EU-IST Project IST-2003-506826 SEKT
SEKT: Semantically Enabled Knowledge Technologies

5.1.1 Knowledge Access on the Semantic Web

I. Thurlow (BT)
N. Kings (BT)
J. Davies (BT)

Abstract

Surveys of relevant research and commercial knowledge access tools were undertaken to ensure that research and development in the SEKT project takes into account and extends previous relevant work.

This document provides an overview of the set of technology surveys undertaken for the SEKT knowledge access workpackage in the areas of: (A) search and browse; (B) knowledge sharing; (C) visualisation and organisation of information; (D) user profile construction; (E) natural language generation; and, (F) knowledge repurposing. For each area, a brief overview of the survey and the associated recommendations are given.

The detailed technology surveys, in each area, are presented as Annexes to this document.

Keyword list: knowledge access, technology survey.

Workpackage:	Knowledge Access
Nature of deliverable:	Report
Dissemination level:	PU
Contractual date of delivery:	30th June 2004
Actual date of delivery:	3rd August 2004

SEKT Consortium

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

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

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

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

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

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

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

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

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

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

Sirma AI EOOD (Ltd.)
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768, Fax: +359 2 9768 311
Contactperson: Atanas Kiryakov
E-mail: naso@sirma.bg

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

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

Executive Summary

The explosion of information and the desire for making better use of corporate knowledge demands radically new technology and processes. One of the main objectives of the Semantically Enabled Knowledge Technologies (SEKT) project[1] is to develop the next generation of tools and technology that will help modern businesses use corporate knowledge to their advantage. The knowledge access workpackage within the SEKT project aims to develop context-aware tools that enable users to access this semantically annotated knowledge.

Pre-existing relevant technology was surveyed. In order to identify the limitations of this current technology and the opportunities afforded by the use of Semantic Web technology, a number of research themes have been identified with regard to the knowledge access tools to be developed in SEKT.

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

Table of Contents

Executive Summary **3**

Table of Contents **4**

1 Introduction **5**

2 SEKT Knowledge Access **6**

2.1 The Knowledge Access Vision 6
2.2 SEKT and Knowledge Access 8

3 Technology Surveys: Overview of Findings **13**

3.1 Search and Browse 13
3.1.1 State of the Art 13
3.1.2 Opportunities for Research and Development 14
3.2 Knowledge Sharing 15
3.2.1 State of the Art 15
3.2.2 Opportunities for Research and Development 16
3.3 Visualisation and Organisation of Information 18
3.3.1 State of the Art 18
3.3.2 Opportunities for Research and Development 19
3.4 User Profile Construction 20
3.4.1 State of the Art 20
3.4.2 Opportunities for Research and Development 21
3.5 Natural Language Generation 23
3.5.1 State of the Art 23
3.5.2 Opportunities for Research and Development 24
3.6 Device and Content Repurposing 25
3.6.1 State of the Art 25
3.6.2 Opportunities for Research and Development 27

4 Conclusions **29**

5 Glossary **30**

6 References **31**

1 Introduction

Surveys of relevant research and commercial tools were undertaken to ensure that SEKT research and development builds upon previous work. This document provides an overview of the set of technology surveys undertaken for the SEKT knowledge access workpackage.

Section 2 gives an overview of the SEKT vision for knowledge access.

An overview of the content of each technology survey is given in Sections 3.1 to 3.6.

The full technology surveys are provided as Annexes of this document. Each survey gives an overview of relevant research and technology, and identifies key areas for further research:

- Annex A: search and browse.
- Annex B: knowledge sharing.
- Annex C: visualisation and organisation of information.
- Annex D: user profile construction.
- Annex E: natural language generation.
- Annex F: knowledge repurposing.

2 SEKT Knowledge Access

2.1 The Knowledge Access Vision

Corporate knowledge is becoming an increasingly important asset. The explosion of information and the desire for greater efficiency and effectiveness in the application of corporate knowledge demands radically new technology and processes. One of the main aims of the SEKT project is to develop the next generation of technology and tools that will help modern businesses use corporate knowledge to their best advantage. The approach taken in the SEKT project is to make extensive use of technology that augments sources of information with meta-data in the form of semantically enriched annotations.

Ontologies, and associated meta-data, will provide the mechanism needed to integrate knowledge resources that are dispersed across multiple repositories into a coherent corpus of interrelated knowledge. Systems that semi-automate the creation and maintenance of ontologies and meta-data will be developed. These systems will be based upon foundational research in four main areas: the semi-automatic generation of ontologies, the semi-automatic extraction of meta-data, ontology and meta-data management, and ontology mediation. Each of the technologies will be realised as software modules that will inter-operate through the use of emerging WWW standards, for example RDF [1], RDFS [2], and OWL [3].

New applications, developed from the knowledge access tools and components developed within SEKT, will exploit ontologies and semantic annotations, enabling people to access, find and ultimately share relevant corporate knowledge in a context-sensitive and more meaningful way. People will be able to access these applications using a variety of devices operating over a range of access mechanisms, e.g. some users will access the applications using a PC connected to a corporate LAN, whilst others will access the applications using a PDA operating over a wireless connection.

Knowledge access applications will take into account the user’s current activities, their interests and preferences, and the capabilities of the access device being used in the following ways:

- Knowledge discovery technology and components, such as semantically enriched indexes and information clustering technology, will enable relevant information to be retrieved from the mass of information stored on an organisation's Intranet.

5.1.1 Knowledge Access on the Semantic Web

- Search engines will combine free-text search with a capability to exploit semantic meta-data in the search and browse process. Search engines will provide a single point of access to both annotated and unannotated information, revealing implicit structures, and allowing a controlled level of inference when seeking answers to queries.
- Semantically aware search agent components will offer improved relevance and timeliness of knowledge delivery, for example enable relevant information to be discovered, filtered, and delivered to people when it is of most use to them.
- Knowledge agents will exploit the interrelationships between selected pieces of knowledge, and put otherwise isolated knowledge into a meaningful context.
- Visualisation technology will enable retrieved information to be viewed and explored in an intuitive way.
- User specific, context aware profiles will be built by components that monitor a person's interactions with the knowledge access applications, corporate information sources, and repositories.
- Natural language generation techniques will be used to transform the retrieved knowledge into text and enable that knowledge to be delivered at the appropriate level of detail.
- Content repurposing modules will render knowledge into an appropriate format for the user’s current context. For example, a summary of a document could be provided for SMS delivery, or a set of meta-data facts could be rendered into natural language suitable for delivery as a text document.

