D I S S E R T A T I O N

Shared Semantic Context in Lifetime Knowledge Archive

ausgeführt zum Zwecke der Erlangung des akademischen
Grades eines Doktors der technischen Wissenschaften
unter der Leitung von

O.Univ.-Prof. Dipl.-Ing. Dr.techn. A Min Tjoa
Institute of Software Technology and Interactive Systems,
Vienna University of Technology, Austria
und
Prof. Dr. Günther Pernul
Department of Information Systems, University of Regensburg,
D-93040 Regensburg, Germany

eingereicht an der
Technische Universität Wien
Fakultät für Informatik

von

Khalid Latif
Matrikelnummer: 0327347
Simmeringer Hauptstrasse 215/314, A-1110 Vienna

Wien, September 2007 _____

Zusammenfassung

Die starke Verbreitung digitaler Technologien in unserer Informationsgesellschaft hat die soziale Kommunikation sowie das persönliche Informationsmanagement signifikant beeinflusst und verändert. Persnliche Informationen kommen heute aus verschiedenen Quellen und werden strukturell sowie semantisch immer komplexer. Die Verwaltung dieser digitalen "Erinnerungen" ist eine interdisziplinäre Herausforderung wobei viele Aspekte von flexibler Softwarearchitektur bis hin zu Automation von heterogenen verteilten Datenspeichern bedacht werden müssen. Die Lösung die in dieser Arbeit vorgeschlagen wird ist eine Mischung aus persönlichem Dokumentenmanager sowie einem strukturiertem Annotationsframework mit einem semantischem Speicher.

Daten aus vielfältigen digitalen Quellen sowie deren Bedeutung werden in einer mehrfachen Ontologie erfasst. Diese Ontologie ist in Schichten geteilt, basierend auf dem jeweiligen Kontext und dem semantischen Verständnis. Jede diese Schichten ist lokal vollständig in seiner Abdeckung der Begriffsbildung, was die Bildung individueller Module erlaubt. Persönliche Informationen werden durch Verbindungen, Dialoge und soziale Interaktion beeinflusst. Für Ontologien digitaler Erinnerungen wird ein solcher gemeinsamer Kontext durch die Wiederverwendung und Rekombination von Ontologien erreicht. Solche Basisontologien kombiniert mit geeigneten Fragen führen zu wechselseitigem Verständnis und Interoperabilität zwischen unterschiedlichen Terminologien.

Das ungeheure Ausmaßder Informationen die sich im Laufe eines Lebens ansammeln stellt eine ernste Herausforderung bezüglich der Verständlichkeit dar. Um diese Verständlichkeit zu verbessern, schlagen wir die Verwendung eines strukturierten Annotationsframeworks vor. Dieses ist auch notwendig um die Informationen von verschiedenen Medien in einem gemeinsamen Modell zu vereinen. Das Framework bietet Informationen zu Daten wie deren zeitlich-räumliche Ausdehnung, beteiligte Agenten und deren Handlungen sowie semantische Informationen wie Wer, Was, Wann und Wo. Die Annotation selbst wird automatisiert vorgenommen wobei auf Techniken aus

der Textananlyse zurückgegriffen werden. Die Verständlichkeit wird weiters durch die Unterteilung des Informationsraums in mehrere Sammlungen erhöht. Diese Sammlungen sind die Grundlage zur Bildung assoziativer Spuren, einem wesentlichen Werkzeug zum Verwalten von digitalen Erinnerungen. Spuren werden in dieser Arbeit aber auch als eine Art Meilensteine oder Landmarken verwendet. Sie Entstehen aus dem Informationsraum des Benutzers, und Elemente die Eigenschaften oder Bedeutungen mit diesen Landmarken gemeinsam haben werden mit diesen assoziiert und verbunden.

Um die Möglichkeiten eines solchen "Erinnerungsarchivs" zu demonstrieren wurde ein erweiterbares, service-orientiertes Framework entwickelt. Der Speicher wird ber spezielle Module befüllt. Die Hauptaufgabe des Frameworks basiert auf semantischer Datenanreicherung und nachfolgender Speicherung in dem semantischen Speicher. Innovative Visualisierungenstechniken wurden zur effizienteren Navigation in diesem Archiv eingefhrt. Die vorgeschlagene night-sky Visualisierung ermöglicht ein besseres Verständnis der dahinterliegenden Daten. Experimente mit verschiedenen Datensammlungen haben gezeigt, dass mit Hilfe der night-sky Visualisierung, aufbauend auf einer konzeptbasierten Suche, das Finden und Lernen in großen Sammlungen erleichtert wird.

Abstract

The emerging pervasiveness of digital technologies within our information society have significantly revolutionized social communication and personal information management. Personal information is now characterized by the fact that it originates from heterogeneous sources and is becoming more and more complex in structure and semantics. Managing these digital 'memories' of a lifetime is a multi-disciplinary challenge, involving all facets of semantic enhancements, flexible software architecture, and automation among heterogeneous and distributed data sources. Consequently, the solution proposed in this thesis is a blend of personal document management, structured annotation framework, and a semantic triple store.

Semantics of a diverse range of digital memories and also their associations are captured in a multifold ontology. The ontology is partitioned into layers based on the information context and the semantic insight. Each layer is locally complete in its coverage of conceptualization which allows easy maintenance of growing individual modules. The personal information emerges through connections, dialog and social interaction. For the ontology of digital memories, such a shared context is achieved by reusing an existing foundational ontology as a base. The reuse of foundational ontologies, through guided questions, can facilitate mutual understanding and interoperability among varying terminologies.

The enormity of the lifetime information poses a serious challenge in terms of comprehensibility. To improve the comprehensibility, we suggest making use of a structured annotation framework. A structured annotation framework is also necessary to bring together information extracted from diverse media types into an integrated model. The framework provides the semantic insight of life-items related to their spatio-temporal location, involved agents & their activities, and semantic content labels; corresponding to who, when, where, and what. The process of annotation is automated using named entity recognition and text mining. Comprehensibility is further improved by partitioning the knowledge space into collections. These collections lay the foundation for constructing associative trails – an essential feature for

managing digital memories of lifetime, and an important aspect of the storytelling. Trails are also constructed in our work by means of landmarks. They emerge from the user's knowledge space. Items that share meaning or physical similarity with the landmark become associated with it and selection of the landmark activates linked items, and vice versa.

Finally, an extensible service-oriented framework is developed for demonstrating the capabilities of the lifetime store. Data is fed into the store using a number of dedicated data acquisition modules. The core functionality of the framework is based upon semantic data enrichment and subsequent ontological storage. Innovative visualization techniques are introduced to effectively navigate in the lifetime archive. The proposed night sky visualization facilitates better understanding of the underlying data by exploiting the overview and details-on-demand interaction technique. Our experiments with different data sets are testament to the hypothesis that sky visualization, on top of concept-focused associative search, can make it easy to locate, link, and learn from even a huge repository.

Acknowledgment

First and foremost, I would like to thank my adviser Prof. Dr. A Min Tjoa, who has been my inspiration and the role model for how to do research. He has always given the advice and direction while still giving me the freedom to explore new horizons. I am incredibly fortunate to have had him as my adviser.

It were Andreas Rauber and Edgar Weippl who guided me in the absence of Prof. Tjoa. My thanks to both of them. Thanks are also due to my colleagues at IFS and all the members of DynamOnt & SemanticLIFE project, I had the privilege to work with. Even though the list is very long, I especially like to mention Shuaib Karim and Amin Anjomshoaa for their close cooperation. A special thanks is reserved for Maria Schweikert & Michael Schadler, who make everything at the IFS run smoothly. They guided me throughout the ins and outs of IFS requirements.

And finally, and not to be forgotten, I thank my family for their prayers. Especially my wife Sadia who rather tolerantly endured my late sittings in the institute. Without their support things might have been different from now.

This work was supported by grants from Higher Education Commission of Pakistan, ASEA-UNINET (ASEAN-EU Academic University Network), and also by the DynamOnt project which is funded by the Austrian Government's FIT-IT Research Initiative on Semantic Web under the contract 809256/5512.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

In 2003, UK Computing Research Committee (UKCRC) started a Grand Challenge initiative to discuss possibilities and opportunities for the advancement of computing research. A year after, a new grand challenge – Memories for Life – was announced (Fitzgibbon and Reiter, 2004), which stated:

> "When computers first appeared, they were used to help people perform numerical calculations and clerical tasks. More recently they have also been used to help people communicate. But there is another potential use of computers which is now emerging, namely as a tool to help people store, search and interpret their [digital] 'memories'... such as emails, digital photographs and Internet telephone calls. People are capturing and storing an ever-increasing amount of such memories, with new types of information constantly being added..."

Managing lifetime memories has been an open ended endeavor ever since its inception by Vannevar Bush (1945). Half a century ago he coined the idea of lifetime memory store – *memex*. In his own words *memex* should be "a device in which an individual stores all his books, records, and communications... an enlarged intimate supplement to his memory." As a matter of fact, declaring Memories for Life a grand challenge by UKCRC has augmented it as a multi-disciplinary problem.

The need of memex like system becomes intensified because of an intriguing trend of digitizing lifetime information and a prolific attitude of communities and individuals toward archiving. For example, Gordon Bell digitized about two decades of his life under the MyLifeBits project (Bell and Gemmell, 2007; Gemmel et al., 2003). His archive has grown to 300,000 items of audio & video files, emails, web pages, and presentations. Recently,

Thai government officials announced that more than 60,000 searchable items including speeches, photographs and official documents from 1934 through 2007 related to King Bhumibol Adulyadej, Queen Sirikit, and other royal family members will be published online (Payne, 2007). Collections of thousands of photographs are common these days because of increased usage of digital cameras. For example, *webshots* community portal claims to have 397 million photos[1] from its members – a count which is growing[2].

1.1 Personal Knowledge Management

Acquisition, organization, and retrieval of information by an individual is commonly defined as personal information management (Boardman, 2004). Over the past decade, the capabilities of storage devices have exceeded the ability of currently available personal information management systems to effectively handle a large amount of data from different sources and varying structures by far.

Personal information management has recently emerged as a multi disciplinary research area. It is being investigated in various fields ranging from information retrieval to human-computer interaction (PIM-SIGIR, 2006). Consequently, the focus has changed from unadorned retrieval to learning and personal growth. Researchers and practitioners from both academia and industry are investigating different knowledge management tools, such as semantic wikis (Oren et al., 2006), for guiding the individuals to *locate, link, and learn* from personal information.

1.1.1 Lifetime Stores and Semantic Desktops

The theoretical and cognitive issues concerning the management of personal knowledge are elaborated by (O'Hara et al., 2006). On the application development front, numerous projects and tools have emerged recently, though most projects are not proven for production use. We do not intend to present a comprehensive state of the art survey of all the projects. Only some of the representative software endeavors for realizing memex are discussed below.

LifeStreams uses a time-ordered stream of documents, as a substitute of the conventional file and directory view in the current desktops (Freeman and Gelernter, 1996). Stream filters are used to organize, locate, and monitor incoming information items. Other systems such as Haystack (Adar et al.,

[1]This figure is taken from http://community.webshots.com on July 05, 2007.

[2]On another photo sharing website *flickr.com* it was observed that around one thousand photos are uploaded in each minute.

1999), SEMEX (Dong and Halevy, 2005), and ScienceOrganizer (Wolfe and Keller, 2005) enable the structuring of the information and allow the user to navigate the associations. Both Haystack and SEMEX focus on the individual. On the other hand, ScienceOrganizer is a collaborative knowledge management and information structuring tool for distributed project teams.

Another effort similar to SEMEX is the iMeMex project which provides richer integration with the desktop operating system through WebDAV interface (Dittrich et al., 2005). Pragmatically, it integrates different data sources from the desktop such as emails, web-pages, and personal documents using a uniform data model (Dittrich and Salles, 2006). Out of the box desktop integration and search solutions are also available from the major software vendors, such as Google Desktop[3] and Windows Desktop[4]. The later allows the integration of shells to expose extensions in the original functionality. Phlat, for example, supports tagging of the desktop items (Cutrell et al., 2006).

A new shift in the personal desktops is the *Semantic Desktop*. These systems, such as Gnowsis (Sauermann et al., 2006) and IRIS (Cheyer et al., 2005), exploit the building blocks of the Semantic Web technologies (such as ontologies and inference) to manage personal information available on the desktop. Most of these systems extract the information from the desktop and import into a repository that also houses ontology for pre-defined entities such as persons, projects, and publications.

MyLifeBits is the most salient memex realization effort under Microsoft Research (Gemmel et al., 2002). It is a lifetime store of articles, books, letters, memos, photos, presentations, videos, and voice recordings (Gemmel et al., 2003). Continuous archival of digital life data is one of the major goal of this project. So far, however, it does not focus on semantically structuring or exploiting the resulting pool of data beyond mere retrieval.

1.1.2 Trails and Semantic Context

The storage technology has made it cheap to digitize and archive nearly everything; ergo, selecting a particular item from the haystack of lifetime information poses a serious challenge. In the words of Vannevar Bush, "[I]t involves the entire process by which man profits by his inheritance of acquired knowledge." He also noted that:

> "[The human mind] operates by association[s]. With one item in its grasp, it snaps instantly to the next that is suggested by the

[3]http://desktop.google.com

[4]http://www.microsoft.com/windows/desktopsearch/default.mspx

> association of thoughts, in accordance with some intricate web of trails. . . "

The World Wide Web is a realization of these *associative trails*. The links in the hypertext build the trails of web-pages. In their present form, the hyperlinks are different from the *association of thoughts* as they are created by the author of the web-page and not by the reader (O'Hara et al., 2006). Consequently, innovative retrieval techniques are needed for selection in the lifetime personal store by exploiting associative indexing and semantic linking.

