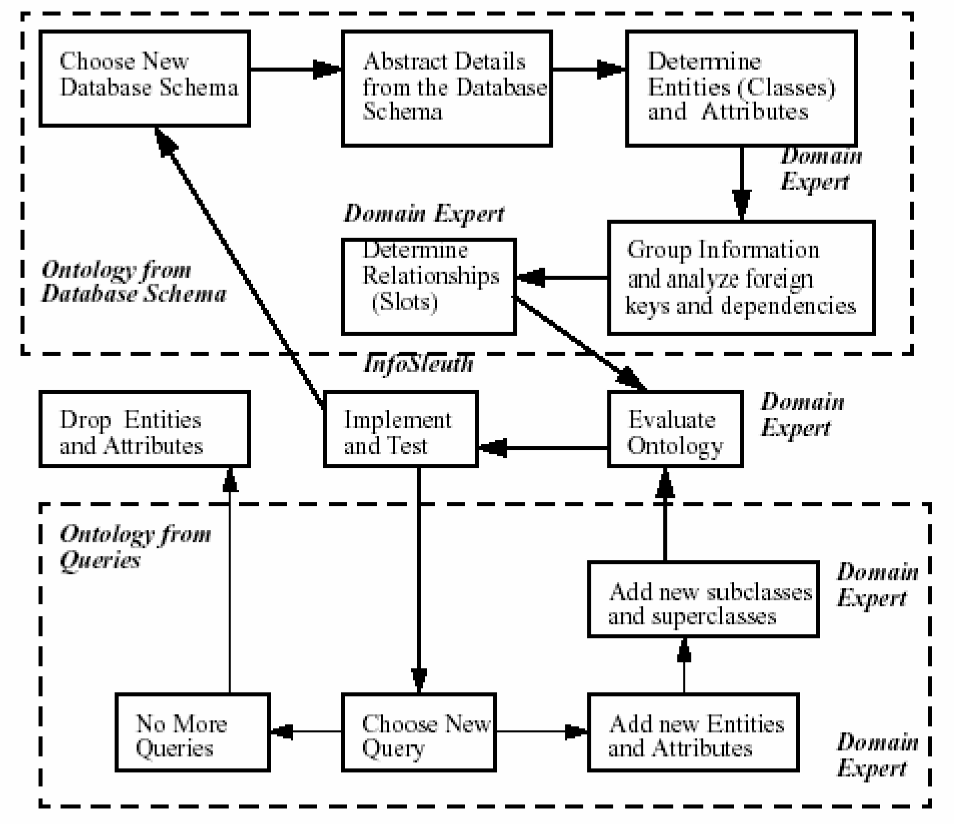

Choose New Database Schema
Abstract Details from the Database Schema
Determine Entities (Classes) and Attributes
Domain Expert
Domain Expert
Ontology from Database Schema
Determine Relationships (Slots)
Group Information and analyze foreign keys and dependencies
InfoSleuth
Drop Entities and Attributes
Implement and Test
Evaluate Ontology
Domain Expert
Ontology from Queries
Add new subclasses and superclasses
Domain Expert
No More Queries
Choose New Query
Add new Entities and Attributes
Domain Expert

Figure 10. Iterative process for building an ontology from database schemas,

The approach has the following steps to perform the process of building an ontology from database schemas:

1. *Abstraction of details*. The goal of this step is to abstract out irrelevant details and determine important information inside the databases. There are two types of abstraction: *abstraction of details related to the data organization and abstraction of details related to local keys.*
2. *Grouping information in multiple tables*. In many database schemas, information about a particular entity is spread across multiple tables. To group this information together the assumption used is that if the primary key of a table appears as a foreign key in another table.
3. *Identified relationships*. To perform this activity, the foreign keys and object identifiers are used to infer relationships between entities and at this point, a new entity-relation schema containing only the most relevant information for the target domain. The taxonomy of the underlying ontology is elicited by means of reverse engineering.
4. *Incorporation of concepts into the ontology suggested by new database schema*. The content of the new database schema, created by means of the abstraction process described in the previous steps, induces the addition of new concepts into the domain ontology. This process is carried out by comparing the mappings established between the concepts in the ontology and entities in the database. If there are some entities that have not been linked with the concepts in the ontology, and the user considers that they are important, then a new concept is created for each important entity.
5. *Defining mappings between elements of the ontology and the underlying database schema.*
6. *Refining the ontology based on user queries.* Based on the user queries, it is possible to add or delete attributes corresponding to an entity and relationships between entities, and to create new entities as a subclass or super-class of already existing entities in the database.

