EU-IST Project IST-2003-506826 SEKT

SEKT: Semantically Enabled Knowledge Technologies

D12.4.1 Technology state of the art review, version 1

John Davies, Paul Warren (BT)

With annexes authored by:

Ian Thurlow, Nick Kings (BT)

Dunja Mladenic, Marko Grobelnik (Jozef Stefan Institute)

Kalina Bontcheva, Diana Maynard, Hamish Cunningham (University of Sheffield)

Atanak Kiryakov, Marin Dimitrov, Dimitar Manov (SIRMA AI EOOD)

Peter Haase, York Sure, Denny Vrandecic, Marc Ehrig, Christoph Tempich
(University of Karlsruhe)

Zhisheng Huang, Frank van Harmelen, Annette ten Teije, Perry Groot, Cees Visser
(Vrije Universiteit Amsterdam)

Jos de Bruin, Francisco Martin-Recuerda (DERI)

Abstract

This report provides an overview of the state of the art in semantic knowledge technologies. Specifically, the report describes: ontology creation; metadata extraction; ontology evolution and maintenance; reasoning with ontologies; ontology merging and mapping; and knowledge access. Also included is a review of the state of the art in ontology engineering methodologies.

Keyword list: semantic knowledge technologies, semantic knowledge management, Semantic Web, ontology creation, metadata extraction, ontology evolution, ontology merging, ontology mapping, knowledge access.

WP12 Exploitation and dissemination

Report	PU
Contractual date of delivery 31/12/04	Actual date of delivery 10/02/05

SEKT Consortium

This document is part of a research project partially funded by the IST Programme of the Commission of the European Communities as project number IST-2003-506826.

British Telecommunications plc.
Orion 5/12, Adastral Park
Ipswich IP5 3RE
UK
Tel: +44 1473 609583, Fax: +44 1473 609832
Contact person: John Davies
E-mail: john.nj.davies@bt.com

Empolis GmbH
Europaallee 10
67657 Kaiserslautern
Germany
Tel: +49 631 303 5540
Fax: +49 631 303 5507
Contact person: Ralph Traphöner
E-mail: ralph.traphoener@empolis.com

Jozef Stefan Institute
Jamova 39
1000 Ljubljana
Slovenia
Tel: +386 1 4773 778, Fax: +386 1 4251 038
Contact person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si

University of Karlsruhe, Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 6592
Fax: +49 721 608 6580
Contact person: York Sure
E-mail: sure@aifb.uni-karlsruhe.de

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1891
Fax: +44 114 222 1810
Contact person: Hamish Cunningham
E-mail: hamish@dcs.shef.ac.uk

University of Innsbruck
Institute of Computer Science
Technikerstraße 13
6020 Innsbruck
Austria
Tel: +43 512 507 6475
Fax: +43 512 507 9872
Contact person: Jos de Bruijn
E-mail: jos.de-bruijn@deri.ie

Intelligent Software Components S.A.
Pedro de Valdivia, 10
28006
Madrid
Spain
Tel: +34 913 349 797
Fax: +49 34 913 349 799
Contact person: Richard Benjamins
E-mail: rbenjamins@isoco.com

Kea-pro GmbH
Tal
6464 Springen
Switzerland
Tel: +41 41 879 00
Fax: 41 41 879 00 13
Contact person: Tom Bösser
E-mail: tb@keapro.net

Ontoprise GmbH
Amalienbadstr. 36
76227 Karlsruhe
Germany
Tel: +49 721 50980912
Fax: +49 721 50980911
Contact person: Hans-Peter Schnurr
E-mail: schnurr@ontoprise.de

Sirma AI EAD, Ontotext Lab
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768 303, Fax: +359 2 9768 311
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Vrije Universiteit Amsterdam (VUA)
Department of Computer Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 444 7731, Fax: +31 84 221 4294
Contact person: Frank van Harmelen
E-mail: frank.van.harmelen@cs.vu.nl

Universitat Autonoma de Barcelona
Edifici B, Campus de la UAB
08193 Bellaterra (Cerdanyola del Vall` es)
Barcelona
Spain
Tel: +34 93 581 22 35, Fax: +34 93 581 29 88
Contact person: Pompeu Casanovas Romeu
E-mail: pompeu.casanovas@uab.es

Executive Summary

This report provides an overview of semantic knowledge technologies, categorised as: ontology creation; metadata extraction; ontology maintenance and evolution; reasoning, in particular here we discuss reasoning in the presence of inconsistencies; ontology merging, mapping and translation; accessing ontologically-represented knowledge; and the use of methodologies to create ontologies.

Ontology creation (SEKT WP1)

The report firstly looks at ontology creation and concludes that there are a number of research themes where knowledge discovery and text mining can be extended to assist ontology creation:

- The use of knowledge discovery for ontology learning in automatic or semi-automatic mode.
- In particular the use of 'active learning' and 'semi-supervised learning' where the available background knowledge can not be easily integrated into a fully automatic approach.
- The extension of techniques previously applied to unstructured text to other forms of data, e.g. multimedia, signals, graphs and networks.
- The development of dynamic ontologies to describe data which is evolving over time.
- The development of scaleable techniques to deal with large volumes of data.

Metadata extraction (SEKT WP2)

The report next looks at ontology-based information extraction as the basis of metadata extraction and finds two major challenges:

- Explore the synergies between ontology generation methods from data mining (SEKT workpackage 1), HLT, and methods from ontology and metadata management (workpackage 3).
- Developing hybrid adaptive information extraction tools, combining rule-based and machine learning approaches and using reasoning services, to perform entity tracking within and across documents.

In addition, there is a need to develop corpora and metrics for evaluating the performance of information extraction tools used specifically to annotate content relative to ontologies.

Ontology maintenance and evolution (SEKT WP3)

For ontology maintenance and evolution, the key research issues are:

- Language-independent ontology evolution.
- Ontology change request specification, where a declarative language would allow reasoning about interactions between ontology changes and constraints.
- Ontology dependency, where ontology mapping, merging, alignment and integration needs to take account of multiple related ontologies.

Reasoning with inconsistent ontologies (SEKT WP3)

There are two ways to reason in the presence of inconsistency. On the one hand, one can attempt to repair any detected inconsistencies. Alternatively, one can avoid the inconsistency, by selecting a consistent subset of the theory. In this case, the key issue is the selection function used to select the consistent sub-theory. SEKT is investigating both approaches.