Although the focus in managing digital memories of a lifetime is the individual, but the ultimate intention is helping an individual be more effective and work better in groups and corporations. Karl Mannheim has put a strong emphasis on the relationship between human thoughts and the social context (Kettler, 1967). On the similar lines, we argue that *personal information is never an individual product and that it emerges through connections, dialog and social interaction. So the resulting personal information should preferably be placed in a shared semantic context.*

1.2 Research Question

There are several specific questions regarding the philosophical, social, theoretical and technical aspects of realizing *shared semantic context for associative personal information*. Among those, the questions which are prime focus of this thesis include:

- Is it possible to accurately model the semantics of a diverse range of digital memories and also their associations? How Semantic Web technologies can help in this regard?

- Manual building of trails for thousands, or even hundreds, of items is a diligent task. How can we realize an efficient and productive system of construction of trails and how much automation in constructing associative trails is possible by exploiting the semantic insights of the contents?

- Which contextual information and other annotations should be stored along with the actual memories and can these be acquired automatically or do they need to be manually entered?

- Personal memories have a strong relationship with the social interactions. Can we guarantee shared semantic context for items in the networked environment?

- How can we easily and effectively retrieve useful information from the stockpile of digital objects spanning a human lifetime? Additionally, how can we overcome the comprehensibility problem and semantics overload in the lifetime knowledge box, presumably containing millions of items.

1.3 Contribution and Thesis Organization

This thesis is an attempt to establish shared semantic context in realization of trails and associative indexing in the personal knowledge box. *Why the term 'knowledge box'?* The WordNet lexical database has 10 noun entries for the word 'box'. One of them is described as: "private area in a theater or grandstand where a small group can watch the performance". We have used the term 'box' in two senses. First of all, the lifetime knowledge box facilitates the user to watch (or navigate) his/her own performance (that is to say, memories and experiences) of the lifetime. Such a knowledge box lets you know, for example, what did you presented in a specific conference, who was session chair, and where you went with the post-conference guided tour. Secondly, the knowledge box refers to the digital archive built through the lifetime capture of these memories.

More specifically, this thesis embodies answers to the questions listed in the previous section. The proposed system is a blend of personal document management, hypermedia information space, and semantic triple store. Machine learning is also applied to different user tasks. Firstly, the DynamOnt approach for dynamic knowledge construction is explained in Chapter 3. Then a layered semantic enrichment model is proposed to master the heterogeneity in lifetime personal information (see Section 3.3). Each layer plays a specific role and going upward in the layer hierarchy increases the semantic insights of the items. The bottom most layer provides structuring of the item contents. In the middle lies a uniform information context model. It acts as a grounding principle to manage semantics overload and improves the comprehensibility in the lifetime knowledge box by revealing lightweight semantics corresponding to who, when, where, and what (see Chapter 4). The top layers expose the shared axiomatic context of individual items and their relationship with others (see Section 4.3) which is necessary to build trails.

Finally, a software architecture pattern for semantic enrichment is introduced. This pattern is used to develop an extensible service oriented framework for demonstrating capabilities of the networked knowledge box (see Chapter 5). Innovative visualization techniques are presented to effectively

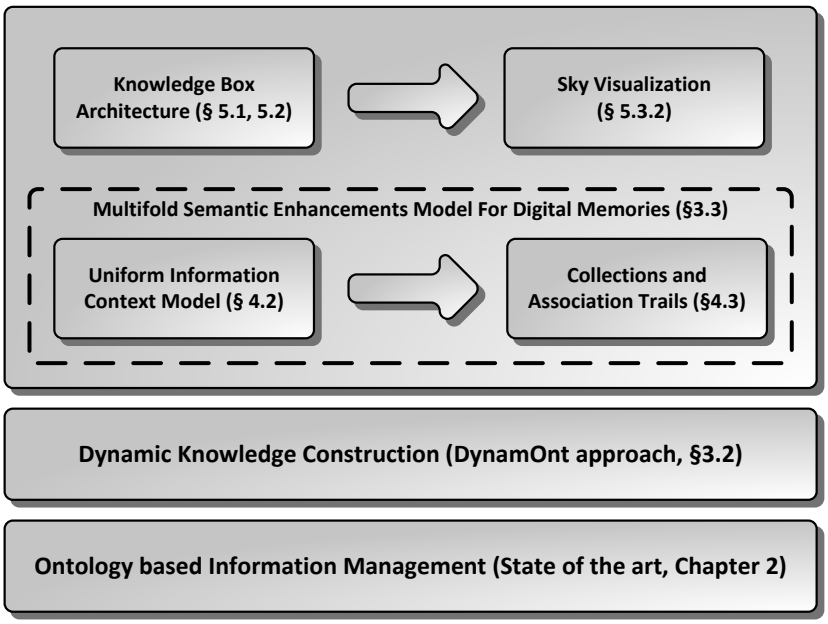


Figure 1.1: Overview of the thesis.

explore memories of life (see Section 5.3). Figure 1.1 shows the overview of the thesis and also depicts the flow.

The proposed solution focuses on the computational aspects of managing and using personal digital archives. This also draws a boundary and limitation of the research presented in this thesis that it does not deal with human cognition. In a way, this work is limited to digital memories. Managing the real cognitive memories may require an integrated effort employing a large scale infrastructure, such as the Blue Brain project (Markram, 2006). We have also not touched on social implications of *memex* like systems such as the followings:

- There might be a number of events, or memories in general, that one likes to forget. As though having them permanently stored may cause distress and can disturb the social behaviors and relationships.

- Dependence on software systems can decrease human ability of recall. A common phenomenon is forgetfulness of phone numbers due to frequent use of digital address books which are usually embedded in mobile phones. A valid question could be that storing and then consulting memories of lifetime is adding the problem or is a solution.

- Memex could be seen as increased surveillance - the notion of “big

brother (software system) watching you." Such a software system can provoke questions regarding privacy but are not discussed in this thesis.

The next chapter provides theoretical background and summarizes relevant research work in the area of ontology based information management. The limitations in the prior work lay the foundation of this thesis. Chapters 3 and 4 provide theoretical building blocks of this thesis and the implementation details along with the navigation and visualization support are present in the Chapter 5.

Chapter 2

Ontologies and Information Management

This chapter serves two sole purposes. First of all, it summarizes fundamental concepts related to use of ontologies in information retrieval and information management, and automatically extracting metadata. An understanding of these concepts is necessary for the development of the rest of the thesis. Secondly, it highlights the "open slots" in the previous work and lays down the theoretical foundation for this thesis.

2.1 Ontologies and Semantic Web

In contrast to the original Web of hypertext documents, the Semantic Web is characterized as a web of data (Herman, 2007). The Semantic Web has four building blocks: *Ontologies* are shared specifications of the domain of discourse and explain precisely what the data means. *Schema annotations* explicate how to map concepts from different ontologies. *Rules* interpret how new data can be formally derived out of the existing data. And finally, *agents and tools* use these components together to realize different scenarios and applications.

The Semantic Web promises to bring structure to the Web through common formats for integration and interchange of data drawn from diverse sources (Berners-Lee et al., 2001). The URI (Berners-Lee et al., 2005) is the cornerstone of the data interchange in the Semantic Web. For instance, the concept *Mina Bazar*[1] could be represented as:
http://pakistan.gov.pk/culture/festival/MinaBazar

[1]It is a cultural family festival common in Pakistan.

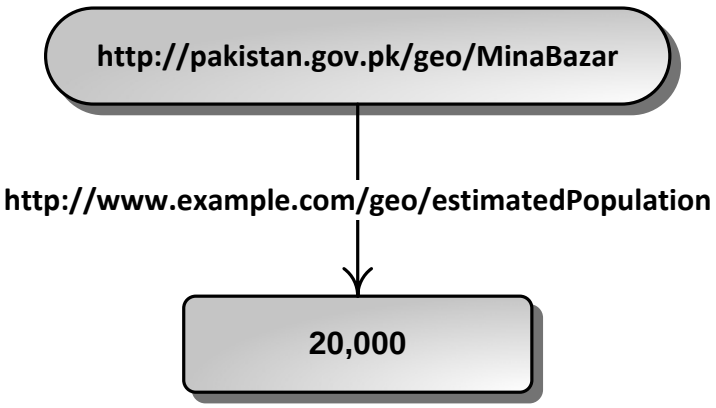


Figure 2.1: Example of an RDF statement.

Each URI is located within a *namespace* and semantically different resources having similar names can co-exist in different namespaces. For example, *Mina Bazar* is also a small town in Balochistan province of Pakistan and may be represented under the following namespace:
http://pakistan.gov.pk/geo/MinaBazar

In February 2004, The World Wide Web Consortium (W3C) released the Resource Description Framework (RDF) as a W3C Recommendation (Klyne and Carroll, 2004). RDF is a language that provides basic syntax to represent information and to share data in the Web. Its abstract data model is directed labeled graph (Hayes, 2004). Each statement in RDF is called a triple of *subject*, *predicate* (or property), and *object*. For example the natural language statement "Mina Bazar has about 20,000 dwellers" is represented in Figure 2.1 in the abstract RDF graph model. The same statement could also be serialized as follows:

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix pak:    <http://pakistan.gov.pk/geo/> .
@prefix geo:    <http://www.example.com/geo/> .

pak:MinaBazar rdf:type geo:SmallTown ;
  geo:estimatedPopulation "20000" .
```

RDF graphs could be retrieved using SPARQL, an SQL-like querying language for RDF. For example, we can search for *all towns with an estimated population of less than 30000, arranged in the ascending ordered of the count of dwellers.*

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
PREFIX geo:    <http://www.example.com/geo/> .

SELECT ?town ?population
WHERE {
  ?town rdf:type geo:SmallTown .
  ?town geo:estimatedPopulation ?population .
  FILTER (?population < 30000)
}
ORDER BY ASC(?population)
```

The back bone of any RDF infrastructure is a triple store for persistent storage of RDF statements. There are already some efforts regarding designing an efficient infrastructure to store triples or quads (triples with additional context) and to provide inference capabilities over it (Alexaki et al., 2001; Beckett and Grant, 2002; Broekstra et al., 2002; Ma et al., 2004; McBride, 2002b; Volz et al., 2003; Wood et al., 2005). A number of studies have been carried out to evaluate the performance of these solutions (Beckett, 2002; Guo et al., 2004; Harth and Decker, 2005; Lee, 2004).

RDF Schema (RDFS) extends the RDF vocabulary for building taxonomies and describing light-weight semantics (Brickley and Guha, 2004). For instance, the concept *SmallTown* from the previous example can be annotating as a *subclass* of *GeographicRegion* and the range of the property *estimatedPopulation* could be set to integers. The extended version of the previous example takes the following shape:

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix pak:    <http://pakistan.gov.pk/geo/> .
@prefix geo:    <http://www.example.com/geo/> .

geo:SmallTown rdf:type rdfs:Class ;
  rdfs:subClassOf geo:GeographicRegion .

geo:estimatedPopulation rdf:type rdf:Property ;
  rdfs:domain geo:GeographicRegion ;
  rdfs:range xsd:long .
  rdfs:comment "Estimated count of dwellers."