Knowledge can only reach its full potential value as useful knowledge if it is available to the right people at the right time. Collaboration services and components will provide the core functions that support semantically aware knowledge sharing tools. The knowledge sharing tools will automatically disseminate relevant shared knowledge directly to those users in a community who have a need for it. As more items are shared, a high quality knowledge resource, for the community, will be built up over time.

Provided with an ontology meeting the needs of a particular community of practice, the knowledge sharing application will arrange knowledge assets into the predefined conceptual classes of the ontology. The conceptual classes will be defined such that they exhibit the resemblance and differences essential to enable the knowledge worker to draw conclusions about relationships and implications. Having the means to organise knowledge assets according to a given ontology, the knowledge sharing application will provide the knowledge worker with effective means for browsing the documents within the resource. In addition, community ontologies tend to evolve over time and the knowledge access tool will be designed to detect such changes and intelligently change the ontology to reflect the changing conceptualisation of the community.

Ontologies will also allow the knowledge-sharing application to facilitate knowledge exchange among knowledge workers in a community, since meaningful interaction can only occur when people share a common interpretation of the vocabulary used in their communication. Furthermore, the knowledge sharing application will allow different communities of practice to access knowledge via the ontology that is most appropriate to them.

Ontology-supported search functions will enable users to find relevant information at a time convenient to them. Engineers and salespeople, for example, may have a very different perspective on how different pieces of knowledge relate to one another. The challenge is therefore to present a natural and intuitive interface onto a knowledge repository for each class of user, giving them their own individualised views onto the repository.

The SEKT vision for improved knowledge access will be facilitated through research and development of six key complementary technologies: (A) search and browse; (B) knowledge sharing; (C) visualisation and organisation of information; (D) user profile construction; (E) natural language generation; and, (F) knowledge repurposing.

2.2 SEKT and Knowledge Access

The SEKT knowledge access tools and components will build upon research and development in the fields of search and browse, knowledge sharing, visualisation and organisation of information, user profile construction, knowledge generation, and knowledge repurposing. A set of context-aware tools and supporting components that enable people to access semantically-annotated knowledge will be developed. Figure

5.1.1 Knowledge Access on the Semantic Web

1 shows a knowledge management functional architecture and indicates a mapping to the high level SEKT systems architecture, shown in Figure 2.

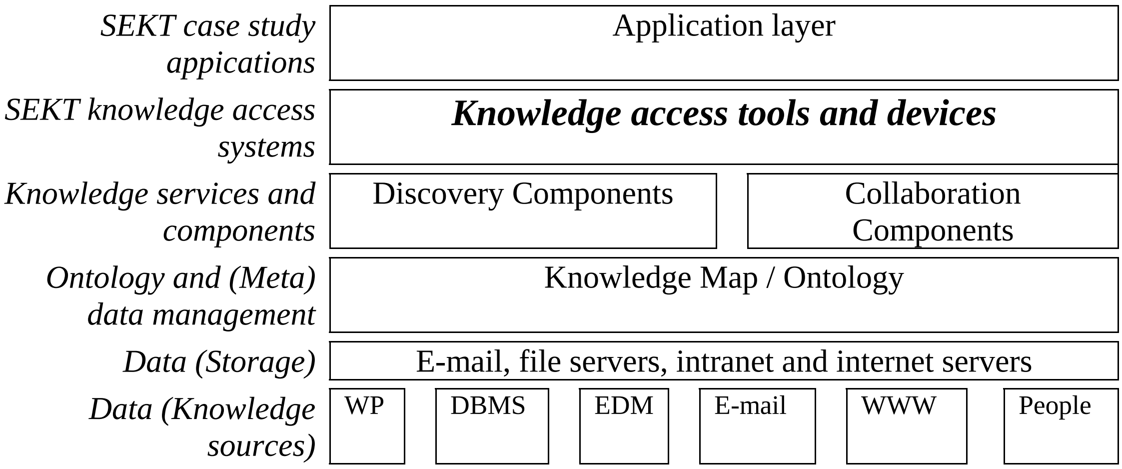

Figure 1. SEKT Knowledge Management functional overview (based upon Ovum [4]).

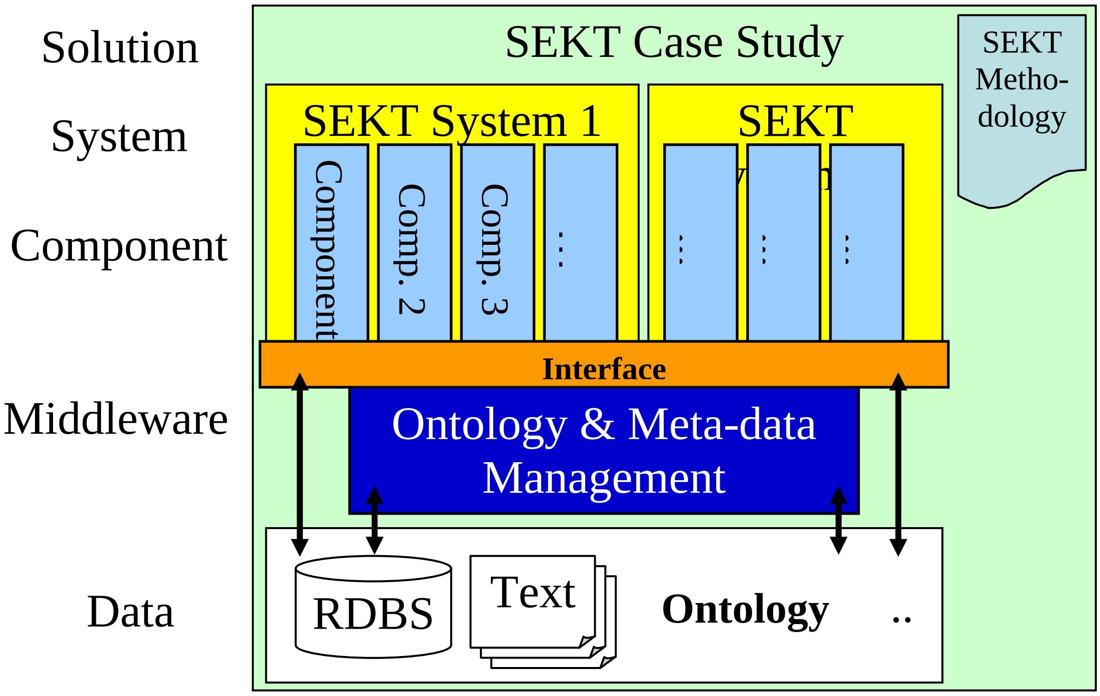

Figure 2. SEKT high level architecture.