Ontology merging and aligment (SEKT WP4)

Ontology merging and mapping is motivated by a number of use cases: instance mediation, which comprises instance transformation, instance unification and query rewriting; ontology merging; and creating ontology mappings. A number of methods, tools and systems have been developed for various of these use cases. Some of these are applicable to ontologies expressed in RDFS, and one is applicable to ontologies in OWL. None of the approaches satisfy all the demands made by the Semantic Web. On the Web, the use of a global ontology, into which other ontologies are mapped, is not realistic. However, the one-to-one approach is not expected to scale, because of the number of mappings which would need to be maintained. Therefore SEKT will investigate a hybrid approach, where islands cluster around influential domain ontlogies. Within these islands there would be a 'global' ontology, into which local ontologies are mapped. One-to-one mappings would then exist between the global ontologies. SEKT will also take account of approaches used for database integration, which can overcome some of the scalability issues with large sets of instances.

Accessing knowledge (SEKT WP5)

For the purposes of the state-of-the-art survey, knowledge access is divided into a number of basic tasks:

- Searching and browsing, where semantic indexing and searching can be used.
- Knowledge sharing, where metadata from a document can be used to precisely target that document at relevant communities.
- Knowledge visualisation and organisation, where the aproach proposed is to translate from the knowledge ontology to a visualisation ontology specifically designed to aid visualisation.
- User profile construction, to generate a profile which then acts as a filter on the general ontology to give a personal view.
- Natural language generation, where structured data in a knowledge base is expressed in natural language.
- Device and content repurposing, where the aim is to deliver a functional presentation of a web page on any access medium or device. The de-facto standard in this area is the W3C RDF-based CC/PP standard, in particular the UAProf application.

Ontology engineering methodologies (SEKT WP7)

An ontology engineering methodology needs to comprise procedures for three activities: ontology management; ontology development; and ontology support. It is argued that argumentation visualisation is particulary appropriate to address the requirements of SEKT, and the survey includes the main aspects of this topic.

Contents

List of annexes

1 Introduction

This report describes the state-of-the-art in semantic knowledge technologies, specifically the use of ontologies. It looks at how:

- ontologies can be created;
- ontologies can be populated, i.e. how instances can be extracted from documents to populate a given ontology;
- ontologies, and the associated instances can be maintained and evolved;
- reasoning can be undertaken with ontologically-represented knowledge;
- ontologies can be merged, or mapped to allow translation between them;
- ontologically-represented knowledge can be accessed;
- and how methodologies can be used to aid the creation of ontologies.

Each of these domains corresponds to a section of the report (sections 3 to 9). In each case there is an associated annex, providing more detail. The annexes associated with sections 5 to 9 have also been issued as SEKT deliverables, and are included here for ease of reference. The annexes associated with sections 3 and 4 have not been independently issued.

2 Definitions and scope

By knowledge technologies we mean here the technologies and processes used to increase the benefits achieved from knowledge in an organisation or beyond. The organisation might be a corporation, a government body, a university, or any grouping of individuals who come together to achieve a common goal. In fact, whilst the technology discussed here, and being developed in the SEKT project, is certainly initially aimed towards organisations, its applicability is essentially wider. The immediate goal of SEKT is the development of semantic technologies for organisational intranets. However, the scope of this study is wider, covering the application of semantic technology to the Web itself, i.e. the Semantic Web.

A frequently quoted definition of an ontology is that given by Gruber: “An ontology is a specification of a conceptualization”[i]. In simpler English, this means a formal way of representing reality, e.g of representing classes, the relationships between classes, and how objects are associated with classes. When quoting this definition the point is often made that the primary purpose of an ontology is to share knowledge, and hence an ontology is a specification of a *shared* conceptualization. For a more detailed discussion of the nature of ontologies, see the classic text on knowledge representation by Sowa[ii].

Mathematically, an ontology can be defined as 4-tuple <C, R, I, A>, where C is a set of concepts, or classes; R is a set of relations; I is a set of instances; and A is a set of axioms. This definition is particularly useful, for example, when discussing mappings between two different ontologies, when we need to be clear which elements of one ontology are being mapped into which elements of a second. Some writers are more limited in their definition of an ontology, specifically excluding the instances. In any case, regardless of the formal definition used, it is often useful in practical implementations to store the classes and the relationships separately from the instances. The latter is sometimes known as the ‘knowledge base’.

A specific form of ontology is that known as a taxonomy. The emphasis here can be summed up by: “especially … including a hierarchical arrangement of types in which categories of objects are classified as subtypes of more abstract categories, starting from one or a small number of top categories, and descending to more specific types through an arbitrary number of levels”[iii]. Effectively, taxonomies are hierarchical classification systems, as used in biology to classify plants and animals. There are currently a number of commercially-available taxonomy systems, aimed at applications within the enterprise. Empolis, http://www.empolis.com, a member of the SEKT consortium, has products incorporating taxonomy software. Other ompanies providing such systems include Verity, http://www.verity.com; Autonomy, http://www.autonomy.com; and Entrieva, http://www.entrieva.com. Commercial products are now becoming available based on ontologies, e.g. from Empolis, Semagix (http://www.semagix.com) and Aduna (http://aduna.biz).

3 The creation of ontologies

In Annex 1, the state-of-the art in knowledge discovery and text mining in the context of the Semantic Web and knowledge management is described. This is the key research area investigated in SEKT workpackage 1.

Knowledge discovery is viewed as a research area with several subfields. Two general definitions covering the field are:

- Studying the design and analysis of algorithms for making predictions about the future based on past experiences (from http://www.learningtheory.org/)
- A process which aims at the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information from data in large databases. (Fayad et al. 1996)[iv]

Text mining addresses the more specific issue of processing and analysing textual data. Text mining is an interdisciplinary area that involves at least the following key research fields:

- *Machine Learning and Data Mining* (Mitchell 1997[v]; Fayyad, et al., 1996; Witten and Frank 1999[vi]; Hand, et al. 2001[vii]) which provides techniques for data analysis with varying knowledge representations and large amounts of data,
- *Statistics* and statistical learning (Hastie, et al. 2001[viii]) which in the context of text mining contributes data analysis (Duda, et al. 2000[ix]) in general,
- *Information Retrieval* (Rijsberg 1979[x], Mani and Maybury 1999[xi]) providing techniques for text manipulation and retrieval mechanisms, and
- *Natural Language Processing* (Manning and Schutze 2001[xii]) providing the techniques for analyzing natural language. Some aspects of text mining involve the development of models for reasoning about new text documents based on words, phrases, linguistic and grammatical properties of the text, as well as extracting information and knowledge from large amounts of text documents.