pak:MinaBazar rdf:type geo:SmallTown ;
  geo:estimatedPopulation "20000"^^xsd:long .
```

RDFS is useful, but does not solve all the possible requirements such as building complex classes, for example, as unions of two or more existing classes. Declaring disjointness or equivalence of classes and restricting a property range when used for a specific class is also not possible within RDFS. The Web Ontology Language (OWL) is another W3C Recommendation (Patel-Schneider et al., 2004) which brings expressive and reasoning powers of Description Logics (DL) to the Semantic Web. The basic description logic, $\mathcal{A}$ttributive $\mathcal{L}$anguage with $\mathcal{C}$omplements ($\mathcal{ALC}$), allows statements to be made with the following constructions:

$Concept$	Unary predicates
$Role$	Binary predicates
$\neg C$	Negation
$C \sqcap D$	Conjunction
$C \sqcup D$	Disjunction
$\exists R.C$	Existence of a role R value-restricted to be filled by concepts of type C
$\forall R.C$	All roles of type R value-restricted to be filled by concepts of type C

Other description logics expressiveness[2] is denoted as follows:

$\mathcal{S}$	An abbreviation of $\mathcal{ALC}$ with transitive roles
$\mathcal{H}$	Role hierarchy
$\mathcal{O}$	Nominals (enumerated classes of object value restrictions)
$\mathcal{I}$	Inverse properties
$\mathcal{F}$	Functional properties
$\mathcal{N}$	Unqualified number/cardinality restrictions (includes $\mathcal{F}$)
$\mathcal{Q}$	Qualified cardinality restrictions
$(\mathcal{D})$	Use of data types

OWL is used to publish and share ontologies, supporting advanced Web search and knowledge management. Three species of OWL are defined, namely Lite, DL, and Full. OWL-Lite supports building taxonomic ontologies and allows simple cardinality constraints on properties. It corresponds to the description logic $\mathcal{SHIF}$. OWL-DL supports maximum expressiveness while retaining computational completeness and decidability. For this reason it doesn't support non-DL constraints, and corresponds to the description

[2]Description logic complexity navigator is a useful tool by Evgeni Zolin for playing with the complexity of reasoning and could be found under:
http://www.cs.man.ac.uk/~ezolin/dl/

RDF/OWL element	Protégé-OWL	Jena
(model)	OWLModel	OntModel
owl:Ontology	OWLOntology	Ontology
owl:Class	OWLNamedClass	OntClass
owl:Class	*OWLAnonymousClass*	isAnon()
owl:Restriction	*OWLRestriction*	Restriction
owl:unionOf	OWLUnionClass	UnionClass
rdf:Property	RDFProperty	OntProperty
owl:ObjectProperty	OWLObjectProperty	ObjectProperty
owl:FunctionalProperty	isFunctional()	FunctionalProperty

Table 2.1: Comparison of Protégé-OWL and Jena API.

logic $\mathcal{SHOIN}(\mathcal{D})$ (Horrocks et al., 2003). OWL-Full has no expressiveness restrictions, but also doesn't guarantee decidability and other computational properties.

Number of open source APIs and tools are available for working with web ontologies (Tjoa et al., 2006). Most prominent are Jena (McBride, 2002b), Sesame (Broekstra et al., 2002), and Protégé[3]. Interestingly, Jena and Protégé-OWL API are closely integrated[4] as the later reuses various services of the former. It is also possible to convert Protégé *OWLModel* to Jena's *OntModel* (see Table 2.1 for an overview of equivalences).

The benefits of ontology-driven knowledge management and reasoning are well established (Davies et al., 2003). The automation that can be achieved by explicitly annotating resources with ontologies is astonishing, ranging from automatic discovery of web services (Fensel et al., 2002; McIlraith et al., 2001) to accessibility improvements for the user interfaces (Karim et al., 2007). Compared with the existing Web, the Semantic Web research is still in its infancy (Economist, 2007a). Most of the existing information on the existing web pages and personal documents on the desktop lacks explicit semantic annotations. Manually annotating the bulk of millions of web pages would take years. Consequently, the researchers are investigating new ways for automatically extracting the meanings from the data using ontologies.

2.2 Ontology-driven Information Processing

The current work in ontology-driven information processing could be classified in four categories: 1) *ontology learning* – learning concepts and their rela-

[3]Protégé-OWL API could be used separate from its editor.

[4]http://protege.stanford.edu/plugins/owl/jena-integration.html

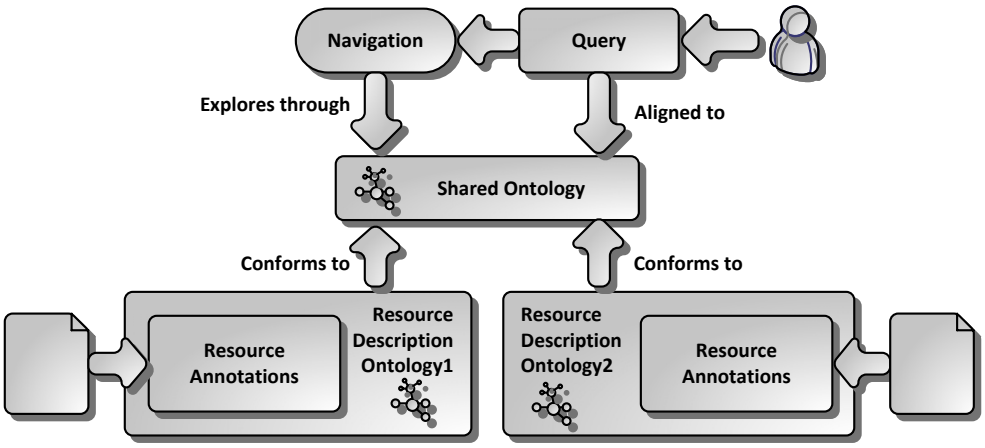


Figure 2.2: Ontology-based information retrieval and navigation model using shared ontology.

tions to build ontologies; 2) *instance learning* – populating ontology instances by extracting metadata from textual contents; 3) *document management* – information retrieval using formal or lightweight ontologies as a guide to manage documents, to categorize resources, or to extract semantic associations; and finally 4) *semantic annotations* – manual or semi-automatic semantic annotations to existing resources such as images, and web-pages. These categories are not mutually exclusive and ontology applications may combine features from multiple categories to achieve a certain goal. Document management, for example, may exploit automatic semantic annotations to reduce the cognitive load off the user. Figure 2.2 represents ontology based information retrieval model for documents by exploiting semantic annotations. The presented model combines data-access via shared ontology approach (Uschold and Jasper, 1999) and ontology-based retrieval model by (Garcia and Sicilia, 2003).

Due to the relationship of ontologies with the domain of discourse, they are mostly developed by the domain experts (or knowledge workers). Different ontology learning methods by using information retrieval techniques are present in research literature (Maedche, 2002; Gómez-Pérez et al., 2003, p.157-163). Most of these efforts only extract hierarchical relations (Park, 2004; Cimiano et al., 2003) for focused domains. Automatically learning a formal axiomatized ontology is still not possible to best of our knowledge.

Filling in the instance data for ontologies involve human labor. The task of populating the ontology is more a craft than science. For example, think of ontology for tourism. One can't go around finding each and every hotel even for a specific region and instantiate the relevant concepts; no mention of the

work needed for the whole world with all the concepts involved in tourism. The process could reasonably be automated by extracting metadata from the natural language text present in the existing web documents, yellow pages, and other domain corpus. Automating the process would provide a smooth transition from the current Web to the Semantic Web. Different research projects have tried to solve this problem with different perspectives. (Handschuh et al., 2002) used a *wrapper-induction* based information extraction system, Amilcare, to extract relational metadata. Amilcare uses a set of manually annotated documents and a learning algorithm to induce a wrapper that annotates documents by inserting XML tags around items to be annotated (Ciravegna, 2001a). The users map Amilcare XML tags to ontology concepts, and the system automatically converts the XML annotations into RDF annotations. Embley (2004) used extraction ontologies, essentially the conceptual-model of wrappers, to extract ontological data.

The OntoSophie system is based on *supervised learning* and therefore learns extraction rules from annotated text and then applies those rules on new documents for ontology population (Celjuska and Vargas-Vera, 2004). The important part of the entire cycle is the user who accepts, rejects, or modifies newly extracted instances to be populated. The first task in the extraction process is to identify important entities and slot values for a particular class in a document. In the next step, it is determined whether the constructed instances described by slot values are correct and whether it should be fed into the class in the ontology or not. Hidden Markov Model based information extraction is also used for extracting ontological data (Valarakos et al., 2004).

In contrast to machine learning based techniques, the Artequakt project links a knowledge extraction tool with *lexicon* and domain ontology to guide information extraction. The extraction tool, first of all, searches online documents and extracts knowledge about artists that matches the given classification structure. The aim is to automatically identify entity relationships which are useful in populating the ontology (Alani et al., 2003). The populated ontologies are maintained in a structured knowledge base. Knowledge extraction is further enhanced using a lexicon-based term expansion mechanism that provides extended ontology terminology. OntoGenie also used a linguistic ontology to convert unstructured data from the Web to structured knowledge (Patel et al., 2003a). It generates ontology instances from unstructured text in a semi-automatic fashion. Unlike other machine learning based solutions, which first extract metadata from the text and then the mapping of extracted metadata is performed with the domain ontology, OntoGenie does the mapping of the concepts in domain ontology into WordNet as a very first step. The mapping is performed by canonizing the English terms defining

the concepts from the domain ontology. This step is crucial as many terms in WordNet may map onto the same concept from the Ontology. For example, the concept *university* in WordNet has more than one sense, such as an *organization*, a *body of students & faculty*, and even a *construction*. A graphical user interface is provided for the domain expert to select the right sense for the automatically discovered mappings. Web pages for the particular domain are then retrieved and parsed word by word. Each word is canonized and compared with the mapped concepts in WordNet. Once OntoGenie has retrieved all the relevant words and their mappings, the relationships that hold between them are extracted. This is done by assuming that a set of newly discovered concepts in predetermined locus around the base concepts are related. Other ontology-based information extraction efforts either use *extraction agents* (Sheth and Ramakrishnan, 2003), or process *semi-structured contents*, such as Wikipedia, for extracting ontology instances (Auer and Lehmann, 2007; Stuckenschmidt and van Harmelen, 2001). The quality of extracted relations in the later case is directly proportional to the quality of structuredness in the contents. Most of the existing efforts in ontology-based information processing either involve strong user interaction for confirmation of extracted results or do not go beyond miniature examples.

In addition to ontology learning and population, researchers are working on number of other ontology-based applications (Uschold and Jasper, 1999). Analogous to web search engines for finding web documents over the web, different ontology search solutions are also needed (Ding et al., 2004; Patel et al., 2003b). The need arises implicitly from the reuse aspect of ontologies, where someone with lesser domain knowledge can search existing domain ontologies from the web and can reuse (parts of) them. Visualizing the retrieved ontologies is another issue. Researchers have proposed graph based visualization solutions (Storey et al., 2001; Tzitzikas and Hainaut, 2006) and cluster maps (Fluit et al., 2003) for navigating the ontology contents in an intuitive way, but the area is still open for exploiting the use of other models and clustering solutions for larger ontologies having thousands of concepts.

In the next section we have elaborated some of the major issues in ontology based information retrieval and have discussed the strategies adopted in different research projects.

2.3 Issues in Ontology-based Information Retrieval

There are a number of issues which drive the research in Ontology-based Information Retrieval (OntoIR), such as how to efficiently mine relationships and how to measure *precision* and *recall* of an ontology learning system. In the following we only have discuss those issues which are considered to be crucial for the *Personal Knowledge Box*.

2.3.1 Knowledge Acquisition – Data Source

The underlaying corpus is a very important parameter for measuring the quality of the results and comparing it with other similar efforts in the area. Unfortunately there is no widely used benchmark corpus for harmonizing the research in ontology-based information retrieval. Only DMOZ Open Directory Project (ODP) provides a taxonomy in RDF format[5]. In most of the projects studied, different search engines were used as a source for acquiring web pages of a particular domain.

Although search engines have greatly enhanced information access and their precision has improved a great deal, still their results may include irrelevant documents. Artequakt project used a filtering mechanism. It applied a vector similarity measure to compare search engine results with exemplars which are taken from trusted sites[6] and selected only those with similarity above a certain threshold. For the personal knowledge box, however, the individual user has the right to make archival decision. This has adverse effect on the underlying retrieval system. Personal information is characterized by heterogeneous and multi-genre information objects. In contrast, the current breed of information extraction solutions work well with single genre corpus and tuning their performance for diverse contents is itself a challenge. We have presented a detailed analyses and implementation details to resolve this issue in Section 5.2.2.

2.3.2 Annotations & Information Extraction Rules

Many researchers have used the wrapper induction techniques to automatically extract knowledge from the web. Their approach is mostly based on

[5]http://rdf.dmoz.org/

[6]e.g. Web Museum site which provide short artist biographies http://www.ibiblio.org/wm/paint

Word	LexCat	SemCat	Action
The	Art		
keynote	Noun		
speech	Noun		
at	Prep		**stime**
4	Digit		
PM	Other	timeid	
will	Verb		

Table 2.2: A tagging rule with associated NLP knowledge.

predefined templates and pattern-based extraction rules. Web pages, however, have varying structures and plenty of formatting styles. And to cover every structure variation needs not just learning but labor. A more promising approach is to use shallow parsing for recognizing syntactical semantic relations (Hammerton et al., 2002). Shallow parsers have the advantage of high speed and robustness, necessary to apply information extraction to a large number of unstructured documents.