The approach has been used to build the EDEN ontology based on different environmental databases such as CERCLIS3 [Cerclis, 1999], ITT [Itt, 1999], ERPIMS [Erpims, 1999] and HAZDAT [Hazdat, 1999].

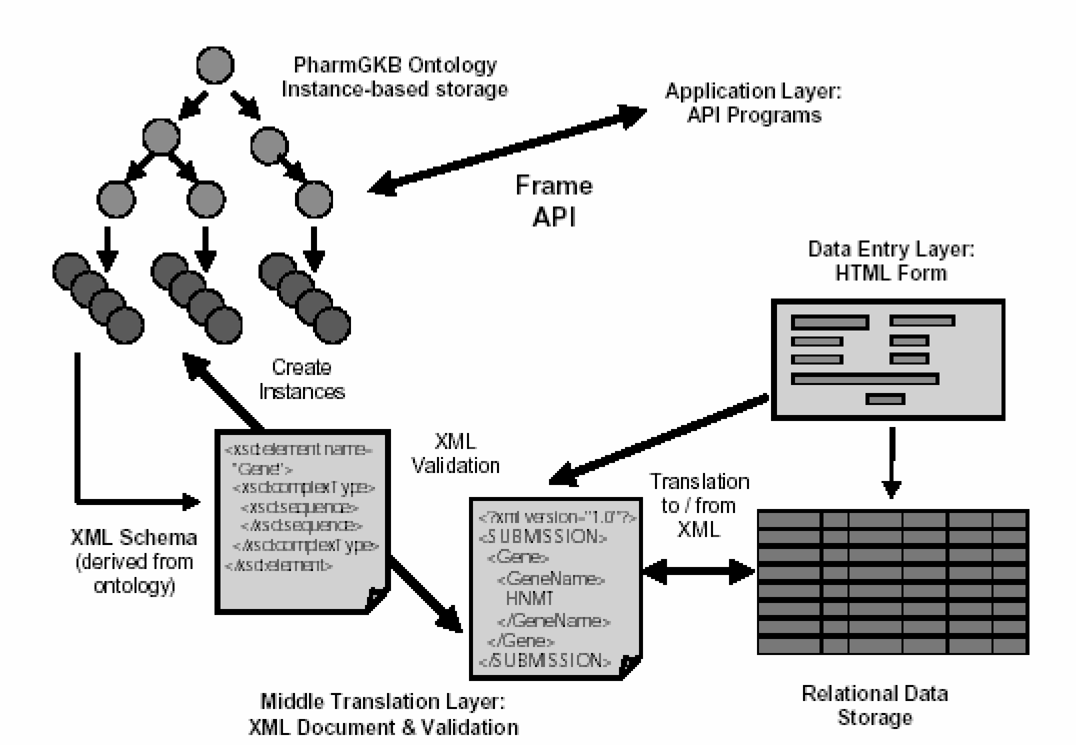
Main techniques used: mappings, reverse engineering
Reuse other ontologies: information not available in papers
Source: schemas of domain specific databases.
Main goal: create and refine an ontology from domain databases
Domain in which it has been applied: environmental information
Tool associated: EDEN system
Evaluation of the knowledge learnt: user validates the whole process
Relevant ontologies built following it: EDEN ontology
URL: information not available in papers.
References

