Towards Corporate Semantic Web: Requirements and Use Cases

Technical Report TR-B-08-09

Gökhan Coskun, Ralf Heese, Markus Luczak-Rösch, Radoslaw Oldakowski, Ralph Schäfermeier and Olga Streibel

Freie Universität Berlin
Department of Mathematics and Computer Science
Networked Information Systems

1st August 2008

GEFÖRDERT VOM

Towards Corporate Semantic Web: Requirements and Use Cases

Gökhan Coskun Ralf Heese Markus Luczak-Rösch
Radoslaw Oldakowski Ralph Schäfermeier
Olga Streibel

Freie Universität Berlin
Department of Mathematics and Computer Science
Networked Information Systems
Königin-Luise-Str. 24-26
14195 Berlin, Germany
coskun,heese,luczak,oldakowski,schaef,streibel@inf.fu-berlin.de

Technical Report TR-B-08-09

1st August 2008

Abstract

In this report, we introduce our initial vision of the Corporate Semantic Web as the next step in the broad field of Semantic Web research. We identify requirements of the corporate environment and gaps between current approaches to tackle problems facing ontology engineering, semantic collaboration, and semantic search. Each of these pillars will yield innovative methods and tools during the project runtime until 2013.

Corporate ontology engineering will improve the facilitation of agile ontology engineering to lessen the costs of ontology development and, especially, maintenance. Corporate semantic collaboration focuses the human-centered aspects of knowledge management in corporate contexts. Corporate semantic search is settled on the highest application level of the three research areas and at that point it is a representative for applications working on and with the appropriately represented and delivered background knowledge.

We propose an initial layout for an integrative architecture of a Corporate Semantic Web provided by these three core pillars.

Contents

1 Introduction **3**
- 1.1 Project Background . . . 3
- 1.2 Research Areas . . . 4
  - 1.2.1 Corporate Ontology Engineering . . . 5
  - 1.2.2 Corporate Semantic Collaboration . . . 5
  - 1.2.3 Corporate Semantic Search . . . 5
- 1.3 Requirements for Corporate Semantic Web Applications . . . 6
  - 1.3.1 Costs . . . 6
  - 1.3.2 Benefits . . . 6

2 Use Cases **8**
- 2.1 Explorative Interviews and Individual Use Cases . . . 8
- 2.2 Generic Use Cases . . . 10

3 Corporate Ontology Engineering **12**
- 3.1 Motivation . . . 12
- 3.2 Requirements Analysis . . . 14
  - 3.2.1 Requirements of Corporate Ontology Engineering . . . 14
  - 3.2.2 Requirements of Ontology Versioning in Corporate Contexts 16
  - 3.2.3 Requirements of Ontology Modularization and Integration in Corporate Contexts . . . 17
- 3.3 State-of-the-Art . . . 20
  - 3.3.1 Ontology Engineering Methodologies . . . 21
  - 3.3.2 Ontology Versioning . . . 23
  - 3.3.3 Ontology Modularization . . . 24
- 3.4 The Corporate Ontology Lifecycle Methodology – COLM . . . 25
  - 3.4.1 Innovative Ontology Versioning with COLM . . . 27
  - 3.4.2 Modularization and Integration Dimensions of COLM . . . 27
- 3.5 Conclusion and Outlook . . . 28

4 Corporate Semantic Collaboration **30**
- 4.1 Collaborative Tool for Modeling Ontologies and Knowledge . . . 31
  - 4.1.1 A Light-weight Ontology Engineering Tool . . . 32
  - 4.1.2 Tool Support for Ontology Engineering . . . 34
- 4.2 Knowledge Extraction by Mining User Activities . . . 36
  - 4.2.1 Related Work . . . 36
  - 4.2.2 Requirements . . . 40
- 4.3 Conclusion And Outlook . . . 42

5 Corporate Semantic Search **44**
5.1 What is Semantic Search 44
5.1.1 Search on the Web 44
5.1.2 Features of Semantic Search 47
5.1.3 Research Directions in Semantic Search 48
5.1.4 Corporate Semantic Search 48
5.2 Search (for Complex Relations) in Non-semantic Data 50
5.2.1 Trend Mining as the Future Search Task 51
5.2.2 Case Study and Business Process Reporting 51
5.2.3 Conceptualization of a Method for Trend Mining 54
5.3 Semantic Search Personalization 57
5.3.1 Users and Requirements Analysis 58
5.3.2 Acquisition of User Profiles 61
5.4 Conclusion and Outlook 62

6 Conclusion and Outlook **63**

A Work Packages **65**

B Acknowledgement **66**

Chapter 1

Introduction

Companies overwhelmed with heterogeneous data from their intranets and with information from the Internet seek innovative approaches for managing and utilizing knowledge required for their business processes. In this regard, Semantic Web offers promising solutions for many lines of business. To facilitate those solutions, proper computational background knowledge is needed. Ontologies are this fundamental artifact in Semantic Web applications. They are suitable means for flexible, scalable, and cost-effective data integration and interoperability in information systems.

Nonetheless, the global deployment of the Semantic Web vision still remains unfulfilled, facing some unresolved problems like scalability, broader adoption of commonly shared ontologies, and trust issues. However, since the corporate world offers a controlled environment, many of these current dilemmas do not arise there: information can generally be trusted, adoption of common ontologies can be enforced more easily, and there are much looser requirements regarding scalability.

By focusing on the application of Semantic Web technologies within a controlled corporate environment we contribute to the further maturing of those technologies. Furthermore, we aim at providing enterprises with scientific and application oriented solutions for improving their competitive advantages through enhanced knowledge management of semantically rich data. The early adopters of these solutions will demonstrate incentives for further corporations to follow and thus may flow into a broader realization of a global Semantic Web.

1.1 Project Background

Corporate Semantic Web is funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions. The project took up work at the beginning of February 2008 as part of the Networked Information Systems working group[1] at the Free University of Berlin. In the next six years twelve work packages will provide a concerted framework which aims to establish economically beneficial adoption of Semantic Web technologies in corporate environments.

[1] http://www.ag-nbi.de/

The group is supported by a Scientific Review Panel (SRP) and an Industry Review Panel (IRP)[2].

1.2 Research Areas

In the era of the information society a long-lasting competitive advantage of business organizations greatly depends on the ability to create, manage, and effectively use corporate knowledge. Semantic Web Technologies offer new possibilities for enhanced integration of heterogeneous business data, information discovery as well as advanced automation of sophisticated tasks. [12, 57] The realization of Semantic Web applications, however, requires semantically rich formalization of business data based on commonly shared and well-defined concepts in form of ontologies. In a corporate setting, the process of creating and utilizing ontologies occurs in a collaborative manner, involving individuals playing different roles within business enterprises and having various degrees of domain knowledge. Once created, ontologies serve as a building block for realizing improved search and navigational functionality according to personalized user profiles.

Consequently, the focus of our research is put on the outlined components of Semantic Web applications for enterprises: **corporate ontology engineering**, **corporate semantic collaboration**, and **corporate semantic search**. In the broad field of Semantic Web research the three pillars of Corporate Semantic Web are placed as shown in Figure1.1.

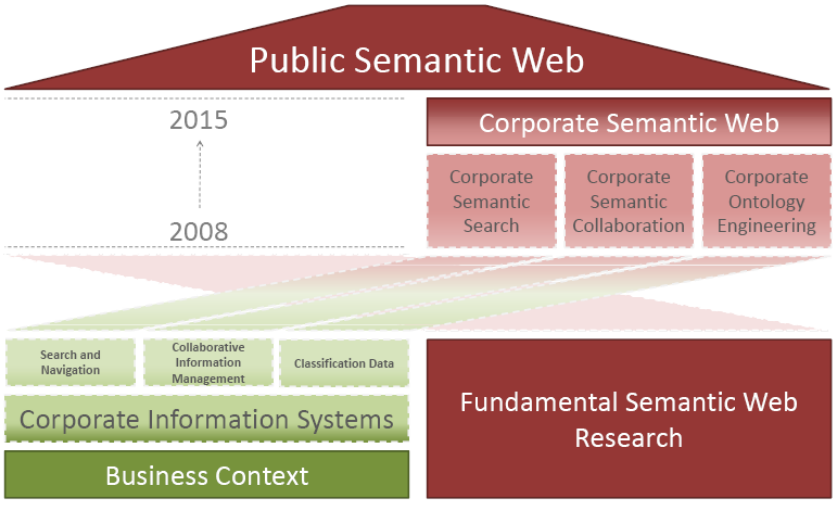

Figure 1.1: Placement of Corporate Semantic Web in research

[2]http://www.corporate-semantic-web.de/review-panel.html

1.2.1 Corporate Ontology Engineering

Ontologies build the central prerequisite for corporate semantic search and other ontology-based applications. Current approaches address ontology engineering in Web-scale settings and cover aspects of ontology lifecycle management and support, human-centered ontology engineering, and agile knowledge engineering, to name just a few key areas. However, the technological foundations do not tackle the problems and needs of corporate settings in a holistic manner. Furthermore, the economic dimensions of the process side are insufficiently researched, which play a major role in business contexts.

Our approach to a corporate ontology engineering scenario concentrates on **ontology modularization and integration (WP 9)**, **ontology versioning (WP 10)**, **ontology cost estimation models for corporation (WP 11)**, and **ontology evaluation (WP 12)**.

1.2.2 Corporate Semantic Collaboration

In corporate semantic collaboration we research methods and tools to model and evolve knowledge collaboratively and share it in a company. Current tools support collaboration only on selected datatypes, e.g., group calendar and files (document archives). An exchange of data between different applications is cumbersome and difficult. To address this problem we employ ontologies as a flexible, scalable, and cost-effective means for integrating data.

Because of the corporate context, we pay attention to the disparate skill level of ontology users and the lack of time to conceptualize knowledge or to annotate data explicitly. To reach an efficient tool support respecting this aspect, **knowledge extraction by mining user activities (WP 5)**, **collaborative tools for modeling ontologies and knowledge (WP 6)**, **dynamic access to distributed knowledge (WP 7)**, and **evolution of ontologies and knowledge by collaborative work (WP 8)** will be focused on.

1.2.3 Corporate Semantic Search

In our research on semantic search in a corporate context, we focus on the development of methods and tools that allow enhanced search and navigation in heterogeneous business data by means of utilizing semantically rich conceptualizations of relevant business domains. We aim to provide individuals playing various roles within business organizations with personalized views on corporate knowledge. This is realized by delivering search results with respect to their individual search context. An ontology based approach, when compared with traditional keyword-based ones, allows the enhancement of search quality and, moreover, opens up new opportunities for an even richer analysis of business data, for example in the form of trend recognition.

In realizing our approach to corporate semantic search we especially concentrate on **search in non-semantic data (WP 1)**, **search personalization (WP 2)**, **multimedia search (WP 3)**, and **search contextualization (WP 4)**.

1.3 Requirements for Corporate Semantic Web Applications

In order to be adopted within corporate environment Semantic Web applications must provide solutions to perceived problems or methods to exploit perceived opportunities. Mere innovation is not enough.

From the corporate perspective the introduction of Semantic Web applications must result in tangible gains like expansion of business, a wider set of business opportunities, or cost reduction of current business processes. This can be realized by providing a superior level of service. Moreover, in order to gain acceptance within enterprises, Semantic Web applications should quickly evolve into something perceived as indispensable, conferring benefits on their users without extra costs or steep learning curves [71]. Although, there are evident opportunities for knowledge-based tasks or enterprises to improve their performance once information sources are integrated and more intelligent information processing is automated, a cost-benefit analysis is in any case essential [33].

1.3.1 Costs

There are different kinds of costs enterprises are facing when planning to embrace Semantic Web technologies [71]. This requires resources for the development of smart ontology formalisms that are representationally adequate but also, which is even more challenging, their population with content of sufficient depth to provide utility in a real-world scenario [34]. Such a process, for most organizations, is associated with steep learning curves and is therefore very costly. Moreover, the migration costs include resources to support the annotation of legacy data, much of which, in corporate context, is stored in relational databases. There is a risk for many enterprises that all of those efforts may generate hefty sunk costs which may prove an extensive barrier to future change.

The ONTOCOM approach ([16],[85]) is a cost-estimation model adopting ideas from software engineering cost-estimation to ontology engineering. However, ONTOCOM does not pay attention to the higher agility of ontology engineering compared to software engineering. The authors themselves conclude in [83] that the future of ontology engineering will have to address those agile aspects focusing the work of domain experts – not ontology engineers – in the process, which will influence the cost-oriented perspective very strong.

1.3.2 Benefits

The adoption of Semantic Web technologies within business enterprises requires discernible benefits for those organizations. Such benefits may arise from better service quality leading to expansion of business, a wider set of business opportunities, or optimization of business processes. Whereas Semantic Web applications developed in academic research are mostly concerned with long-term benefits arising from network effects, business organizations are rather interested in short-/medium-term and individual gains, independent of any future network effects. This, however, does not mean that network effects should be overlooked.

The areas within organizations which would benefit most from the adoption of Semantic Web technologies are data integration and semantic search [11, 57], which, as argued, could be accommodated with technologies for knowledge extraction, ontology development and mapping. An integrated information system would be able to manipulate data from heterogeneous corporate sources and use background knowledge represented in ontologies to make inferences that were not possible before.

Moreover, semantically rich data can be matched against statements representing personalized profiles to generate recommendations or targeted products. Since customers tend to be more loyal to personalized services [33], such increase in service quality is likely to create another benefit in form of a competitive advantage.

Chapter 2

Use Cases

In this chapter we describe the industrial use cases which were identified during the kick off phase of the project Corporate Semantic Web. We start with an introduction on how we proceeded to focus on real world use cases of the cooperation partners which can be generalized to a large scale of various companies. We subsume all specific use cases and bring them together into three of those generic use cases which will be addressed by future work of the project.

2.1 Explorative Interviews and Individual Use Cases

To facilitate research which is guided by real industrial use case scenarios, each of the companies from the industrial review panel was visited once during the kick-off phase of the project. The researchers presented fundamental facts and related chances of Semantic Web technologies in industrial applications. Afterwards a discussion with leading members of the companies was held which should help to identify relevant use cases. All use cases in this section are described on a very low detailed level, because the goal was not to solve specific problems of single companies but to find out which are the main and generic barriers for companies to raise the effort spent on research and development on Semantic Web technologies.

The Trouble ticketing Use Case

The projektron GmbH is the software producer of a Web-based project management solution. The customers are from various industrial sectors. A central component of the product is a trouble ticketing system which is also in use at the projectron GmbH itself.

A major problem of the ticketing system is that the forms of the frontend and the backend contain free text fields at most. Differences in the terminology of the internal employees, the customers and the dedicated terminology of the software for core concepts yield communication problems and thus raise the time spent on identification of problems. Consider this example: the term task, which is used in the system, may be translated to job or issue in the description

of a trouble ticket and an internal employee is not able to match these terms immediately.

We expect a large benefit of Semantic Web technologies in this case regarding the computation of related and similar trouble tickets for the backend user. To reach this, the existing classifications have to be translated into ontologies which have to evolve consistently with the software system versions.

Intelligent Media Use Case

The Condat AG is a big IT service provider with a broad range of offers. A core sector is the media and television sector where the Condat AG provides solutions for broadcast and program management. In general the company addresses problems of information personalization, integration, and distribution.

With the upcoming MPEG-7 standard and the chance to annotate multimedia sequences in combination with EPG, IPTV, and program management systems it is possible to improve the categorization of contents and topics.

In an initial step Semantic Web technologies can be used to provide a matching between EPG and personal user profiles. As a result of that it is possible to compute user specific recommendations.

Collaborative Ontology Building Use Case

The Semtation GmbH is a company which produces and sells a software extension for Microsoft Visio which supports ontology-based modeling of business processes. An additional background model in form of an ontology facilitates consistence checking for collaboratively built process models.

To build the initial model an ontology engineer creates a corporate domain model. The model is refined by a team of corporate domain experts which discuss design decisions and term proposals. Even though an ontology is used for the background model, the modeling which the domain experts perform is limited to maintenance of a taxonomy. They do not have to deal with relations and properties. However, a typical process of collaborative ontology engineering runs up, which is not structured or supported by any tools for finding and documenting the consensus, yet.

Since ontologies are already in use in the background of the SemTalk software, we expect an impact by a simplification and structuring of the ontology refinement. For this purpose a tool for ontology editing is needed, which allows unskilled people to edit ontology primitives directly.

Innovative Search Use Case

The neofonie GmbH is a company which focuses research and development in the field of search technologies. As an innovative sector neofonie identified trend mining methods and methods for search for complex relations in texts.

Search in huge and heterogeneous data repositories or on the Web is a challenging task. It is impossible to ensure that semantically enriched data exists immediately in companies when Semantic Web technologies are adopted in any possible way. Thus, hybrid approaches will become relevant for efficient and satisfying search in corporate environments.

An initial step in the sector of trend mining with those systems was recently made. We expect from the used technologies, that they provide benefit for acquiring and integrating knowledge for corporate ontologies beyond this aspect as well. Yet, such generic hybrid search components are missing.

Knowledge Integration Use Case

The EsPresto GmbH is an IT service provider which focuses the corporate knowledge management sector. As a core offer EsPresto uses Web 2.0 technologies such as blogs and wikis to facilitate flexible, dynamic, and distributed knowledge management. As a core component a search engine indexes information from all of the systems in use.

EsPresto offers a tagging component as part of the Web 2.0 knowledge management portfolio. This system is already enabled to make use of ontologies for tag recommendations. The core problem, which makes this approach less successful compared to conventional tagging, is the fact that the costs for the development of an appropriate individual ontology for each customer are too high. The benefit of Web 2.0 technologies are the low initial investments thus they may be used in a small project in a company for testing purposes until they are scaled on the whole company.

In this case a solution is needed which allows the agile evolution of ontologies. By this it may be possible to release a very small and possibly generic prototype ontology which is in use in a part of a company before it has to be scaled to represent the whole corporate knowledge.

Further Cooperation Partners

Since it is a major goal of the project to build up a cooperation and innovation network for semantic technologies in the eastern German regions, various efforts have been made to raise the initial circle of the five partner companies from the industrial review panel. As a result of information published online, the Merck Serono KGaA contacted the project. At this point both parties prepare jointly a project proposal for the concrete cooperation.

Furthermore, the Ontonym GmbH, a spin-off organization of the Freie Universität Berlin, became a full cooperation partner. The core business of Ontonym is a semantic search service. Based on ontologies, query expansion and matching methods are performed to improve the quality of search results. The ontologies are evolved manually by an evaluation of queries which contain unknown terms. Ontonym will be the evaluation use case for our innovative ontology engineering methodology COLM.

2.2 Generic Use Cases

At the beginning of this chapter we explained what we wanted to achieve by the interviews with the cooperation partners of the project. The goal was to find generic use cases which can be used to show which gaps have to be closed to enable a broader success of Semantic Web technologies for small and mid-sized companies. We introduced at least four individual use cases which have influenced our definition of three global and generic use case which will be briefly

described in this section – the product improvement use case, the usability use case, and the knowledge integration use case.

The Product Improvement Use Case

Small and mid-sized companies face two problems. First, they are not very familiar with Semantic Web technologies yet. Second, they often do not have very flexible capabilities of investing in research and development of innovative technologies because their core products have to be maintained and supported. We recognized a lack of components and methods to integrate ontologies in existing software and a gap between the orientation of ontology engineering methodologies and the needs of running software lifecycles.