Amilcare, S-CREAM's text analysis component, also uses a shallow parser for preprocessing the text. The processed text is later annotated by applying the rules induced during its training phase based on LP2 algorithm. The algorithm is a wrapper induction methodology (Ciravegna, 2001b) that unlike other wrapper induction approaches, uses linguistic information in the rule generation process (see Table 2.2). Additionally the precision and recall of rules can be tuned by experts without major retraining. The output of the process is either a single XML tagged document or a list of XML tagged text snippets.

In contrast, the OntoSophie system assigns unique XML tag to each slot within any class of the ontology. For a particular document in the training set, the user selects a specific class from the domain ontology and annotates the text with relevant tags. The annotated documents are then processed to separate the sentences into noun phrases, verb phrases and other high-level constituents. As a next step it learns extraction rules related to some of the class in the domain ontology. The rule is fired if a sentence or its part satisfy all the constraints. The rule confidence is influenced by coverage and error measures. OntoSophie promises high precision by making use of rule confidence.

Web documents have limitless vocabularies, structures, and composition style to represent approximately the same content, even they may use different expressions or linguistic structures. So other than a shallow parser, Arte-

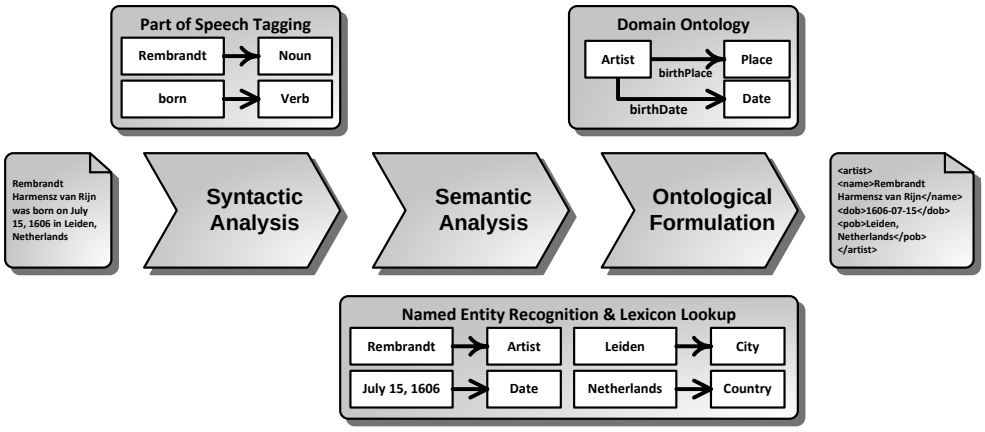


Figure 2.3: Knowledge extraction process in Artequakt.

quakt uses an ontology coupled with WordNet[7] for term expansion. The fed documents are first divided into phrases and grammatically related phrases are then grouped by the Apple Pie parser[8]. As a next step it uses ANNIE[9] and WordNet (Fellbaum, 1998) to identify entities (see the graphical representation of the process in Figure 2.3). WordNet's lexical chains (synonyms, hypernyms and hyponyms) were used to reduce linguistic variations among extracted entities and ontology terms.

2.3.3 Mapping Extracted Metadata

Ontologies represent domain knowledge through explicit formalization and specification of the concepts and their corresponding relationships (Gruber, 1995). One of the major issues in OntoIR is extracting these relationships and aligning them to appropriate classes in the ontology. Conventionally, the task of discovering relations is done via morphologically determining the verbs and the relationships to nouns (Craven et al., 1998). The approach works fine for simple toy classes but fails to produce good results in real world ontologies which are rather complex in the sense they have large number of concepts and relationships among each others and their automatic discovery may result into *mapping conflicts*.

Two types of conflict situations are evident in OntoIR. One in which more than one value for one property could be extracted (*value conflict*), and another in which values for different properties of different entities are

[7]A general purpose lexical database http://www.wordnet.princeton.edu
[8]www.cs.nyu.edu/cs/projects/proteus/app
[9]The text extraction component of GATE (Cunningham et al., 2002)

extracted (*type conflict*). For the first situation, OntoSophie simply makes use of confidence value of the rule. The value extracted by a rule with high confidence value is preferred over the others. The second situation is much more complex than the first as it is very important to determine which classes the new instances should be fed into. To resolve the issue the user maybe provided with all the extracted possibilities while automatically preselecting those that are believed to be strongly accurate. S-CREAM makes use of explicit discourse representation for mapping tagged output of Amilcare to the target graph structure of the ontology. The idea is to identify different discourse entities – *centers* – and to associate relevant attribute values with them, known as *centering theory* (Grosz and Sidner, 1986; Strube and Hahn, 1999). S-CREAM uses a lightweight single-template version of the centering theory in which only one type of tag is determined to introduce a new discourse referent and every other pair of tag name and tag value is attached to this entity as an attribute. This is further helped with ordering information and some additional rules for complex models. This user-driven ontology based annotation approach, however, can't reliably identify complex relationships. For example, for the following extracted statements:

ISWC	*instOf*	*Conference*
ISWC	*city*	*Athens*
ISWC	*keynote*	*Thomas Gruber*
ISWC	*topic*	*Social Semantic Web*

the target graph structure should ideally take the following form:

ISWC	*instOf*	*Conference*
ISWC	*locatedAt*	*Athens*
Athens	*instOf*	*City*
ISWC	*features*	*Keynote*1
*Keynote*1	*instOf*	*KeynoteSpeech*
*Keynote*1	*deliveredBy*	*Thomas Gruber*
*Keynote*1	*topic*	*Social Semantic Web*

Another similar approach, used in OntoGenie project, is to flexibly assume a set of concepts discovered in predetermined locus around the concepts to be related, known as *principle of locality*. The overall process is iterative where parts of ontology are filled and instances of intermediate nodes are

kept blank. Such blank nodes may be filled on while analyzing other web pages of related domain. To better understand the idea, consider an instance of *University* and an instance of *Country*. A relationship could be assumed to hold between them even if there is no information about an intermediate node, *City* or *State* in this case. Compared with other techniques, principle of locality result in higher recall by discovering largely disconnected knowledge instances and then linking them by information discovered from other resources.

WordNet is used intensively in different information extraction systems. It is also a hallmark of many OntoIR systems such as OntoGenie and Artequakt. To disambiguate the concept mapping to WordNet, OntoGenie implementation provides a graphical user interface to the domain expert to select the right sense for automatically discovered mappings. Artequakt bypasses the need for such defined mappings through lexical chains and the relationships are determined through linguistic bindings of entities. This approach lessens the user intervention in the metadata mapping process.

2.3.4 Measuring Semantic Similarity

Revealing associations by transforming unstructured contents into formal representations requires a deep analysis of the text and is generally considered very difficult task. New tools can capitalize on the advantages of the Semantic Web technologies to build formally valid and logically correct interconnected knowledge space. One of the major issues in doing so requires measuring similarity among unstructured text and concepts from the ontologies. There are different ways to compute similarity[10] and usage of a particular measure depends on the intended purpose and the target objects to be compared.

Cosine similarity is a famous content similarity measure (Rijsbergen, 1979). It is also exercised in calculating semantic similarity of concepts based on the frequency of their occurrences in the domain text (Wang et al., 2003), semantic similarity of documents (Meziane and Rezgui, 2004), and also for extracting taxonomies from text (Cimiano et al., 2003). The Jaccard coefficient is another content similarity measure, and is conventionally used to measure similarity of documents based on the term vectors (Chakrabarti, 2003, p.68-70). It could also be used to measure taxonomic similarity for concepts by replacing term vectors with the corresponding set of ancestors of the target concept (Maedche and Zacharias, 2002). If $H(c) = \{\acute{c} \in \mathcal{O} \mid c \sqsubseteq \acute{c}\}$ is the set of ancestors of the concept c from the ontology $\mathcal{O}$ then the taxonomic

[10] A catalog of semantic similarity measures intended for ontology alignment is presented by (Euzenat et al., 2004).

similarity of two concepts $c_1, c_2 \in \mathcal{O}$ is defined as follows:

$$\sigma(c_1, c_2) = \frac{\mid H(c_1) \cap H(c_2) \mid}{\mid H(c_1) \cup H(c_2) \mid} \tag{2.1}$$

This measure doesn't reveal significant quantitative similarity for concepts belonging to different ontologies which are aligned with a common upper level ontology. In such a case only the distance from the lowest common parent could be used in the similarity measure along with other indicators. Other set theoretic measures are also investigated for semantic similarity such as Tversky (1977) model of similarity. (Rodríguez and Egenhofer, 2003) used normalized form of Tversky's model to measure similarity between entities based on their feature descriptions including synonyms, semantic neighbors, attributes, and parts. If A & B are corresponding feature description sets of the entities c_1 & c_2, and $\alpha(c_1, c_2)$ represents relative importance of their non-common features (such that $0 \leq \alpha \leq 1$) then their semantic similarity is measured as follows:

$$\sigma(c_1, c_2) = \frac{\mid A \cap B \mid}{\mid A \cap B \mid + \alpha(c_1, c_2) \mid A/B \mid + (1 - \alpha(c_1, c_2)) \mid B/A \mid} \tag{2.2}$$

Menczer (2005) used a probabilistic semantic similarity measure for items ϕ_1 and ϕ_2, both placed under the concepts c_1 and c_2 respectively from the same ontology $\mathcal{O}$. If c_0 represent the lowest common parent of concepts c_1 & c_2, and $Pr[c]$ represents prior probability that any item is mapped to the concept c, then the normalized semantic similarity between the items ϕ_1 and ϕ_2 is computed as under:

$$\sigma(\phi_1, \phi_2) = \frac{2 \log Pr[c_0]}{\log Pr[c_1] + \log Pr[c_2]} \tag{2.3}$$

A prominent effort in the ontology-based semantic association is the SemDis[11] project. Different aspects of modeling (Sheth et al., 2003), ranking (Aleman-Meza et al., 2005), discovery (Sheth and Ramakrishnan, 2003), and query (Anyanwu and Sheth, 2003) of semantic associations are covered in their work. A special focus is drawn towards property based associations. Two concepts c_1 and c_n are semantically associated if there exists a path in the graph structure of the ontology i.e. $c_1, p_1, c_2, p_2, c_3 \ldots c_{n-1}, p_{n-1}, c_n$ where c_i are concepts and p_i are properties. An interesting extension to this work

[11] http://lsdis.cs.uga.edu/projects/semdis/

might be to assign salience weights to the properties and then to calculate the propagated score of the association. (Khan et al., 2004) used propagated score of related concepts to measure the similarity. (Jeh and Widom, 2002) also used scores to measure similarity of objects in a graph structure based on their related similar objects.

In contrast to properties, (Walker, 2003) used the notion of generalizations to measure semantic similarity for concepts with at least one common parent. Finding an exhaustively complete set of generalizations is very much subjective and depends on the parent concept. If $g(\acute{c}, c)$ represents the set of generalizations of the concept $c \sqsubseteq \acute{c}$ then the semantic distance $\sigma(c_1, c_2)$ between two concepts c_1 and c_2 is the minimum number of generalizations to be ruled out to get an exact match with respect to a common parent $\acute{c}$, such that $c_1 \sqsubseteq \acute{c}$ and $c_2 \sqsubseteq \acute{c}$.

$$\sigma(c_1, c_2) = min[\mid g(\acute{c}_i, c_1) \mid + \mid g(\acute{c}_i, c_2) \mid \bullet \quad \forall \acute{c}_i(c_1 \sqsubseteq \acute{c}_i \wedge c_2 \sqsubseteq \acute{c}_i)] - 2 \tag{2.4}$$

A new paradigm in information management and association discovery is human computing or social computing where individuals contribute their piece of the solution to solve a scientific puzzle. Games (von Ahn, 2006) and collaborative bookmarking[12] are prominent examples of social computing. GiveALink[13] is a new kind of search engine based on collaborative bookmarking. Machine learning algorithms are applied to discover semantic associations between the bookmarks. For instance, if many users place the same pair of web sites in the same category in their bookmarks file, then GiveALink ranks the two sites as closely related. The aggregate data shapes the ontology for the Web. (Menczer, 2005) highlighted that a user-centric model of semantic relationships can harness the individual user's context.

2.3.5 Domain Ontology

The current Web has enormous amount of information in the form of unstructured web documents, lacking explicit semantics. The Semantic Web, which is an enrichment of the existing Web, makes use of ontologies to represent semantics. A smooth transition toward the Semantic Web is need of the time where all the information will ultimately be machine understandable, paving the way for automatic association discovery and establishing trails in

[12]http://www.del.icio.us

[13]www.givealink.org

the information. The complexity involved in managing ontologies is high, which causes hindrance in realization of the dream to a large scale.

Automatically extracting relevant information and finding its correct alignment within the axiomatic space of the target ontology is not an easy task. Complexity in implementing a solution for this problem is directly proportional to the complexity of the domain ontology, potentially containing a lot of concepts and relationships between them.

So far the researchers are working on automating the process of semantic annotation for particular domains such as tourism and collaboration in research. Artequakt used parts of CIDOC[14] ontology to represent artists' personal and work detail. OntoGenie framework was tested with a very simple University Ontology having concepts such as *University*, *Country* and *State*; covering very few attributes.

Existing efforts in this area do not have exhaustive coverage of their domains of discourse. It is worth mentioning that just research collaboration includes a lot of concepts ranging from project management to personal information management. A generalized solution which works for every domain is a dream, any solution that work for every aspect of just one domain has also not been achieved.

2.4 Dynamic and Growing Ontologies

Web ontologies are rarely static. The changes are influenced from different directions such as correcting errors, adding new axioms, or even by improving the domain model. These changes have very deep side effects. The ontology-driven applications, web-pages, and agents might depend on the target ontology, and any change in the ontology contents (axioms) can potentially effect their behavior.