- Kashyap, V. (1999). Design and Creation of Ontologies for Environmental Information Retrieval. Twelfth Workshop on Knowledge Acquisition, Modelling and Management Voyager Inn, Banff, Alberta, Canada. October, 1999.
- Cerclis. (1999). Superfund Data: The Comprehensive Environmental Response, Compensation, and Liability Information System (CERCLIS[25])
- Itt. (1999). Innovative Technologies Treatment Database, Environmental Protection Agency
- Edr. (1999). Environmental Data Registry, http://www.epa.gov/edr
- Erpims. (1999). Environmental Resources Program Information Management System, http://www.resdyn.com/erpims

7.2.3 Rubin and colleagues' approach

This approach [Rubin et al., 2002] proposes to automate the process of filling the instances and their attributes' values of an ontology using the data extracted from external relational sources. This method uses a declarative interface between the ontology and the data source, modelled in the ontology and implemented in XML schema. The process allows the automatization of updating the links between the ontology and data acquisition when the ontology changes. The method has been tested in a pharmacogenetics domain, using the PharmGKB ontology and linking this ontology with data acquisition from a relational model of genetic sequence data. The approach needs several components (see figure 10): *an ontology* (containing the domain concepts and the relations among them), *the XML schema* (is the interface between data acquisition and the ontology), and an *XML translator* (to convert external incoming relational data into XML when it is necessary).

[25] http://www.epa.gov/enviro/html/cerclis/cerclis_overview.html



```
<xsd:element name=
 "Gene">
 <xsd:complexType>
  <xsd:sequence>
  </xsd:sequence>
 </xsd:complexType>
</xsd:element>
```

```
<?xml version="1.0"?>
<SUBMISSION>
 <Gene>
  <GeneName>
   HNMT
  </GeneName>
 </Gene>
</SUBMISSION>
```

Figure 11. Model for data acquisition in PharmGKB

The proposed steps are:

1. *Create the ontology model for the domain.*
2. *Creating the XML Schema*. In order to generate an XML schema from the ontology, there must be a convention for naming and organizing the XML elements and attributes. Once the ontology is built and the constraints on data values are declared, the XML schema is sufficiently determined, and it can be written directly from the ontology.
3. *Data acquisition*. Data acquired from external relational data sources must be put into an XML document that uses the syntax specified by XML schema. This is a direct mapping from columns in a relational table to the appropriate elements in the XML schema.
4. *Ontology evolution and propagating changes.* The approach proposes to use two kinds of slots in the ontology design for automating of updating the XML schema. One of these is the XML schema, and the other is an administrative slot. If the change in ontology structure affects only to the administrative slot, then there will be no change in the XML schema or data acquisition. In other case, a new XML schema must be created with the same procedure than in the step 2.

Main techniques used: mappings
Reuse other ontologies: PharmGKB ontology[26]
Source: databases
Main goal: create the instances
Domain in which it has been applied:
Tool associated: has been used Protégé
Evaluation of the knowledge learnt: not proposed
Relevant ontologies built following it: PharmGKB

[26] http://www.pharmgkb.org/

URL: http://www.nigms.nih.gov/funding/pharmacogenetics.html
References

- Rubin D.L., Hewett M., Oliver D.E., Klein T.E., and Altman R.B. (2002). *Automatic data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML.* In: *Proceedings of the Pacific Symposium on Biology*, Lihue, HI, 2002 (Eds. R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale and T.E. Klein).

7.2.4 Stojanovic and colleagues' approach

The approach presented here has been developed within the WonderWeb[27] project and is part of the OntoLiFT prototype (integrated in KAON Workbench) whose main aim is to extract light ontologies from resources such as XML Schema or relational database schemata.

This approach [Stojanovic et al., 2002] tries to build light ontologies from conceptual database schemas using a mapping process. To carry out the process, it is necessary to know the underlying logical database model that will be used as source data.

The approach has the following five steps to perform the migration process.