The research activities within SEKT are primarily concerned with the upper levels of the architecture shown:

- Application Layer

  Applications are built from one or more knowledge access tools, for example, a digital library portal may be constructed from search, browse, knowledge sharing, expertise location and visualisation tools. The scope of SEKT Work Package 5,

5.1.1 Knowledge Access on the Semantic Web

"Knowledge Access", is restricted to providing search & browse, knowledge sharing, and visualisation & organisation tools, underpinned by user profiling and natural language generation components. WP6, "Integration", is concerned with creating end-user applications from components and systems.

People will access their applications from a variety of devices over a range of access mechanisms, e.g. some users will access the applications from their PC over a corporate LAN, whilst others will access the applications via a PDA over a restricted bandwidth wireless connection. The modular approach to building applications from underlying components and services allows the applications to be easily tailored to different user needs, and different information requirements. The applications provide the user interface that brings the underlying services together in a consistent and easily accessible way.

- Knowledge access tools and devices

  The knowledge access tools (search and browse, knowledge sharing, and visualisation) are built from the underlying knowledge discovery and collaboration services/components, e.g. a knowledge sharing tool may be built by integrating alerting, and user profiling and document analysis components.

- Discovery Components

  The discovery services and components provide the core functions that enable relevant information to be retrieved from an organisation's information and knowledge sources. Typical discovery components include semantically-enhanced search engine technology that “crawls” and indexes information sources in a distributed, heterogeneous environment, semantic analysis components (e.g. document analysis), document clustering components, text mining components and natural language generation (from underlying ontologies).

  Other discovery services or components include those for linguistic analysis (e.g. a word stemmer), profile matching components, and domain specific thesauri services that can be used, for example, to reformulate a user supplied query. Components that monitor a person's interactions with their applications in order to build user specific, context aware profiles will also be provided in this layer. Search agent components (push technology) will allow relevant information to be discovered, filtered, and then delivered seamlessly to people when it is of relevance to them.

- Collaboration Components

  The collaboration services and components provide the core functions that support the knowledge sharing tools. These services and components support the creation and maintenance of communities of interest and make it possible to build tools that applications/people can use to share knowledge within an organisation.

- Knowledge Map / Ontology

  The knowledge map provides the domain specific ontologies[2] and their associated meta-data, and maps the information sources to domain specific views of those sources. Thus, information from a general information source such as a web server, a news feed, or a library catalogue can be presented in a tailored fashion. The domain specific ontology provides a logical overview of the distributed, heterogeneous information resources to the discovery and collaboration tools/components, enabling multiple domain specific searches and sharing services to be constructed, thereby giving people with different information/knowledge requirements alternative views on the same data.

  The visualisation tools/components that help people to visualise/navigate information resources of importance to them also makes use of ontologies (and associated meta-data) to provide user specific views of those resources, enabling different user groups to view the information which is of most relevance to their needs.

- Knowledge Repository

  The knowledge repository provides the functions for the creation, population, maintenance, and mediation of the organisation's information/knowledge through the use of ontologies and their associated meta-data. The information stored ranges from simple user profiles to extensive document classification schemes. Knowledge Discovery technology, e.g. data mining, text mining, and web mining, is used to construct/generate and populate the ontologies semi-automatically. Human Language Technology (HLT) populates these ontologies further by

2 Each application of SEKT will require domain-specific ontologies, but these will be supplemented by general SEKT ontologies which are common to all applications. The SEKT ontologies will be used to represent areas such as user profiles, etc.

extracting meta-data from existing content. The ontologies (and associated meta-data) stored in the knowledge repository are managed, maintained and evolved by ontology management components developed within the SEKT consortium. Reasoning components that enable reasoning across distributed heterogeneous knowledge bases/information resources are also provided in this layer.

3 Technology Surveys: Overview of Findings

Surveys of relevant research and commercial tools were undertaken to ensure that SEKT research and development builds upon previous work. An overview of the content of these surveys is given in the following sections, covering a summary of the state of the art and the identified research and development agenda for SEKT. These summaries inevitably omit some details and the reader is referred to the attached annexes for full information.

3.1 Search and Browse

3.1.1 State of the Art

A number of limitations of current search technology were identified. Corporate search engines based on conventional IR techniques alone tend to offer high recall and low precision. These search engines generally treat search in isolation: the results from a search engine for a given query are identical, being independent of the user submitting the query and the context in which the user made the request. The majority of users do not make use of a search engine's advanced search options [5], even though they are openly available. Furthermore, the majority of users tend not to view more than the first few pages of results returned from the search engine [6]. Although search agents can reduce the overhead of completing a manual search for information, user reluctance at using advanced search functions may prevent such tools being used to their full advantage.

Research and commercial search engines have demonstrated that significant improvements in the relevancy of search engine results can be obtained by augmenting traditional IR matching techniques with algorithms that exploit the hyperlink structure of the WWW [7]. Although the hyperlink structure of a typical corporate Intranet is expected to differ from that of the WWW, as many links are provided simply to ease navigation, it has been shown that similar link-based algorithms can be usefully applied to searching corporate Intranets [8]. Research has also demonstrated that presentation and simple navigation of results enable users to focus on items in categories of interest rather than having to browse through a long list of sequential results, e.g. [9,10,11]. Improved clustering, presentation and navigation of results are now available in many commercial search engines

3.1.2 Opportunities for Research and Development

Search engines that make use of the underlying ontological representation of an organisation's information are expected to improve search further and alleviate the problems associated with conventional search services. Personalised, context-sensitive, domain-specific, search services will be further researched and developed. Rather than attempting to provide better IR algorithms to overcome the problems associated with conventional search engines, Semantic Web technology enabled search engines have tended to augment traditional search by indexing and searching domain-specific semantic annotations, and by exploiting the structure of domain-specific ontologies [12, 13]. A number of search engines are now emerging that use techniques that apply domain-specific knowledge to the indexing, similarity evaluation, results augmentation and query enrichment processes. In addition, ontologies can provide an overarching knowledge structure to unify heterogeneous knowledge sources.