One of the most popular applications of text mining is document categorization. Other applications include document clustering, visualization, search based on the content, automatic document summarization, automatic construction of document hierarchies, document authorship detection, identification of plagiarism of documents, and user profiling

Several areas where knowledge discovery and text mining can be applied in Semantic Web-based knowledge management are identified and discussed:

- Semantic web knowledge management applications typically involve deep structured knowledge represented using ontologies and associated metadata. Since KD techniques are mainly about discovering structure in data, this can serve as one of the key mechanisms for structuring knowledge. We refer to such approaches as “ontology learning” which is usually performed in automatic or semi-automatic mode.
- Fully automatic KD approaches are not always the most appropriate, since often it is too hard or too costly to integrate the available background knowledge about the domain into fully automatic KD techniques. For such

cases the KD approaches such as “Active Learning” and “Semi-supervised Learning” are argued to be more appropriate.

- Knowledge management applications typically operate on unstructured or semi-structured text-based documents. An important property of such data is that it is relatively easily manageable by humans. In the future we may expect applications which also integrate less human –accessible data (e.g. multimedia, signals, graphs/networks). For such situations there will be much more emphasis on automatic or semi-automatic methods offered by KD technologies which are not limited to a specific data representation.
- Data and corresponding semantic structures evolve over time. The consequence is that we need to be able to evolve the ontologies which model the data accordingly – we call these structures “dynamic ontologies”. A KD sub-field called “stream mining” addresses this issue and is discussed.
- Scalability is a central issue in KD and, correspondingly, knowledge management applications often have to deal with large volumes of data. The way in which KD can offer scaleable techniques is discussed.

The annex concludes by overviewing TextGarden, a KD software suite produced by the SEKT project.

4 Extracting the metadata

4.1 Information extraction

Annex 2 describes how information extraction (IE) can be used for semantic annotation. It provides both an overview of the subject and also a comparison of a number of existing systems. This is the subject of research in SEKT workpackage 2.

The goal of IE is to take unseen texts as input and produce fixed format unambiguous data as output. It differs from information retrieval (IR) in that IR finds relevant texts and presents them to the user, whilst IE analyses text and presents only the specific information in which the user is interested.

IE can be viewed as comprising five tasks:

- Named entity recognition (NE), which finds and classifies names such as companies, people and places.
- Coreference resolution (CO) which finds identity relations between entitities in the text. For example, with CO we can find that, in a given text, ‘George W Bush’, ‘the President’, and ‘he’ all refer to the same person.
- Template element construction (TE) which, using CO, adds descriptive information to the results of NE.
- Template relation construction (TR) which finds relations between different entities in the text, e.g. that ‘George W Bush’ *is President of* ‘the USA’.
- Scenario template production (ST) which fits the results of TE and TR into specified event scenarios. An example might be that ‘George W Bush’ and ‘Dick Cheney’ were involved in a ‘presidential inauguration ceremony’.

Each of these five tasks has been the subject of rigorous performance evaluation. The difficulty of the tasks is to an extent dependent on three factors:

- Text type, i.e. whether the text is a newspaper article, text from the Web, or the output of a speech recogniser.
- Domain, i.e. the subject matter of the text and the style in which it is written.
- Scenario, i.e. the particular event of interest, e.g. mergers between companies.

4.2 The five IE tasks

4.2.1 Named entity recognition

NE is the simplest and most reliable IE task. It can be performed with up to 95% accuracy, whilst even humans do not attain 100% accuracy. It is weakly domain dependent, i.e. small changes of domain require some changes to the system and large changes of domain require quite large changes.

4.2.2 Coreference resolution

CO is an imprecise process and, depending on the domain, only 50-60% may be relied upon. This compares with human scores, in comparative tests, of around 80%. However, proper-noun coreference resolution is significantly easier than anaphora resolution. In the former problem we are identifying equivalences between, e.g.

'IBM', 'IBM Europe', 'International Business Machines' etc. In the latter we are seeking to identify equivalence between proper nouns and pronouns, e.g. 'George W Bush' and 'he'. CO systems are domain dependent.

4.2.3 Template element production

TE associates a type with a named entity, e.g. personal name, date, company name etc, and also descriptive information. In one test the best systems scored around 80% whilst humans scored 93%. TE is weakly domain dependent.

4.2.4 Template relation production

This task requires the identification of a small number of possible relations between the template elements identified in the template element task. Examples include:

- an employee relationship between a person and a company,
- a family relationship between two persons,
- a subsidiary relationship between two companies.

The best scores are around 75%. TR is a weakly domain dependent task.

4.2.5 Scenario template extraction

ST is a difficult IE task. The best competitive scores have been around 60%, although human scores can be as low as around 80%. It is possible to increase precision at the expense of recall, i.e. system can be developed to not make many mistakes but at the expense of missing occurrences of relevant scenarios. The ST task is domain dependent and tied to the scenarios of interest to the users.

4.3 Semantics and ontologies

The split of IE into the various tasks above was developed prior to the development of the ideas of the Semantic Web and semantic knowledge technology. However the basic approach can be adapted to the various tasks required by semantic knowledge management.

4.3.1 Semantic annotation

The tasks here are to:

- Annotate and hyperlink named entities in text.
- Index and retrieve documents with respect to the entities referred to.

We imagine the existence of an ontology, with classes and relations, and a knowledge base, with instances of the classes and specific relations between the instances.

Then in the first task we identify entities associated with the ontology's classes, and establish links between these entities and the appropriate instances in the knowledge base, thereby providing both class and instance information about the entities.

The second task is an extension of the classical information retrieval task, except that documents are retrieved on the basis of the entities contained within it, rather than the particular words used. Thus a search for 'George W Bush' might retrieve a document which does not mention him by name, but does refer to 'the President', where the system has previously concluded that 'the President' and 'George W Bush' are equivalent.

4.3.2 Ontology-based information extraction

Ontology-based IE (OBIE) is the technology used for semantic annotation. It uses a formal ontology rather than a flat lexicon or gazetteer structure as is used by traditional IE. Moreover, by linking the extracted entity to its semantic description, it identifies it in a way which means that entities can be traced across documents.