The Usability Use Case

Companies which already use ontologies as appropriate means for knowledge representation in their software products face the problem that cost-effective evolution of the models is not possible without transferring ontology refinement tasks to the customers themselves. That limits the possibility of creating complex and fine granular ontologies because of the limited knowledge of unexperienced users of ontology primitives. Ontology development tools are needed which close this gap between technical terminologies and intuitive human understanding of knowledge models.

The Knowledge Integration Use Case

Since we mentioned that it is one problem to facilitate the adoption of Semantic Web technologies for innovative solutions in parallel to the conventional products of software producing companies it is another huge problem to introduce those ontology-based systems in corporate information systems infrastructures at all. Because of the high effort which is needed to implement corporate domain ontologies the cost-benefit-ratio is unclear and in most cases tackled by the efficiency of conventional developed software systems. Methodologies and tools are needed which allow the cost-effective implementation and integration of ontology-based systems in existing corporate infrastructures without touching the users workflows too much.

Chapter 3

Corporate Ontology Engineering

In this chapter we give an introduction to ontology engineering in general. Afterwards, we raise requirements for this task in corporate contexts and compare the state of the art methodologies against these objectives. We identify ontology versioning and ontology modularization as important areas when starting to apply ontologies in corporate information system infrastructures. Finally, we describe an innovative methodology for corporate ontology engineering, which respects these two aspects. This chapter covers the work packages WP 9 and WP 10.

3.1 Motivation

Semantic Web aims at bringing semantics to the World Wide Web. Adding meaning to plain text requires understanding of concepts. Enabling machines to understand concepts is not possible without a machine-understandable representation of concepts and their relationships. For that purpose ontologies were introduced. Following Gruber [40], an ontology is defined as an "explicit specification of a conceptualization" and "Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner". Recently researched and developed methodologies have addressed the problem of engineering those consensual ontologies in a collaborative way by a group of people who combine technical knowledge about ontology languages, theoretical knowledge about the deduction of explicit and implicit knowledge from descriptions, and practical knowledge about the domain. But, especially the latter aspect of the definition has tackled the broad success of ontologies on the scale of the World Wide Web (WWW), yet. It is very difficult, if not impossible, to reach ontology consensus and force the use of a manageable amount of standard ontologies which allow the computational representation of nearly everything which exists in the world.

There are recently four bottlenecks identified [48] why valuable ontologies are rarely found on the Web. That ontologies and thus Semantic Web applications have not reached industrial application as well, is caused by several other reasons which are cost- and process-oriented problems [84]. Since the Technol-

ogy Roadmap for the Semantic Web [30] introduces "Semantic Web Business Gaps" in general, we combine both arguments mentioned before to raise the following four gaps between corporate contexts and ontology engineering:

The academic orientation gap: Research on ontologies in information systems is a new field, compared to other disciplines of computer science, e.g., research on databases. Tools and use cases are oriented to scientific-formed problems. Only few early adopters in the areas of enterprise information integration, content management, life sciences, and government allow the construction of real-world use cases. But, the formulated goals of lasting developments and advanced outreach to industry [30] rarely take place, yet. That happens because of missing sustained financial bases beyond funding.

Ontologies are more or less treated as the background artifact of Semantic Web applications. That is far from promoting them as flexible means for knowledge representation and data integration and as a result of that the processes and tools rarely follow usability needs for end-users, but address developers with an academic background.

The application maturity gap: In consequence of the *academic orientation gap* only very few tools reach a mature development state and become ready for productive use, such as the Protégé tool[1] or Virtuoso server[2]. If there are such tools, an adequate comparability is missing because only few valuable benchmarks exist for those (e.g. [15]) and transparent standards are not reached yet. This fact yields a strong and reasonable suspiciousness in computationally handling of large-scale ontology-based applications.

The process gap: Beyond the gaps referring to the academic orientation of the research in general and the immaturity of the applications and tools the process gap is an important aspect for missing acceptance of ontologies in the industry. Most of the developed methodologies address ontology engineering for the Web. That means they aim at the development of ontologies for world wide or at least inter-corporate application domains, e.g., [63, 87, 76, 88]. Regarding individual knowledge as a corporate competency with intra-corporate applications and a demand of security for this knowledge is out of scope.

Process-sided the methodologies have developed from ontology engineering as an individual discipline without any application-dependence and an ontology as the output artifact towards a discipline which tries to lessen the gap to software engineering because an ontology rarely comes along without any application. However, the agility of knowledge engineering processes is exceeding the agility of software engineering processes and the whole contextual environment influences the ontology evolution. Methodologies which respect these aspects are needed but not developed so far. Several case studies for existing and well-researched methodologies try to proof the applicability of them [91] but are not adequate regarding real-life problems in the corporate context. The two most promising areas where early adopters use ontologies beneficially are life sciences and health care. However, both areas are representants for domains which con-

[1] `http://protege.stanford.edu/`
[2] `http://virtuoso.openlinksw.com/`

tain a high level of structured or standardized models which is different from most of the other corporate domains.

The cost-benefit-estimation gap: Finally, there is a strong need for companies to estimate the cost-benefit-ratio of an investment in information infrastructures. Since the current methodologies disrespect the increased agility of knowledge evolution and since cost-estimation models are derived from software engineering approaches, this problem is unsolved.

Estimating the cost-benefit-ratio of an ontology-based information system is hard because an individually engineered ontology has to integrate all parts of corporate knowledge while the application which uses that ontology only fulfills some special purposes. The effort for the ontology engineering task is hard to be valorized by only one application. Because of that it is hard to start the application of ontology-based information systems as a step-by-step process, in contrast to conventional information systems, e.g., Web 2.0 tools. A reliable set of standard ontologies and measures for proper ontology selection for special needs and requirements is missing because of a lack of knowledge about evaluating ontologies against specific objectives.

3.2 Requirements Analysis

Previous ontology engineering methodologies aimed at developing ontologies for the World Wide Web, which is a highly dynamic, large-scale, open, and heterogeneous system. These characteristics are huge burdens for the creation and maintenance of ontologies. They exacerbate the design of widely applicable methodologies and tools. Limiting the targeted system to a controlled environment as the corporate environment allows focusing on different aspects of ontology creation and evolution. Thus, a new ontology engineering methodology is required which concentrates less on the general applicability but more on the corporate wide usability and efficiency. The main requirements for such a methodology are derived from the aspects of the corporate environment which are very different to the previous methodologies' targeted systems. Thus, a close look at the corporate environment is necessary in order to get a detailed picture about the context in which the ontology is going to be used, so we can derive requirements which have to be fulfilled by a methodology which aims at ontology engineering for corporate environment.

3.2.1 Requirements of Corporate Ontology Engineering

As mentioned before "Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner". In the World Wide Web it is very difficult to create a shared vocabulary which is agreed by all participants. In contrast, a company consists of domain experts with very specialized knowledge. These experts are communicating personally in meetings and compose joint documents. Domain specific terms can be agreed upon corporate wide which simplifies the process of finding a shared vocabulary. In fact, the understanding of terms are mostly company wide unambiguous. And in case of disagreement between employees the company internal hierarchy allows the definition of terms which has to be accepted by all. In this context an ontology

for a corporate environment is not primarily needed to find a consensus as in the World Wide Web but rather needed to make the already existing and shared understanding of terms explicit. In this regard the most important challenge for corporate ontology engineering is to make employees' expertise and achievements explicit and make this knowledge company wide usable by employees and applications.

However, there is a case in which finding a consensus is also beneficial for a company, because companies might comprise of departments with different competencies, views, and permissions. Although the tasks and targets of departments differ in respect of content there might be some similarities, for example in used methods or processes. Explicit knowledge about the work of departments in form of ontologies would allow well-designed applications to compare the processes and achievements and would also support efficient communication between members with different competences. This would lead to the ability of detecting redundancies and overlappings within the company, which in turn would increase the flexibility and efficiency of the company.

Besides making the expertises of the employees and their work explicit by using ontologies it is also possible to represent knowledge about the company itself and about its employees. This is an additional benefit of using ontologies within a company. Making administrative knowledge as project information and user profiles explicit in a semantic fashion would enable management support by tools and personalized access to corporate information. For example, a manager could be assisted by management tools based upon ontological representation of administrative knowledge by building teams for projects.

Using ontologies in a company

- makes implicitly shared understanding/knowledge explicit,
- supports tools to increase the productivity,
- makes communication between departments more efficient,
- makes work of different departments comparable,
- simplifies administration.

Having discussed the potential benefits of introducing ontologies into a corporate environment it is important to check which requirements have to be fulfilled by new approaches and technologies to be adopted by a company. This is due to the fact that potential benefits are not sufficient arguments for a company because an adoption of a novelty comes with the change of an already existing and running system and brings always an additional cost at least at the very first phase for the introduction of the novelty, which is called the Capital Expenditure (CAPEX).

The incentives which direct a company are very different from the incentives of the World Wide Web, which is the basis for the first Semantic Web vision. Getting better applications as well as services and improving users' experiences drive the evolution of the Internet. It is a kind of a social network in which users are interacting in an uncontrolled environment. Novelties are accepted due to its improving impact on users' experiences. Unlike the World Wide Web companies' incentives are increasing the profit and ensuring their existence. As soon as novelties do not obey these principles they will not be introduced.

Representing knowledge by ontologies for the aforementioned goals presumes their existence. But the development of ontologies is a very cumbersome and time-consuming task. For that reason it is very important to create an ontology development methodology that keeps the corporate context in view and takes account of the aforementioned business incentives by adopting novelties.

The outcome of this are the following requirements for the corporate ontology engineering methodology:

- The influence of the ontology development and maintenance process on the work flow of domain experts have to be minimized to avoid negative influence on their productivity.
- The already existing and running system must not be disturbed.
- The need for ontology engineers have to be minimized to reduce costs.
- The ontology has to evolve with the progress of the company.

3.2.2 Requirements of Ontology Versioning in Corporate Contexts

As a result of the specific characteristics of corporate ontology engineering settings, ontology versioning has to provide specific requirements as well. The three core aspects which have to be fulfilled are:

Application-dependence: Ontologies are not useful without any application. In corporate contexts we can state that an ontology never comes along without any application. Moreover, to allow a beneficial adaption it is used by more than one application at the best. During the development time of an ontology this fact can be controlled by a specific set of competency questions. A version control system for ontologies in corporate settings has to respect changing or new competency questions with reference to changing or new facts.

Contextualized environment: In a company, the information systems infrastructure is broad and heterogeneous. Various specific applications yield various specific data in form of files or databases for example. This data in combination with the workflows the users follow to perform specific tasks build a contextualized setting for the evolving corporate knowledge which should integrate all of these knowledge sources. Thus, the evolution of corporate ontologies depends on the evolution of the whole information systems infrastructure.

Cost-sensitivity: The evolution of ontologies is the concealed cost-driver in ontology-based information systems since ontology engineering methodologies do not provide a lightweight approach to this task. A configurable measure for version necessity detection should allow the control and avoidance of cost-intensive evolution steps.

In the corporate context the need for reliable facts delivered by a knowledge base is a very important aspect. Dynamically evolving knowledge can tackle the consistency of ontologies in a very unpredictable way. Thus, a proper versioning

mechanism for ontologies in the corporate context has to pay a lot of attention to the distribution of consistent versions of the ontologies to all applications in the contextualized environment which make use of them. This problem gets even more stronger when coexisting views of a knowledge domain are in use at the same time evolving unintegratedly. An example evolution of an initial ontology version 1 is depicted in Figure 3.1.

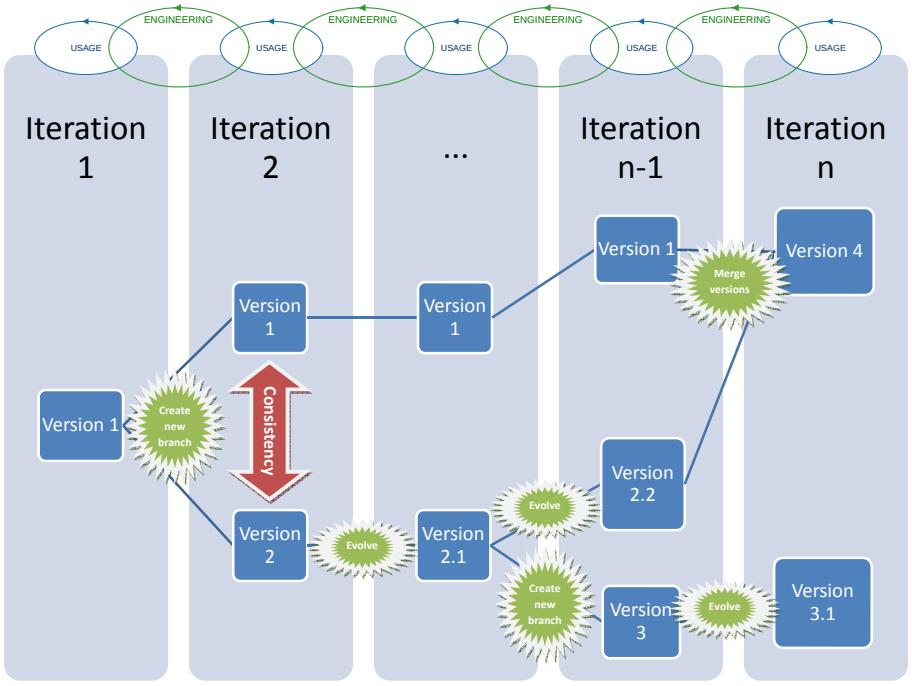

Figure 3.1: Simplified schema of unintegratedly evolving, coexisting ontology versions

3.2.3 Requirements of Ontology Modularization and Integration in Corporate Contexts

In "Understanding Ontological Engineering" [31] Vladan Devedzic mentions the following features as desirable qualities for ontologies:

- decomposable
- extensible
- maintainable
- modular
- tied to the information being analyzed
- universally understood
- translatable

Keeping the corporate context in view not every feature is equally important. The latter two features are not very important for an ontology to be used within a company. In contrast, being maintainable, extensible, decomposable, and modular are very essential. Due to the dynamics of knowledge within a company, extensibility of ontology is necessary to reflect new achievements and make them usable for further processes. In order to keep the operational expenditure (OPEX) as little as possible, easily maintainable ontologies are very important.

A company consists of various competence centers with different expertise. A system which has to represent that knowledge should follow this structure and consist of different parts, which we call modules, with special knowledge representing that expertise. How the different interfaces of competences look like and how cooperation takes place is part of the knowledge of the management, which could be represented by a higher level ontology (as shown in Figure3.2).

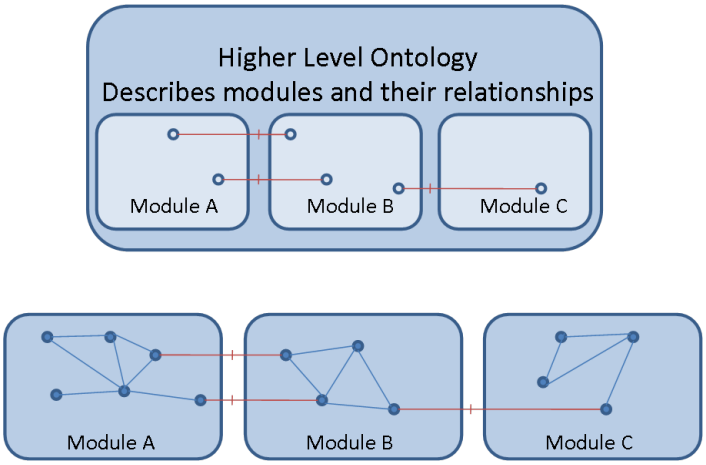

Figure 3.2: Ontology Modularization Structure

In fact, modularization is the most important feature of corporate ontologies, because a modular structured ontology will also have a positive impact on the maintainability and the extensibility. It is easy to understand that maintaining and extending smaller ontologies is much easier than large scale ontologies, which aim at representing the knowledge of a company in an holistic manner.

Due to different understandings of the term modularization in literature there is a need to clarify how this concept is understood in this work. D'Aquin et al. define modularization in [28] as follows: "From an ontology engineering perspective, modularization should be considered as a way to structure ontologies, meaning that the construction of a large ontology should be based on the combination of self-contained, independent and reusable knowledge components."

In contrast, within the NeOn Project[3] Modularization is defined as follows. "Ontology Modularization refers to the activity of identifying one or more modules in an ontology with the purpose of supporting reuse or maintenance. Note: We can make distinctions between: Ontology Module Extraction and Ontology Partitioning. Ontology Module Extraction refers to the activity of obtaining

[3]www.neon-project.org, NeOn Glossary of Activities to references

from an ontology concrete modules to be used for a particular purpose (e.g. to contain a particular sub-vocabulary of the original ontology). Ontology Partitioning refers to the activity of dividing an ontology into a set of (not necessary disjoint) modules that together form an ontology and that can be treated separately."

While d'Aquin et al. [28] describes modularization as a "way to structure" and defines it more as a design principle, within the NeOn Project modularization is seen as a process on an already existing ontology.

Summing up, modularization can be divided into two main streams as illustrated in Figure 3.3. On the one hand it is understood as a *design principle* or an approach on developing ontologies. Similar to component-based software engineering it proposes to design ontologies by breaking down the complexity in smaller pieces and create small self-contained ontologies, called modules, which are used in order to compose more complex ontologies. On the other hand modularization is understood as kind of *ontology process* which can be executed on ontologies to create smaller ones out of it. This understanding is in turn divided into two categories. Partitioning ontologies is a process in which a bigger ontology is divided into smaller pieces like a puzzle. This is mostly done in a formal way by logical means to achieve small self-contained modules which can be put together to compose the original ontology. The second category, called module extraction, is more an application oriented activity which aims at extracting a particular part of an ontology mostly for a special use case. This means, that some content of the original ontology might be lost.

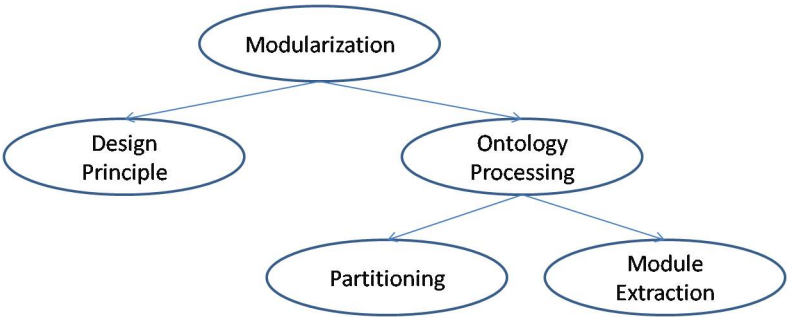

Figure 3.3: Understandings of Modularization

Having clarified the understanding of the term modularization it is reasonable to take a look at the goals for which modularization is used. In [32] the goals of modularization are defined as follows

- ontology reuse
- scalability for information retrieval
- scalability for evolution and maintenance
- complexity management
- understandability
- personalization

From the corporate context perspective all of these goals are very important. Ontology design from scratch and the maintenance of existing ontologies are difficult tasks. For that reason reusing existing ontologies is preferable. But in most cases existing ontologies are not optimized for the targeted domain. They need to be customized. It must be possible to extract relevant parts of an existing ontology and use them as building blocks by creating a useful ontology. There are two concepts for unifying two ontologies to a single one, namely ontology integration and ontology merging [74].

In the NeOn Project[4] ontology integration is defined in the following way "Ontology Integration refers to the activity of including one ontology in another ontology."