Semantically enriched queries enable search engines to retrieve documents indexed with terms that are not present in the query, but which relate to the overall concepts associated with the query. Greater precision can be achieved, for example, by filtering the results with contextual, domain-specific information. Furthermore, Semantic Web search engines that exploit domain-specific ontologies can augment the results from conventional IR search engines with relevant supplementary information. A domain-specific ontology consisting of terms, synonyms for those terms, and relationships between those terms and their synonyms can improve search engine recall. Fully automated systems that facilitate the annotation, indexing, and retrieval of documents with respect to named-entities are being researched and developed.

A person's search context will be influenced by the many different roles, responsibilities and interests they may have. Search engines need to provide more personalised search services that focus on the information needs of an individual user, rather than trying to address the needs of everybody. Although improvements have been made in identifying the user's search context, such as deciding which possible meaning of a word is being used in order to disambiguate a query, further research is required to find better ways to determine and express the user's information needs, and to match those needs to relevant content. There is a need for search services to interpret user intentions and context at the time of the search and to learn user preferences dynamically and adaptively. The identity of the individual, the place, the time, the purpose can all help understand context.

Further research and development of tools that monitor a user's information and application usage in order to build a more accurate profile of a user's information needs should be investigated. The prompt detection of a user's changing information need, e.g. in conjunction with a role change, will be an important factor in ensuring that user queries are augmented with information from the most appropriate search context. Search agents that proactively fetch information on behalf of the user, based on the user's personal interest profile and/or current activity, should be researched further. This kind of seamless knowledge delivery integrates knowledge access more tightly with day-to-day business processes: making access to knowledge as painless as possible for the end-user (which could be key to the uptake of knowledge management technology). The link structures and inter-relationships between user profiles may give further useful information that can be exploited by the search engine. The results of a search should be presented to the user in a form that is easy to view, simple to navigate, and in context with the user's information needs.

3.2 Knowledge Sharing

3.2.1 State of the Art

Knowledge sharing tools combine the functions of locating and distributing information within some form of community[3]. As a user requires information to undertake a task, information relevant to that task can be located. Underpinning knowledge sharing tools is the premise that someone in the user's wider community has already created relevant information (explicit knowledge transfer) or someone is able to provide help or advice (tacit knowledge transfer). More sophisticated knowledge sharing tools use profiles of the users to identify and route information to people, as and when that information is required. Understanding a user's changing context emerges as a central research problem, only partially addressed by today's technology.

Knowledge management is defined as the "systematic application of actions to ensure that an organisation obtains greatest benefit from the information that is available to

[3] Community here should be interpreted broadly – appropriate communities could range from a small project team to much larger organizational units and be more or less formal in nature.

it" [14]. Knowledge here is seen as the experience of the people within an organisation, combined with information gleaned from documents within the organisation, as well as access to relevant reports from the outside world. Warwick [14] suggests that effective knowledge management requires a combination of organisational, social, and managerial initiatives alongside the development of appropriate knowledge access technologies. Knowledge sharing software supports the activities to collate, categorise and distribute information [15].

It is important to state that not all forms of knowledge are treated the same. Nonaka (cited in [14]) formulated a theory of organisational learning that focussed on the transformation of tacit into explicit knowledge and visa versa. Tacit knowledge is held internally by each knowledge worker, and is formed by past experiences, beliefs and values: tacit knowledge informs the decision making process; and, Nonaka suggests that tacit knowledge is the most valuable form. Explicit knowledge is represented by some document, such as a web page or a video, which has been created with the goal of relaying some piece of tacit knowledge from one person to another. Organisational learning occurs as people participate in shared activities, and their knowledge is articulated, making it available to others [16].

The goal for any knowledge management system is to create an environment to facilitate knowledge exchange, through the use of Collaboration and Discovery services [4], as explained in Section 2.2 above. Whatever the modes of communication used, the primary goal of the software is to create group memory and team awareness [17]. To that end, Crossley *et al.* [18] suggest that the knowledge sharing systems must be sensitive to both the information needs of an individual, as well as the overall needs of the community; information should be delivered relevant to a current user's profile.

3.2.2 Opportunities for Research and Development

The purpose of a knowledge sharing tool is to allow information, created by an author in one context, to be used by others in other contexts. This suggests a number of areas where Semantic Web technology could be exploited.

5.1.1 Knowledge Access on the Semantic Web

Traditionally we think of a knowledge-sharing portal to which users' upload documents for sharing. Semantic technology offers a range of functionality here:

- The ability to (semi-)automatically extract meta-data from a document means that the system can more precisely target that document at relevant communities. There is no longer a requirement on the part of the document's author to think about to whom the document should be sent. The system, assuming it knows users' interests, undertakes this function.
- The knowledge of a user's interests, represented as a profile, can be created by the user. However, the profile can also be learned from the meta-data of the documents he or she habitually reads, or even creates.
- With the understanding gained from monitoring a user's information needs, the system can proactively push knowledge, from whatever source, to the user.
- The meta-data associated with the document will relate to one ontology, e.g. that of the author or of a library where it is stored. Through ontology mediation, this meta-data can be recast against the framework of the reader's own ontology. This helps the reader to see the document in his or her own context and, in particular, be able to locate that document at a future date.
- In addition, as a document is read by a varied audience, it can be annotated with information about the readership. This can be partly a conscious action on the part of the reader, but it can also draw on the readers' profiles. In this way the document's meta-data becomes richer as the readership expands. For example, an understanding of the document's existing readership will help target it on new readers, and also may guide a search engine undertaking a search for a third party.
- An extension of this is to use the portal to locate people with particular interests or knowledge, such as identifying an expert on a particular topic, or identifying a community of people with a shared interest. By extending the portal in the manner to facilitate expertise location, the portal supports the sharing of tacit as well as explicit knowledge.

In the context of a company intranet this portal approach is very valuable. However we know that much user knowledge is held on users' desktop machines. In a recent report [19], analysts Frost & Sullivan comment that *"Due to the fact that many [knowledge] resources never make it to the network drive, work is duplicated and productivity is encumbered. Implementing a searchable P2P file sharing system in the* *enterprise space will help to disseminate business critical information and increase productivity."* This suggests extending the portal to give access to knowledge on users' desktops. To do this, the portal would need to mediate between the portal ontology and the ontology employed by knowledge creators: the portal would need to mediate over the user's own file structure.