OBIE corresponds to a combination of NE and CO. Given that the former is much harder than the latter, this makes OBIE a much harder task than NE alone. OBIE poses two main challenges:

- the identification of instances from the ontology,
- the automatic population of ontologies with the new instances in the text.

In the first of these we are associating entities with existing entities in the ontology. In the second, new entities are identified which, whilst corresponding to existing classes in the ontology, do not have associated instances in the knowledge base. In this case new instances are created.

4.4 Challenges in ontology-based information extraction

The outstanding challenge, to be addressed in SEKT, is to develop ontology-based IE which can be configured to provide a service that will annotate any page relative to a particular ontology. Software agents can thereby use IE services to find instances of concepts from their own models, so that the software extracts directly to the user's own ontology.

In detail, the challenges are to go beyond the state-of-the-art by:

- Exploring the synergies between ontology generation methods from data mining (WP1), HLT, and methods from ontology and metadata management (WP3)
- Developing hybrid adaptive information extraction tools, combining rule-based and machine learning approaches and using reasoning services, to perform entity tracking within and across documents.

4.5 Quantitative evaluation

Progress in information extraction has to a significant extent been driven by competitions with published corpora and evaluation metrics, e.g. those associated with the Message Understanding Conferences. However, such existing corpora and tools are not suitable for evaluating IE tools in the Semantic Web context, for a number of reasons. These corpora and metrics only detect very coarse-grained types of entities, without a specific ontology. Moreover, they provide no measure of the accuracy with which references are created between the entities and events in the documents and the instances associated with the target ontology. Corpora and metrics are required for evaluating the performance of IE tools used specifically to annotate content relative to ontologies. This will include:

- Detection of entities and events, given the target ontology of the domain.
- Disambiguation of the entities and events in the documents with respect to instances in the knowledge base, e.g. measuring whether the IE correctly disambiguated ‘Cambridge’ to the correct instance ‘Cambridge, UK’ or ‘Cambridge, MA’.
- The decision as to when a new instance needs to be added to the knowledge base.

An evaluation corpus is required, annotated with the correct ontological class and instances. The corpus needs to be divided into two parts; one for system development and testing; the other for evaluation only.

New metrics are also required. In OBIE correctness now becomes more a matter of degree, compared with classical IE where information is either extracted correctly or not. To illustrate this point consider an entity representing a charity. In an ontology with both the class ‘charity’ and ‘company, the allocation of this entity to the class ‘company’ is wrong, but not as wrong as allocating the entity to the class ‘country’, say.

5 Maintaining and evolving ontologies

Annex 3 provides a comprehensive review of ontology evolution processes and frameworks from the literature. Related work in other relevant fields, like databases and software engineering is examined. The annex proposes a general framework for the evolution of ontologies and investigates the methods for the evolution of OWL-based ontologies and their applications. It explores the formalization of the semantics of change. This takes account of a number of aspects of consistency, including structural consistency, logical consistency, and user-defined consistency.

The key open research issues needing to be addressed are summarised below and discussed in more detail in the annex: This work is undertaken in SEKT workpackage 3.

5.1 Language-independent ontology evolution

Existing ontology evolution techniques tend to rely too heavily on the underlying ontology language. The research goal should be a language-independent ontology evolution system, enabling more general purpose software tools to be developed. The approach proposed here is the declarative modelling of all aspects of ontology evolution.

5.2 Ontology change request specification

A declarative language for the specification of a request for a change would allow ontology changes and constraints to be expressed in a single framework, and thus, would allow reasoning about interactions between changes and constraints. This language would differ from existing ontology query languages, which are only used for retrieval of data from an ontology. It would extend these languages to allow representation of ontology modifications, as well.

5.3 Ontology dependency

Another research issue where much remains to be done is the problem of working with multiple related ontologies. There are various tasks here, such as ontology mapping, ontology merging, ontology alignment and ontology integration.

Ontology mapping relates similar concepts and relations from different ontologies to each other; whilst ontology merging creates a new ontology from two or more existing ontologies with overlapping parts. Each of these dependency forms puts different requirements on the evolution between dependent ontologies. Many of these can be resolved by introducing a meta-ontology that captures relationships between entities from different (object level) ontologies and this is explicated in the report.

5.4 The SEKT focus

In the context of the SEKT project, we are specifically interested in the OWL ontology language, since it is a recommendation of the W3C and offers a richer language and clearer semantics than RDF(S). Therefore the report looks in some depth at ontology management and evolution in OWL, particularly since, being a relatively new language, there is little previous work aimed specifically at OWL.

The focus of ontology management research in SEKT is on developing a tool framework and methods for the evolution of OWL-based ontologies and their application for data integration scenarios in the presence of heterogeneous evolving data sources. The report discusses an approach to formalizing the semantics of change for the OWL ontology language (in particular for OWL DL and its sublanguages), embedded in a generic process for ontology evolution. Our formalization of the semantics of change allows the definition of arbitrary consistency conditions – grouped under structural, logical, and user-defined consistency – and the definition of resolution strategies that assign resolution functions to ensure these consistency conditions are satisfied as the ontology evolves. This flexibility allows support for various fragments of the OWL-DL language.

This leads us to the development of a software prototype, *evOWLution*, based in the KAON2 infrastructure which is currently being developed in the EU IST DIP[1] project. evOWLution implements the results of the described original research done in the SEKT project on the methods for a consistent evolution of OWL ontologies.

A second prototype developed and described here is *dlpconvert*. It is a tool to convert an OWL encoded ontology, that lies within the DLP fragment, to another syntactic representation. The DLP fragment has certain computational advantages and is expressive enough to represent the large majority of existing ontologies. This work is important in the context of ontology management since it is a further step towards language independent modelling of knowledge bases, leading to greater interoperability and increased tool availability.

[1] http://dip.semanticweb.org/

6 Reasoning with inconsistent ontologies

One of the key features of the Web (and by extension the Semantic Web) is that it is characterized by scalability, distribution, failure-tolerance and multi-authorship. All these characteristics may introduce inconsistencies into the Semantic Web. An overview is given here. Annex 4 provides more detail on this. Reasoning in the presence of inconsistency is a topic of research in SEKT workpackage 3.

There are two main ways to deal with inconsistency. One is to diagnose and repair it when we encounter inconsistencies. Schlobach and Cornet[xiii] propose a non-standard reasoning service for debugging inconsistent terminologies. Another one is to simply avoid it and to apply a non-standard reasoning method to obtain meaningful answers. In this report, we will focus on the latter.