Pinto et al. refer to ontology integration and ontology merge in [75] as two different ontology reuse processes. The meaning of both terms is distinguished by the following definitions "Merge is the process of building an ontology in one subject reusing two or more different ontologies on that subject. In a merge process source ontologies are unified into a single one. In a merge process source ontologies are truly different ontologies and not simple revisions, improvements or variations of the same ontology. In an integration process one can identify in the resulting ontology regions that were taken from the integrated ontologies. Knowledge in those regions was left more or less unchanged."

The possibility to identify parts within the resulting ontology which were taken from modules is a valuable aspect of integration which allows for future decomposition which was identified as a desired quality of ontologies. Additionally, in case of new versions of original ontologies which were previously modularized, it is easier to integrate these updates because of the possibility to separate the modules, on which those updates have an effect.

To sum up, a methodology for creating corporate ontologies must respect the following requirement from ontology modularization and integration perspective:

- modular design
- knowledge discovery
- knowledge selection
- module extraction
- ontology / module integration / merge
- maintenance and extension on module level

3.3 State-of-the-Art

The range of ontology engineering methodologies widened during the last years. The approaches mostly differ in details referring to the composition of ontology engineering and application development, the range of users interacting in ontology engineering tasks, and the degree of lifecycle support. While some

[4]www.neon-project.org, NeOn Glossary of Activities to references

methodologies assume ontology experts or at least knowledge workers with little technical experience, others also address users with no technical experience at all.

This section presents a brief overview about recent and well-known ontology engineering methodologies. A range of approaches for ontology versioning and ontology modularization have been developed, which we introduce afterwards as well.

3.3.1 Ontology Engineering Methodologies

Our overview will briefly introduce the approaches of *METHONTOLOGY*, On-To-Knowledge (*OTK*), *DILIGENT*, *RapidOWL*, and the *NeOn methodology*.

METHONTOLOGY [63] is a concept-oriented approach to build ontologies from scratch, reuse existing ontologies or re-engineer knowledge. It has adopted the central structure of its process from the IEEE standard for software engineering, which was assumed as being more mature than any new developed process for this special purpose. That yields the three central activities of management activities (scheduling, control and quality assurance), development (specification, conceptualization, formalization, implementation, and maintenance), and support (knowledge acquisition, evaluation, integration, documentation, configuration). These activities underrun a cyclic lifecycle which allows for the iterative release of evolving ontology prototypes.

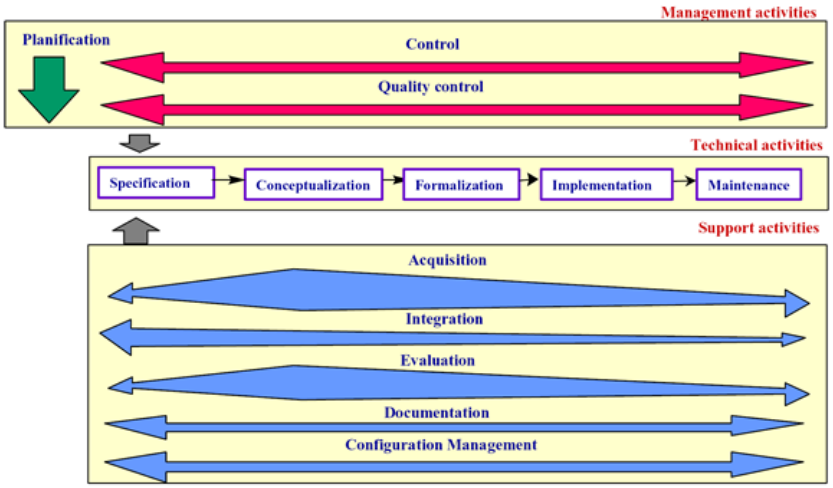

Figure 3.4: The methodology development process [99]

OTK [87] is less concept-oriented because it has an application-dependent focus on ontology engineering. That means software engineering and ontology engineering processes should be aligned. Thus participants which are not very familiar with ontologies appear as stakeholders in early phases of the process for

identification of use cases and competency questions. Mindmaps are proposed as lightweight means for these tasks. *OTK* introduces a centralized as well as a distributed strategy for ontology maintenance but lacks a detailed description or evaluation of both strategies.

A totally loosely controlled process, which respects the high distribution of ontology engineering in web-scaled settings and the totally disparate skill level of process stakeholders is introduced by *DILIGENT* [76]. An ontology consensus is reached by an argumentation-based approach following a dedicated argumentation model (the DILIGENT argumentation ontology). Every individual is free in adapting the central ontology consensus locally and modify this adaption for its own purposes. The evolution of the consensual model is depending on these local adaptions. A lifecycle is underrun in *DILIGENT*, which enables an iterative evolution of the central consensual ontology while the detailed process phases (build, local adaption, analysis, revision, local update) concentrate on reaching this human-centered consensus by argumentations about concept.

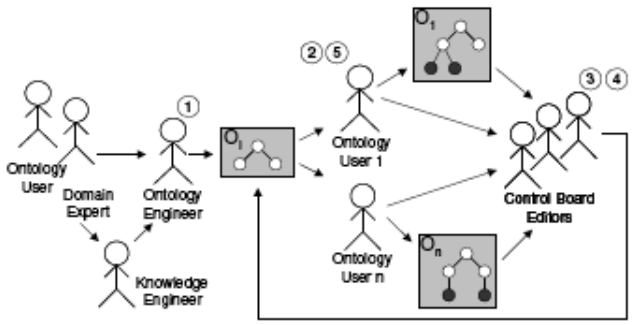

Figure 3.5: The DILIGENT setting of roles and functions [76]

Since software engineering practices have been adopted for ontology engineering the well-thought approach of agile engineering has come into focus of research in this field as well. In [6] *RapidOWL* is introduced as an idea of agile ontology engineering. Auer proposes a paradigm-based approach without any phase model. *RapidOWL* is designed to enable the contribution of a knowledge base by domain experts even in absence of experienced knowledge engineers. However, the view on ontology usage is limited to the rapid-feedback, which is nonspecific referring to the stakeholder who gives it and how it is given.

The most recent developed ontology engineering methodology is the *NeOn methodology* [88]. Starting from the assumption that former methodologies provide detailed process description but lack the appropriate "style and granularity" as it is known from software engineering methodologies. The *NeOn methodology* attends to facilitate guidelines for building individual ontologies by reuse and re-engineering of other domain ontologies or knowledge resources and for plugging in continuously evolving ontologies. Addressed target groups of the methodology are software developers as well as ontology practitioners which should be enabled to build ontology networks by use of ontology building platforms (e.g. NeOn Toolkit, Protégé, or TopBraid Composer). The definition

of process phases and activities is accompanied by a description of purposes, inputs and outputs, actors involved, methods, techniques, and tools used for their execution.

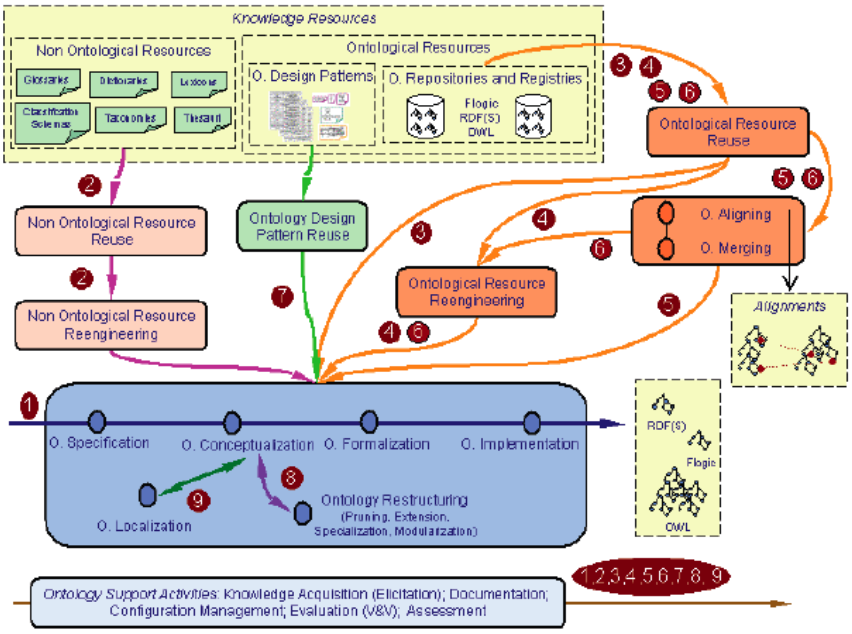

Figure 3.6: Ontology development processes in the NeOn methodology [88]

3.3.2 Ontology Versioning

Ontology versioning is defined as "the ability to manage ontology changes and their effects by creating and maintaining different variants of the ontology" [52]. Adopting this definition the core tasks for proper ontology versioning come clear as being (1) change management, (2) ontological diff and (3) transparent maintenance of coexisting ontology versions.

An early approach towards ontology version was mentioned in [46]. Helflin and Hendler's technique solved the problem of maintaining coexisting older versions of an ontology by simply maintaining each version as a separate web page, which means that each version has a specific unique URL. To enable the detection of backward-compatibility between the versions the BACKWARD-COMPATIBLE-WITH field was specified. That contains the danger that even inofficial revisions are stated as backward-compatible with even though no one has specified it as original.

Noy and Musen describe the problem of comparing two ontology versions which is strongly different from textual comparison of software source codes and needs the definition of a structural diff between ontologies [67]. This structural diff describes which frames have changed between two versions and in-

cludes"not only what has changed but also some information on how the frames have changed".

Since the structural diff only takes the ontology language non-specific semantics into account the semantic diff as introduced in [95] addresses the language specific semantics. By extending the SemVersion tool with individual semantic diff implementations a flexible version control is possible which provides not only straightforward evolution of ontologies but branch and merge operations and conflict detection between different versions as well.

To facilitate an expressive and flexible way of describing and accessing ontology versions, ontology metadata vocabularies, such as OMV [44], are used. Thus it is possible to process metadata about the language non-specific conceptualization and metadata about a concrete implementation. Ontology metadata can contain provenance information about the process of reaching the ontology consensus as well.

3.3.3 Ontology Modularization

Regarding modularization in the sense of ontology processing, there are different approaches within literature which are divided into three categories by [82]: query-based methods, network partitioning, and extraction by traversal.

The first category is called *query-based methods*, that is an approach which was inspired from the database field. Ontological queries similar to SQL, which is used to formulate a database query, lead to responses in form of sub-ontology segments. The query language SPARQL [77], KAON view [96], and RVL [59] are examples for this approach. Application of query-based methods makes sense for extracting very small pieces for one time use, without the need to access the semantics of the original ontology.

Network partitioning is the second category, which follows the idea that any system has the property of near-complete decomposability. In this sense, regarding to ontologies, it must possible to find groups of objects which have a closer relationship to each other than to the other objects. By comparing ontologies to networks of connected nodes, through interpreting class hierarchy as an acyclic and directed graph and the relations as links, it is possible to apply segmentation algorithms from the networking area to ontologies. In contrast to the other approaches, this approach is only based upon the ontology and its structure itself, without any other input. Additionally it considers the ontology as a whole and partitions it completely by trying to keep the original semantic as complete as possible. Structure-based partitioning [86], Automated Partitioning using E-connections [37], and Snark [58] are some examples for this approach.

The third category is called *extraction by traversal*. It is similar to the second approach because it interprets the ontology as a network, too. The difference is, that this approach starts at a particular concept, which has to be given as an input, and extracts a module, which represents a focused view on the given concept. The original ontology stays untouched, but a new sub-ontology will be created. In that sense it is more an extraction than a partitioning. Two examples for this approach are [69] and [14].

3.4 The Corporate Ontology Lifecycle Methodology – COLM

In the last sections we presented four gaps between industrial needs and currently reached status of ontology engineering research, requirements of corporate ontology engineering, an overview of well-researched state-of-the-art methodologies for ontology engineering, and methods and tools which tackle the challenges of ontology modularization, integration and versioning. The following table shows which methodology fits which part of our requirements:

Requirements	METH	OTK	DIL	ROWL	NeOn
Workflow integration	-	-	-	-	-
Independant system integration	+	+	+	+	+
Engineering reduction	-	-	-	+	-
Context aware evolution	-	-	-	-	-

We conclude that the requirements of small- and mid-scale enterprises to ontologies and ontology engineering cannot be fulfilled by existing methodologies, methods and tools. Moreover, we state that two central aspects for cost-sensitive engineering are workflow integration and evolution which both are not provided by any of the existing methodologies. Thus, we derive a new point of view on ontology engineering processes in corporate settings, which we introduce by the presentation of an innovative ontology lifecycle in this section.

From our perspective the evolution of knowledge is the basal entity of ontology lifecycles and that it is strongly dependent on the usage by unexperienced people with lack of time to note, annotate, or feedback explicitly. Recent ontology engineering methodologies focus well the initial build of ontologies and present process models for iteratively evolving ontologies. However, they rarely present any agile processes, methods or tools focusing the latter. The weight influence of knowledge evolution to the cost-benefit-ratio of knowledge-based systems is a non-rejectable fact in the agile context of knowledge management. Corporate knowledge evolves indefinitely.

Altogether, we conclude the need of a new ontology lifecycle model for ontologies in corporate contexts. The model should allow an intuitive understanding of raising costs per iteration. Because of the change in the environment complexity from Web-scale to corporate-scale, we assume that it is possible to converge ideas of agile software engineering and ontology engineering. But, for this purpose it is necessary to change the definition of what is assumed as being agile.

Recent approaches such as RapidOWL focus the agile paradigms *value*, *principle*, and *practice* as a development philosophy. That accompanies agile software engineering as it is intended in the *Agile Manifesto*[5]. This focus is limited to engineering tasks, while usage is factored out and does not present or follow any lifecycle model at all. It comes clear, that changing requirements of an application, which uses the knowledge base, over time are only one agile aspect influencing ontology prototype evolution. Another is the dynamic of knowledge referring to the evolving dimensions of data and user activities depending on processes and workflows.

[5] `http://agilemanifesto.org/`

Our proposed methodology – The Corporate Ontology Lifecycle Methodology (COLM)[6] is in a very early stage. Until now we have developed eight phases of a two-part lifecycle which is depicted in Figure 3.7. They refer either to the outer cycle as creation/selection, validation, evaluation, evolution/forward engineering or to the inner circle as population, deployment, feedback tracking, and synchronization. The outer cycle represents pure engineering, which is an expert-oriented environmental process. The inner constitutes the ontology usage, which is a human-centered concurrent process. Indeed, the fact, that other methodologies agree on the existence of pre- and post-development processes, such as documentation or support, is not explicitly shown in our lifecycle model, yet. But, since the approach aims at the construction of an architecture which provides a cost-sensitive evolution of knowledge in form of ontologies, we think that these activities are part of the engineering cycle. That means they are either provided by a contractor (purchase) or a dedicated corporate unit (creation/selection). Both possibilities will be described in detail as part of future work.

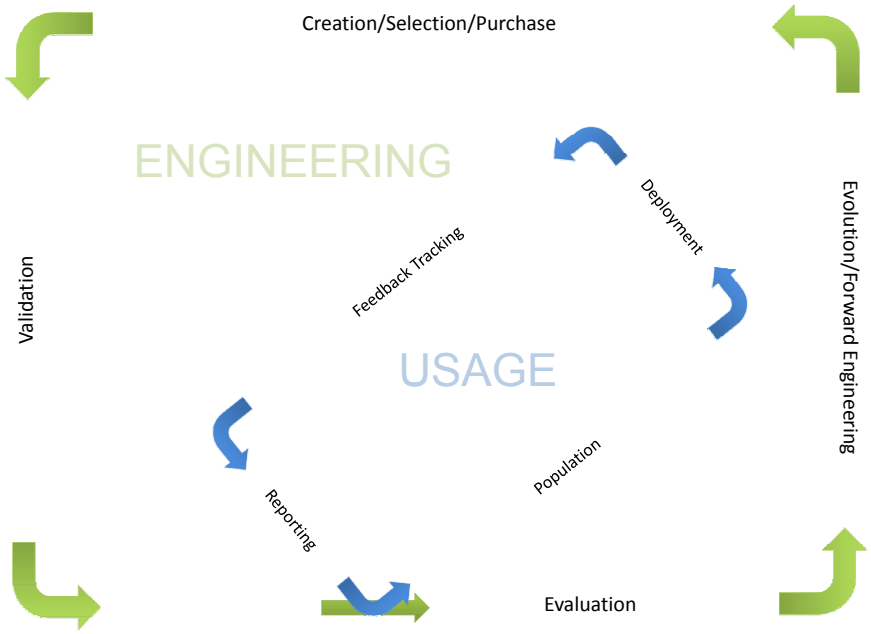

Figure 3.7: The Corporate Ontology Lifecycle

Starting the process at *creation/selection* means to start the knowledge acquisition and conceptualization, to re-use or re-engineer existing ontologies, or to commission a contractor to develop an ontology. The result of this phase is an ontology, which is *validated* against the objectives. At the intersection point between the engineering and the usage cycles the ontology engineers and the domain experts decide whether the ontology suits the requirements or not. If this is approved the ontology is *populated*, which means that a process for in-

[6] `http://www.corporate-semantic-web.de/colm-lifecycle-methodology.html`

stance generation from structured, semi-structured and unstructured data runs up. The ontology is *deployed* and in use by applications. Throughout the whole *feedback tracking* phase, formal statements about users feedback and behavior are recorded. A *reporting* of this feedback log is performed at the end of the usage cycle. That means that all feedback information, which were collected until a decisive point, are analyzed respecting internal inconsistencies and their effects to the currently used ontology version. The usage cycle is left and the knowledge engineers *evaluate* the weaknesses of the current ontology with respect to the feedback log. This point may also be reached, when the validation shows that the new ontology is inappropriate to the specification. The lifecycle closes with the *evolution/forward engineering* of the ontology by engineers and domain experts.

The innovative approach towards agile ontology engineering allows an evolution of rapidly released ontology prototypes. We expect from our model to allow an intuitive view to ontology engineering processes and facilitate a cost estimation in the run-up of cost-intensive evolution steps. We reach these improvements by a convergence of ontology engineering and ontology usage controlled by innovative versioning, modularization and integration approaches.

On the whole, COLM addresses the generic knowledge integration use case (Section 2.2) which is relevant for various cooperation partners, e.g. EsPresto GmbH (Section 2.1) and Merck KGaA. This is a joint contribution of the work packages WP 9 – WP 12.

3.4.1 Innovative Ontology Versioning with COLM

COLM addresses ontology versioning in an explicit way, because the focus of the whole methodology is on agile evolution of ontologies. Thus, an innovative ontology versioning component will be the central contribution of work package WP 10. The goal is to provide an approach combining the lifecycle model with methods and tools, which track as much knowledge as possible, so this tracking log can be processed by the versioning component to generate new knowledge and necessarily related ontology metadata about the new ontology version of the evolving prototype. The approach addresses the generic product improvement use case (Section 2.2) because companies, e.g. the Ontonym GmbH, follow such an approach manually without proper tool support.

We will propose an innovative ontology metadata model which enriches the amount of existing approaches with provenance information referring to the knowledge fed back by the tracking log. The detection of the necessity of new ontology versions will depend on a combination of structural and semantic diff, an analysis of the ontology metadata, and the tracking log.