A further step will take us to a genuinely P2P approach [20]. Here a knowledge sharing tool is local to the user's desktop and the user is able to specify a user's interests through their own ontology, as represented by the PC's file structure. Mediation will be needed to translate between that local ontology and the local ontology of the user's colleagues; as well as between the local ontology and that of the corporate desktop.

3.3 Visualisation and Organisation of Information

3.3.1 State of the Art

Intuitive visualisation of knowledge will become more and more important to mitigate information overload problems as users demand easy location and recognition of relevant content. Web visualisation technologies, such as X3D, SVG and Java3D [21,22,23], are reviewed and compared and guidelines for technology selection presented in Annex C. Similarly, an overview of generic interactive graphical structures is provided along with appropriate guidelines.

There are a number of techniques that can be applied to the visualization of complex ontologies. These techniques, however, need to be applied with careful thought in the design of the user interface (UI). It is important to choose correctly the shapes and graphical properties of the elements that appear in the UI. This topic is particularly important in the case of three-dimensional visualizations. Some variety must exist, but it is preferable in general to use consistent shapes for the same type of information elements, and use colours or transparency levels to differentiate them.

Changing state is one of the most studied topics in the design of interfaces for information visualization applications. When there is a change at a general level, such as zoom or clustering, it is important to preserve context and keep visualizing the information in an adequate way. This is achieved by modifying the representation of the previous scene, reducing it or reformatting it, in order to provide a reference to the last state.

Real world metaphors are used in many applications to aid user understanding. The use of rooms, galleries, maps or planetary systems to represent relations among data can be useful in providing an intuitive interface. However, it is necessary to take into account that there are applications where the user needs to visualize nodes located in separate groups, and it should be able to visualize the different groups, or at least to change from one to another in a easy manner.

The presence of graphical elements for help, or *indicators* about how to manage the application should be realistic and clear, and the scene should not be overly complex. It is important to make the user notice the presence of certain elements, without breaking the overall scene atmosphere. Some indicators whose use is recommended are:

- Point of view changes.
- State indicators.
- Context locators.
- Browsing aids: visited paths, example paths.

3.3.2 Opportunities for Research and Development

Semantic Information Visualization is a recent trend that integrated experience from previous visualisation research for the specific purpose of the Semantic Web initiative. Since the Semantic Web is built upon the structuring of information (via ontologies), there are clear opportunities for information visualization.

However, some new problems have appeared in the area of Semantic Web content visualization and access. Since this content was originally designed for machine processing, its visualization often leads to unintuitive results. Based on the survey of existing applications and technologies, the use of specific technologies for converting this content into a more human readable one is forseen. For that purpose the concept of a visualization ontology [24] responsible for translating the machine-oriented data into a user-friendly visualization is introduced. Using a visualization ontology, both the application developer and the user can define how to represent the information and how to interact with it.

3.4 User Profile Construction

3.4.1 State of the Art

Personalised information delivery based on user and document profiling is an important step in making computers helpful and accessible to a wider community. Software that learns from people's behaviour when accessing knowledge repositories and the underlying ontologies will enable software agents to build up personal profiles of the user’s interests, roles, and context, i.e. current activity, user preferences, and device capabilities. Other knowledge access tools, e.g. the knowledge sharing tool, can use the profile to resolve a user’s context when searching, sharing and managing information.

Annex D gives an overview of the approaches to derive user profiles in the context of Semantic Web applications. The topic of user profiling for the Semantic Web is introduced. The process of collecting and pre-processing data from various sources, e.g. Web access logs, is described. An overview of the methods used to profile users and documents based on the content of text documents is presented. Collaborative filtering and web mining techniques are described and compared. In general, the area of user profiling is divided into three main subtopics: (1) Content Based User Profiling, (2) Collaborative User Profiling, and (3) Web Usage Mining for User Profiling.

Each of the three sub-areas provide different point of view to the analysis of data collected from an online human user and used to build a model: the “user profile”. “Content Based User Profiling” deals with the aspect of using the actual content of the text documents viewed by the user (typically from the web) which are used to build a model of content topics characteristic for a particular user. These models are further used to rank the content (in different contexts) for recommendations. “Collaborative User Profiling” deals with the aspect of calculating user communities with similar interests. The model of a user or a community consists from the most significant activities performed by them. The underlying assumption is that an individual user can benefit by being informed about the activities of other users with similar interests. “Web Usage Mining for User Profiling” is specialized for the analysis of web usage data coming usually from web-server log-files. Crucial questions are concerned with user identification and behavior modeling in the relation to the web-site structure (pages and links).

3.4.2 Opportunities for Research and Development

One of the key concerns of the Semantic Web is the generation and exploitation of ontologies and associated meta-data in various applications. In a sense, user profiles serve as models for providing a personalised level of semantic annotations: the user profile acts as a filter on a more general ontology to give a personal view. Furthermore, the profile of the author of a document could be used to help annotate the document. The particular interests of the author might, for example, influence how the document is classified or how entities within the document are annotated. The meta-data associated with an entity "Jaguar" might be different if we know the author of a document works within the motor industry, rather than if we know the author has an interest in South America. User models are built by automatic or semi-automatic means using machine learning and data mining methods. One of the most important challenges for building user-models is an efficient semi-automatic mode where only limited amount of human time is available for providing answers to different questions.

In real-world situations we would like to reduce the amount of that manual work needed by most of the approaches to automatic document filtering, categorization, user profiling, information extraction, text tagging. For instance, in document profiling for automatic document categorization, we start with a set of documents where each document is assigned to some categories based on its content. Using “unlabeled data” and “bootstrapping learning” are two directions that enable important reduction in the required amount of hand labeling and can potentially be useful for semantic Web applications.

The “unlabelled data” approach to document categorization proposed in [25,26] can be described as combining Expectation Maximization and the Naive Bayesian classifier. This approach uses a small number of labelled documents to classify further documents into, say, one of twenty news groups or classify the document into a type such as “student page” or “technical page”. Initially, a classifier is trained with only labelled documents. The trained classifier is then used to assign probabilistically-weighted class labels to all unlabeled documents. A new classifier is trained using all the documents and iterate until the classifier remains unchanged. The final result heavily depends on the quality of the labels assigned to the small set of hand labelled data: it is much easier to hand label a small set of good quality examples, rather than a larger set of medium quality examples.