For reasons explained above[2], the report considers ontology specifications for which their logical foundations are based on Description Logics and OWL, the Web counterpart of Description Logics. Semantically, description logics can be considered as a subset of first order logic. The classical entailment in logics is *explosive*: any formula is a logical consequence of a contradiction. Therefore, conclusions drawn from an inconsistent knowledge base by classical inference may be completely meaningless.

The report proposes a general framework for reasoning with inconsistent ontologies and suggests how an inconsistency reasoner can be developed for the Semantic Web. The general task of an inconsistency reasoner is: given an inconsistent ontology, return meaningful answers to queries. Reasoning with inconsistency is an old topic in logics and AI. Many approaches have been proposed to deal with inconsistency and these are briefly surveyed with a focus on paraconsistent logics.

The development of paraconsistent logics was initiated to challenge the 'explosive' problem of the standard logics. Paraconsistent logics allow theories that are inconsistent but non-trivial. There are various paraconsistent logics, most of which are defined on a semantics which allows both a proposition and its negation to hold for an interpretation. Levesque's limited inference[xiv] allows the interpretation of a language in which a truth assignment may map both a proposition P and its negation ~P to true. Extending the idea of Levesque's limited inference, Schaerf and Cadoli[xv] propose an approach based on selecting a subset of the language which can be used as a parameter in their framework and allows their reasoning procedure to focus on a part of the theory while the remaining part is ignored

The key issue is then the selection of the most appropriate subset. This is achieved by means of a selection function which selects a consistent subset of an inconsistent theory. The different approaches to selection found in the literature are reviewed and inform our own approach. Briefly, our main idea's key strength is that it is relatively *simple:* given a selection function, which can be defined on syntactic or semantic relevance, we can always select some consistent sub-theory from an inconsistent ontology (s*emantic* relevance has been investigated in computational linguistics).

[2] Standards-based, expressivity and clear semantics.

Standard reasoning is then used on the selected sub-theory to find meaningful answers.

Specifically, the report proposes a pre-processing algorithm and a strategy of inconsistency reasoning processing based on a linear extension strategy, and shows that a linear extension strategy is useful to create meaningful and sound answers to queries, although they may be undetermined and not always maximal.

Development of the selection function by means of a syntactic relevance measure is described and several examples of how the syntactic relevance approach can be used to obtain intuitive reasoning results are given.

Finally, the design and architecture of the DIG-based PION system for reasoning with inconsistent ontologies is presented. DIG – Description Logic Interface - is essentially an API for accessing DL reasoners, see, for example, http://www.sts.tu-harburg.de/~r.f.moeller/racer/interface1.1.pdf

In future work, we will investigate formal properties of selection functions as well as different approaches to selection functions (e.g. semantic-relevance based). We will also investigate different extension strategies as alternatives to the linear extension strategy, in combination with different selection functions (also instance semantic relevance based), and evaluat their performance characteristics.

7 Ontology merging and alignment

Annex 5 provides a state of the art survey of ontology merging and alignment. A brief overview is given here. This is the subject of research in SEKT workpackage 4.

7.1 Overcoming heterogeneity

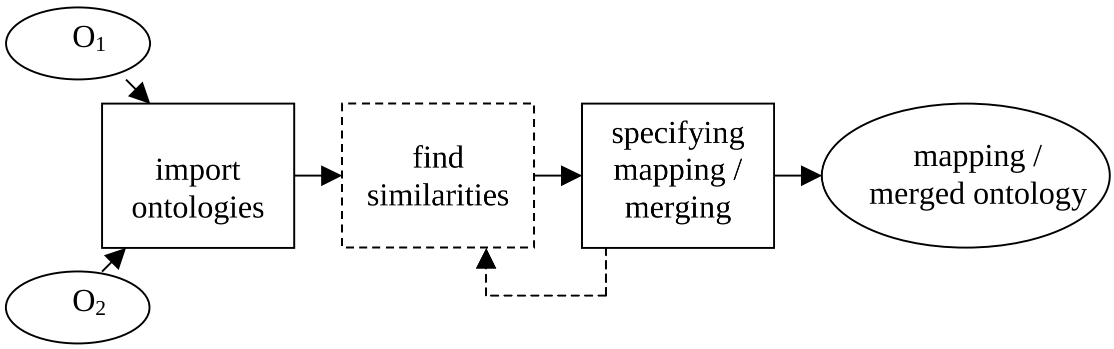
Inevitably, within one domain, a variety of ontologies will exist to describe that domain. More broadly, across related domains there will be overlapping ontologies. The aim of the technology discussed here is to merge ontologies and to enable interoperation between systems using different ontologies.

Indeed the process of creating standardised ontologies, across and between organisations, is lengthy and difficult. The approaches discussed here provide a more flexible alternative to the search for universally accepted standards. These technologies, whilst valuable in the enterprise, are crucial if we are to build Semantic Webs across organisations, and essential if we are to build anything like a global Semantic Web.

There are two different goals which may be adopted, both of which share common approaches. In one, we may wish to *merge* two or more ontologies. For example, this might be done by forming a union or an intersection of the classes and relations. Alternatively, we may leave the ontologies but establish a *mapping* between them. This may be done on a 1-1 basis, or by mapping each of a number of ontologies into a central or global ontology.

Figure 7.1, taken from Annex 5, illustrates both processes emphasising their commonality. The find similarity process may be a manual process, or it may be automated. There may be feedback from the 'specifying mapping / merging' process and 'find similarities', making the combined processes iterative.

Figure 7.1 The ontology mapping and merging process



It may also be necessary to *align* ontologies. In this case the ontologies are kept separate, but at least one of the original ontologies is adapted such that the conceptualization and the vocabulary match in overlapping parts of the ontologies.

The report discusses in detail the various sorts of mismatches which can occur. It also compares the merits of 1-1 mappings between ontologies with the use of a global ontology into which all pertinent ontologies are mapped. A solution in both cases is the use of mediators to map between ontologies. In the former case each ontology has a mediator which needs to be aware of each other mediator. In the latter case one mediator is able to map between the global ontology and each of the local ontologies.

The remainder of this section follows the ontology merging and aligning state-of-the-art report in discussing the use cases which are crucial for ontology mediation in the Semantic Web; a framework for evaluating different approaches; a survey and comparison of such approaches; and finally some conclusions.