3.4.2 Modularization and Integration Dimensions of COLM

Modularization and integration of ontologies is in different ways integrable into the lifecycle. As illustrated and mentioned previously COLM facilitates cost estimation in the run-up of cost-intensive evolution steps. The modularization and integration dimensions of COLM helps to decrease CAPEX as well as OPEX and supports to realize ontology adoption to the corporate environment by keeping the incentives of a company in view. In work package WP 9 we will realize this in the following way:

At the very beginning of COLM during the creation/selection phase it is reasonable to look for existing ontologies for the sake of reusability, because developing ontologies from scratch is a very cumbersome and time-consuming task. Expecting candidate ontologies which perfectly fit into the targeted system is unrealistic, some customization will be necessary in order to adapt candidate ontologies and make them useful. At this point modularization and integration are important mechanisms to allow reusability even if the candidate ontologies are not usable in their original form. The possibility of extracting only relevant parts of existing ontologies and integrate them in order to achieve a useful ontology decreases CAPEX drastically and makes ontology application realistic and really attractive for companies.

Modularization during the lifetime of ontologies is also possible. This can be done based upon diverse aspects. The closed and controlled corporate environment allows obtaining information as relevance of concepts and relationships regarding departments and application. It also enables to observe the evolution of the ontology and allows to identify vague parts which change very often and well-established parts which change not so often. Following criteria such as relevance of concepts or their probability to change allows a dynamic mechanism to create modules for increasing the efficiency in case of relevance of concepts and for increasing the reliability w.r.t. probability values regarding potential change of concepts and relations.

In the face of the existence of different independent modules COLM can be understood as the lifecycle for each module and the whole ontology as a composition of the modules. That is, as many lifecycles as existing modules would exist in parallel and an additional one which represents the whole ontology. In fact, modularization and integration is an important aspect to tackle the generic knowledge integration use case (Section 2.2).

3.5 Conclusion and Outlook

In this chapter we introduced corporate ontology engineering as the discipline of the Corporate Semantic Web which provides environmental and core processes for the design, the implementation and the maintenance of proper computational background knowledge. We motivated our research in this field by deriving four gaps, which hindered a broad success of ontologies as suitable means for flexible, scalable, and cost-effective data integration and interoperability in corporate information systems.

After presenting the general requirements of ontology engineering in corporate settings we mentioned specific requirements of two core paradigms in this context – ontology versioning and ontology modularization and integration. Afterwards, we gave a brief introduction into recent developed and well-researched methodologies for ontology engineering, methods and tools for ontology versioning, and ontology modularization and integration concepts. We discovered a lack of a lifecycle which respects the needs of corporate ontology engineering. Thus, we designed the Corporate Ontology Lifecycle Methodology (COLM)[7] our approach towards agile, integrative, and cost-effective ontology engineering.

COLM lets us derive a set of functional requirements, which enables a holistic support for it. Thus, we examined each phase for needed tool support and list

[7] `http://www.corporate-semantic-web.de/colm-lifecycle-methodology.html`

this as follows:

Creation: Access to global repositories of standard ontologies or available contractor for initial ontology development.

Validation: Discussion support and support for collaborative decision making for experts and non-experts.

Population: Tools for (semi-)automatic annotation of data.

Deployment: System for supplying the appropriate ontology version to applications.

Feedback tracking: System for integration of lightweight communication platforms, e.g., forums or feedback forms and automatic recovery of user behavior into a feedback log.

Reporting: System, which exports a snapshot of the log at a dedicated point of time, which decides whether a new ontology version is needed and how the consensual new knowledge looks like.

Evaluation: Validation and reasoning tools which enable an evaluation of the log snapshot referring to the actual working ontology version.

Evolution: System which allows the evolution of ontologies (e.g. creation of views, coexisting branches or just new versions).

As a result of this it is necessary to examine existing tools along the new functional requirements raised. We aim at finding the appropriate tool(s), which suit the process integrative. The central artifact of an architecture which implements COLM will be an innovative versioning component to facilitate the agile evolution of ontologies in a contextualized corporate setting.

Chapter 4

Corporate Semantic Collaboration

As previously mentioned, ontologies provide a shared understanding of a domain of interest. In a corporate setting, ontologies describe terms and their interrelation relevant to business context. Thus, they can be used to support communication within and between companies and reflect business processes. As businesses are dynamic, for example new departments are established or contract partners are changed, an ontology or larger parts of it may become obsolete. In such cases ontology engineers will be responsible to update the ontologies.

However, there are also examples in business that only result in smaller changes of the ontology, e.g., add a new concept for annotating documents. In these cases it is time-consuming and expensive to involve ontology engineers. Furthermore, the user has to wait for the ontology engineer to model the new concept, before he/she can use it. From our point of view, a better approach is to enable the user to make smaller changes to the ontology.

In section 3.4 we described the new Corporate Ontology Lifecycle Methodology (COLM) which especially considers the role of users in the ontology modeling process. As a complementary task we will develop a tool in the work package *Collaborative Tool for Modeling Ontologies and Knowledge* that implements the lifecycle of COLM. Therefore, we collected requirements (section 4.1) for a lightweight ontology engineering tool that can easily be operated by inexperienced users.

With a ontologies being self-contained repositories for business related terms and their interrelations, their advantages can only be leveraged if the knowledge kept inside them is referenced from outside, e.g. by adding annotations to electronic documents that reference a certain concept in an ontology. Furthermore, in order to exploit all advantages of ontologies, they should reflect all business relevant terms, i.e. they should be up-to-date. Keeping the ontology up-to-date and annotating electronic artifacts are both unpopular and time-consuming tasks for employees. Therefore, we investigate approaches to collect knowledge from user activities automatically, e.g., compiling all documents written for a project by annotating them with the project name. In section 4.2 we analyze the requirements for such approaches.

4.1 Collaborative Tool for Modeling Ontologies and Knowledge

While the term ontology is widely understood as a shared and formal specification of a conceptualization [40], the term knowledge is vaguely defined. In [56] Liebowitz compiled several definitions of knowledge which are relevant to the topic of knowledge management (KM). In the context of the project Corporate Semantic Web we will use the definitions of Woolf and Turban:

- Knowledge is organized information applicable to problem solving. [97]
- Knowledge is information that has been organized and analyzed to make it understandable and applicable to problem solving or decision-making. [92]

Both definitions consider knowledge as information for solving problems, a situation that regularly occurs in companies. For example, a person joining a company, a department, or a project team has to get familiar with his/her new environment and tasks. This includes simple activities such as getting a pencil or mailing a letter.

The knowledge of a company does not only comprise electronic artifacts, e.g., documents and images, but also the experiences, skills, and know-how of the employees. Organizing this information systematically is known as knowledge management with the goal to get the right information to the right people at the right time. Before reaching this goal information has to be collected, compiled, and integrated. Furthermore, documents have to be annotated with metadata which is typically a manual and time-consuming task [45]. A knowledge management system generally to manages great quantities of documents. It has to "know" about the content of the documents to effectively support employees in their work. As an example application we discuss searching of documents in Chapter 5.

The development of Semantic Web standards such as the Resource Description Framework (RDF) [10] or the Web Ontology Language (OWL) [9] lay the foundations to build vocabularies with a well-defined semantics. As far as the annotator and the annotation consumer actually agree upon the vocabulary and its meaning documents can easily be found in the knowledge management system. Difficulties arise if they do not. For example, one of our industrial partner told us that departments of the enterprise do use different vocabularies (see Chapter 2). From our viewpoint ontology engineers cannot accomplish the task to model a vocabulary that reflects the terminologies of all departments. Rather, the employees dealing with the documents and being familiar with the topic should be enabled to bring their perspective of the domain into the vocabulary.

As a consequence, the employees need a light-weight ontology engineering tool (shortened: ontology editor) for collaboratively developing their vocabulary. With *light-weight* we want to emphasize that the tool can easily be used by persons unskilled in ontology engineering. The need for a light-weight ontology engineering tool is also supported by a study conducted by Simperl et al. [84] who investigated the development of ontologies for commercial as well as academic applications for a wide range of domains. They discovered that only a small fraction of the interviewees had received ontology engineering training a priori.

Furthermore, we interviewed industrial partners of the CSW project, namely EsPresto AG selling knowledge management solutions and Ontonym GmbH operating semantic query expansion and text comparison services. From the ontology distributors' point of view, an ontology must either be needed by a greater number of companies or be expandable by the customers to be cost-effective. The latter point means that the ontology distributor develops only a base ontology and provides a user-friendly tool for modifying this ontology.

In the following section we identify requirements for a light-weight ontology editor. Afterwards, we describe currently available tools for modeling ontologies and investigate to what extent they meet our requirements.

4.1.1 A Light-weight Ontology Engineering Tool

In Section 3.4 we described the new methodology for ontology engineering COLM. Besides investigating the theoretical foundations of ontology engineering we also focus on an appropriate tool support. In the following we describe a set of requirements helping us to categorize and evaluate existing ontology engineering tools (cf. Section 4.1.2). Afterwards, we describe the requirements of a collaborative tool for modeling ontologies and knowledge in a corporate context.

We distinguish between the requirements related to the user interface (frontend) and the ones related to technical issues and functionality (backend). Requirements of the first group influence the "look and feel" of the tool, thus, having immediate impact on the effectiveness of the tool and the acceptance by its users. The backend requirements focus on issues that have to be considered while implementing the tool. The backend comprises the system that is responsible to provide the data for the frontend. Although not visible in the frontend they also affect how a user experiences the tool, e.g., efficiency of the algorithms and support for modeling tasks. Some of backend requirements may be induced by requirements of the user interface.

Besides the above categories some requirements are indirectly associated with the ontology engineering tool which are not covered by this report. For example, before the editor can be deployed in a company, it has to be introduced to the future users, e.g., by giving a tutorial.

User Interface

In connection with the user interface we discuss the following requirements on the tool: customizable of the user interface, adaptability to workflows, adequate visualization, and support of communication and collaboration.

In [18] Braun et al. pointed out that the compelling simplicity of Web 2.0 applications should be transfered to ontology engineering, e.g., lowering the barriers: informal, light-weight, easy-to-use, and easy-to-understand. To transfer this idea to the corporate context the ontology editor should provide a *customizable user interface*. On the one hand, we can reduce the complexity of the user interface by considering the experience of a user. For example, the editor could automatically decide on the basis of existing relationships if a new concept is a subclass or an instance of a given concept, so the user does not have to know about subclasses and instances. Typically, there will be predefined perspectives for the various user groups of a company, e.g., a

simple one for unskilled persons. On the other hand, we can provide a context-oriented perspective by adapting the level of detail, e.g., hiding functionality and information that are not needed in the current context of the user.

In connection with that we demand of the ontology editor that it is *adaptable to workflows* of a company. As stated in [84] current methodologies of ontology engineering implement a rigid workflow instead of providing a set of method assemblies. For example, users may differ in their approach of modeling knowledge in cause of their experiences or current projects.

To provide a tool that supports users effectively in their tasks we have to think of an *adequate visualization* of the information and knowledge contained in the knowledge base. Referring to [22] we have to consider the different background of the recipients and make a trade-off between an overview and details that need to be communicated. Therefore, the ontology editor should follow the idea of "overview, zoom & filter, details-on-demand"[1].

A main intention of an ontology is to develop a shared understanding of a domain of interest. Hence, we require from the ontology editor that it should *support communication and collaboration* between its users. The ontology editor should provide both synchronous (instant messaging) and asynchronous (comment, discussion) communication. Using these functionalities users can make comments about concepts and their relationships and discuss modeling issues. Thus, they contribute to a *shared* understanding of the domain and the evolution of the ontology.

Technical Issues and Functionality

Specifying a user-friendly interface is only one aspect of developing an ontology engineering tool. Before a user can work with the editor we have to address some issues in the backend: ACID properties, inconsistencies, automatic documentation of the ontology, and security aspects.

The conceptual design of a component that manages the ontologies is a main issue of realizing a collaborative tool for ontology engineering. A requirement for the management of the ontologies is that the backend implements the well-known *ACID properties* – atomicity, consistency, isolation, and durability (ACID) – from database management systems [38]. Since the ontology editor is a multiuser system these properties guarantee, besides others, that any user can edit or use the ontology as if no other user is using them, e.g., a user will never see any intermediate results. Realizing this property is difficult, because changes made by a person may affect the complete ontology, e.g., moving a larger hierarchy of concepts to another place in the ontology. Thus, other changes have to be rejected until a consistent state has been reached.

Versioning and modularization of ontologies are also issues of the backend. Since these aspects of ontology management have already been described in Sections 3.4 we do not go in detail at this point.

Ontologies should reflect the real world to be useful, therefore, the ontology editor has to *deal with inconsistencies* of the ontologies. Inconsistencies arise for example from the different perspectives on a domain, because people have different background knowledge (user context) or work in different departments of a company. A strategy to reduce the possibility of inconsistencies is to start

[1] Talk of Ben Shneiderman at the Centre for HCI Design, City University London, 09/2007.

with a fixed set of concepts that cannot be changed – a kind of upper ontology. Additionally, the system should point to possible conflicts, e.g, duplicate concept labels, and help the user to resolve them.

The ontology editor should *automatically document the ontology* [84]. For instance, it extracts definitions of concepts, comments, discussions, and annotations from the ontology and transforms them into a human readable document. As a kind of documentation the system should also *keep track of provenance information.* For example, it stores information about changes of the ontology (history of who did what) and which other information sources are/were connected to the ontology.

Finally, we want only to mention *security aspects* as they are an issue to most systems. The ontology editor should have to provide a user management, so that users can only access and modify the parts of the ontology which they are authorized for. In contrast to ontology editors being available on the Web we do not need special mechanism to protect the system against vandalism in a corporate environment, because logging of the user actions would reveal the originator.

4.1.2 Tool Support for Ontology Engineering

In this section we give an overview of existing ontology engineering tools and investigate which requirements they fulfill. Hereby, we distinguish between desktop and Web applications. Ontology engineering tools implemented as desktop applications have to be installed locally on some PC and the user has to connect to a server to work with ontologies. In contrast, Web-based tools use a web browser to interact with the user – prominent examples are wiki-like ontology editors.

Desktop tools

Protégé [68] is the most accepted tool for ontology building. Its appearance is similar to software development environments. Protégé is rich in function and language support and very scalable because of its extensibility. Since Protégé contains collaborative components it is possible to develop consensual ontologies in a distributed fashion using lightweight access to the process by discussion and decision making about proposed changes. This feature does not respect any roles or permissions. Versioning control is enabled on ontology level, but not on conceptual level, enriched by the annotations from the structured argumentation. Any abstraction from technical terms is missing.

To sum up, Protégé is a very useful tool for engineering ontologies in a team of experts with a lack of lifecycle support in a usage-oriented architecture.

SWOOP [50] is a desktop environment for ontology engineering, which is a bit straightforward at the expense of functionality. The representation of the concepts allows a web-browser-like navigation and is a bit intuitive for non-experts. A search form supports quick searches on the recently used ontology or at least all ontologies stored. Quick reasoning support is implemented in the same fashion. However, there is no abstraction from technical primitives enabled. By definition of remote repositories, it is possible to commit versions of ontologies.

Altogether, SWOOP is a tool for ontology engineering tasks for experts and well-experienced users. It has its strengths in quick and intuitive navigation in and search on ontologies but lacks functional flexibility and lifecycle support.

Web-based tools

OntoWiki [5] is a php-based wiki-like tool for viewing and editing ontologies. It is setting up on pOWL which makes use of the RAP API[2]. OntoWiki[3] allows administration of multiple ontologies (called knowledge bases) and provides inline editing as well as view-based editing. As an abstraction from conceptual terms OntoWiki includes an alternative visualization for geodata (Google Maps) and calendars auto-generated from the semantic statements stored. However, a general abstraction from technical primitives (e.g. class, subclass, SPARQL, etc.) in the user front-end is missing. Altogether, it allows only one single view for all users and does not respect any roles or permissions. Changes to the conceptualized knowledge have to be done manually. The ontology history is concept-oriented not ontology-oriented and implemented as known from wiki systems.

We conclude that OntoWiki is an ontology engineering tool and a knowledge base for experienced users with an academic background and that it does not support lifecycle management.

Ikewiki [80] implements the semantic wiki-idea and focuses annotation of wiki-pages and multimedia content. It is possible to generate an alternative graph visualization for the context of each annotated page. However, Ikewiki does not support any abstraction from technical primitives for users with less experience in the field of ontologies. Restricted views referring to roles or permissions are not provided. The ontology history is concept-oriented not ontology-oriented and implemented as known from wiki systems.

We summarize about Ikewiki, that this tool addresses familiar wiki users with technical experience which do not need any control of the conceptualization and lifecycle support.

SOBOLEO [98] supports the Web-based collaborative engineering of SKOS ontologies resources[4] and the annotation of web resources using concepts from an SKOS ontology. In contrast to the wiki-based ontology editors, it uses a proprietary user interface using modern Web technologies to simplify its usage. For example, a user can drag a concept from the ontology and drop it onto a web resource to annotate the resource with that concept.

To summarize, SOBOLEO is a light-weight editor for SKOS ontologies. Although the authors write in [98] that "annotations are maintained collaboratively", they do not explain how collaboration is realized, e.g., how do they handle conflicting changes occurring at the same time. From the description in the paper it is unclear, if they implement a certain methodology and how evolution of an ontology takes place.

[2] `http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/`
[3] `http://aksw.org/Projects/OntoWiki`
[4] http://www.w3.org/TR/2008/WD-skos-reference-20080609/

Summary

Our experience concludes that there exists a strong distance between the recently accepted approaches and the needs of our ontology lifecycle. The tools either have an engineering-oriented perspective, which deals with the ontology application- and user-independently, or they reckon the conceptualization on an application level for knowledge management without respecting unfamiliar users. The latter is emphasized if we note that the barriers of wiki-syntax for users without any technical background are underestimated.

4.2 Knowledge Extraction by Mining User Activities

As mentioned in the introduction to this chapter, manually annotating digital artifacts is a time-consuming and unpopular task. Therefore, it seems natural to automate this process, i.e. to harvest metadata that are already explicitly or implicitly contained in an artifact or that can be derived from the way a user interacts with the artifact.

Collaborative systems appear to be a suitable candidate for this task, since they offer creation and revision of digital artifacts and user interaction in a defined way. More often than not, a digital artifact like a document can, among other means, be described by the process it belongs to. If, for example, an employee wants to take a week off and files the appropriate application form to her boss, the nature of the document she uses is determined by the name of the process "application for leave". Backed by a meaningful ontology, facts that are closely related to that term can be derived and ease later retrieval of that document, for example by the keyword "vacation", even though it might not be explicitly used anywhere in the document itself.

Knowledge derived from automatically collected metadata may in turn be used to automate processes. Imagine an incident ticket system scenario where the processes "report error/create issue", "work on issue" and "resolve issue" are defined. In certain cases it might be likely that the same issue is reported by numerous customers over and over again. For that case, a database with frequently asked questions and a fourth process "create FAQ entry from frequently resolved issue" are established. Although the wording in each issue report belonging to a set of equal incidents may differ, annotations about the context of each incident may help automating the process of consolidating similar incidents and constructing an FAQ entry. Automated metadata generation would unburden the customer service technicians from the task of manually annotating each issue report.

In the following sections, we examine existing works in the field of automated metadata generation based on users' actions on digital artifacts. Then, based on previous meetings with our industrial partners, we identify and formulate requirements for metadata generation in a collaborative environment.

4.2.1 Related Work

The majority of recent works that aim at automating the process of generating or harvesting metadata from user activity can be found in the field of Semantic

Desktops and user profiling on the web. While the Semantic Desktop is not a sufficient solution in a collaborative environment, some of the ideas behind it might be worth a glance and turn out to be applicable in this context.