Bootstrap learning to classify Web pages exploits the fact that most Web pages have some hyperlinks pointing to them. Each Web page can, therefore, be described either by its content or by the content of the hyperlinks that point to it. Initially, a small number of documents are hand labelled, and this set is used to make predicted classes for new documents. A small number of the documents for which the prediction was the most confident are added to the set of labelled documents and the process is repeated. The set of labelled documents is increased through a series of iterations, and the approach is successful as long as the initial labels gave good coverage of the problems space. This approach was proposed in [27] and supported by the computational learning theory in the proposed Co-Training theorem.

Recent potentially relevant work includes also mining the extracted data [28], to collect information about different companies from the Web. As Web documents are naturally organized in a graph structure (via hyperlinks), there are also research efforts on using that graph structure to improve document categorization [29], to improve Web search and visualization of the Web. The usefulness of these methods in semantic Web context is a key area for future research which we intend to address in SEKT.

User modelling approaches relate to the Semantic Web technology in several ways:

1. In the first view, user modelling serves as a building block and as a technology for annotating information units (e.g. documents) with user-specific meta-data.
2. The second view is to use ontologies for modelling user profiles. This could be seen as
    a) Using ontologies as a background-knowledge when constructing user-profile, or
    b) A user-model being represented as an ontology (using ontology construction techniques).

The last approach (2b) is the most challenging one and is still under research in different contexts.

3.5 Natural Language Generation

3.5.1 State of the Art

The aim of Natural Language Generation (NLG) is to produce natural language text, tailored to the presentational context and the needs of the target reader, from structured data in a knowledge base. User profiles can be used in conjunction with NLG techniques to select appropriate presentation strategies, e.g. deliver short summaries to a user's WAP phone, or a present a longer text if the user is accessing the application using their desktop PC. NLG techniques can be used to present the same data to different users, depending on their levels of expertise, e.g. simple terminology can be used to explain unknown terms to the novice user, while more complex terminology and text style can be presented to the expert user. NLG can provide automated documentation of ontologies and knowledge bases.

NLG can play a number of roles, when applied to the Semantic Web, but NLG systems that are specifically targeted towards Semantic Web ontologies have started to emerge only recently. For example, there are some general purpose ontology verbalisers for RDF and DAML+OIL [30] and OWL [31]: these systems are based on templates and follow closely the ontology constructs, e.g., *"This is a description of John Smith identified by http://... His given name is John..."* [31]. The advantages of Wilcock's method [31] is that the approach is fully automatic and does not require a lexicon. A more recent system which generates reports from RDF and DAML ontologies is MIAKT [32]. In contrast to Wilcock's approach, MIAKT [32] requires some manual input (lexicons and domain schemas), but on the other hand it generates more fluent reports, oriented towards end-users, not ontology builders. It also uses reasoning and the property hierarchy to avoid repetitions, enable more generic text schemas, and perform aggregation.

On the other end of the spectrum are sophisticated NLG systems such as TAILOR [33], MIGRAINE [34] and STOP [35] which offer tailored output based on user/patient models. Systems like Wilcock's [31] and MIAKT [32] tend to adopt simpler approaches, exploring generalities in the domain ontology, because their goal is to lower the effort for customising the system to new domains. Sophisticated systems, while offering more flexibility and expressiveness, are difficult to adapt by non-NLG experts. For example, experience in MIAKT showed that knowledge management and Semantic Web ontologies tend to evolve over time, so it is essential to have an easy-to-maintain NLG approach.

The ONTOGENERATION project [36] explored the use of a linguistically oriented ontology (the Generalised Upper Model (GUM) [37] as an abstraction between generators and their domain knowledge base. The project developed a Spanish generator using systemic grammars and KPML [38]. The main difference from the MIAKT system [32], which also uses a similar approach, comes from the number of concepts and relations used to abstract the generator from the concrete domain ontology. MIAKT uses only 4 basic properties, in order to make it easier for non-linguists to carry out this task. The size and complexity of GUM make this process more difficult for non-experts. In general, there is a trade-off between expressivity and the number of linguistic constructs in the ontology. Therefore the more lightweight MIAKT approach is mainly suitable for applications where more schematic texts are sufficient and the goal is to have non-linguists being able to customise the generator for new domains.

3.5.2 Opportunities for Research and Development

The approach that will be focussed upon is applied (or shallow) NLG systems (e.g. MIAKT), the goal of which is to generate text from domain knowledge using computationally efficient and robust NLG algorithms.

The reason behind this choice comes from the fact that Semantic Web ontologies tend to have hundreds or thousands of concepts, thus requiring efficient methods for their processing. The main perspective of this survey is on the applicability of the methods to ontology-based generation and their “user-friendliness”, i.e., how easy or difficult it is for a non-NLG expert to modify the system's behaviour, for example, by providing new lexicalisation of concepts or modifying a text plan.

Another strand of relevant research has looked into the problem of building portable and customisable NLG systems from relational databases [39]. The ILEX approach requires a formal definition of domain knowledge as a taxonomy or an ontology and a mapping of ontology items to their lexicalisations. In the case of Semantic Web ontologies, the information about domain types and data types of the slot fillers is already formally specified, unlike in databases. Work on ILEX [39] is mainly focused on low-cost methods for providing adaptivity and generation of comparisons, some of which are also likely to be relevant in Semantic Web context.

3.6 Device and Content Repurposing

3.6.1 State of the Art

The aim of device independence is to deliver a functional presentation of a web page on any access medium or device [40]. Devices typically include PCs, PDAs, WAP phones and printers. Identification of the characteristics of the target device is a key requirement. The de-facto standard in this area is the W3C RDF-based CC/PP standard[4], and in particular the UAProf application, supported by most mobile phone vendors.

There are four broad approaches to device independence:

- Static Adaptation – the data is held in different formats produced by hand for each target device.
- Client Adaptation – the client browser is responsible for interpreting data streams in a device independent format
- Server Adaptation – the device characteristics are communicated to the server as part of the request, and the server responds with an appropriate data stream.
- Proxy Adaptation – a software agent mediates between the server and client, receiving data in one format from the server, and reformatting for delivery to the client.

In practice, successful adaptation is likely to combine all of these approaches. Manual authoring will be necessary to provide the optimum display for different devices. The details of line breaks and exact position are best left to the client. Proxies are likely to be used to convert between different protocols, such as HTTP and WAP. Only the server can provide support for very thin clients, and make efficient use of bandwidth.