The changes in the ontology makes it "dynamic" in many ways. Heflin and Hendler (2000) defined dynamic ontologies as those *evolving over time* and being developed in a dynamic and heterogeneous environment such as the Web. It also refers to the fact that Communities of Practice (CoP) over time develop a common understanding based on joint interpretation of a shared terminology (Gahleitner et al., 2005). The terminology is dynamic in the sense of constantly growing in scientific discourse and being revised over time. Weinstein and Alloway (1997) have mentioned *growing* ontology

[14]CIDOC is a conceptual reference model to represent an ontology for cultural-heritage information, developed by ICOM/CIDOM document standards group by aggregating existing information sources to one coherent package.
http://cidoc.ics.forth.gr/index.html

as dynamic ontology where the agents add new concepts at runtime. As the concerned CoP develop a deeper understanding of domain knowledge, the terminology might face structural and semantic refinements, thus *moving from a loosely clustered terminology to a semi-formal and sometimes even formal ontology.*

No matter whether it is intended for small groups or for large CoP, ontology development has to allow personal views. Users should be enabled to personalize the ontologies while still being able to communicate with other members of CoP about such ontologies with the aim of converging on a common core ontology for their specific purposes.

So far we have taken two different views on personal knowledge management and ontologies. In the first chapter we explained, in the light of Mannheim's thesis of *social context*, that personal information is not an island. It is strongly driven by social interactions, dialog, and associations. On the other hand, the specification of conceptualization changes over time as the individuals develop deeper understanding of the extracted concepts. The annotations evolve and become rich in semantics. Such an *evolving model has to converge on a shared ground.* Both views are indispensable for realizing a lifetime personal knowledge management framework. In the next chapters we will explain how dynamic ontologies could be developed and managed by reusing concepts from foundational ontologies, and further used in the context of personal knowledge management to achieve shared semantic context.

Chapter 3

Building Dynamic Ontologies

Automatic semantic matchmaking is a challenge for the Semantic Web in general and for automatically establishing associative trails from the personal information. Currently, ontologies are developed mainly to implement software systems that focus on specific problems, without considering the ontology reuse and alignment aspects. The focal point of such ontologies is their usability and not the soundness of axiomatic theories.

The trade-off between usability and formality is a difficult one. On the one hand, formality comes with increased complexity, making it hard for current inference tools to interpret the semantics. On the other hand, lightweight taxonomic ontologies grounded on best practices and developed by reusing fragments of foundational ontologies can achieve formality without compromising usability.

This chapter is organized as follows: First of all an outlook on ontology reuse and foundational ontologies is presented followed by an analysis of different risk factors that might hinder the reuse of ontologies. The methodology for building dynamic ontologies is summarized in Section 3.2. The interaction intensive, question driven approach for semantics interpretation and term alignment is elaborated in Section 3.2.3. And finally we discuss the details of ontology for digital memories.

3.1 Ontology Reuse

Ontology integration, alignment[1], and reuse are at the heart of Semantic Web vision. Constrained by strong philosophical foundations and varying schools

[1]We have alternatively used the terms ontology alignment and ontology mapping. For comprehensive details on ontology matching, mapping, and alignment, refer to (Kalfoglou and Schorlemmer, 2003) and (Shvaiko and Euzenat, 2005).

of thought, ontology development is still a difficult task for *non-ontologists*. One critical issue is to identify philosophical standing of the ontology. Making this determination at the outset of the development process can potentially constrain thinking, leading to an inadequate or incomplete definition, and ultimately may prove to be formally wrong (Uschold, 1996) in satisfying a particular scenario. Reusing an existing ontology as a base could provide taxonomic and axiomatic contexts for the ontology.

Ontology reuse in turn requires ontology mapping. (Kalfoglou and Schorlemmer, 2003) defined an ontology as a pair:

$$\mathcal{O} = \langle V, A \rangle \tag{3.1}$$

where V is the vocabulary (concepts and relationships) and A specifies axioms (the intended interpretation of vocabulary). They described *total ontology mapping* from $\mathcal{O}_1 = \langle V_1, A_1 \rangle$ to $\mathcal{O}_2 = \langle V_2, A_2 \rangle$ as a morphism $f(V_1) \mapsto V_2$ such that all interpretations satisfying $\mathcal{O}_2$'s axioms also satisfy $\mathcal{O}_1$'s translated axioms i.e. $A_2 \models f(A_1)$. On the other hand, they introduced *partial ontology mapping* from $\mathcal{O}_1 = \langle V_1, A_1 \rangle$ to $\mathcal{O}_2 = \langle V_2, A_2 \rangle$ using a sub-ontology $\acute{\mathcal{O}}_1 = \langle \acute{V}_1, \acute{A}_1 \rangle$ where $\acute{V}_1 \subseteq V_1$ and $\acute{A}_1 \subseteq A_1$ such that there is a total mapping from $\acute{\mathcal{O}}_1$ to $\mathcal{O}_2$, i.e. $f : \acute{V}_1 \rightarrow V_2$

Similarly, we have defined *reuse of axiomatic context* of an ontology $\mathcal{O}_f = \langle V_f, A_f \rangle$ in another ontology $\mathcal{O} = \langle V, A \rangle$ as morphism $f(\acute{V}) \mapsto V_f$ such that $\acute{V} \subseteq V$ is root/top level vocabulary of the ontology $\mathcal{O}$ i.e.

$$\forall\, V_j \in V\, \exists\, \acute{V}_i \in \acute{V} \quad (V_j \sqsubseteq \acute{V}_i) \tag{3.2}$$

3.1.1 Reusing Foundational Ontologies

Recent efforts to realize the Semantic Web have accelerated research on the development of ontologies (Herman, 2006), a development that has been progressing slowly ever since the early efforts of the ancient Greek philosophers[2]. In the area of Information Systems & Softwares, the use of ontologies is very diverse, ranging from simple metadata description in software configuration[3] to build a *lingua franca* for resolving the terminological and conceptual incompatibilities between information networks of varying archetype and different provenance (Nejdl et al., 2002; Smith, 2003). Depending upon the

[2]See 'History of Ontology' website http://ontology.buffalo.edu/history.htm maintained by Barry Smith (Accessed July 13, 2007).

[3]Such as in Mozilla Firefox: http://www.mozilla.org/rdf/doc/ (Accessed July 13, 2007).

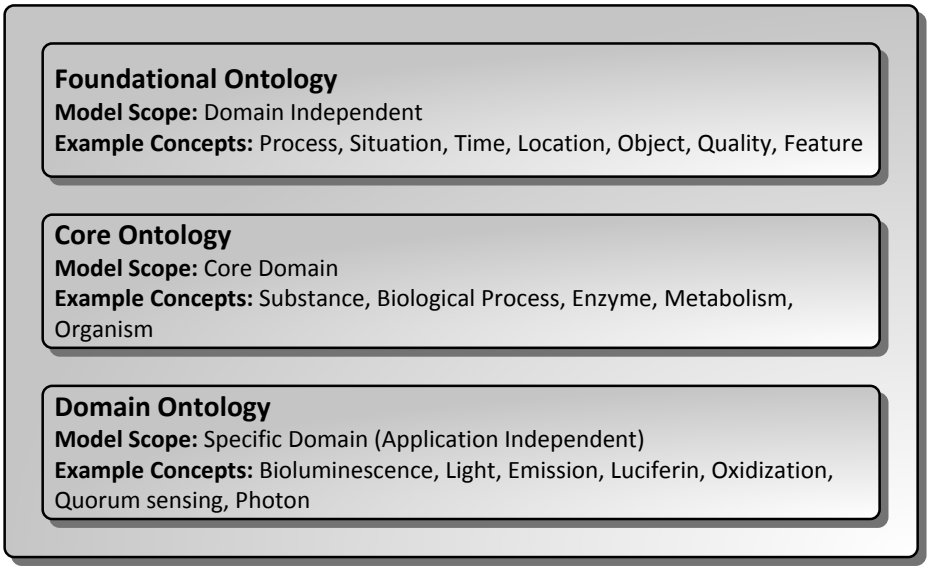


Figure 3.1: Classification of ontologies based on domain of discourse.

coverage of concepts and the model scope, the ontologies are categorized as foundational and domain ontologies (Guarino, 1998). Figure 3.1 exhibit these categories with example concepts. The work presented in (Behrendt et al., 2005; Guarino, 1997) serve as reference point for an in-depth understanding of the usage of formal and foundational ontologies.

In the ontology building process, identification of the concepts and patterns that should be modeled in the ontology is the most important and critical question. Adopting a high level view from upper ontologies provides an enormous jump start in answering this question (Masolo et al., 2003). Ontologies are catalysts for knowledge sharing (Chandrasekaran et al., 1999; Edgington et al., 2004) and mediation in heterogeneous environments (Lyttleton et al., 2005). The reuse of foundational ontologies can further facilitate mutual understanding and interoperability among ontologies that vary otherwise.

Recently, different ontology development methodologies have emerged (Jones et al., 1998), some of which advocate the reuse of concepts or *patterns* from upper level (foundational) ontologies (Gahleitner et al., 2005; Gangemi, 2005; Damjanović et al., 2007). In contrast to the foundational ontologies, a *global ontology* has worldwide knowledge which is not necessarily true for every foundational ontology. We do not intend to use global ontology, but a carefully built formal upper level ontology. After an exhaustive study of different upper-level ontologies (Behrendt et al., 2005), we decided to use

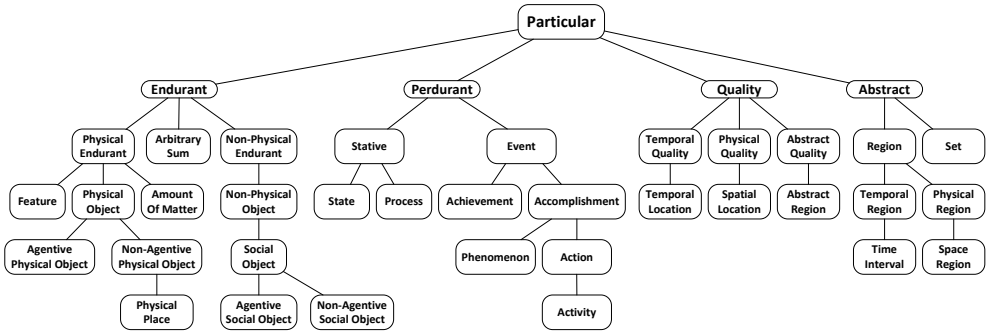


Figure 3.2: Fragment of top level of DOLCE foundational ontology.

DOLCE for our work as it is based on sound axiomatic theories and is available in OWL-DL.

3.1.2 DOLCE and OntoWordNet

Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) is part of the foundational ontology library (Masolo et al., 2003) from the WonderWeb project. It is an ontology of *particulars* having four top level concepts: *endurant*, *perdurant*, *quality*, and *abstract* (Gangemi et al., 2002). It aims at "capturing the ontological categories underlying natural language and human commonsense." Figure 3.2 depicts an excerpt of the top level of DOLCE vocabulary. Some of the basic categories and relations defined in this ontology are:

- Endurants (objects) and Perdurants (occurrences)
- Qualities (properties) and Quals (values)
- Constitution and Participation
- Parthood and Temporary Parthood
- Dependence and Spatial Dependence

In addition to the above categories, DOLCE ontology makes use of *multiplicative* approach, that is to say, different entities can co-exist in the same space and time (Masolo et al., 2003, p.8). This notion is explained with an example of vase and the amount of clay.

> "[T]he vase does not survive a radical change in shape or topology, while, necessarily, the amount of clay does. According to

> the multiplicativist, these must be different entities that are co-located: the vase is constituted by an amount of clay, but it is not an amount of clay. Certain properties a particular amount of clay happened to have when it was shaped by the vase-master are considered as essential for the emergence of a new entity – the vase. In language and cognition, we refer to this new entity as a genuine different thing: for instance, we say that a vase has a handle, but not that a piece of clay has a handle."

On the other hand, WordNet is a lexical database (Fellbaum, 1998) and is used extensively by ontology authors to ground their ontologies (Niles and Pease, 2003). To benefit from its coverage of terminology, it was aligned with DOLCE ontology (Gangemi et al., 2003). The work resulted in a major restructuring of WordNet's top level concepts. For example, the WordNet synset ⟨*Process*, *PhysicalProcess*⟩ has hypernym (a.k.a. is a *kind of*) *PhysicalEntity*, whereas OntoWordNet classifies it as *Phenomenon*. WordNet's verb classes, compared to nouns in OntoWordNet, were investigated by (Gomez, 2001) for semantics interpretation.

3.1.3 Ontology Design Patterns

The work related to Conceptual Ontology Design Patterns (CODePs) complements DOLCE ontology (Gangemi, 2005). CODeP are essentially *interconnected fragments* of this foundational ontology.

There are two core ontology patterns that also depict two of the major categories in DOLCE. The first pattern elaborates participation of *objects* in *occurrences*, known as *Participation Pattern* (see Figure 3.3). The *Description-Situation Pattern* exposits the classification of *situations* with the help of *descriptions* (Gangemi, 2005). Reusing ontology patterns is beneficial for many reasons. Some of them are stated below:

Modularity: Ontology patterns are interconnected fragments from the foundational ontologies. Exercising the same pattern in the target domain ontology entails effective reuse. Ontology reuse further allows separation of concerns and development of modular ontologies.

Axiomatic Context: Reusing well established solutions is one way to have a shared axiomatic context to the ontologies at lower levels.

Ontology Matching: Grounding axioms in formal *principles* and reusing agreed upon vocabulary modeled by the community paves the way to on-the-fly ontology matching.

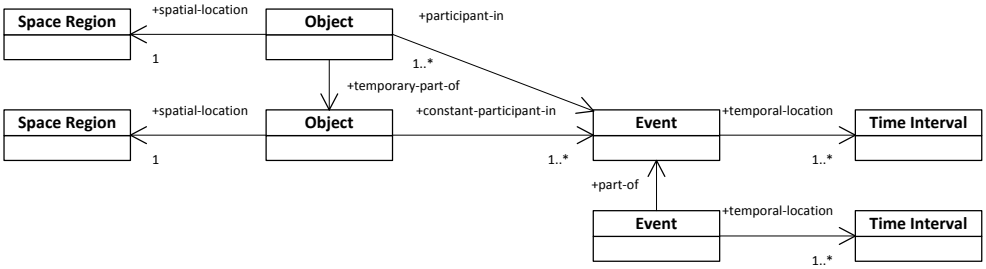


Figure 3.3: Participation pattern from DOLCE.

3.1.4 Risks in Ontology Reuse

Indecision in interpretation of the semantics of concepts from foundational ontology may hinder the domain ontology development, especially in terms of the reuse of patterns and principles of the foundational ontology. We anticipate the following possible factors that may arise and hinder the domain ontology development process while using an upper level ontology, and give a brief account of their analysis against DOLCE and OntoWordNet.

Abstraction Level

The upper ontologies do not reach down to the domain level. Use of OntoWordNet can effectively resolve this issue. For example, *Photosynthesis* is described as *synthesis of compounds with the aid of radiant energy* in WordNet. Its hypernym[4] chain leads to *PhysicalEntity*, which adds to the confusion in finding correct axiomatic space for this concept in DOLCE. In contrast, OntoWordNet has redefined *Photosynthesis* and is aligned with *Phenomenon* in DOLCE. Thus, the OntoWordNet mappings helped us find its correct alignment with DOLCE, which would otherwise be difficult due to abstraction in upper level foundational ontologies, in this case, DOLCE.

Formality Level

We can't find a sufficient set of semantic descriptors to ontologize the domain terminologies in coherence with upper level ontology - therefore failing to align the domain vocabulary with the upper level ontologies. The upper level ontology may have some formal philosophical assertions that cannot be matched with the specific domain ontology or which do not match the purpose and scope of the ontology. Although this issue can be resolved, to some extent, by using OntoWordNet vocabulary and its mapping with DOLCE,

[4]Hypernym is the more general class of another synset.

more examination of the issue and further research is needed. For example, modeling "chess game" using DOLCE vocabulary is difficult due to issues involving integration of Description Logics with Constraint Programming formalism.

Monolithic View

Although adopting a high level view from a single monolithic ontology is easier from the modeling point of view, it may still hinder the cause of interoperability. Major ontology players, such as John Sowa, advocate the use of multiple foundational ontologies coupled with mappings *to move along the lattice of different ontological commitments*[5]. The work done in (Masolo et al., 2003) is a good reference point for easy and rigorous comparisons among different ontological approaches; concluding that the most important challenge for the Semantic Web is careful isolation of fundamental ontological commitments and their formal relationships. The use of lexical semantics could be investigated to further develop a system of automatically discovering semantic equivalence between two different foundational ontologies.