Semantic Desktop Systems

The most elaborated candidates in the field of Semantic Desktops that deploy some mechanism for automatic metadata generation are *iris*[5][25], the *NEPOMUK*[6] Semantic Desktop Framework[39], NEPOMUK's reference implementation *gnowsis*[7][78], and, although limited compared to the others, Apple's Spotlight search[8]. iris and gnowsis both follow the approach of implementing system hooks that intercept certain events caused by a user's action on a file or by communication. In [26], Chirita et al. propose a technological approach for picking up information available during those events, as well as suitable ontologies for turning this information into inferable knowledge for three contexts they consider the most important in the field of desktop information management:

Email An event is triggered when new mail arrives. Besides the textual content of the email, information from the *subject* field, the *sender* field, the *reply_to* field, the date the email was sent and the *comments* field are turned into RDF and used for annotating files attached to the mail, if present. The underlying ontology proposed by Chirita et al. provides, for example, information about people and means for relating an email address to a person, taking into account that one person can own multiple email addresses. As a result, complete context information is stored along with a file that arrived by mail, and a document can be found by the name of the person who sent it or a keyword from the subject of the mail it arrived with (see Figure 4.1).

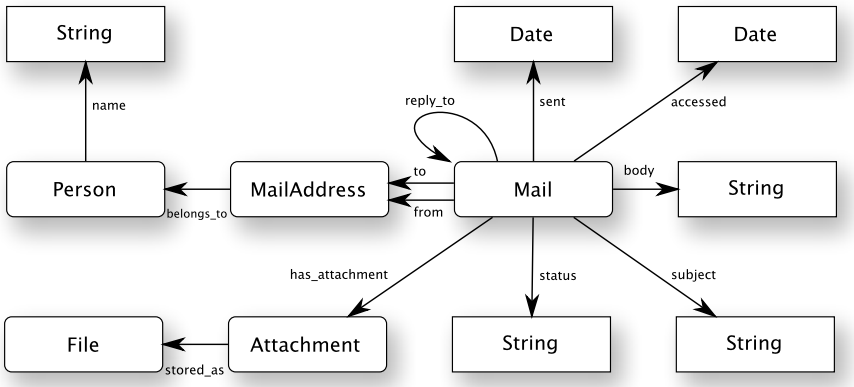

Figure 4.1: An ontology for annotations from email, proposed by Chirita et al.

[5] http://www.openiris.org
[6] http://nepomuk.semanticdesktop.org
[7] http://www.gnowsis.org
[8] http://developer.apple.com/macosx/spotlight.html

Web History Chirita et al. suggest that documents which are downloaded from the web be enriched by context information that can be extracted from the web history. That is in particular titles of visited pages, names of links the user followed, and search terms used in known search engines like google.com or CiteSeer.

File Hierarchy As a third source for meta information, Chirita et al. consider the location in the file system where the user chooses to save a particular document [26]. When the system is notified that a file has been saved, it splits its file path into its individual components. Context metadata provided by file and directory names are enriched using WordNet[64], a lexical reference system which contains English nouns, verbs, adjectives and adverbs organized into synonym sets, each representing one underlying lexical concept. As also used for semantic searching (see chapter 5.1.2), different relations link the synonym sets. The following additional relationships are considered (see Figure 4.2):

- *Hypernym*: Designates a class of specific instances. For example, *vehicle* is a hypernym to *automobile*.
- *Holonym*: Designates the superset of an object of which it is a part or a member. For example, *automobile* is a holonym to *wheel*, and *Europe* is a holonym to *Germany*.
- *Synonyms*: A set of words that are interchangeable in some context. E.g., *automobile* and *car* are synonyms.

If, for example, a document is stored in a folder named "Bielefeld", the system can deduce, by resolving "Bielefeld"'s holonym relations using WordNet or a similar ontology, that "Bielefeld" is part of "Germany".

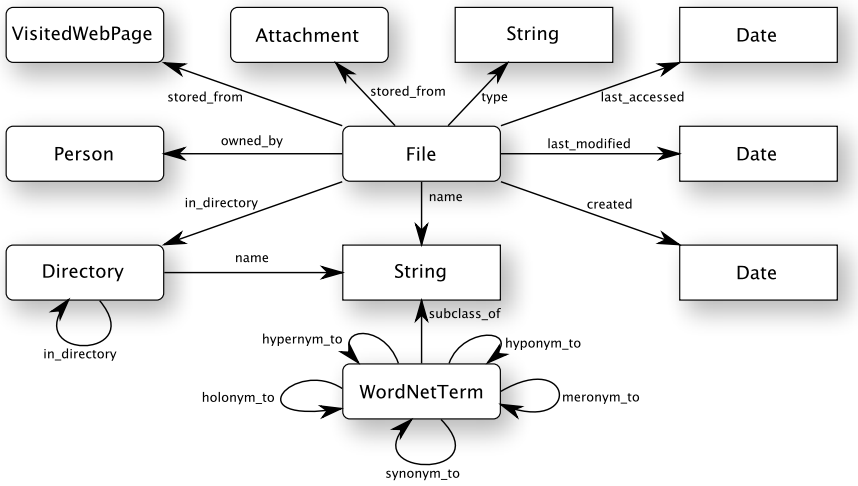

Figure 4.2: An ontology for annotations from user actions on files, proposed by Chirita et al.

Applied in combination, and backed by a sufficiently conclusive ontology, the above approaches can produce significantly enriched metadata that allow for enhanced search queries like:

- Find the document a coworker from the CR department sent me last autumn about the customer in Germany that produces recycled paper.
- Find the document about tax law in Indonesia that I downloaded last week while I was searching the web for Hotels in Jakarta.

We think that the above mentioned concepts and technologies for single user desktop systems can, to some extent, be deployed in collaborative environments like networked groupware systems or web based collaboration tools like BSCW[9] or Microsoft's Sharepoint[10]. Exchanging messages between users works in a way similar to the traditional email system, and the mode of managing files is similar in both desktop and web based collaboration systems, i.e. files are stored within a hierarchy of folders. However, groupware systems offer a lot more sources for metadata, like author profiles, as well as department and project information. Furthermore, well defined processes are a fundamental characteristic of collaborative systems, which generic use of a desktop system are missing by nature. Also, groupware systems provide means for tracking document usage (by individuals and thereby by departments or in the context of a particular project). In [23], Cardinaels et al. suggest a document annotation and indexing system by means of considering usage context information. As valuable metadata may be extracted from these sources, we think that the concepts used in semantic desktop environments are not a fully satisfying solution in a collaborative context.

Environmental Context

Another rather technical approach is to generate metadata from measurable context information at the time a file is created. Modern digital cameras, for example, are equipped with additional sensors or receivers for geographical coordinates, daylight, or other environmental parameters which are stored in the metadata section of the taken image files. By the time of writing, this context data is mainly used in location based picture sharing services, like Google Earth where users can publish their pictures and relate them to the location where they were taken.

User Tracking / Group Modeling

While the above approaches concentrate on retrieving metadata about digital artifacts, others focus on gathering information about users of a system by observing their activities with the aim to provide intelligent recommender systems by deducing the user's general interest within the system's domain. While this technology is broadly used by shopping sites such as amazon.com, they cannot easily be deployed in environments with focus on collaboration, because their heuristics rely on a large number of users and, in contrast to shopping

[9] `http://www.bscw.de/english/index.html`
[10] `http://www.microsoft.com/sharepoint`

sites, collaborative environments normally have only few users and a comparatively large and heterogeneous quantity of knowledge to manage. In [6], Auer et al. propose a document recommender system for group collaboration based on recent user queries that adapts to the characteristic needs for collaborative environments with a low user count. A similar work has been conducted in the field of e-learning. In [49] Jung et al. construct a model and heuristics for discovering social networks and common interests in e-learning systems by analyzing user profile interlinking, blog entry backtracking, and communication channels. In [66] Najjar et al. define a framework and an XML schema for collecting and expressing attention metadata that is collected from users' actions inside e-learning environments and when using their web browsers, like web sites visited and search terms used, and derive strong user profiles and user interests. An ontology for modeling interrelations between online communities and their users has been developed by the *sioc* (Semantically-Interlinked Online Communities) project group[11].

User Role Modeling

While the above approaches concentrate on a perspective on users themselves and their interests within a domain, other works exist that put more emphasis on social graphs and especially roles that users fill within that graph. In [27] Culotta et al. propose a system for automated address-book population, expert-finding, and social network analysis from email inboxes. The system observes a user's email account and retrieves information about each sender from the web using conditional random fields. By following links to other persons and applying the same techniques, it constructs a social network around the user. The authors claim their system provides capabilities for alleviating form entry of contact and social linkage information, finding experts in large companies or communities, automatically recommending additional experts in a certain field, finding persons who are hubs or authorities in a particular field, and finding the best path to a particular person.

While many works in the field of automated user profiling and role modeling exist, leveraging their results require rich and robust ontologies that relate a person's activities and roles to skills. Some skill ontologies have been developed, e.g. in the context of the project *Knowledge Nets*[12] at the AG NBI. We plan to further investigate the potentials of user role modeling and a possible combination with skill knowledge.

4.2.2 Requirements

According to what has been discussed during the kick-off meetings with our industrial partners, we attempt to formulate a set of requirements for our work in the field of automated collaborative ontology engineering.

Data Integration

In section 4.2.1, we mentioned environmental context as a provider for metadata. While by now, this information is mostly used for geotagging pictures shared by

[11] `http://sioc-project.org`
[12] `http://wissensnetze.ag-nbi.de/`

online communities and for private picture collections, it can also be matched and related to business context information and thereby ease data integration. For example, pictures taken on a construction site may be automatically related to a construction project.

No User Interaction

As implied by the term "automated", an important requirement is that the process of metadata recognition and generation and ontology population be transparent to the user and does not interrupt his workflow. Previous attempts to establish semantic enrichment of documents by manual annotation have turned out to be error prone, expensive and poorly accepted by end-users. We believe that automated enrichment of documents with metadata without any impact on a user's workflow but with significantly better results when it comes to retrieving knowledge, can be a crucial incentive for end-users to use collaborative tools.

Availability of Domain Ontologies

As pointed out in Section 4.2.1, the lack of fitting ontologies prevents new heuristics for information gathering from being used in productive settings. Our aim is clearly not to provide a super intelligent tool that builds ontologies from scratch. While technological details remain to be determined, it is already clear that the envisioned approaches rely on deriving metadata from existing domain knowledge. The availability of domain ontologies describing domain specific vocabulary is therefore required.

Traceability and Controllability

Although transparent, automatic metadata generation and document annotation should be traceable and controllable. Since automated knowledge generation can be error prone to some extent, we consider the facility to review and if necessary correct auto-generated metadata indispensable.

Conflict Resolution

Furthermore, working on digital artifacts in collaborative environments with automated annotation enabled may lead to contradictionary annotation rendering the underlying knowledge base inconsistent. Means for detecting inconsistencies is thus an important requirement. Automated checking for knowledge base consistency is provided by reasoners, but performance can become an issue with large knowledge bases. Some reasoners already offer incremental reasoning, verifying only newly created statements, but at the time of writing, this feature remains subject to research [42]. It remains to be investigated at which level and by which means conflict detection is practical and reasonable.

We extend the requirement for keeping user interaction at a low level to the problem of conflict resolution. However, there might be situations in which it is impossible to solve a conflict automatically. Sometimes apparently conflicting information even does not necessarily have to be a conflict but can be due to missing context information somewhere else in the knowledge base. For example, while a project ontology may be designed to relate exactly one leader to a

particular project, different departments involved in that project may appoint an additional department internal responsible. In this case, the appropriate action would not be conflict resolution but to make the ontology more expressive.

Privacy

Another important problem that has to be tackled is privacy. It is possible if not common that professional communication by email or chat contains private notes, if the sender and recipient know each other personally. The senders and recipients are not likely to agree that their private communication turns into annotations that are visible to anybody. Privacy becomes even more important if a public document, for example sent as an email attachment, is accompanied by non-disclosed information within the body of the mail. The document could be annotated with sensitive information from the mail body which would then become visible to everyone who receives a copy of the document. It is therefore required that the process of metadata harvesting and annotation be controllable by any party that is involved in the process. Since this clashes with the requirement for transparency, a compromise has to be found. A conceivable manner for providing transparency without interrupting the user in his workflow could be to issue a silent notification each time a document has been annotated automatically. If the user feels sensitive information could have been added to the document, she can choose to review what has been added and optionally delete sensitive parts. Another possible approach, in combination with the aforementioned could be to establish a list of stopwords which should never be used for annotation. This list would of course have to be adapted to every customer's special needs. Altogether, this is still subject to further investigation and the ideas mentioned above should only serve as an initial impulse.

User Motivation

One of our industrial partners' employees have expressed fear that the envisioned degree of automation in the process of knowledge collection might render their jobs useless and that they might become dispensable. These particular employees' task is to retrieve answers to coworkers' or customers' questions from a library. While an automatically constructed FAQ as described in the scenario in Section 4.2 could indeed take over this task, there will still be a demand for knowledge workers who supervise this process and maintain the knowledge base. A social requirement is thus to motivate employees to use such a system by ensuring and underlining that the process of automated knowledge generation may relieve employees from repetitive tasks and reduce response times, while it will at any rate generate new challenges and need for personnel to tackle them.

4.3 Conclusion And Outlook

While the lack of methodologies for ontology engineering and management constitutes one of the principal obstacles for deployment of semantic web techniques in corporate settings, another important problem is how to fill existing ontologies with life, i.e. with individuals. In this chapter, we introduced requirements for tools and techniques that fulfill this task in collaborative environments. We

focused on automation, taking into account user activities, processes and workflows.

We plan to investigate means for deploying techniques developed for Semantic Desktop systems in collaborative tools like groupware systems or content/document management systems. We will focus on process automation, traceability and controllability. We will also follow the approach of building and exploiting user group and role models from users' activities and combining thereby derived knowledge with knowledge expressed in skill ontologies like the one developed in the context of the project *Knowledge Nets* at the AG NBI (see section 4.2.1.

Chapter 5

Corporate Semantic Search

As business oriented research concentrating on the Semantic Web organized within corporate structures, Corporate Semantic Web covers a wide spectrum of innovative scientific and application oriented solutions for research problems in a corporate context. In the previous chapters, we explained two fundamental pillars of the Corporate Semantic Web research. In this chapter, we introduce the last research pillar. Based on advantages of Corporate Ontology Engineering and Corporate Ontology Collaboration, the Corporate Semantic Search brings new ideas and methods for realizing high quality search tasks in the corporate context. This chapter covers the work packages WP 1 and WP 2.

Regarding the variety of *Semantic Search* definitions, we start this chapter by presenting some search engine services that we explored on the WWW. In this short overview of the selected search services on the Web, we concentrate on presentation of some aspects of the search. Furthermore, we discuss the evolution from "simple" search to semantic search. Summarizing the features of semantic search in Section 5.1.2, we close the characterization of semantic search in Section 5.1.3 by describing current research directions of semantic search research.

In Section 5.1.4 we explain the constitutive components of Corporate Semantic Search. With regard to the specified components, we introduce in the last sections of this chapter our research issues and contributions to the research on corporate semantic search.

5.1 What is Semantic Search

5.1.1 Search on the Web

Many search services are available on the Web. Probably the most known one is Google[1] search. Whilst the search query is based on precisely formulated keywords and the users trusts the statistics, the search results users can get from Google can satisfy their needs or can help for further search. When one wants to know the answer for a question formulated in natural language or cannot specify the search query precisely, the common search services offered

[1] `http://www.google.com`

by Google, Yahoo[2], MSN[3], etc. seem to deliver only partially satisfying results.

An interesting solution for handling natural language questions is presented by *social* search engines, i.e. `www.answers.yahoo.com` or the German `www.wer-weiss-was.de`. Hence, this solution needs some kind of expert community for dealing with questions. These "experts" are simple Web users and therefore the search results are limited to their knowledge. Another kind of social search is offered by Twitter[4]. On `http://search.twitter.com/` one can search for news written or linked by Web users that are members of Twitter. Here, search results are based on latest information that people are talking about on the Web. And these results are also limited only to this information.

Very powerful solution for improved search on the Web is offered by Powerset[5], Hakia[6], True Knowledge[7] and Cognition[8]. Hakia's and Powerset's search is based on NLP techniques and semantic search technology. Both seem to deliver interesting results to search queries based on natural language questions. Also the demo-videos about search feasibility of True Knowledge[9] and Cognition[10] engines are showing very promising search services that are almost comparable to a kind of expert systems for search. True Knowledge calls its search engine an Answer Engine and Cognition announces the use of "Semantic NLP technology that understands word and phrase meanings in modern computer applications".

Besides of answering the complex search queries like natural language questions, there are other interesting aspects of search engines, i.e. the visualization of results and user support in specifying search query. Exalead[11] offers next to its search result list small previews of the results. This visualization feature helps the user to distinguish between the interesting and irrelevant information not only by reading the URL of the source but also by having a look on the visual structure of the source. Quintura[12] uses the tag cloud of semantically similar words in order to help specifying more clearly what one means by searching for a given keyword. Since the Web 2.0 users can deal with a tag cloud intuitively, this solution is better than presenting cluster results next to its result list as offered by Clusty[13].

Summarizing here discussed aspects of the search we can define following issues relevant to the search problem: search query form, the quality of results, user involvement and search result visualization. However, what does the *semantic* search mean and for which of discussed aspects is semantics relevant? Before giving a systematic description of semantic search in the next section, we introduce two future forecasts for the Web search which can help to understand the semantic search as a next-level search approach in the evolution of Web. As shown in Figure 5.1.1, semantic search relies on Semantic Web and

[2] `http://www.yahoo.com/`
[3] `http://www.msn.com/`
[4] `http://twitter.com/`
[5] `http://www.powerset.com/`
[6] `http://www.hakia.com/`
[7] `http://www.trueknowledge.com/`
[8] `http://www.cognition.com/`
[9] `http://www.trueknowledge.com/`
[10] `http://www.cognition.com/info/videodemo.html`
[11] `http://www.exalead.com`
[12] `http://quintura.com/`
[13] `http://clusty.com/`

on natural language search. Following to talk[14] given by Nova Spivack[15] at "The Next Web Conference 2008" , semantic search is a search that satisfies its user *more* than the common search based on keywords and used nowadays. Furthermore, semantic search is a step between Semantic Web, called Web 3.0, and the following Web 4.0, called by him the "Intelligent Web".