Of these four approaches, static adaptation is generally regarded as resource intensive and a ‘last resort’, and client adaptation is outside the scope of this project. Proxy adaptation is typically concerned with providing a general service applicable to multiple independent web sites. Server adaptation is specific to a particular

[4] http://www.w3.org/Mobile/CCPP/

application. Annex F concentrates on server adaptation, although the distinction between server and proxy adaptation is somewhat blurred; software accepting output from SEKT software and adapting it for a client could be located either on the same server, or implemented as a separate specialist web service.

Some of the key requirements of a device independent platform are:

- Translation into a suitable language/representation e.g. HTML, WML.
- Selection of appropriate content e.g. long description, short description.
- Pagination of data to suit the target device.
- Provision of navigation links through the data structure.
- Provision of suitable input data structures.
- Image processing that takes account of bandwidth and device capabilities.

As a general principle in computer science, it is important to separate out the independent, orthogonal aspects of a problem in order to remove redundancy and reduce maintenance costs In the context of device independence it is useful to distinguish between content, layout, structure and style:

- Content refers to the raw data to be displayed.
- Style refers to such aspects as colour, font and background images.
- Layout refers to the geometric relationship of different components of the page.
- Structure refers to semantic relationships between different elements; these are likely to form the basis of navigational links, and may be used to suggest layout.

Selection of style is a relatively trivial matter since the advent of Cascading Style Sheets (CSS), a styling standard widely supported by commercial software, and selection of content based on device characteristics is also straightforward. The main challenge is to make use of the document structure to provide a suitable layout and this is the main focus of current research.

Most current commercial systems and proposals are based on the idea of adding extra semantic information to the document, usually in the form of extra HTML style mark-up. The semantics of the extra mark-up are built into the software, and are used to select content, produce a suitable layout in a suitable language, select appropriate multimedia file formats and so forth. The commercial MobileAware and Volantis systems both take this approach, as do the DIWAF, CONSENSUS and PRINCESS projects. Each of these defines its own small subset of extra markup tags extending

XHTML. The extended languages are variously called MobilityTags [41], DIML, RIML and so forth.

3.6.2 Opportunities for Research and Development

As part of the SEKT project, a device independent framework prototype application has been developed which generalises these various techniques. Instead of an extended mark-up language, the prototype defines a data model and a Java API. The basic data structure is a tree where each node consists of a set of attributes and a list of subtrees. Data structures can be built using the API, but are also sufficiently close to the XML model to allow XML documents to be read in directly. This data structure is then matched up against a template that allows data elements to be selected and embedded within a text document.

The net result is a system similar to XSL Transformations - more declarative and simpler but rather less powerful. This transformation framework is embedded in an architecture that makes CC/PP information available to the author. CC/PP attributes can be used both to select a particular layout template, and to select alternative data elements within a template. The framework can be used to replicate each of the applications mentioned above, as well as to experiment with different semantic constructs.

All of the systems and proposals mentioned above rely on the manual tagging of data with semantic information that can be used to select an appropriate layout. These systems are ‘unintelligent’; the user defines both the tags and the layout and there is no attempt to search for a suitable layout within some solution space.

There are a number of ways in which Semantic Web-based techniques could be used to enhance the adaptation:

- It might be possible to define a meaningful ontology of devices. This could identify a richer set of device classes, which would allow the author to write more sophisticated layouts for more precisely targeted devices. Such a classification system could also use inheritance to infer device attribute from parent classes.
- The systems above use device profiles to select content and layout. This could be extended by including personal profile information. Such a profile might give higher priority to certain information, for example it might give high priority to messages from certain people and this information could be used to enhance the layout of a page.

- A search engine might identify semantic elements on a page and automatically mark up the results list accordingly. For example, it might be possible to distinguish between the author and subject of a document.
- The results of WP5's knowledge generation task will be used to modify content appropriate to the device at hand. For example, knowledge summarisation could be used to generate a sufficiently concise summary of a document to send to a mobile phone via an SMS service. Similarly, a chunk of text could be generated to describe an ontological class, its attributes, place in an ontology, etc. and this text delivered to a user who has expressed an interest in that topic (class).

These areas are currently under investigation. The first target is to build a device ontology. This should be relatively straightforward, since the standard device profile description is based upon RDF, though consideration will be giving to recasting the CC/PP work into OWL. The next step is to include personal profile information alongside device information, that is, to move beyond UAProf to more general composite capability/preference profiles.

4 Conclusions

An overview of the vision for knowledge access within the SEKT project has been presented. Surveys of relevant research and commercial tools were undertaken to ensure that research and development in the SEKT project builds upon previous work.

The following technology areas were surveyed: (A) search and browse; (B) knowledge sharing; (C) visualisation and organisation of information; (D) user profile construction; (E) natural language generation; and, (F) knowledge repurposing. This paper has given an overview of the content of the technology surveys. The reader is encouraged to read the surveys, presented in the Annexes of this document, to gain further insight into these complementary technologies.

5 Glossary

API	Application Programming Interface
BT	British Telecommunications PLC
CC/PP	Composite Capabilities/Preference Profiles
DAML+OIL	DARPA Agent Markup Language (DAML) + Ontology Inference Layer (OIL)
DBMS	Database Management System
EDM	Electronic Document Management
HLT	Human Language Technology
HTTP	HyperText Transfer Protocol
IR	Information Retrieval
LAN	Local Area Network
NLG	Natural Language Generation
MIAKT	Medical Imaging and Advanced Knowledge Technologies
P2P	Peer-to-Peer
PDA	Personal Digital Assistant
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	RDF Schema
SEKT	Semantically Enabled Knowledge Technologies
SMS	Short Messaging Service
SVG	Scalable Vector Graphics
UAProf	User Agent Profiling
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol
WML	Wireless Markup Language
WP	Work Package
WWW	World Wide Web
XML	Wireless Markup Language
XSL	Extensible Stylesheet Language

5.1.1 Knowledge Access on the Semantic Web

6 References

1 http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ Last accessed on 28th July 2004.

2 http://www.w3.org/TR/rdf-schema/ Last accessed on 28th July 2004.

3 http://www.w3.org/TR/owl-features/ Last accessed on 28th July 2004.