1. *Capture information* from a relational schema through reverse engineering. The process considers relations, attributes, attributes types, primary key, foreign keys and inclusion dependencies present in the relational database model. The mapping process tries to preserve as much information as possible from the database schema.
2. *Analyse the obtained information* to built ontological entities by applying a set of mapping rules. These rules specify the way to migrate elements present in the database model into the ontology, including the constraints upon the database schema where these are present. There are rules for concept creation (the general assumption is that each relation is converted into a concept), for inheritance (creates an inheritance relationship if an inclusion dependency between two relations exists and concepts for both relations), and for relations [Stojanovic et al,. 2002; Behm et al., 1997]. These rules are applied in that order to create the ontology incrementally.
3. *Schema translation.* In this step the ontology is formed by applying the rules mentioned in the previous step. The translation process should create all ontological entities, organise concepts into the taxonomy, detect auxiliary relations in the original schema and remove redundant information.
4. *Evaluate, validate and refine* the ontology.
5. *Data migration.* The objective of this step is the creation of ontological instances based on the tuples of the relational database. It has to be performed in two phases: first, the instances are created, and in the second phase, relations between instances are established.

Main techniques used: mapping techniques.
Reuse other ontologies: not proposed
Source: relational schemas
Main goal: maps the relational schema with a conceptual schema
Domain in which it has been applied: information not available in papers
Tool associated: OntoLiFT.
Relevant ontologies built following it: information not available in papers.
Evaluation of the knowledge learnt: it is carried out by a domain expert.
URL: **http://wonderweb.semanticweb.org/publications.shtml**
References

[27] http://wonderweb.semanticweb.org/

- Behm, A.; Geppert, A.; Dittrich, K. (1997). On the Migration of Relational Schemas and Datato Object – Oriented Database Systems. In Procceding of the 5th Int. Conference on Re-Technologies for Information Systems (Klagenfurt, December 1997), pp, 13-33.
- Stojanovic, L.; Stojanovic, N.; Volz R. (2002). Migrating data-intensive Web Sites into the Semantic Web. Proceedings of the 17th ACM symposium on applied computing (SAC), ACM Press, 2002, pp. 1100-1107.
- Volz, R.; Oberle, D.; Staab, S.; Studer, R. (2003). OntoLiFT Prototype. IST Project 2001-33052 WonderWeb Deliverable 11.

7.3 Tools for ontology learning from relational schemata

We did not find any relevant tool to perform the ontology learning process from relational schemata at the time when this deliverable was written.

7.4 Conclusions

In this section, we have presented an overview of the most important methods used to perform the ontology learning process from relational schema. The next table is a summary of all them.

- The methods presented here aim to build an intermediate schema to allow extracting knowledge from the source schema and put into the target one. This process is mainly manual, and it is based on learning mapping techniques.
- There are not any relevant tool to perform the ontology learning process from this kind of source.

Table 7. Summary of ontology learning methods from relational schema.

Name	Main goal	Main techniques used	Reuse other ontologies	Sources used for learning	Tool associated	Evaluation	Bibliography
Johannesson's method	To map a relational schema with a conceptual schema	Mappings techniques	No	Relational schemas	Information not available in papers	User	Johannesson 1994
Kashyap's method	To create and refine an ontology	Mappings techniques Reverse engineering	Yes	Schemas of domain specific databases	EDEN	User	Kashyap 1999
Rubin and colleagues' approach	To create ontological instances	Mappings techniques	Yes	Relational schema of a database	Information not available in papers	User	Rubin et al., 2002
Stojanovic and colleagues' approach	To create ontological instances from a database	Mappings Reverse engineering	No	Schemas of domain databases	OntoLift (KAON)	User	Stojanovic et al. 2002

8 Conclusions

This deliverable presents an overview of the most relevant ontology learning methods, techniques, and tools taken as input different sources like text, semi-structured data, dictionaries etc. Moreover, a summary of relevant tools, if available, to carry out each learning approach has been done.

The general conclusions of this overview are:

- Ontology learning is a suitable process:
    - to accelerate the knowledge acquisition process necessary to build an ontology from scratch,
    - to reduce the time required to enrich an existing ontology,
    - to speed up the construction of ontologies to be used for different purposes in the Semantic Web. All the methods and tools presented in this deliverable allow to reach these main goals.

The main lacks for all the methods and tools presented in this overview are that there are not integrated methods and tools, that combines different learning techniques and heterogeneous knowledge sources with existing ontologies to accelerate the learning process. There are not a methods or techniques for evaluating the accuracy of the learning process either.