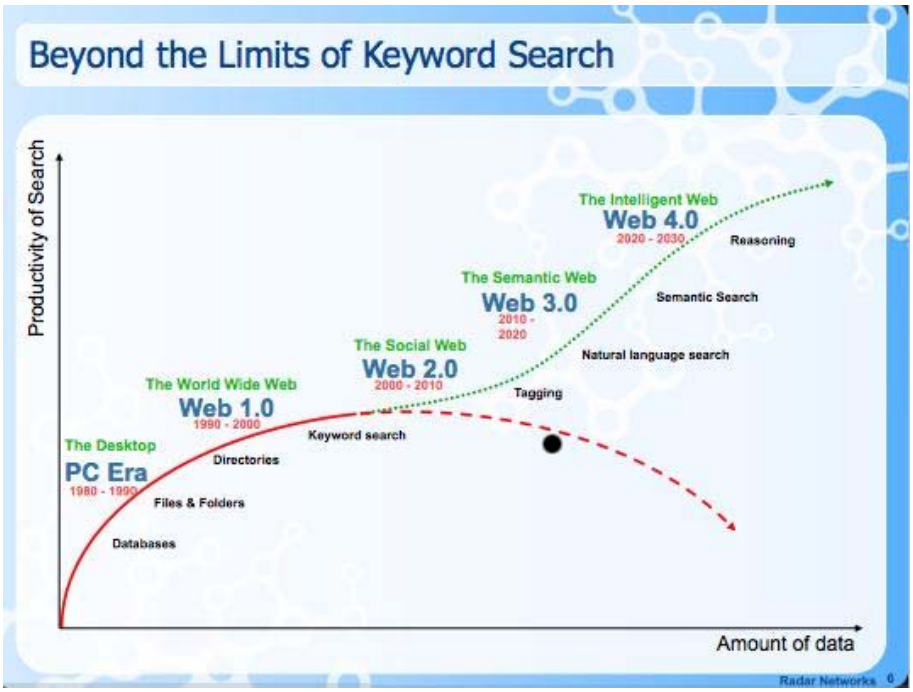

Figure 5.1: from Nova Spivack Semantic Web Talk at The Next Web Conference 2008, April 2008

According to Peter Norvig forecast for the next 50 years, semantic search could be placed as the first step in development of "digital intermediary" offering us suggestions and refinements and delivering an annotated answer as the search result. *"Today, 12 years into the era of search engines, we still have not made good on Brant's boast. Search engines deliver relevance but knowledge requires human work.*

In 50 years the scene will be transformed. Instead of typing a few words into a search engine, people will discuss their needs with a digital intermediary, which will offer suggestions and refinements. The result will not be a list of links, but an annotated report (or a simple conversation) that synthesizes the important points, with references to the original literature. People won't think of "search" as a separate category - it will all be part of living."[16]
by Peter Norvig[17]

[14] `http://thenextweb.org/2008/04/03/nova-spivack-the-semantic-web-as-an-open-and-less-evil-web/`

[15] The author of this talk is a Web 2.0 participant and the founder of a Semantic Web application Twine `http://www.twine.com/`

[16] from issue 2578 of New Scientist magazine, 18 November 2006, page 50

[17] Director of Research at Google Inc

5.1.2 Features of Semantic Search

In the first section we introduced aspects relating to the search in general and we gave an intuitive view on semantic search. In order to define more concrete the characteristic of semantic search, we want to refer to some work on classification of semantic search approaches.

In [62] several semantic search approaches have been classified due to the following classification categories:

1. architecture
2. coupling
3. transparency
4. user context
5. query modification
6. ontology structure
7. ontology technology

In his survey Mangold defines the semantic search referring to semantically supported document retrieval. The author doesn't consider information retrieval based on: *hybrid semantic approach* combining ontology based search with keyword based search [13] or *pure semantic approach* based on ontology search [93] which are other possibilities of defining semantic search. However, we outline his classification criteria since we find them helpful for characterizing semantic search features in general.

The first classification category, architecture, is defined by the same features for semantic search engines as for the non-semantic ones: *stand alone search* or *meta search*. Essential for the semantic search is the second category, the coupling which relates to coupling between documents and ontologies. This can be realized by two approaches: *tight coupling* where the meta data of documents explicitly refer to concepts of a specific ontology or *loose coupling* where the documents are not committed to any available ontology. The transparency feature is divided in three types: *transparent*- the system appears as an ordinary search engine, *interactive*- recommendation system, where the user is asked for interaction, and *hybrid*- combination of interactive and transparent behavior. Considering the user context category, there is a distinction between *learning* kind of user context where the context is extracted from user interaction dynamically, and the *hard-coded* approach. The query modification refers to the semantic modification of user queries and often plays the central role in the semantic search engines. The variants of query modification are divided in: *manually modification* (i.e. used by Quintura, see also Section 5.1.1) where the modification possibilities are presented to the user by an appropriate part of an ontology, *query rewriting* that can be realized by *augmentation, trimming* and *substitution* (see [62] for more details on this) and *graph-based* that is based on tight coupling. The last two categories for classifying semantic search approaches are referring to ontology structure and technology. The differences in ontology structure between semantic search approaches are centralized around properties used in ontology. They determine the semantic reusability of ontologies, whereas the ontology technology focuses on technological reusability.

5.1.3 Research Directions in Semantic Search

Since this characterization is based on semantic search approaches published in 2000-2006 (see [62] for more details), in the following we want to present the general research directions in semantic search as described in [60] in 2005.

Augmenting traditional keyword search with semantic technologies: In the augmentation of traditional keyword search with semantic techniques many approaches are related to the query modification that was already mentioned in the previous section. The modification can be based on query expansion or query constriction using additional knowledge from ontologies. Using ontology, the search keywords are matched to the concepts included in it. In particular, the additional knowledge about:

- synonyms[18]
- meronyms
- hypernyms
- homonyms
- polysems

is helping to constrain or to broaden the search query. An other way of keyword search augmentation is realized by matching the search keywords with the ontology concepts in order to improve the relevance of the delivered search results, not to expand the query.

Basic concept location: This research direction deals with the task of efficiently locating instances of core semantic data types. Basic concept location refers to the coupling category described in the section 5.1.2

Complex constraint queries: The front-end problem of search based on ontology graphs are often the queries. Usually, web users are not able to spend their time formulating a very complex query, even if it should help them to find the information they need. Research on complex constraint queries deals with this problem searching for solutions that may help users in creation of query patterns as intuitively as possible.

Problem solving: This research direction addresses all kinds of use cases where there is a problem description and a knowledge base from which we can infer the solution for the given problem.

5.1.4 Corporate Semantic Search

Since we gave a description of essential search aspects and possible meanings of semantic search in Section 5.1.1 and we outlined characterization of it in Section 5.1.2 as well as classification of the research directions in semantic search

[18] see Section 4.2.1 on page 38 for definitions

in the previous section, we describe in this section the restrictions and characterization of semantic search while using the corporate context. In the first, we complete the issues defined in Section 5.1.1: search query form, quality of search results, user involvement, search results visualization by two additional aspects: search objective and search data set. Whereas search objective refers to the problem solving research direction as described in Section 5.1.3, search data set addresses the hybrid or pure semantic approach mentioned in the beginning of Section 5.1.2. Following to the defined issues, we distinguish important elements of the corporate semantic search as listed below:

- **Data:** The corporate data on the intranet includes, among others, natural language based information, numeric data and multimedia data. In general, we distinguish between structured, semi-structured, and unstructured natural language based information which means:
  - Ontologies are structured information since they provide formalized knowledge
  - XML-data (i.e. texts stored in XML), Wikis, Blogs, e-news belong to the semi-structured textual information
  - Chats, e-mails, white papers, technical reports are the unstructured information

  Apart from that, there is information based on numeric data. In general we consider the use of structured and semi-structured information as the primary one for the search process [19]. Following to [61], additional benefit in the realization of the intranet based search lies in applying structured corporate information to the unstructured one in order to achieve better search results. With respect to the different data formats the corporate semantic search research aims at optimizing the access to corporate information using Semantic Web technologies.

- **Users:** We assume, that the users of search in a corporate world are experts in different domains. It implies two facts: they *know* what they are searching for which means that they can formulate their search queries more precisely than the common Web users, and *their knowledge is* often *limited* to the corporate context. This limitation refers often to the context of their own area of expertise which means that they expect high quality of search results within own domain and they need support in the cross-domain search. Regarding the needs of users, we are concentrating on the personalization of search in the corporate semantic search research.

- **Problems:** Searching on the WWW includes not only the search for documents. Search for people (`http://www.yasni.de/`), events (`http://www.zvents.com/`), or similar pictures (`http://www.picollator.com/`) belong also to the search "problems" which we call search objectives. We assume that the search objectives in the corporate context are related to the corporate problems. Besides the common search for information, there is a need for search for products, experts, customer support solutions or complex relations like forecasts or trends. With respect to the search

[19] The use of multimedia data is planed for the second phase of the project

objectives, we concentrate on the context of the search in the corporate semantic search research.

- **Methods:** The difference between search methods used for search in corporate world and search on the WWW lies in the fact of trust and knowledge. Whereas the Web represents an open and therefore less predictable and trustable area, the corporate world offers more reliable data, more trustable users, more condensed knowledge. Following to that, the corporate world is an appropriate environment to demonstrate new search approaches. Since corporate environment serves as a proper subset of the WWW, the methods are further applicable to the Web.

There are lots of different possibilities for realizing the semantic search task in the corporate context. With regard to the characterizations introduced in the previous sections, in our work we are concentrating on the following research issues:

- Search for complex relations in non-semantic data (WP 1)
- Personalization of Search (WP 2)

5.2 Search (for Complex Relations) in Non-semantic Data

The systematic integration of formalized semantic relations and expert knowledge in the process of search allows for deep semantic analysis of the available information. Semantic search in non-semantic data, in this context, refers to the different forms of semantic search in corporate and business applications that aim at bringing additional benefits using solutions based on combination of Semantic Web technologies with the standard statistic methods usually applied to the standard search task. As mentioned in the previous section, in this part of research on corporate semantic search we are concentrating on search for complex relations in the non-semantic data. Addressing the research direction of *problem solving* in semantic search and based on the achievements of *augmenting traditional keyword search with semantic technologies*, we focus on the search for trends.

Determination and early detection of emerging trends can be retrieved from numeric data as well as from texts. Research projects like GIDA[20] , and TREMA [21] have shown that there is a huge demand particularly in the financial domain for the research on and development of useful trend mining methods which are able to include analyzes of textual information in the process of trend recognition.

With regard to the corporate hybrid information systems and using the multimodal corporate data, an adequate trend recognition method will be developed for the recognition of temporal changing patterns in semi-structured textual information sources relevant to the chosen market objective. The key objective of this work package will be to develop a method for searching for and learning

[20]Description online: `www.computing.surrey.ac.uk/ai/gida`
[21]`http://www.projekt-trema.de`

of complex relations from semi-structured information. Our contribution lies in the innovative way of knowledge integration.

5.2.1 Trend Mining as the Future Search Task

Predictive Intelligence is becoming more and more important for companies. As shown in the IT-Trend studies for 2008 by Capgemini [22], this is a growing part of Business Intelligence task for many IT-companies. We assume that the search for complex relations, in particular, the Trend Mining based on textual information will be playing a huge role for the corporate world in the next decades. Regarding the related work, we are in particular concentrating on the gaps in research on trend detection in textual data as shown in [54] [55]. Using the novel idea of knowledge acquisition by applying the collaborative tagging system, we make a contribution to knowledge integration. Furthermore, we are focusing on the work done in the research projects: GIDA [36][2] and its follower, TREMA[23]. These projects concentrated on the fusion of multimodal market data in order to mine trends on financial markets (GIDA, TREMA) and in market research (TREMA). They provide us with our research direction. Similar to TREMA, we are using the advantages of Semantic Web technologies in order to support the textual trend recognition. The difference lies in our idea of applying an Extreme Tagging System [89] additionally to the trend ontologies.

5.2.2 Case Study and Business Process Reporting

Considering the problems of increasing the company competitiveness and improvement of strategic planing, we could identify two different cases that show the needs for automatic trend mining from textual information. So far, the use cases[70] identified relate to the financial markets and to the market research. Both case studies are based on the previous project research. Since the case study research done in the TREMA project is relevant to our research problem, we decided to adapt and proceed these case studies. In the following, we introduce both cases and comment on the case study for our application idea at the end of this section.

Finance Markets:[24] One possible application area is the financial market domain. Traditionally, world events as reported in financial and economic news are important indicators for the strength and consistency of the value of currencies, such as the Euro, and other trading instruments. Such reported events can have far reaching implications to the value of investments made by institutions. Other examples of market moving news are the reported performance of a company or a key speech by a leading political figure or a company chief executive. A clear correlation of financial market fluctuations with structures of name and movements in the news can be observed. Crucially, this correlation is causative: the information precedes the fluctuations. Information about critical events may be presented through a variety of media. Conventional information sources are newswire systems such as Reuters, but also online news portals play a more and

[22] `http://www.at.capgemini.com/m/at/tl/IT-Trends_2008.pdf` figure 21, s. 31 Business Intelligence: Competitive Advantages

[23] Project website: `www.trema-projekt.de`

[24] Many thanks to JRC Capital Management Consultancy & Research GmbH

more important role. These systems deliver several thousands of pieces of information every day to the desktop of the investor. It is the task of the investor to filter through these data, supported by keyword based filters of the site, for the relevant headlines. While headline information may be of immediate value, the required information such as predictions, expectations and other indications of change may not be as immediately obvious. These content features of the full news text are our main concern. Furthermore, the first indications of changes in events are likely to be evidenced at the source of the event. And this may be found somewhere else - in a company's ad-hoc announcement, or even in a blog or discussion forum. The need to extract the significant information from the massive quantity available is now agreed to be a prime need for all forms of knowledge and information management in business. The strategic and timely delivery of such content in a form that (human or mechanical) decision makers will be able to react upon can be considered as a significant requirement for information systems.

Market Research:[25] The objectives of market research projects are to identify market trends as well as the analysis of buyer preferences and behaviors in high tech markets.

Huge amounts of quantitative and qualitative data are generated in the process of market research that must be triangulated and analyzed. In order to outline future market potentials, prognoses are made based upon the collected market data.

The limitations of the current working methods lie in the fact that large amounts of qualitative data have to be analyzed and related to quantitative data. Currently the classification, generalization, and interpretation of qualitative data takes place manually.

The market studies include two main types of questions:

- quantitative questions (scaled questions): single choice, or multiple choice questions
- open ended questions which can be divided in two types: results of primary research (in particular reasons or motivations, comments about ratings or rankings, strengths and weaknesses etc.)and results of secondary research based upon an explorative Internet research in order to analyze specific market trends.

Primary research: Open ended questions are systematically integrated into a market research questionnaire, complementing the quantitative questions. The processing of those open ended questions includes the following steps:

- Collection (of respondents' opinions)
- Back translation in a common language: English or German if relevant
- Categorization
- Aggregation and statistical analysis of categories (frequencies and percentages)

[25]Many thanks to metrinomics The Market Feedback Company

Categorization involves the analysis of customer comments. Firstly, a coding list is developed based upon of 20-30% of available comments. This list is usually project specific and includes all relevant categories used for the analysis of the open ended questions. After the export of all relevant comments into an Excel table, the relevance of each comment is then identified manually. Each comment is allocated one or several pre-defined categories. In addition and, if relevant, the comment will be defined as *positive* or *negative*. If necessary, additional categories are added to the coding list.

In the last step, a count is made of the frequency with which a particular comment is mentioned in each category.

In some cases such as *buying intention*, the added value lies in the connection between quantitative and open ended (qualitative) responses.

For example: Buying intention below 5 on a scale of 1 to 10 (1= very low buying intention; 10= very strong buying intention): The screening and counting of relevant comments gives a picture of the main reasons for deciding in favor or against the product purchase.

In parallel, on the basis of the quantitative responses, a classification of respondents is made as well as an analysis of purchasing motives. This is done with

- Correlation analysis
- Decision tree analysis (this is the preferred method due to the non linearity of data)

Rules are derived out of decision trees and each respondent is categorized according to a specific rule. A procedure which would combine the rule categorization with the assignment to a specific open ended category would be very valuable because the buyer profiles would be complemented with qualitative inputs and therefore would be more transparent.

Secondary research: This is an explorative process which is used in certain types of studies where a broader input and orientation is needed in addition to the study data. The secondary research can be divided into the following stages:

- Definition of main study objective
- Development of hypotheses that provide a structure for the information research and a focus for the research based upon the required and relevant information. A hypothesis can be answered with a 'yes' or a 'no'.
- Collection and screening of Internet links
- Qualitative and explorative expert interviews (option)
- Aggregation

Types of secondary information can be (examples):

- User and buying experiences
- Reports about products or markets
- Test reports
- Predictable or unpredictable events

- Regulations and laws
- Sales channels

The trend mining application should allow the increase of process efficiency in the analysis of qualitative research data by supporting the following tasks:

- The analysis of market and buying patterns by processing a broad amount of text based information. This can be generated through interviews (primary research) or Internet research (secondary research). This includes the automation of the categorization process of open ended questions and the filtering of relevant information.
- trends identification; connection and comparison of qualitative trends with trends calculated upon quantitative study data
- market prognosis on the basis of identified trends

Both cases imply the including of the WWW information for the trend mining. However, we are considering them focusing only on the "closed world" assumption. That means that the case for financial markets will be limited to a given company and its interaction with the market. In the first, we aim to concentrate on the internal corporate textual information and numeric data in order to develop an appropriate trend mining solution. In the following section, we introduce our work on conceptualization of a trend mining method.

5.2.3 Conceptualization of a Method for Trend Mining

Since we deal with a complex kind of semantic search which is in particular based on finding and learning of complex relations in textual information, our method conceptualization is so far based on following definitions:

- **Qualitative Data:** The qualitative data is represented by texts and semi-structured information.
- **Simple Hybrid Information System:** Since the qualitative and quantitative data provide hybrid properties of the corporate information system, we refer in this work to a simple hybrid information system. This is simply a system providing information based on qualitative and quantitative data We can find many examples of simple hybrid information systems in different areas of application, i.e. in financial market analysis, market research, concurrence monitoring etc. Tasks such as strategic planning, decision support, early emergency detection, and trend recognition are parts of those application areas and can be supported by intelligent data analysis.
- **Trend:** Due to the work described in [65], where text based trend analysis is presented through the example of topic trends, texts streams are analyzed with regard to the following tasks in topic analysis:
  - Topic Structure Identification: learning a topic structure in a text stream
  - Topic Emergence Detection: detecting the emergence of a new topic

– Topic Characterization: identifying characteristics of topics

In order to analyze trends, we have to define what is a *trend*. Since we aim firstly to originate our trend recognition process in the numeric data, we will treat the given text stream in a similar way as we might a data stream. With regard to the trend analysis based on time series, the analysis process consists of four major *components* or *movements* for characterizing time-series data [43]. We refer to the *long-term movements* that can be visualized by a *trend curve*. Based on the trend curve generated over quantitative data, we identify *time segments* for those long-term movements that can have positive or negative trend values ("ups" and "downs" on the market). Correlating this segments to the news stream, we identify a priori three trend classes: positive, negative and neutral class and divide the news stream in the 3-category text corpus. Analyzing text corpus, we will search for specific, so called *trend-indicating* keywords and statements. Trend-indicating keywords from the financial market domain are i.e. *cut*, *concern*, *recession*, etc. These simple keywords are subject to what we call trend indicating *language patterns*.

When analyzing text corpus, we are concentrating on trend indicating language structure and on the characterization of this structure. Firstly, we propose to divide the identification of trend indicating language patterns in the non-semantic feature extraction and in semantic feature annotation.

- **Trend Ontology:** The trend ontology should be used as the knowledge base about trend knowledge that can be applied in different phases of the trend mining process. Accordingly, the following application fields of the trend ontology in trend recognition process can be identified:

  – the support of feature extraction for the learning process

  – the support of information retrieval

  – the support of semantic-based classification and clustering

  Furthermore, the trend ontology has to be defined as a knowledge model containing:

  – the meta-level knowledge about the domain relevant concepts

  – common keywords used in the given domain

  – knowledge about trend indicating terms, relations and rules in the domain

  Finally, the trend ontology defines:

  – important concepts of financial markets in given language: tradable instrument, stock, derivative, star analyst

  – partially defines concepts semantic fields,i.e. trader, selling, buying, institute, ...