4 Ovum Ltd (1999) Knowledge Management: Building the Collaborative Enterprise.

5 Jansen, B. J., Spink, A., and Saracevic, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management. 36(2), 207-227.

6 Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. March 2002. From E-Sex to E-Commerce: Web Search Changes. Computer. March 2002. 107-109.

7 Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. 7th International World Wide Web Conference, 14-18 April 1998, Elsevier, Computer Networks and ISDN Systems, vol.30, no.1-7 pp 107-17.

8 Xue, G., Zeng, H., Chan, Z., Ma, W., Zhang, H., and Lu, C. Implicit link analysis for small Web search. SIGIR’03, July 28-August 1, 203, Toronto, Canada.

9 Finding more useful information faster from web search results, Wu, Y. F, Shankar, L., Chen, X, CIKM’03, November 3-8, 2003, New Orleans, Louisiana, USA.

10 Chen, H., and Dumais, S. Bringing Order to the Web: Automatically Categorizing Search Results. CHI ‘2000, The Hague, Amsterdam.

11 Dumais, S., Cutrell, E., and Chen, H. Optimizing Search by Showing Results in Context. SIGCHI’01, March 31-April 4, 2001, Seattle, WA, USA.

12 Davies, J., Weeks, R., and Krohn, U. QuizRDF: Search technology for the Semantic Web.

13 Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003) “KIM – Semantic Annotation Platform” in 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 834-849, Springer-Verlag Berlin Heidelberg 2003. Web page visited May, 2004.

14 Marwick, A.D. (2001) “Knowledge management technology” in IBM Systems Journal, Vol. 40, No. 4, pp814-830. Web page visited May, 2004: http://www.research.ibm.com/journal/sj/404/marwick.html

15 Shneiderman, B. (2003) Leonardo’s Laptop: Human Needs and the New Computing Technologies. The MIT Press, London, England.

16 Wenger, E. (1999) Communities of Practice: Learning, Meaning and Identity. Cambridge University Press, UK.

17 Udel, J. (2004) “The New Social Enterprise” in InfoWorld. Vol. 26, No. 13, pp. 47,50-52, Mar 29, 2004.

18 Crossley, M., Kings, N.J. and Scott, J.R. (2004) “Profiles – Analysis and Behaviour” in Location and Personalisation: Delivering Online and Mobility Services. Ralph, D. and Searby, S. (editors). The IEE, London.

19 Frost and Sullivan (2003) "U.S. Enterprise Class Peer to Peer". Networking Technology Assessment. A297-74. San Jose, CA, USA.

20 http://swap.semanticweb.org

21 http://www.web3d.org/x3d.html

22 http://www.w3.org/Graphics/SVG/

23 http://java.sun.com/products/java-media/3D/

24 J. Contreras, V.R.Benjamins, J.Prieto, D.Paton, S.Losada, D.Gonzalez. Duontology: an Approach to Semantic Portals based on Domain and Visualization Ontology. May, 2003, KTWeb: http://www.ktweb.org/

25 Nigam, K., McCallum, A., (1998). Pool-Based Active Learning for Text Classification, Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98, 1998.

26 Nigam, K., McCallum, A., Thrun, S., and Mitchell, T., (2001). Text Classification from Labeled and Unlabeled Documents using EM, Machine Learning Journal.

27 Blum, A, Mitchell, T., (1998). Combining Labeled and Unlabeled Data with Co-training, In COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers.

28 Ghani, R., Jones, R., Mladenic, D., Nigam, K., Slattery, S., (2000). Data Mining on Symbolic Knowledge Extracted from the Web, In KDD-2000 Workshop on Text Mining, 2000.

29 Cohen, W., Singer, Y., (1999). Context-sensitive learning methods for text categorization, in ACM Transactions on Information Systems, v17 pp 171-173.

30 Wilcock, G. and Jokinen, K (2003) “Generating Responses and Explanations from RDF/XML and DAML+OIL”. In Knowledge and Reasoning in Practical Dialogue Systems, IJCAI-2003. pp 58-63, Acapulco.

31 Wilcock, G. (2003) “Talking OWLs: Towards an Ontology Verbalizer”. In Human Language Technology for Semantic Web and Web Services, ISWC’03. pp109-112, Sanibel Island, Florida.

32 Bontcheva, K. and Wilks, Y. (2004) “Automatic Report Generation from Ontologies: the MIAKT approach”. In Ninth International Conference of Natural Language to Information Systems (NLDB’2004).

33 Cécile L.P. (1998) “Tailoring object descriptions to a user’s level of expertise”. In Computational Linguistics. Vol. 14, No. 3, pp64-78, September. Special Issue on User Modelling.

34 Buchanan, B.G., Moore, J.D., Forsythe, D.E., Cerenini, G., Ohlsson, S. and Banks, G. (1995) “An intelligent interactive system for delivering individualized information to patients”. In Artifical Intelligence in Medicine. Vol. 7, pp117-154.

35 Reiter, E., Robertson, R. and Osman, L. (2003) “Lessons from a Failure: Generating Tailored Smoking Cessation Letters”. In Artificial Intelligence. Vol. 144, pp41-58.

36 Aguado, G., Bañón, A., Bateman, J.A., Bernardos, S., Fernández, M., Gómez-Pérez, A., Nieto, E., Olalla, A., Plaza, R. and Sánchez A. (1998) “ONTOGENERATION: Reusing domain and linguistic ontologies for Spanish text generation”. In Workshop on Applications of Ontologies and Problem Solving Methods, ECAI’98.

37 Bateman, J.A., Magnini, B. and Fabris, G. (1995) “The Generalized Upper Model Knowledge Base: Organization and Use”. In Towards Very Large Knowledge Bases, pp60-72.

38 Bateman, J.A. (1997) “Enabling technology for multilingual natural language generation: the kpml development environment”. In Journal of Natural Language Engineering. Vol 3, pp15-55.

39 O’Donnell, M., Knott, A., Oberlander, J. and Mellish, C. (2000) “Optimising text quality in generation from relational databases”. In Proceedings of the International Natural Language Generation Conference (INLG’00). Pp133-140.

40 W3C Working Group (2003) Device Independence Principles. Note 01 September 2003. http://www.w3.org/TR/2003/NOTE-di-princ-20030901/ Last accessed on 30th July, 2004.

41 MobilityAware (2004) Introduction to the Mobility Tag Set. http://developer.mobileaware.com/tutorials/intro.jsp Last accessed on 30th July, 2004.