7.2 Use cases

There are three broad use cases to consider:

- instance mediation, which in turn breaks down into instance transformation, instance unification, and query rewriting;
- ontology merging;
- creating ontology mappings.

The first two of these are dependent on, i.e. make use of, the third.

7.2.1 Instance mediation

Instance mediation is the process of reconciling differences between the two instance bases, each described by an ontology.

In instance transformation the task to be performed is the transformation of an instance of a source ontology to an instance of the target ontology, where both the original and transformed instance provide information about the same real-world object.

In instance unification we have two instances of an ontology and first wish to determine whether they both refer to the same real-world object. If they do, we then wish to unify them into a newly created instance. In some cases we may be able to specify exact conditions which unambiguously specify whether the two instances refer to the same object. In other cases we may use a similarity measure which expresses the probability of their referring to the same object.

Instance transformation and instance unification are often required in query rewriting. In query rewriting we have an application which employs one ontology and wish to query a data source employing a different ontology. In order to do this, the query is first rewritten in the data source ontology. After execution of the query, the results are transformed back to the application ontology and unified with the local instances in that ontology.

7.2.2 Ontology merging

In ontology merging two source ontologies are merged into one target ontology based on the source ontologies. The source ontologies are then replaced by the target ontology. Alternatively, the source ontologies may remain along with the mappings to the merged ontology. In the former case the instance stores must be merged. In the latter case the source ontologies can maintain their instance stores, with instance transformation and unification taking place during run-time.

7.2.3 Creating ontology mappings

The previous use cases require that a mapping be created between the source and target ontologies. The first stage in doing this is to establish similarity between the entities, e.g. concepts and relations, in the ontologies. This can be done either manually or automatically. Having established the similarities they can be used as the starting point for specifying the mappings; a process which is in general semi-automatic.

7.3 The evaluation framework

The different approaches to ontology mediation described in the survey are all compared against the following characteristics:
- Ontology language;
- Mapping language, which may be the same as that used for the ontologies;
- Mapping patterns. Investigating the use of mapping patterns is a large part of the ontology mediation work in SEKT. It is therefore instructive to determine where existing methodologies use such patterns.
- Automation support.
- Applicability to our use cases.
- Implementation, i.e. supporting tools which includes both tools to support the user in creating the mapping, and tools which do run time mediation.
- Experience reported.

In the survey itself this framework is used to describe various approaches. The approaches are divided into two categories:
- Methods and tools for ontology matching, merging and mapping.
- Data integration systems which use ontologies. Unlike the first category, these are comprehensive in the sense that they typically include different types of functionality.

In addition, some specific techniques are described for which it is not appropriate to use the above framework.

7.4 Comparison of methods

The survey in the report divides the approaches into ‘methods and tools’ and ‘integration systems’. Of the former, tools created for the Semantic Web support RDFS. One tool surveyed (PROMPT) supports OWL. Another category of tool which performs only ontology matching have their own internal representations. By their nature the integration systems do not focus on inter-operability with other ontology tools.

In some cases the mapping language used is the same as the ontology language. In other cases a specific mapping language is used. The types of mapping often present in specific mapping languages can be seen as elementary mapping patterns. However the use of mapping patterns is, as yet, not well developed.

The tools and systems provide varying degrees of automation. Some ontology matchers are completely automated. However, they only perform one step in the overall mapping process. Mapping and merging of ontologies is often an interactive process, with the user free to adopt the suggested action or ignore the suggestion.

Most of the experiences reported in the literature are really toy problems; real experiences with ontology mapping and ontology-based information integration are lacking. In part this is because there are not yet many ontologies on the Web, although there are some experiences with real-life data sources, such as bibliographic sources.

7.5 Conclusions

None of the approaches surveyed fit all the criteria for ontology mediation on the Semantic Web. Those developing Semantic Web ontology mediation systems can learn from the data integration systems which provide services for query answering over distributed heterogeneous data sources. However, these are currently within the more controlled environment of the enterprise. On the Web the use of a global ontology, into which other ontologies are mapped, is not realistic. However, the one-to-one integration approach is also not expected to scale, because of the number of mappings which would need to be maintained. A hybrid approach seems likely, where islands cluster around influential domain ontologies. Within these islands there would be a ‘global’ ontology, into which local ontologies are mapped. One-to-one mappings would then exist between the global ontologies. SEKT will initially adopt this approach. SEKT will also take account of database integration approaches, which can overcome some of the scalability issues with large sets of instances.

8 Accessing knowledge

Knowledge management can be broadly defined as the more effective and efficient management and use of an organisation's knowledge. Knowledge, to be useful and effective, needs to be delivered to the relevant people at the right time in the most appropriate format. This imples a requirement that knowledge repositories are capable of being searched, browsed, shared, and represented and visualised in various formats. Of course, this is as much true of knowledge represented semantically using ontologies as it is true of knowledge represented purely as textual documents.

For the purpose of this state of the art survey, and for our work in SEKT generally, we divide knowledge access into a number of basic tasks. Some of these, such as searching and browsing, are directly part of the process of accessing knowledge. Others, such as user profile construction, support this process. These tasks are the subject of research in SEKT workpackage 5.

The basic tasks are:
- searching and browsing
- knowledge sharing
- knowledge visualisation and organisation
- user profile construction
- natural language generation
- device and content repurposing

In this report we provide an overview of the whole area. Annex 6 provides more detail. The annex is taken from SEKT deliverable D5.1.1[xvi], which itself has detailed annexes for each of the six areas.

8.1 Searching and browsing

In general, corporate search engines, based on conventional information retrieval techniques, tend to offer high recall but low precision. Moreover, they do not take into account the context in which the user makes a query, i.e. for a given search string they return the same responses to all users. Although they offer various advanced features, these are not used by the majority of users.

Commercial Web search engines have demonstrated that conventional IR techniques can be augmented with algorithms which exploit the hyperlink structure of the Web. Similar techniques are now being demonstrated as being relevant to intranet search. At the same time, improved clustering, presentation and navigation of results are now available in many commercial search engines.

A number of search engines are now emerging which index and search domain-specific semantic annotations and exploit the structure of domain specific ontologies. This can achieve greater precision whilst at the same time augmenting the results with supplementary information from the dataset of instances associated with the ontology. Such a dataset is termed a knowledge base.

Search engines need to provide more personalised search services which focus on the information needs of an individual user, taking into account that an individual user may have several roles. This is an area where further research is required, e.g. to find better ways to express the user's information needs. Further research is also required into search engines which proactively fetch information on behalf of the user.