  – semantic fields are based on part_of, belongs_to and defines relations

  – partially defines trend-indication of every important concept

- **Extreme Tagging System:** Extreme Tagging System (ETS) is an extension of collaborative tagging systems and it allows for collaborative construction of knowledge bases. An ETS offers a superset of the possibilities of collaborative tagging systems in that it allows to collaboratively tag the tags themselves, as well as relations between tags. Unlike previous research on emergent semantics of collaborative tagging systems, ETS are not destined to exclusively produce hierarchical ontologies but strive to allow the expression and retrieval of multiple nuances of meaning, or semantic associations. The production of relevant semantic associations can then be automatically controlled through social network regulation mechanisms.

As shown in the Figure 5.2, our method concept includes the following steps:

1. Collecting
2. Preprocessing
3. Learning
4. Correlation

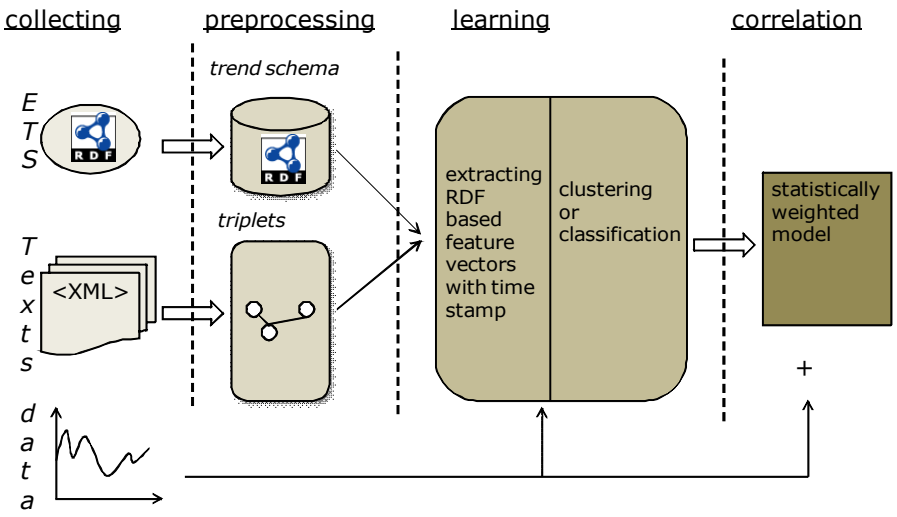

Figure 5.2: Learning and correlation framework

Since this is an early-staged concept, we aim to focus on the following parts of the trend recognition process:

- **Collecting:** which means extracting trend schema based on RDF/S-data from ETS (using RDFS) and extracting Subject-Predicate-Object triplets from text. Important for the extraction is the time stamp.
- **Preprocessing:** which refers to preprocessing of semi-structured information regarding the time stamp and expert knowledge. Important for the extraction is the time stamp.
- **Learning:** which refers to the learning of complex relations by extracting RDF/S-based vectors and using clustering or classification methods

5.3 Semantic Search Personalization

In the context of software systems, personalization refers to the development of models and methods for capturing and representing information about users including preferences, goals, needs, knowledge, capabilities, environment, behavior, devices used, etc., and exploitation of this information to tailor the system's behavior to the needs of an individual user and his expectations [53]. In the Web context, Baldoni et al. define the personalized access to Web data as *"the process of supporting the individual user in finding, selecting, accessing, and retrieving Web resources (or meaningful sub-sets of this process)"* [7]. Personalized search techniques use information stored in user profiles, additionally to the user's current search or query, to estimate the user's wishes and select the set of relevant information.

Previous approaches to personalization on the Web, researched since the nineties of the last century, mostly focused on *adaptive hypermedia systems* [20] and *recommender systems* utilizing Web mining techniques (based on the content as well as usage of Web resources) [21]. The former, however, face the problem of re-usability in a system-independent manner and only work well on a fixed set of documents defined at the design time of the system, whereas the latter require a critical mass of data for the underlying machine learning algorithms to produce results of satisfactory quality. Those limitations arise from the fact that traditional Web resources are mainly designed for humans to read. Personalization techniques, on the other hand, rely on machines which must be provided with knowledge in a machine-interpretable format. This is the reason why Semantic Web technologies provide the most appropriate environment for realizing personalization.

The ultimate goal of the Semantic Web [12] is to offer end users enhanced possibilities to benefit from electronically stored information which is given a well-defined machine-understandable meaning, thus better enabling computers and people to work in cooperation. Most of the research in the Semantic Web, so far, has focused on the first part of the vision i.e. formalisms, languages, reasoning, and appropriate technologies for semantically rich representation of data. Since these technologies provide an environment capable of allowing enhanced integration of heterogeneous data, information discovery as well as advanced automation of sophisticated tasks, a growing interest is being raised for their potential in realizing user-centered applications providing personalization functionality [4, 17]. Semantic-based techniques not only provide software systems with a more precise understanding of the application domain, formalized in ontologies, but can also be used for a richer representation of the user related information itself. Examples of current research on application areas of personalization based on Semantic Web technologies include recommender systems [3], adaptive hypermedia systems [24], eLearning [47, 73], peer-to-peer networks [41], information portals [19], Web Service discovery and composition [1], just to name a few.

Since the main focus of the Corporate Semantic Web research project aims at the application of Semantic Web technologies within enterprise environment, we investigate, in the following sections, the potential for semantic personalization in the corporate context. First, we identify the potential users who may benefit from personalized access to corporate information and analyze the conditions under which personalization makes sense. As next, we briefly describe our

approach to user profile acquisition.

5.3.1 Users and Requirements Analysis

As already argued in Chapter 1.3, from the corporate perspective the introduction of Semantic Web applications must result in tangible gains like expansion of business, a wider set of business opportunities, or cost reduction of current business processes. In this section we first investigate how enterprises may benefit from personalization targeted at various users in the corporate context. As next, we analyze the situations where personalization is expected to show its benefits as well as identify the key requirements.

Users

The application of personalization techniques provides the greatest benefit in environments characterized by user diversity w.r.t. their preference, goals, needs, and knowledge. As far as business organizations are concerned such conditions can clearly be observed. In corporate environments personalization can be targeted at external (customers or third parties like business partners, investors, etc.) or internal users (employees).

Customers. The huge overload of information on the Web makes it difficult for customers to find information relevant for their purchase decision or products/services fulfilling their requirements. Searching a catalog containing hundreds of thousands of items can be frustrating and unproductive. As shown in [72], online users have limited patience for locating material in a large information space that does not provide effective guidance. Since customers tend to become more loyal to services which can be customized to their liking [33], personalization, by providing better service quality than the opposition, may generate competitive advantage attracting more customers thereby increasing revenues. Semantic Web technologies, at this point, have a great potential in enhancing personalization, as well-annotated knowledge sources and product/service descriptions can be matched against RDF statements about consumer preferences to create recommendations or improve navigation based on ontologies describing relationships between items.

External third parties. Depending on the nature of the relationship to the user (business partner, investor, supplier, etc.), which can be represented in a corporate ontology, they can be provided with different personalized views on corporate information within existing intranet, extranet, as well as public corporate Websites. Delivering external parties a personalized access to relevant information may increase productivity and business efficiency by lowering transaction costs of business-to-business co-operation.

Employees. According to analyst firm IDC, knowledge workers spend from 15% to 35% of their time searching for information [35]. Moreover, due to the vast amounts of distributed and heterogeneous corporate data they often do not find all the relevant information. Within business organizations users require different kinds of information depending on:

- department (accounting, marketing, Human Resources, etc.)

- role (e.g. CEO, project manager, administrative assistant, etc.)
- experience/knowledge (e.g. expert, novice)
- goals (e.g. professional training, solving a technical problem, etc.)
- context of the current task (e.g. project-oriented)

An enterprise ontology representing the structure and employees of a business organization would provide the foundations for semantic user profiles which can be matched against document metadata as well as ontologies formalizing corporate knowledge thus enabling personalization within companies. Personalized access to enterprise information based on user context would not only reduce search time and costs but also increase productivity by providing the right information at the right time and in the right format.

When to Personalize?

There exist versatile possibilities for the personalization of search. Vallet et al. [94] identified the situations in knowledge-driven media services where personalization is expected to show its benefits, which may also apply in corporate environments:

Large amount of available content. Often the search result contains more content items than the user can process, even after cutting down their number with queries and conditions. Especially when the filtering conditions, if present, do not provide sufficient discrimination between relevant content (e.g. user browsing by category "project name" finds hundreds of different kinds of documents related to a particular project). In this situation, a personalized user-context-based rank can be used to scatter and sort results as well as provide an individual view on corporate information. For example, an accountant would see financial information like invoices or purchase statements, whereas a Human Resource manager would receive information on employees involved, their work experience, skills, etc.

New content available. If the user is not aware of new relevant content, this would require the user to repeat all the former queries to keep up to date. The notification update can either be automatically delivered by the system or triggered by the user himself.

Imprecise user needs. When the user formulates a vague query, such as "show me available information related to division X", personalization can provide automatic criteria for filtering, ranking, and presenting information based on the user corporate profile.

Short available user time, effort/quality tradeoffs. Even if the user may achieve best search results manually or by doing successive incremental queries, in some situations due to time constraints or other priorities, he might be willing to relax his goals in order to achieve quicker and easier results.

Initializing a search session. Personalization can provide an initial information set ("page zero") adapted to the user context, which would serve as a starting point for further information requests.

Those situations described above may often overlap (e.g. user works under time pressure, has imprecise needs and too much information is available) making personalization even more valuable feature [94].

Requirements

Based on the specific characteristics of the corporate environment and on additional feedback acquired during the meetings with our industrial partners we identified the key requirements for a beneficial realization of personalized access to enterprise information.

1. **User Diversity.** Personalization functionality on the *corporate semantic web* must be implemented and applied to deal with great user diversity. A prerequisite for successful personalization is an adequate semantically rich representation of users as well as their current work context. Since, each user can perform different roles and tasks throughout the time, these dynamic aspects of user characteristics have to be taken into account by providing means for efficient profile management.

2. **Integration.** Especially in the case of knowledge workers who handle information from various information sources including:

   - corporate portals
   - organizational data from shared file repositories
   - enterprise ontologies
   - public Websites
   - documents stored on the local computer
   - contacts from a local address book

   a personalized search should provide individual view on all the relevant information for which the user has gained access rights. This, however, requires semantic representation of the enterprise knowledge in from of ontologies as well as semantic markup of corporate documents. Within the scope of our research project these topics are investigated in the areas *corporate ontology engineering* and *corporate semantic collaboration*.

3. **Transparency and Control.** Since modeling of user context is a quite complex task, personalization techniques provide "good guesses", but cannot aim at delivering perfect answers. Therefore, the user must have the option to inspect and edit his profile so as to be able to amend wrong system guesses or even to turn personalization off. This is especially crucial in order to avoid the risk of latency, when results are shown based on historical data which no longer apply to the current user context (e.g. tasks or project the user is no longer involved in) because the user profile has not/could not been automatically updated.

4. **(Semi-)automated Profile Generation and Update.** The advantages of personalized search should not be outweighed by too much effort put on each individual user to explicitly specify and update his profile. Too obtrusive personalization would not only cost time but may lead to decrease in user acceptance. Hence, a successful application of personalized search must provide means for (semi-)automated profile creation and update.

5. **Performance and Scalability** is another key requirement regarding search engines in general.

5.3.2 Acquisition of User Profiles

Each user adapted service or application requires a user profile to perform personalization accordingly. Additionally, adaptive systems need some form of representation of their domain in order to provide the foundations for user modeling and reasoning. Corporate ontologies have the potential to fill this role. [51] identify three important roles ontologies may play in user modeling:

- defining user models
- providing a vocabulary of metadata for objects and concepts of a particular domain
- supporting user interfaces to the user model (graph visualization)

Respectively, we propose an ontology-based structured profile approach [90] with three layers of specification:

1. The lowest layer holds personal information of who the user is (name, gender, education, etc.) as well as stable characteristics of the user position and role within the company (e.g. department, job title, role). This information can be acquired from the enterprise ontology, representing the organizational structure and employees, without the need for the user to specify it explicitly. Moreover, the enterprise ontology may contain generalized group profiles for different groups of corporate members (e.g. specification of relevant information sources and access rights for different corporate departments, types of documents like invoices, press releases, etc., they mostly search for).

2. The middle layer describes dynamic aspects of the user context within the enterprise which does not change frequently (e.g. participation in projects, skills, social network among corporate members). This layer of the user profile can also be populated with information represented in the enterprise ontology.

3. At the top layer we represent temporary and short-lived user context. This information is automatically collected based on the observation of the current work context of the user. At this point, some techniques researched in the field of the Semantic Desktop [79], [29] may be applied. Schwarz et al. [81] describe an approach to capture user activity and represent user context and goals, based on four different levels of abstraction:

- **Workspace level** representing the operating system and applications. Here, various events (e.g. mouse clicks, typing, starting of applications) are being observed.
- **User action level** containing user actions (e.g. creating a new document) inferred from a series of workspace events.
- **Task concept level** captures user goals (e.g. writing project proposal), which are inferred from a series of user actions required, and represents them in an ontology of user goals.
- **Process level** connects to the organizational processes which are explicitly modeled in the ontology representing the company.

A profile for an individual user is created by the aggregation of user characteristics from all three layers.

5.4 Conclusion and Outlook

Semantic search offers new possibilities for enhanced access to corporate knowledge by utilizing ontologies of the applications domain and the corporation itself. In this chapter, we gave an overview of different aspects of search and briefly described the main research directions as well as identified the key factors and requirements of search within enterprise context. Our research in this field mainly concentrates on utilizing innovative semantic search techniques to facilitate deep analysis of available information by analyzing complex relationships in non-semantic data (i.e. trend mining) as well as on providing users with personalized access to corporate information.

In the research on search for complex relations in non-semantic data, we will focus on our approach of Learning and Correlation Framework. In particular, we plan to concentrate on collecting, preprocessing and learning stages in the search process. These steps are important for integrated knowledge retrieval from non-semantic corporate data. In the field of sematnic search personalization, in the next step, we will focus on means for efficient profile management as well as different aspects of the realization of personalized search, i.e. narrowing search based on user profiles (query refinement, query rewriting), content rating, and personalized visualization and navigation (e.g. graph navigation, faceted browsing).

Chapter 6

Conclusion and Outlook

In this report, we introduced our initial vision of the Corporate Semantic Web as the next step in the broad field of Semantic Web research. We identified requirements of the corporate environment and gaps between current approaches to tackle problems facing ontology engineering, semantic collaboration, and semantic search. Each of these pillars will yield innovative methods and tools during the project runtime until 2013.

Personal interviews with industrial cooperation partners yielded individual Semantic Web use cases which were used to conclude three genric use cases. The latter were identified as relevant generic aspects for small- and mid-sized companies to refrain from Semantic Web technologies and named as **(1) product improvement use case, (2) usability use case, and (3) knowledge integration use case**.

Corporate ontology engineering will improve the facilitation of agile ontology engineering to lessen the costs of ontology development and, especially, maintenance. We introduced an innovative lifecycle methodology called COLM for this purpose. A more detailed description and an integrative architectural design will be part of closest future work. We aim at the development of a new ontology versioning approach which implements COLM.

Corporate semantic collaboration focuses the human-centered aspects of knowledge management in corporate contexts. Therefore, we aim at the development of appropriate tools to support light-weight collaborative ontology editing and an improvement of the automation in knowledge acquisition and evolution processes. On the whole this pillar will provide partial tool support for COLM.

Corporate semantic search is settled on the highest application level of the three research areas. At that point it is working on and with the appropriately represented and delivered background knowledge to enable individually conditioned search results. But moreover, the focus on innovative and hybrid techniques for mining trends from textual data opens new chances for knowledge acquisition and integration as well.

On the whole, all three parts will be put together to an integrative Corporate Semantic Web. The outreach of the results will be promoted and supported by models for the cost-benefit-estimation of semantic technologies following an integrative architecture which is depicted in Figure 6.1.

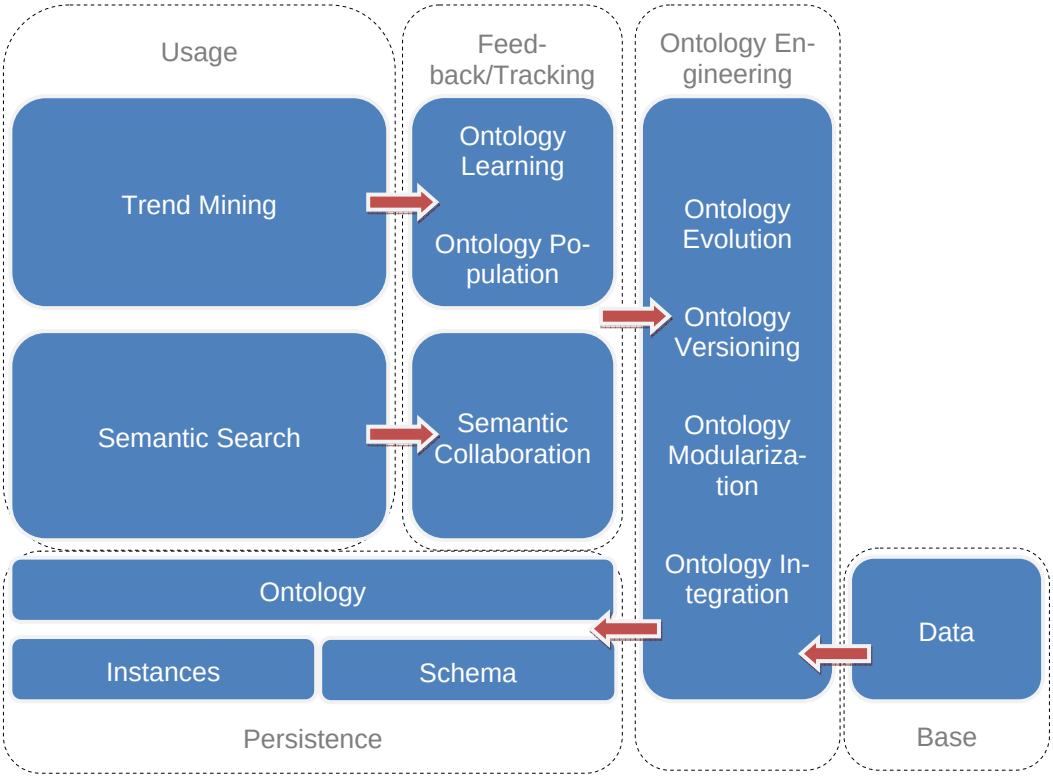

Figure 6.1: Initial concept of an integrative architecture of a Corporate Semantic Web

Appendix A

Work Packages

Work package 1	**Search in non-semantic data**	02/08-01/11
WP 1 Task 1.1	Analysis of business processes and reporting, knowledge retrieval from business processes and reporting	02/08-05/08
WP 1 Task 1.2	Concept of method for knowledge retrieval from non-semantic corporate data	05/08-07/08
Work package 2	**Search personalization**	02/08-01/11
WP 2 Task 2.1	Requirements analysis: Semantic search user identification, their features and information needs	02/08-05/08
WP 2 Task 2.2	Concept of method for automatic user profile generation for personalized search	02/08-05/08
Work package 5	**Knowledge extraction by mining user activities**	02/08-01/11
WP 5 Task 5.1	Analysis of corporate information exchange	02/08-05/08
WP 5 Task 5.2	Desing of a semantic collaborative tool for the acquisition of implicit knowledge about employees	05/08-07/08
Work package 6	**Ontology- and knowledge modelling using collaborative tools**	02/08-01/11
WP 6 Task 6.1	Requirements analysis: for different use of knowledge in corporate context	02/08-05/08
WP 6 Task 6.2	Design of a tool for collaborative modelling of corporate ontologies	05/08-07/08
Work package 9	**Ontology modularization and integration**	02/08-01/11
WP 9 Task 9.1	Design of a methodology for ontology modularization and module integration	02/08-07/08
Work package 10	**Ontology versioning**	02/08-01/11
WP 10 Task 10.1	Design of a methodology for corporate ontology engineering, which respects the needs of multiple coexisting ontology versions.	02/08-07/08

Appendix B

Acknowledgement

This work has been partially supported by the "InnoProfile-Corporate Semantic Web" project funded by the German Federal Ministry of Education and Research (BMBF).