Semantic Enrichment

The ontological enrichment of terminologies is possible, but the *semantic expressiveness of the resulting knowledge representation remains vacuous.* Prior research by (Smith and Rosse, 2004; Gangemi, 2005) has proved that ontological enrichment of terminologies using well established principles from foundational ontologies not only contributes to semantic expressiveness but also helps in achieving on-the-fly ontology matching.

3.2 DynamOnt Methodology

Ontologies are catalyst for sharing knowledge between automated agents. Terminologies, on the other hand, are similar to ontologies but their audience are humans rather than being interpreted by computer applications. They focus more on their linguistic and communicative functions and are not fully ontologized (Budin, 2003). Creating ontologies in a dynamic way requires an ontology development methodology which allows graceful migration from weakly structured terminologies to highly structured axiomatic theories, and to transform linguistic expressions into formal ontological statements. This is a difficult task and is mostly executed by ontology experts.

[5]http://suo.ieee.org/email/msg03804.html

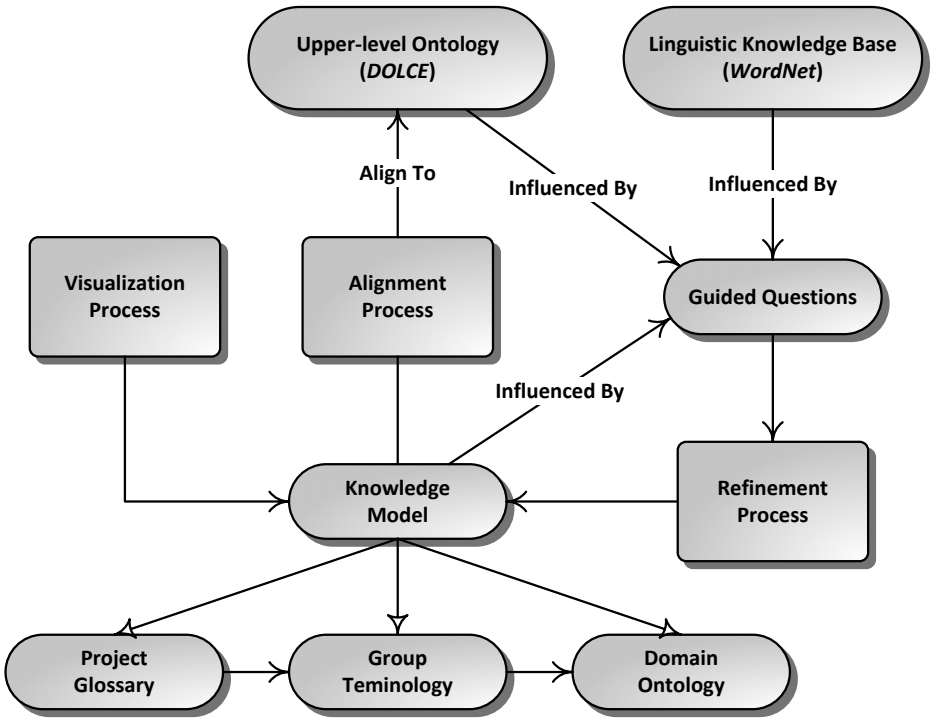


Figure 3.4: Schematic overview of DynamOnt methodology

3.2.1 Methodology Overview

DynamOnt takes the challenge of creating ontologies away from dedicated ontology designers. It aims at developing a comprehensive methodological framework for the development and maintenance of dynamic ontologies that are semantically rich, formally sound, and highly interoperable (Gahleitner et al., 2005). Figure 3.4 shows a schematic overview of the DynamOnt process. The modeling process addresses various forms of knowledge models at different levels of formality, starting with the glossaries (textual descriptions of terms), taxonomies (hierarchically structured terms), thesauri (interrelations of terms), and ontologies (axiomatized theories).

The users begin the ontology development process by creating a glossary, either starting from existing collection of terminology or by creating new ones. Over time, additional relations and attributes are added and the glossary gradually expands in size and complexity. The consistency of the knowledge model is maintained through the alignment and the refinement processes. In contrast to automatic ontology building from glossaries (Park, 2004), DynamOnt reckon on the knowledge of domain experts. Guided questions lead the users to a more structured knowledge model. Using upper level

ontologies, such as DOLCE, DynamOnt guides the user by asking questions and automatically detects possible inconsistencies or errors in the ontology.

Furthermore, the refinement process is influenced by linguistic knowledge bases. The newly created ontology is aligned with OntoWordNet by automatically adding links whenever possible and selectively prompting users where required. An adequate visualization helps the user to better understand the given knowledge model. As a result, the DynamOnt system lead to tightly coupled ontology-centric content repository that will offer large improvements in productivity of individuals to manage their personal knowledge and content resources.

3.2.2 Phases of Process Model

The DynamOnt model comprise of nine phases. Phases 1, 2 & 3 are dedicated to requirements engineering and validation, 4, 5, 6 use ontology design methods and the phases 7, 8, 9 are about implementation of the target system. Comprehensive details of these phases could be found in (Gahleitner et al., 2006; Gruber et al., 2007), we have only summarized the phases below:

Phase 1 – Identify the Problem

The starting point of the DynamOnt methodology is a decision to work on a specific problem situation in a knowledge intensive environment that may cover more than one domain. This decision starts the ontology development project with its first function to identify the problem. Usually domain experts will be assigned to describe the situation and to bring in new ideas for solving the problem at hand. The likely output of this step is informal descriptions of the problem in natural language.

Phase 2 – Structure the Problem

The second phase describes the problem from a user perspective. Collaboration is crucial in order to get more information on the problem that may cover complementary dimensions and/or conflicting views. Multiple user scenarios from different experts on the same topic could be helpful to get a broader view.

Phase 3 – Identify Purpose and Scenarios

The third phase addresses mutual understanding of the goals of the project, identifying the purpose, and eliciting a description & usage of the intended

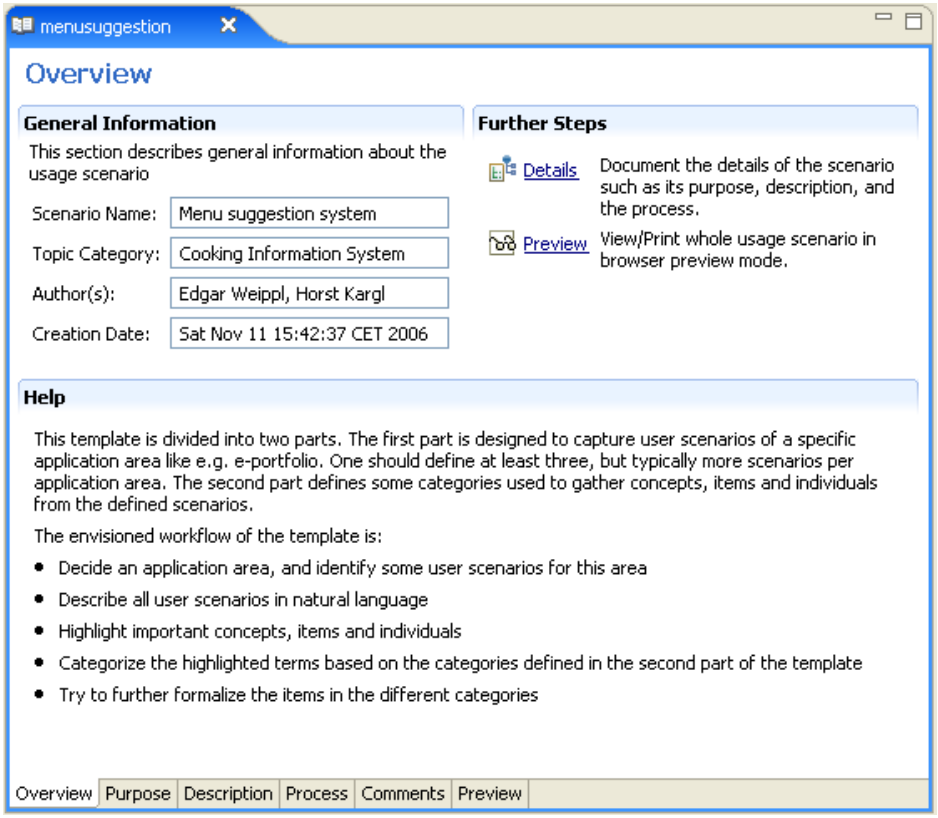


Figure 3.5: The usage scenario editor.

ontology. The ontology usage scenarios are structured and include different sections, such as *problem statement*, *purpose*, and *process description*. Describing the purpose and problem statement at early stages of ontology building is in line with different ontology development methodologies (Gruber, 1995; Jones et al., 1998). The knowledge workers are helped and guided by templates and GUI forms for creating the usage scenarios (see Figure 3.5).

Phase 4 – Identify Main Concepts

Within this phase, the individual and collaborative efforts of the various domain experts lead to an initial list of important terms/concepts and relations for different areas. The knowledge workers are further supported by templates and forms in order to do their research in a way that can lead toward better formalization of the conceptual models. Term extraction is applied to the domain corpus including documents and existing terminology databases. Selected terms are further analyzed and domain experts are guided in aligning these terms with DOLCE.

Phase 5 – Create Non-formal Models

The goal of this phase is to create non-formal models for domain concepts such as agents, roles and tasks, which are interrelated through attributes and relations. The models reflect not only the initial list of domain knowledge but also the classification according to the formality and abstraction level which will focus on different parts of the models as well as allow better integration of existing external models. This approach leads to mutual understanding of complex models. The guided questions derived from DOLCE provide a formal but transparent basis for the domain experts' negotiation. The detail of terminology alignment using guided questions is presented in Section 3.2.3.

Phase 6 – Knowledge Design

The inputs for this phase are mainly the non-formal models and the classification according to the expressiveness dimension. The classification helps to decide, which parts of the ontologies has to be formalized to a certain degree. The non-formal models are used not only to produce the formal model but also as input for several aspects of the software design and the community design. And in a lot of cases one will ignore this phase if there is no need for a fully formal model. Therefore one could argue for a clear separation between the non-formal and the formal model. Nonetheless, it is recommended to keep the separation between the models. Major revisions of the models will be done by the experts within the non-formal model and then transferred to the formal model; minor revisions could be done within the formal model itself.

Phase 7 – Community Design

The acceptance within the main user communities is an important factor for the success of the model and the system. These user communities could be "internal" domain experts (analogous to the developers in software engineering setting), as well as external user communities of the system. The acceptance of the formal model and the system can be raised in two ways: firstly, by introducing the model/system to the users in trainings or workshops and secondly by adapting existing business processes according to inputs of the resulting formal model.

Phase 8 & 9 – Software Design & Implementation

Writing software specifications (phase 8) and implementing the target knowledge driven application (phase 9) are two major tasks in any software engi-

neering process model but are beyond the scope of the DynamOnt project. Nevertheless, we integrated both phases to provide a complete process model starting from knowledge modeling to a fully developed software system.

3.2.3 Question Driven Terminology Alignment

Reusing DOLCE is sometimes difficult because of ambiguities in the correct interpretation of domain terminology. For example the concept *conference* could easily be confused by the ontology author with either *an event* or *an assemblage*, which are disjoint concepts in DOLCE. Such ambiguities cause problems in many scenarios such as for the intelligence analysis to monitor and prevent terrorist activities (Economist, 2007b). Our work focuses on reusing CODePs (Gangemi, 2005) by aligning domain terminology to correct DOLCE classes. The users are guided through a question driven mechanism to disambiguate any confusion in interpretation of their terminology (Latif et al., 2007), in line with DOLCE.

To demonstrate the strengths of the proposed methodology we have used variants of "IFIP working conference" problem. It has been used traditionally in the area of semantic data modeling in series of IFIP conferences on Comparative Review of Information Systems Design Methodologies (CRIS) held in Netherlands[6] and also in the work of (Yang, 1993; Krogstie, 1995). Different parts of the problem are elaborated throughout this thesis to illustrate many modeling situations. We intend to cover all aspects of the *conference* starting from the call for papers (CFP) to participation in the conference. We believe this is a very comprehensive example and has blend of all colors in the lifetime of a scientist.

Cognitive Support

Although the current breed of ontology management tools have made it a lot easier to build new ontologies, it is difficult to reuse concepts from existing ontologies. The main reason for this are ambiguities in semantics' interpretation of language (Harel and Rumpe, 2004) and concepts that are biased by philosophical orientation, domain nuance, and design constructs introduced at the time of their modeling. In line with the interaction paradigm – *knowledge should be confirmed by experience of actual perceptions that determine knowledge* (Goldin and Wegner, 2006) – ontologies should be built by human experts. Rather than automating the alignment task in ontology development, systems should be built for supporting human experts in alignment and reuse in the ontology building process (Falconer et al., 2006).

[6] http://www.informatik.uni-trier.de/~ley/db/conf/cris/

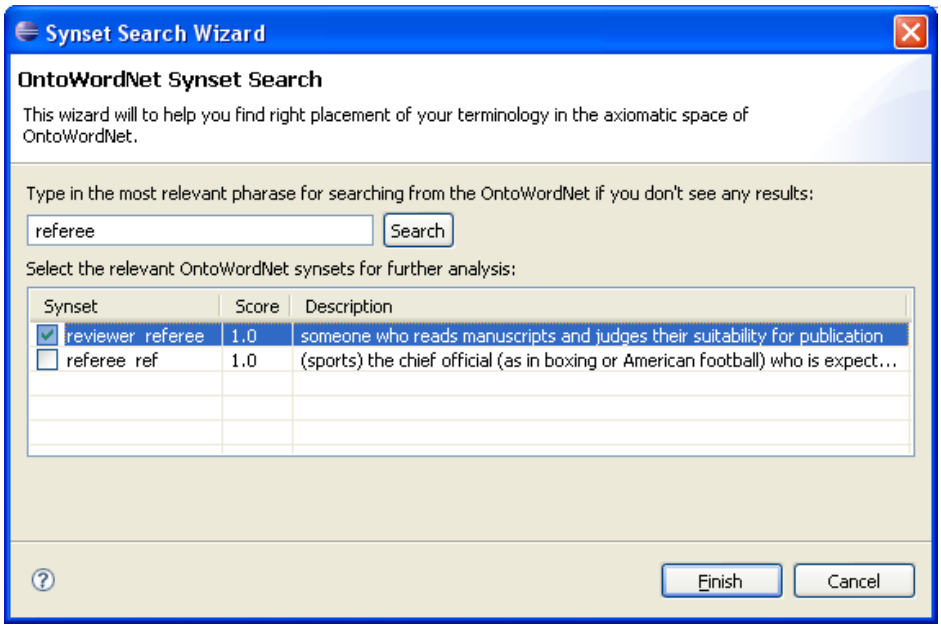


Figure 3.6: Alignment of terms with OntoWordNet.

Our hypothesis is that the concepts harvested during the ontology engineering workflow should be made available to the *knowledge worker* for rationalizing semantics based on the experience of actual perceptions that determine knowledge, guided by the best practices followed in building foundational ontologies and thesauri. Effectively, the terminology is aligned to concepts in foundational ontology, thus reusing its axiomatic context.

In the subsequent sections we introduce a novel approach for realizing this hypothesis, which also provides necessary evidence to prove its expedience in ontology engineering. Our approach focuses on helping user of the system, rather than ontology expert, in eliciting his/her knowledge by aligning the terminology with the foundational ontology. Helping the user through questions and the consequences of their answers, in finding right axiomatic context for the concepts, has resulted in effective semantics interpretation and ontology reuse for achieving shared axiomatic context in the personal knowledge box.

Bottom-up Analysis

The alignment procedure starts with bottom-up analysis. For the domain concept c, its possible mappings $M_c = \{c_{m1}, c_{m2}, \ldots c_{mn}\}$ with OntoWordNet synsets are discovered and presented to the user (see Figure 3.6).

Now, alignment of each selected mapping $c_{mi} \in M_c$ with corresponding

DOLCE class (a *subclass* of *Particular*) is identified by using OntoWordNet taxonomic links and is represented in another set $P_c = \{c_{p1}, c_{p2}, ... c_{pk}\}$ where $f : (c_{mi} \in M_c) \rightarrow (c_{pj} \in P_c)$. The set P_c renders different possibilities for aligning the concept c with DOLCE's axiomatic space. Two special situations may arise and need to be processed. In the first case, elements of P_c may be equal, which means that all OntoWordNet concepts in M_c were aligned to the same DOLCE class. In the second case, P_c might have only one element. The later situation arises when the user selects only single WordNet sense as being relevant for the concept c. In both cases, the concept c is aligned with the first element in P_c without proceeding further. Otherwise, we proceed with the normal flow of the alignment methodology.

For example, consider the concept *Conference*. The noun *conference* has three senses 1) *a prearranged meeting for consultation or exchange of information*, 2) *an association of sports teams*, and 3) *a discussion among participants who have an agreed topic*. These senses are interpreted as *Gathering*, *Organization*, and *AuditoryCommunication* respectively by OntoWordNet and are aligned with *Collective*[7], *AgentiveFigure* and *InformationRealization* from DOLCE. Consequently, DOLCE's class hierarchy is traversed to determine decision points – the places of deviations in the synset alignment with DOLCE for varying senses.

Class Hierarchy as Concept Chains

Each path of the class hierarchy from the alignments in the previous steps is transformed to a concept chain. Concept chains are needed for efficient comparisons and traversing the class hierarchy. A concept chain is a graph like structure of concepts based on the *subsumption* relationship. Concept chains are virtual collections and support navigation through operations, such as *next*, *previous*, which are delegated to the actual taxonomy in the foundational ontology. In general, for an ontology $\mathcal{O}$ a concept chain Ψ_{c_1} for a concept $c_1 \in \mathcal{O}$ is defined as a sequence of ordered pairs $\langle c_i, c_j \rangle$ of concepts such that $c_i, c_j \in \mathcal{O}$ and $c_j \sqsubseteq c_i$. For any sequence that doesn't involve multiple inheritance between the concepts, we can simplify the structure to make up a set of concepts.