8.2 Knowledge sharing

The importance of knowledge sharing and reuse in knowledge management in order to share best practice and prevent duplication of effort has led much interest in supporting communities of practice. Knowledge sharing tools combine the functions of locating and distributing information within communities. The assumption is that someone in the user's community has already created or is aware of relevant information or that someone is able to offer help or advice. User profiles can be used to route information to people, as and when the information is required. Understanding a user's changing context is a central research problem.

SEKT will use semantic technology to develop a range of functionality for enhancing knowledge sharing, including:

- The use of metadata automatically extracted from a document to precisely target that document at relevant communities.
- Enhancing a user's profile through use of the metadata relating to the documents he or she habitually reads or creates.
- Through ontology mediation, knowledge described with one ontology can be viewed by a user in the context of his or her own local ontology.
- The metadata associated with a document can be extended with information from the profiles of the readership. This can help guide a search engine in future.
- A portal can employ user profile information, expressed with regard to an ontology, to identify experts or communities of people with shared interests. We know also that much knowledge is stored on users' desks. By extending the portal with ontology mediation technology, access could be given to knowledge on the users' desktops, described using their various ontologies.

8.3 Knowledge visualisation and organisation

The use of visualisation in the design of the user interface has been considerably studied. For example, one of the most studied topics is that of changing state, e.g. when zooming in or out. Throughout such changes it is important to preserve context, e.g. by modifying the representation of the previous scene in order to provide a reference to the last state. Real world metaphors have been used in many applications. Examples are rooms, galleries, and even planetary systems.

What is important is that there are graphical elements, or indicators, for help about how to manage the applications, and that these are realistic and clear. The user must be aware of these without breaking the overall scene.

There are a number of techniques which can be applied to the visualisation of ontologies. Indeed, since Semantic Web technology is built upon the structuring of information, this offers clear opportunities for information visualisation. However, since the ontologies and metadata associated with semantic knowledge technology are intended for machine processing, visualisation using these ontologies often leads to unintuitive results. The concept of a visualisation ontology is proposed, whereby the machine-oriented data are translated into a user-friendly visualisation. Using a visualisation ontology, both the application developer and the user can define how to represent the information and how to interact with it.

8.4 User profile construction

The use of user profiles is central to the knowledge access tools described elsewhere in this section: the profile is used by other tools to select information of relevance to a given user, to deliver that information in the appropriate format and to identify user interests.

In general, the area of user profiling is divided into three main subtopics:

- Content based user profiling which uses the content of the text documents viewed by the user to build a model of content topics characteristic of that user.
- Collaborative user profiling which identifies user communities with similar interests.
- Web usage mining for user profiling which analyses usage data, usually from web-server log-files.

Once constructed, the user profile acts as a filter on the general ontology to give a personal view. The profile of an author can also help annotate a document. Ontologies themselves can be used in the representation of user profiles, and the use of ontologies in this way is a current research challenge.

8.5 Natural language generation

The aim of natural language generation (NLG) is to take structured data in a knowledge base as input and produce natural language text, tailored to the presentational context and the target reader: e.g. short text for a WAP phone, longer text for a PC, or to suit the user, e.g. simple terminology for a novice and more complex terminology for an expert. In the context of semantic knowledge management, NLG is required to provide automated explanation and documentation of ontologies and knowledge bases

NLG systems which are targetted towards Semantic Web ontologies have started to emerge only recently. There are some general purpose ontology verbalisers for RDF, DAML+OIL and OWL. A simple approach avoids the use of a lexicon. However, more fluency can be obtained through some manual input, e.g. lexicons and domain schemas. Reasoning and property hierarchies can also be used to avoid repetition, enable more generic text schemas, and perform aggregation.

More sophisticated ontology-based systems have been developed for specific domains. However, whilst these are more flexible and expressive, they are difficult to adapt by non-NLG experts. Since knowledge management and Semantic Web ontologies tend to evolve over time, an easy to maintain approach is needed.

Moreover, Semantic Web ontologies tend to have hundreds or thousands of concepts. For this reason, in SEKT we shall focus on an applied (or shallow) NLG system, the goal of which is to generate text from domain knowledge using computationally efficient and robust NLG algorithms.

Previous work on knowledge generation from relational databases has looked at low cost methods for providing adaptivity and generation of comparisons. This work is likely to be relevant to work on Semantic Web applications.

8.6 Device and content repurposing

The aim of device independence is to deliver a functional presentation of a web page on any access medium or device, e.g. PC, PDA, WAP phone and printer. The de-facto standard in this area is the W3C RDF-based CC/PP standard, and in particular the UAProf application, which is supported by most mobile phone vendors.

There are four broad approaches to device independence:
- Static adaptation, where the data are held in different formats produced by hand for each target device. This approach is necessarily resource intensive.
- Client adaptation, where the client browser is responsible for interpreting data streams in a device independent format.
- Server adaptation, where the device characteristics are communicated to the server as part of the request, and the server responds with an appropriate data stream.
- Proxy adaptation, where a software agent mediates between the server and the client, receiving data from the server and reformatting for delivery to the client.

Client adaptation is outside the scope of this project. SEKT concentrates on server and proxy adaptation. The distinction between them is somewhat blurred. Software accepting output from SEKT software and adapting it for a client could be located either on the same server, or implemented as a separate specialist web service.

Most current commercial systems and proposals are based on the idea of adding extra semantic information to the document, usually in the form of extra HTML style mark-up, which typically extends XHTML. The semantics of the mark-up are built into the software, and are used, for example, to select content, produce a suitable layout in a suitable language, and select appropriate multimedia file format.

Within SEKT a device independent framework prototype application has been developed which generalises these techniques. CC/PP attributes can be used both to select a particular layout template, and to select alternative data elements within a template.

All of these systems rely on the manual tagging of data with semantic information which can be used to select an appropriate layout. There is no attempt to search for a suitable layout.

There are a number of ways in which Semantic Web technology could be used to enhance the approach, e.g.:
- An ontology of devices would allow the writing of more sophisticated layouts for more precisely targetted devices.
- As well as using device profiles to select content and layout, personal profile information could also be used. For example, this would enable a higher priority to be given to messages from certain people, and this might influence the layout of a page.

- As already discussed, knowledge generation could be tailored to the device, e.g. knowledge summarisation could be used to generate a concise summary for a mobile phone via an SMS service.