Bibliography

[1] C. Abela and M. Montebello. Predicts: A personalised service discovery and composition framework. In *Proceedings of Semantic Web Personalization Workshop (SWP06)*, 2006.

[2] Khurshid Ahmad. Events and the causes of events. In Lee Gillam, editor, *Proceedings of the Workshop on Making Money in the Financial Services Industry, at the 6th International Conference on Terminology and Knowledge Engineering*, 2002.

[3] Sarabjot S. Anand, Patricia Kearney, and Mary Shapcott. Generating semantically enriched user profiles for web personalization. *ACM Trans. Internet Techn.*, 7(4), 2007.

[4] L. Aroyo, V. Dimitrova, and J. Kay. Proceedings of Workshop on Personalization on the Semantic Web (PersWeb05), held in conjunction with 10th International Conference on User Modeling, UM2005. 2005.

[5] Sören Auer, Sebastian Dietzold, and Thomas Riechert. Ontowiki - A tool for social, semantic collaboration. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 736–749. Springer, 2006.

[6] Sören Auer and Heinrich Herre. Rapidowl - an agile knowledge engineering methodology. In Irina Virbitskaite and Andrei Voronkov, editors, *Ershov Memorial Conference*, volume 4378 of *Lecture Notes in Computer Science*, pages 424–430. Springer, 2006.

[7] Matteo Baldoni, Cristina Baroglio, and Nicola Henze. Personalization for the semantic web. In Norbert Eisinger and Jan Maluszynski, editors, *Reasoning Web*, volume 3564 of *Lecture Notes in Computer Science*, pages 173–212. Springer, 2005.

[8] Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors. *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer Science*. Springer, 2008.

[9] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein.

Owl web ontology language reference. http://www.w3.org/TR/owl-ref/, February 2004.

[10] Dave Beckett. RDF/XML Syntax Specification (Revised). http://www.w3.org/TR/rdf-syntax-grammar/, 2004. W3C Recommendation.

[11] T. Berners-Lee. Semantic web: Where to direct our energy?, 2003. Keynote speech at the 2nd International Semantic Web Conference (ISWC2003), USA. http://www.w3.org/2003/Talks/1023-iswc-tbl/.

[12] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, May 2001.

[13] Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfrachni, and Daniela Petrelli. Hybrid search: Effectively combining keywords and semantic searches. In Bechhofer et al. [8], pages 554–568.

[14] Mehul Bhatt, Carlo Wouters, Andrew Flahive, J. Wenny Rahayu, and David Taniar. Semantic completeness in sub-ontology extraction using distributed methods. In Antonio Laganï, Marina L. Gavrilova, Vipin Kumar, Youngsong Mun, Chih Jeng Kenneth Tan, and Osvaldo Gervasi, editors, *ICCSA (3)*, volume 3045 of *Lecture Notes in Computer Science*, pages 508–517. Springer, 2004.

[15] Chris Bizer and Andreas Schulz. Berlin SPARQL Benchmark (BSBM). http://www4.wiwiss.fu-berlin.de/bizer/BerlinSPARQLBenchmark/, 2008. Website seen on 2008-06-01.

[16] Elena Paslaru Bontas and Malgorzata Mochol. Towards a cost estimation model for ontology engineering. In Rainer Eckstein and Robert Tolksdorf, editors, *Berliner XML Tage*, pages 153–160, 2005.

[17] M. Bouzid and N. Henze. Proceedings of the International Workshop on Semantic Web Personalization, co-located with the 3rd European Semantic Web Conference, Budva, Montenegro, June 12, 2006.

[18] Simone Braun, Andreas Schmidt, Andreas Walter, Gabor Nagypal, and Valentin Zacharias. Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference (WWW 07), Banff, Canada*, 2007.

[19] Ingo Brunkhorst and Nicola Henze. User awareness in semantic portals. In *Proceedings of the International Workshop on Personalization on the Semantic Web PerSWeb'05*, 2005.

[20] Peter Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129, 1996.

[21] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[22] Remo A. Burkhard. Learning from architects: The difference between knowledge visualization and information visualization. In *Proceedings of the eigth International Conference on Information Visualization (IV04)*, July 2004.

[23] K. Cardinaels, M. Meire, and E. Duval. Automating metadata generation: the simple indexing interface. *Proceedings of the 14th international conference on World Wide Web*, pages 548–556, 2005.

[24] F. Carmagnola, F. Cena, C. Gena, and I. Torre. A multidimensional approach for the semantic representation of taxonomies and rules in adaptive hypermedia systems. In *Proceedings of the Workshop on Personalisation on the Semantic Web: PerSWeb '05*, 2005.

[25] A. Cheyer, J. Park, and R. Giuli. Iris: Integrate, relate. infer. share. In *1st Workshop on The Semantic Desktop. 4th International Semantic Web Conference, p. 15, Nov 2005.* DTIC Research Report ADA454793, 2005.

[26] P.A. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl, and R. Paiu. Activity based metadata for semantic desktop search. *Proceedings of the 2nd European Semantic Web Conference*, 2005.

[27] Aron Culotta, Ron Bekkerman, and Andrew Mccallum. Extracting social networks and contact information from email and the web. In *In CEAS-1*, 2004.

[28] Mathieu d'Aquin, Anne Schlicht, Heiner Stuckenschmidt, and Marta Sabou. Ontology modularization for knowledge selection: Experiments and evaluations. In Roland Wagner, Norman Revell, and Gïnther Pernul, editors, *DEXA*, volume 4653 of *Lecture Notes in Computer Science*, pages 874–883. Springer, 2007.

[29] Stefan Decker, Jack Park, Dennis Quan, and Leo Sauermann. The semantic desktop - next generation information management & collaboration infrastructure. proc. of semantic desktop workshop at the iswc, galway, ireland. volume 175 of *CEUR Workshop Proceedings ISSN 1613-0073*, November 2005.

[30] Alexandre Delteil, Roberta Cuel, and Vincent Louis. Knowledge web technology roadmap. Technical Report 44, University of Trento, 2007.

[31] Vladan Devedzic. Understanding ontological engineering. *Communications of the ACM*, 45(4):136–144, 2002.

[32] Paul Doran. Ontology reuse via ontology modularization. In *KnowledgeWeb PhD Symposium 2006 (KWEPSY2006)*, 2006.

[33] L. Downes and C. Mui. Unleashing the killer app. *Harvard Business School Press*, 2000.

[34] J. Ellman. Corporate ontologies as information interfaces. *IEEE Intelligent Systems*, 10 No.1:79–80, 2004.

[35] Susan Feldman. Special idc report. enterprise search technology: Information desaster and the high cost of not finding information. *Portals Magazine*, 12, 2003.

[36] L. Gillam, K. Ahmad, S. Ahmad, M. Casey, D. Cheng, T. Taskaya, P.C.F. Oliveira, and P Manomaisupat. Economic news and stock market correlation: A study of the uk market, 2002.

[37] Bernardo Cuenca Grau, Bijan Parsia, Evren Sirin, and Aditya Kalyanpur. Automatic partitioning of owl ontologies using e-connections. In Ian Horrocks, Ulrike Sattler, and Frank Wolter, editors, *Description Logics*, volume 147 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.

[38] Jim Gray. The transaction concept: Virtues and limitations. In *Proceedings of the 7th International Conference on Very Large Data Bases*, pages 144–154, 1981.

[39] T. Groza, S. Handschuh, K. Moller, G. Grimnes, L. Sauermann, E. Minack, M. Jazayeri, C. Mesnage, G. Reif, and R. Gudjonsdottir. The NEPOMUK Project-On the Way to the Social Semantic Desktop. *3rd Int. Conf. on Semantic Technologies (I-SEMANTICS)*, 2007.

[40] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[41] P. Haase, M. Ehrig, A. Hotho, and B. Schnizler. Personalized information access in a bibliographic peer-to-peer system, 2004.

[42] C. Halaschek-Wiener, B. Parsia, and E. Sirin. Description logic reasoning with syntactic updates. *Proc. of ODBase2006*, 2006.

[43] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc, 2006.

[44] Jens Hartmann, Elena Paslaru Bontas, Raúl Palma, and Asunción Gómez-Pérez. Demo - design environment for metadata ontologies. In York Sure and John Domingue, editors, *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 427–441. Springer, 2006.

[45] Randall Hauch, Alex Miller, and Rob Cardwell. Information intelligence: metadata for information discovery, access, and integration. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 793–798, New York, NY, USA, 2005. ACM.

[46] Jeff Heflin and James A. Hendler. Dynamic ontologies on the web. In *Proc. of AAAI/IAAI 2000*, pages 443–449, 2000.

[47] Nicola Henze. Personalized e-learning in the semantic web. *International Journal of Emerging Technologies in Learning (iJET)*, 1(1), 2006.

[48] Martin Hepp. Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing*, 11(1):90–96, 2007.

[49] J.J. Jung, I. Ha, S. Ghose, and G. Jo. Collaborative User Tracking for Community Organization on Blogosphere: A Case Study of eLearning@ BlogGrid. *LECTURE NOTES IN COMPUTER SCIENCE*, 4256:276, 2006.

[50] Aditya Kalyanpur, Bijan Parsia, Evren Sirin, Bernardo Cuenca Grau, and James A. Hendler. Swoop: A web ontology editing browser. *J. Web Sem.*, 4(2):144–153, 2006.

[51] Judy Kay and Andrew Lum. Ontology-based user modelling for the semantic web. In *Online Proceedings of the UM (User Modelling) 2005 Workshop on Personalisation on the Semantic Web (PerSWeb)*, pages 11–19, 2005.

[52] Michel A. C. Klein. Versioning of distributed ontologies. Technical Report Del 20, Vrije Universiteit Amsterdam, december 2002.

[53] A. Kobsa. Generic user modeling systems. *User Modelling and User-Adapted Interaction Journal*, 11: 49-63, 2001.

[54] April Kontostathis, Leon Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer-Verlag, 2003.

[55] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series, 2000.

[56] Jay Liebowitz. *Knowledge Management Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 1999.

[57] A. Lïger, J. Heinecke, L. Nixon, P. Shvaiko, J. Charlet, P. Hobson, and F Goasdouï. *Semantic Web for Business: Cases and Applications*, chapter Semantic Web take-off in an Industry Perspective. IGI, 2008.

[58] Bill MacCartney, Sheila A. McIlraith, Eyal Amir, and Tomïs E. Uribe. Practical partition-based theorem proving for large knowledge bases. In Georg Gottlob and Toby Walsh, editors, *IJCAI*, pages 89–98. Morgan Kaufmann, 2003.

[59] Aimilia Magkanaraki, Val Tannen, Vassilis Christophides, and Dimitris Plexousakis. Viewing the semantic web through rvl lenses. In Dieter Fensel, Katia P. Sycara, and John Mylopoulos, editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*, pages 96–112. Springer, 2003.

[60] Eetu Makela. Survey of semantic search research. In *Proceedings of the Seminar on Knowledge Management on the Semantic Web*. Department of Computer Science, University of Helsinki, 2005.

[61] Christoph Mangold. *Konzepte und Realisierung einer kontextbasierten Intranet-Suchmaschine*. Doctoral thesis, University of Stuttgart, Faculty of Computer Science, Electrical Engineering, and Information Technology, Germany, November 2007.

[62] Christoph Mangold. A survey and classification of semantic search approaches. *Int. J. Metadata Semant. Ontologies*, 2(1):23–34, 2007.

[63] Mariano Fernandez and Asuncion Gomez-Perez and Natalia Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, pages 33–40, Stanford, USA, March 1997.

[64] G. Miller et al. WordNet: An Electronic Lexical Database. *Communications of the ACM*, 38(11):39–41, 1995.

[65] Satoshi Morinaga and Kenji Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 811–816. ACM, 2004.

[66] J. Najjar, M. Wolpers, and E. Duval. Attention Metadata: Collection and Management. *WWW2006 Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection. Edinburgh, Scotland (23 May 2006), http://www. cs. kuleuven. ac. be/najjar/papers/www2006. pdf*, 2006.

[67] Natalya F. Noy and Mark A. Musen. Promptdiff: A fixed-point algorithm for comparing ontology versions. In *National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, pages 744–750, Edmonton, Alberta, Canada, July 2002.

[68] Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.

[69] Natalya Fridman Noy and Mark A. Musen. Specifying ontology views by traversal. In Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors, *International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, pages 713–725. Springer, 2004.

[70] Streibel Olga. Xml-clearinghouse report 17: Xml-technologies and semantic web for trend mining in business applications. Technical report, Freie Universität Berlin, XML-Clearinghouse Project, 2007.

[71] K. OïHara, H. Alani, Y. Kalfoglou, and N. Shadbolt. *Agent Systems in Electronic Business*, chapter Features for Killer Apps from a Semantic Web Perspective. IGI Global, 2007.

[72] Jonathan Palmer. Designing for web site usability. *Computer*, 35(7):102–103, 2002.

[73] Jianguo Pan, Bofeng Zhang, Shufeng Wang, and Gengfeng Wu. A personalized semantic search method for intelligent e-learning. *ipc*, 0:11–14, 2007.

[74] Elena Paslaru-Bontas. *A Contextual Approach to Ontology Reuse: Methodology, Methods, and Tools for the Semantic Web*. Dissertation, Freie Universitaet Berlin, 2007.

[75] Helena Sofia Pinto and ao P Martins Jo˙A methodology for ontology integration. In *K-CAP '01: Proceedings of the 1st international conference on Knowledge capture*, pages 131–138, New York, NY, USA, 2001. ACM.

[76] S. Pinto, C. Tempich, S. Staab, and Y. Sure. *Semantic Web and Peer-to-Peer*, chapter Distributed Engineering of Ontologies (DILIGENT), pages 301–320. Springer Verlag, 2006.

[77] Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. Recommentation, W3C, January 2008.

[78] L. Sauermann, G.A. Grimnes, M. Kiesel, C. Fluit, H. Maus, D. Heim, D. Nadeem, B. Horak, and A. Dengel. Semantic Desktop 2.0: The Gnowsis Experience. *EPOS*, 6:7, 2006.

[79] Leo Sauermann, Ansgar Bernardi, and Andreas Dengel. Overview and outlook on the semantic desktop. In Dennis and Leo Sauermann, editors, *Proceedings of the 1st Workshop on The Semantic Desktop at the ISWC 2005 Conference*, 2005.

[80] Sebastian Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *WETICE '06: Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pages 388–396, Washington, DC, USA, 2006. IEEE Computer Society.

[81] S. Schwarz and T. Roth-Berghofer. Towards goal elicitation by user observation. In Andreas Hotho and Gerd Stumme, editors, *Proceedings of the LLWA 2003*, pages 224–228, Karlsruhe, oct 2003. AIFB Karlsruhe, GI.

[82] Julian Seidenberg and Alan Rector. Web ontology segmentation: analysis, classification and use. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 13–22, New York, NY, USA, 2006. ACM.

[83] Elena Paslaru Bontas Simperl and Christoph Tempich. Ontology engineering: A reality check. In Robert Meersman and Zahir Tari, editors, *OTM Conferences (1)*, volume 4275 of *Lecture Notes in Computer Science*, pages 836–854. Springer, 2006.

[84] Elena Paslaru Bontas Simperl and Christoph Tempich. Ontology engineering: A reality check. In Robert Meersman and Zahir Tari, editors, *OTM Conferences (1)*, volume 4275 of *Lecture Notes in Computer Science*, pages 836–854. Springer, 2006.

[85] Elena Paslaru Bontas Simperl, Christoph Tempich, and York Sure. : A cost estimation model for ontology engineering. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 625–639. Springer, 2006.

[86] H. Stuckenschmidt and M. Klein. Structure-based partitioning of large class hierarchies. In *In Proceedings of the 3rd International Semantic Web Conference*, 2004.

[87] York Sure and Rudi Studer. On-to-knowledge methodology — expanded version. On-To-Knowledge deliverable 17, Institute AIFB, University of Karlsruhe, 2002.

[88] Mari Carmen Surez-Figueroa, Guadalupe Aguado de Cea, Carlos Buil, Klaas Dellschaft, Mariano Fernndez-Lpez, Andrs Garca, Asuncin Gmez-Prez, German Herrero, Elena Montiel-Ponsoda, Marta Sabou, Boris Villazon-Terrazas, and Zheng Yufei. D5.4.1. neon methodology for building contextualized ontology networks. NeOn Project Deliverable D5.4.1., UPM, FEBRUARY 2008.

[89] Vlad Tanasescu and Olga Streibel. Extreme tagging: Emergent semantics through the tagging of tags. In Liming Chen, Philippe Cudré-Mauroux, Peter Haase, Andreas Hotho, and Ernie Ong, editors, *ESOE*, volume 292 of *CEUR Workshop Proceedings*, pages 84–94. CEUR-WS.org, 2007.

[90] Clïudio Teixeira, Joaquim Sousa Pinto, and Joaquim Arnaldo Martins. User profiles in corporate scenarios. *iciw*, 0:614–619, 2008.

[91] Christoph Tempich, Elena Paslaru Bontas Simperl, Markus Luczak, Rudi Studer, and Helena Sofia Pinto. Argumentation-based ontology engineering. *IEEE Intelligent Systems*, 22(6):52–59, 2007.

[92] Efraim Turban and Louis E. Frenzel. *Expert Systems and Applied Artificial Intelligence*. Prentice Hall Professional Technical Reference, 1992.

[93] Victoria S. Uren, Yuangui Lei, and Enrico Motta. Semsearch: Refining semantic search. In Bechhofer et al. [8], pages 874–878.

[94] D. Vallet, Ph. Mylonas, M. A. Corella, J. M. Fuentes, P. Castells, and Y. Avrithis. A semantically-enhanced personalization framework for knowledge-driven media services. In *Proceedings of IADIS WWW/Internet Conference(ICWI '05), Lisbon, Portugal, October 19-22, 2005.*, 2005.

[95] Max Völkel and Tudor Groza. Semversion: Rdf-based ontology versioning system. In *Proceedings of the IADIS International Conference WWW / Internet 2006 (ICWI 2006)*, 2006.

[96] Raphael Volz, Daniel Oberle, and Rudi Studer. Views for light-weight web ontologies. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 1168–1173, New York, NY, USA, 2003. ACM.

[97] H. Woolf, editor. *Webster's new world dictionary of the American language*. G. and C. Merriam, 1990.

[98] Valentin Zacharias and Simone Braun. Soboleo - social bookmarking and lightweight ontology engineering. In *Proceedings of the International Workshop on Social and Collaborative Construction of Structured Knowledge (CKC)*, 2007.

[99] scar Corcho, Mariano Fernndez-Lpez, Asuncin Gmez-Prez, and Angel Lpez-Cima. Building legal ontologies with methontology and webode. In V. Richard Benjamins, Pompeu Casanovas, Joost Breuker, and Aldo Gangemi, editors, *Law and the Semantic Web*, volume 3369, pages 142–157, 2003.