$$\Psi_{c_1} = \{c_r, c_{r-1}, \cdots c_1\} \tag{3.3}$$

such that the concepts $c_i, c_{i+1} \in \Psi_{c_1}$ satisfy

[7] $Collective \equiv Collection \sqcap \forall\, member.Agent$

$$c_i \sqsubseteq c_{i+1} \quad \forall\ 1 \leq i \leq r-1 \tag{3.4}$$

From the previous *Conference* example, the first two concept chains for the initial two senses are as follows:

$$\begin{aligned} \Psi_{c_1} &= \{\mathit{Particular}, \mathit{SpatioTemporalParticular}, \mathit{Endurant} \\ &\quad \mathit{NonPhysicalEndurant}, \mathit{NonPhysicalObject}, \mathit{SocialObject}, \\ &\quad \mathit{NonAgentiveSocialObject}, \mathit{Collection}, \mathit{Collective}\} \end{aligned} \tag{3.5}$$

$$\begin{aligned} \Psi_{c_2} &= \{\mathit{Particular}, \mathit{SpatioTemporalParticular}, \mathit{Endurant} \\ &\quad \mathit{NonPhysicalEndurant}, \mathit{NonPhysicalObject}, \mathit{SocialObject}, \\ &\quad \mathit{AgentiveSocialObject}, \mathit{AgentiveFigure}\} \end{aligned} \tag{3.6}$$

The third concept chain is complicated as it involves multiple inheritance. First of all lets take a look at how DOLCE has modeled the concept *InformationRealization*.

$$\begin{aligned} \mathit{InformationRealization} \equiv\ & \mathit{PhysicalRealization} \sqcap \\ & \exists\, \mathit{realizes}.\mathit{InformationObject} \end{aligned} \tag{3.7}$$

where as

$$\begin{aligned} \mathit{PhysicalRealization} \equiv\ & (\mathit{PhysicalEndurant} \sqcup \mathit{PhysicalQuality} \sqcup \\ & (\mathit{Perdurant} \sqcap \exists\, \mathit{participant}.\mathit{PhysicalEndurant}) \sqcup \\ & (\mathit{Situation} \sqcap \exists\, \mathit{settingFor}.(\\ & \quad \mathit{PhysicalEndurant} \sqcup \mathit{PhysicalQuality} \sqcup \\ & \quad (\mathit{Perdurant} \sqcap \exists\, \mathit{participant}.\mathit{PhysicalEndurant})))) \\ & \exists\, \mathit{realizes}.\mathit{NonPhysicalObject} \end{aligned}$$

From the concept chains, it is evident that alignments contradict at the point of further classification of *SpatioTemporalParticular*, *Endurant*, and *SocialObject*. Their classification as being an *Event* or *SocialObject* is sorted out and the user is guided in deciding the correct alignment for the target domain ontology. In the subsequent sections we will explain the question and answer model for resolving such ambiguities in alignment.

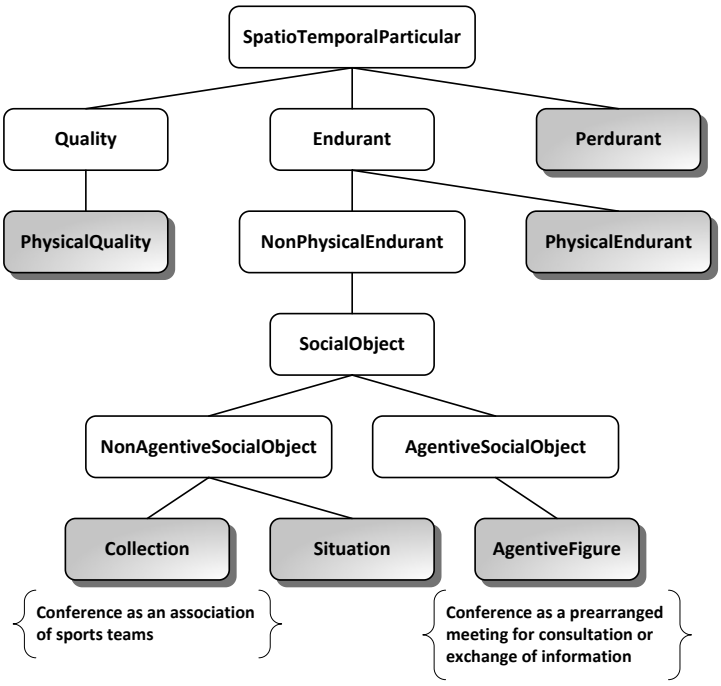


Figure 3.7: Fragment of DOLCE taxonomy and alignment for different senses of *Conference*

Game Theoretic Perspective

In game theory, a game tree is a graphical representation of a game and provides information about the strategies and the order of moves (Morris, 1994). The game tree consists of nodes, which are points at which players can take actions and are connected by edges, which represent the actions that may be taken at that node. The root node represents the first decision to be made. Every set of edges from the root node through the tree eventually arrives at a terminal node, representing an end to the game. Each terminal node is labeled with the payoffs earned by the player if the game ends at that node.

The question driven alignment methodology is, in a way, similar with the game tree approach. The bottom up analysis constructs a kind of *game* tree with the terminal nodes annotated with the payoffs for alignment with the corresponding OntoWordNet concept. If the term alignment game ends, the term in context is ultimately aligned with the corresponding DOLCE concept in the chain following the terminal OntoWordNet concept.

Slot	Explanation
q-for	Each question corresponds to a specific DOLCE class referred to by this slot. *Example: .../ontologies/ExtendedDnS#agentive-figure*
description	This slot describes the body text of the question and is taken, for the most part, from DOLCE's description of the class. Some modifications are made in order to make it easily understandable for the domain experts. This description also includes a variable $concept$ which is replaced with the user term. *Example:* Do you consider *Concept* to have roles within a society or community and hence can act like a physical agent?
hint	It provides an exemplar to help the domain expert in answering the question. *Example:* This might be true for an organization as it can have the plan to promote or regulate some activities by means of the powers conferred to it by some legal system and is executed by means of the physical agents that act for the organization.

Table 3.1: Details of question model.

Guided Questions and Answers

Question answering using Semantic Web technologies is not a new idea. (Aroyo et al., 2006) demonstrated the utility of ontology-driven dialogs to acquire domain knowledge. PowerAqua made use of distributed semantic contents to answer user queries in natural language (Lopez et al., 2006). Our approach begins the other way around entirely – the *DynamOnt system asks questions of the knowledge worker rather than the user asking questions.* Although the questions are posed in natural language, consequences of their answers are first semantically described.

Questions correspond to concepts in the foundational ontology, in our case, DOLCE. The structure of the question model, with an example for the concept *AgentiveFigure*, is explained in Table 3.1. It is worth mentioning that harvesting competency questions from DOLCE turned out to be the most difficult task in the implementation of our methodology. Initially we only modeled 45 questions for different DOLCE classes, for the most part, those mentioned in CODePs (Gangemi, 2005).

Four possible answers are permitted for each question to declare the con-

sent for aligning the user term with the DOLCE concept referred to by the question. The answers include 1) *Agree* – relevant DOLCE concept, 2) *Partially agree* – agreement with some uncertainty, 3) *Partially disagree* – disagreement with some uncertainty, and 4) *Disagree* – the DOLCE concept is not a right match for the user term. Options 2 and 3 are included to incorporate weaker notion of (dis-)agreement. Each answer is weighted symmetrically, that is, an agreement or disagreement gets equal weight. (see Table 3.2 for details). In addition, relative weights are allowed by introducing variables α as a factor of agreement and β as a factor of uncertainty. Initially they are set to 3 and 0.4 respectively.

Answer Choice	**Weight**	**Defaults**
Agree	$+\alpha$	+3.0
Partially Agree	$+(\alpha \times \beta)$	+1.2
Partially Disagree	$-(\alpha \times \beta)$	-1.2
Disagree	$-\alpha$	-3.0

Table 3.2: Answers and their weights.

Top-down Analysis

For each concept chain Ψ_c, a corresponding answer set $\Psi_a^c = \{a_1, a_2, ...a_r\}$ is constructed such that $a_i \in \Psi_a^c$ is a relevance score for class $c_i \in \Psi_c$. The relevance score is computed from users' answers for the questions against related to DOLCE alignment. To start with, elements of the answer set are initialized with zeros.

The concept chains correspond to hierarchical paths in DOLCE. The paths are established after bottom-up analysis of OntoWordNet alignments with DOLCE. In the next phase the concept chains are traversed in reverse order. It is a top-down approach considering the hierarchy of the classes in a concept chain. The domain experts are asked questions about each class excluding those questions that don't necessarily add new knowledge. For the *Conference* example, asking the domain expert if conference could be classified as *SpatioTemporalParticular* doesn't resolve any ambiguity. This strategy reduced the number of questions required to effectively align the term with a DOLCE class. A question against the leaf class is also asked to confirm the alignment (c.f. Figure 3.8). Finally, the user's answer scores for all concept chains are enumerated and the concept chain with the highest score wins the alignment decision. The algorithm for processing two concept chains Ψ_{c_1} and Ψ_{c_2} is presented next:

Algorithm 3.1 (Process Two Concept Chains for Questions) *This algorithm traverses two concept chains Ψ_{c_1} and Ψ_{c_2}, having corresponding answer sets $\Psi_{a_1}^{c_1}$ and $\Psi_{a_2}^{c_2}$, and asks competency questions of the domain expert for aligning the concept c with DOLCE.*

1. / set the pointer to first (root) item in the chain */*
2. $\Psi_{c_1}.MoveFirst()$
3. $\Psi_{c_2}.MoveFirst()$
4. / Skip till contradiction */*
5. ***while*** $\Psi_{c_1}.Current() = \Psi_{c_2}.Current()$ ***do***
6. $\Psi_{c_1}.MoveNext()$
7. $\Psi_{c_2}.MoveNext()$
8. ***end while***
9. ***repeat***
10. / Compute scores and get the concept chain and corresponding answer set with relatively higher score this far*/*
11. $[\Psi_{c_x}, \Psi_{a_x}^{c_x}] \leftarrow ComputeScore([\Psi_{c_1}, \Psi_{a_1}^{c_1}], [\Psi_{c_2}, \Psi_{a_2}^{c_2}])$
12. / Get concept at the current index */*
13. $[cindex, c_x] \leftarrow \Psi_{c_x}.Current()$
14. ***if*** $cindex = -1$ ***then***
15. / We have reached the end of the concept chain but have only achieved partial agreement or no agreement at all. Ask the user to either align with the class corresponding to the concept chain having a relatively higher score or restart the procedure after selecting different senses of c from OntoWordNet.*/*
16. ***break***
17. ***end if***
18. $score \leftarrow AskQuestionFor(c_x, c)$
19. $\Psi_{a_x}^{c_x}[cindex] \leftarrow score$
20. $\Psi_{c_x}.MoveNext()$
21. ***until*** $\sum \Psi_{a_x}^{c_x} < \tau$ */* τ is agreement threshold*/*

Variation in Top-down Analysis

Some concepts are more easily aligned with DOLCE than others. *Referee*, for instance, has three senses as noun in WordNet including *a chief sports official*, *a reviewer*, and *an attorney*. All these senses are aligned with *SociallyConstructedPerson*. The only competency question required was for the leaf class and the alignment was achieved successfully.

An interesting scenario is when there is only a single possible alignment with DOLCE but the domain expert states otherwise. For example, On-

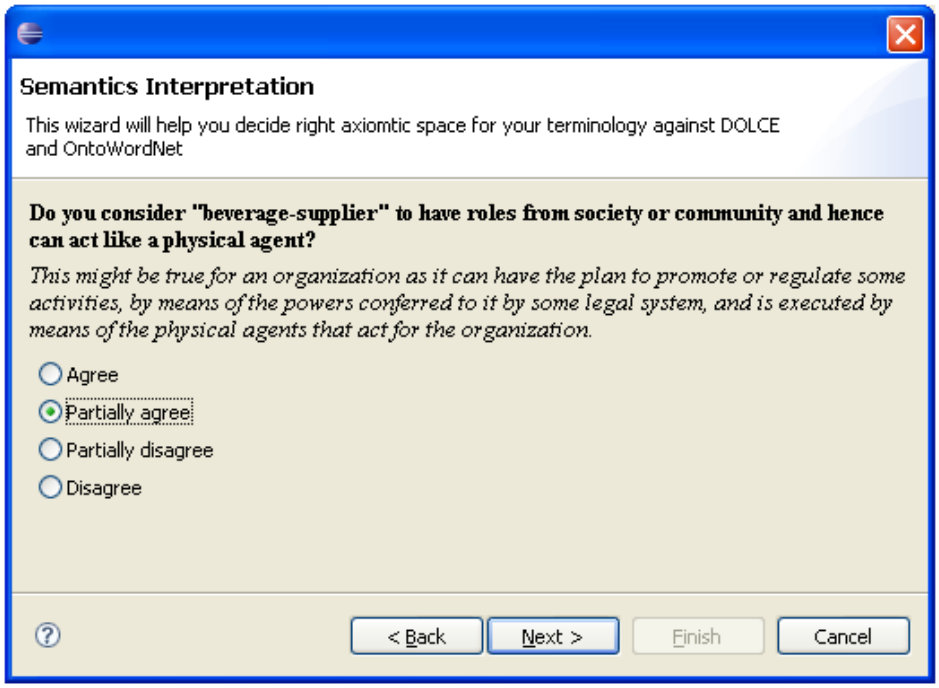


Figure 3.8: A screenshot from terminology alignment wizard.

toWordNet aligns the concept *Recipe* with DOLCE as *Situation*. Its concept chain is as follows:

$$\langle \quad \textit{Situation}, \textit{NonAgentiveSocialObject}, \textit{SocialObject}, \textit{NonPhysicalObject}, \textit{NonPhysicalEndurant}, \textit{Endurant}, \textit{SpatioTemporalParticular}, \textit{Particular} \rangle$$

The domain expert is asked to confirm the alignment with *Situation*. As a consequence of disagreement, a question is asked for each class in a bottom-up way to find out the top most class with which user agrees to align. Alternate classifications of that class are then traversed to find the right match. In the case of *Recipe*, it turned out that the user was more interested in *Recipe* being a *Plan* identified after being asked the question about sub classes of *NonAgentiveSocialobject*.

Web Services for Guiding Alignments

To support ontology building and terminology alignment, the backend of the workbench consists of different web-services. The client interface, on the

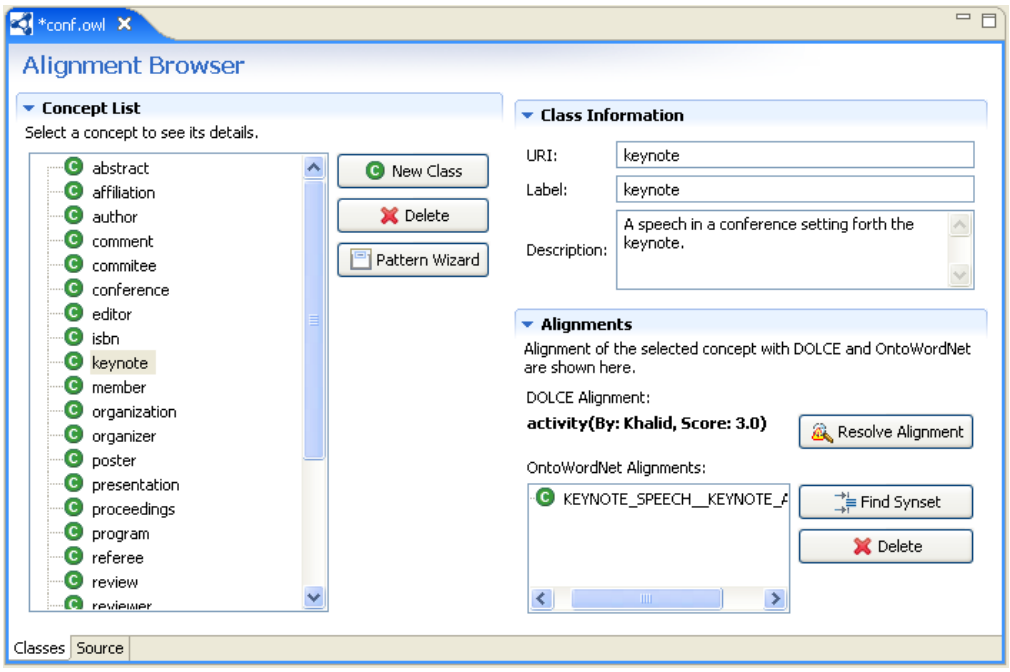


Figure 3.9: Overview of OntoWordNet alignment editor: Alignment of the concept *keynote* with *KeynotSpeech* from OntoWordNet and its mapping with *Activity* is highlighted.

other hand, are developed as an Eclipse Rich Client Platform[8] (c.f. Figure 3.9). DOLCE, OntoWordNet, and WordNet are exposed as web-services. Details of operations supported by the OntoWordNet web-service are given below.

With current Semantic Web frameworks and APIs, such as Jena, it is hard to load both OntoWordNet and DOLCE on one machine along with the DynamOnt workbench because of mammoth memory requirements. For such pragmatic reasons we have deployed these web-services, on a separate machine[9]. To improve efficiency, the OntoWordNet web-service uses Lucene[10] index of the class labels and descriptions. The index is generated using the script mentioned in Figure 3.10. This strategy greatly improved the lookup performance.

The user is allowed to select a concept and align it with the relevant DOLCE class by answering the questions posed by the system. Answer sets for the user are accumulated to match corresponding concept chains.

[8]http://www.eclipse.org/rcp/
[9]http://storm.ifs.tuwien.ac.at:8081/
[10]http://lucene.apache.org

Operation	Description
GetSenses	Given a term returns URI's of all the matching OntoWordNet concept.
GetDescription	Given an OntoWordNet concept this operation returns its detailed description.
GetParent	Returns immediate parent of the OntoWordNet concept.
GetParentsChain	Returns hierarchical chain of parents (subClass relations).
GetAlignment	Returns DOLCE alignment of a WordNet synset from the OntoWordNet mappings.

For example, for three concept chains there will be 3 distinct collections of answer sets. The answer set with the highest accumulated score is adopted for the alignment decision and the concept is aligned with the leaf concept referred to in the concept chain for that answer set.

This far, we have presented a methodology for building formal ontologies by aligning the domain terminology with the foundational ontology. Reusing the axiomatic context of the foundational ontology effectively resolves misconstructions in the domain modeling process. Competency questions guide the user in terminology alignments.

3.3 Building Ontology for Digital Memories

The ontology for digital memories can serve two purposes. On the one hand it could give uniform structure to the metadata and on the other hand it can provide semantic insight of the contents. The difference could be understood with an example of email. An *ontology* for email could either provide the model to structure its header fields or can move one step ahead in modeling a framework for explicitly augmenting the body contents. Here we introduce a modular ontology for semantic enhancements of digital memories. The ontology covers both of the aspects and is developed using the DynamOnt approach which is explained in the previous sections. In phases 1 to 5 of DynamOnt, we covered the first aspect and the next chapter deals with the framework for augmenting the unstructured contents with semantics.

First of all, we highlight how and why digital documents are manifestation of memories. Then we briefly discuss that RDF, being a canonical format from W3C, is better suited for long term preservation of life items. And then, we provide a detailed account of multifold semantic insight approach