9 Ontology engineering methodologies

Core to SEKT is the application of ontologies to knowledge management problems. Central aspects for the methodology therefore include the efficient and effective creation and maintenance of ontologies in settings such as provided by the SEKT case studies, which is the focus of this state of the art report. An overview is provided here and more detail is in Annex 7. Ontology engineering methodologies are the subject of research in workpackage 7.

It is argued that such a methodology should comprise a set of procedures for three activities:

- Ontology management activities
- Ontology development oriented activities
- Ontology support activities

The following requirements on such a methodology are identified:

- *Documenting the history of development*: it is important to document the results after each activity. In a later stage of the development process this helps to trace why certain modelling decisions have been undertaken. It is of course preferable to automate the recording process as much as possible;

- *Clear process*: each of the methodologies surveyed provides some sort of step by step “cookbook”. However, they differ in the requirements on the ontology engineer. We want to minimise the learning curve for less experienced ontology engineers: however, in most cases the methodologies are not fine-grained enough to enable the untrained person to create an ontology from scratch. This is thus identified as an important area for further research;

- *Evaluation measures*: In ontology engineering setting evaluation measures should provide means to measure the quality of the created ontology. This is particular difficult for ontologies, since modelling decisions are in most cases subjective.

It is argued that research on argumentation visualisation is particularly appropriate to address many of the requirements on the envisioned SEKT methodology and the main aspects of this topic are presented.

Existing tools are surveyed, both in the Ontology Engineering and Argumentation areas. This leads to an analysis of past and current research, resulting in the identification of a number of future research directions. It emerges that argumentation visualization is relatively mature from the research perspective.

First attempts have been made to combine findings from argumentation visualization and ontology engineering. Future research directions identified are:

- Identification of the most relevant arguments in ontological discussions.
- Support for synchronous as well as asynchronous discussions.

- Integration of argumentation visualization and ontology visualization.
- Formalisation of ontology discussions.
- The extent to which tracing of arguments enhances comprehensibility of the resulting ontology.
- Usefulness of conflict resolution strategies.
- Application of automated methods to support argumentation.
- Extent to which moderation is required or can be omitted.
- Extent to which argumentation visualization facilitates the ontology evolution process.
- Comparison of systematic distributed ontology engineering and informal approaches.

Existing KM methodologies have a centralized approach towards engineering knowledge structures requiring *knowledge engineers*, *domain experts* and others to perform various tasks such as *requirement analysis* and *interviews*. While the user group of such an ontology may be huge, the development itself is performed by a comparatively small group of domain experts, who *represent* the user community, and ontology engineers who *help structuring*.

However, early analysis of the SEKT case studies indicates that they will be

(i) **decentralized**;
(ii) rely on **evolving ontologies** to adapt to changing environments, and;
(iii) evolve ontologies in a **loosely controlled manner**.

In such decentralized settings working based on traditional, centralized knowledge management approaches becomes infeasible. While there are some technical solutions toward Peer-to-Peer knowledge management systems, traditional methodologies for creating and maintaining knowledge structures appear to become unusable like the systems they had been developed for in the first place. SEKT will work towards a methodology to support such scenarios.

Furthermore, the initial methodology will integrate specific issues of the three core technologies that are explored within SEKT (Knowledge Discovery, Human Language Technology and Ontologies & Metadata Management). This initial version will be applied and evaluated within the case studies. The methodology will be extended by capturing lessons learned and best practices. The resulting methodology will be embodied in an illustrated guidebook for implementing and applying the SEKT technology in different settings to facilitate the take-up and transfer of the technology.

10 Conclusions

In this report we have reviewed the key areas contributing to the development of semantic knowledge technologies. These are also crucial to the development of the Semantic Web. The deliverable demonstrates that members of the SEKT consortium are fully aware of how their individual areas contribute to semantic knowledge technologies; and of the challenges to be met in implementing those technologies. The deliverable also serves as a useful vehicle for increasing interdisciplinary awareness.

SEKT is contributing to the principal research themes in each technology. The report has already indicated which SEKT workpackage is investigating each technology. SEKT is also bringing together these technologies into an integrated platform, capable of being implemented in distributed form and of being scaled to commercial applications. The technologies and the platform are then being tested through three case studies, in workpackages 9 to 11. Feedback from these case studies will not only test out the viability of the technology but also how it should be applied to achieve real benefits.

Bibliography and references

[i] A translation approach to portable ontologies, T.R. Gruber, Knowledge Acquisition, 5(2): 199-220, 1993, http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html
[ii] Knowledge Representation, John F. Sowa, Brooks/Cole, 2000, ISBN 0 534-94965-7
[iii] Webster's 1913 dictionary
[iv] Advances in Knowledge Discovery and Data Mining, Fayyad, U., Grinstein, G. G. and Wierse, A. (editors) MIT Press, Cambridge, MA, 1991.
[v] Machine Learning, Mitchell, T.M., The McGraw-Hill Companies, Inc., 1997
[vi] Data Mining: Practicl Machine Learning Tools and Techniques with Java Implementations, Witten., I.H., Frank, E., Morgan Kaufmann, 1999.
[vii] Principles of Data Mining (Adaptive Computation and Machine Learning), Hand, D.J., Mannila, H., Smyth, P., MIT Press, 2001
[viii] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Hastie, T., Tibshirani, R. and Friedman, J.H., Springer Series in Statistics, Springer Verlag 2001
[ix] Pattern Classification, 2nd edition, Duda, R. O., Hart, P.E. and Stork, D.G., Wiley-Interscience, 2000
[x] Information Retrieval, Rijsberg, C. J., van, Butterworths, 1979
[xi] Adances in Automatic Text Summarization, Mani, I., Maybury, M.T. (editors), MIT Press, 1999
[xii] Foundations of Statistical Natural Language Processing, Manning, C. D., Schutze, H., The MIT Press, Cambridge, MA, 2001
[xiii] Non-standard reasoning services for the debugging of description logic terminologies, Scholoback, S., and Cornet, R., Proceedings of IJCAI 2003, 2003
[xiv] A knowledge-level account of abduction, Levesque, H. J., Proceedings of IJCAI'89, p.p. 1061-1067, 1989
[xv] Tractable reasoning via approximation, Shaerf, M., and Cadoli, M., Aritificial Intelligence: 74:249-310, 1995
[xvi] Knowledge access on the semantic web, SEKT deliverable D5.1.1, I. Thurlow, N. Kings, J. Davies, August 2004