```
// Load Jena model
OntModel model = ModelFactory.createOntologyModel(OntModelSpec.OWL_DL_MEM);
model.read(ownFileStream, “http://www.loa-cnr.it/ontologies/WordNet/OWN");

// Prepare Lucene index writer
File dir = new File(ownIndexPath);
IndexWriter writer = new IndexWriter(indexDir, new StandardAnalyzer(), true);

// Iterate through all the classes
Iterator iter = model.listClasses();
while (iter.hasNext()) {
   OntClass cls = (OntClass)iter.next();
   Document doc = new Document();

   // Prepare concept label and description for indexing
   String interm = getConceptLabel(cls);
   doc.add(Field.Text(FIELD.LABEL, interm));
   String comment = cls.getComment(null);
   if (comment!=null) doc.add(Field.Text(FIELD.DESCRIPTION, comment));

   doc.add(Field.UnIndexed(FIELD.URI, cls.getURI()));

   writer.addDocument(doc);
}

// Close index writer
writer.optimize();
writer.close();
```

Figure 3.10: Building full-text index of OntoWordNet using Lucene.

of modeling the ontology for digital memories.

3.3.1 Problem, Purpose, and Scenarios

The starting point of the DynamOnt methodology is to describe and structure the problem from the user perspective resulting in comprehensive structuring of the problem and detailed description of the ontology usage scenarios.

Documents & Activities as Digital Memories

A document is usually viewed as either *a purposeful and self-contained collection of information* – focusing on information content and exchange such as in business collaborations, or as *a specialization of the record of a happening* with the intension to rationalize the memory of an experience in the digital

form (Smith, 2005). It is implicit in both cases that digital documents could easily be related to the underlying personal memories, albeit not necessarily cognitive memories

For instance, consider the following use case: Research institutions mostly run multiple projects in collaboration with other academic institutions or industrial partners. On the lower side each research project produces 10-20 deliverables and other documents. Many discussions take place among the partners and quite a number of existing research material is studied and analyzed. These documents contain a significant portion of the individuals' memories. The research consortium, as well as individuals, might be interested in organizing the documents based on their knowledge contents and also in keeping track of the relationships between documents for efficient retrieval.

Conclusive of the above scenario and as mentioned by (Czerwinski et al., 2006; Fitzgibbon and Reiter, 2004), digital documents are attributed as memories because they reflect one's thoughts and experiences. But one thing not covered in the user's documents is the memory of an happening such as taking a print-out of a document and committing some changes during a collaborative editing task. In most of the cases, memories of such operations and activities are captured outside the premises of the *personal* documents (such as the versioning system and the printing log with is maintained by the printer). So, we argue for a broader notion of the digital memories which encircles the records of activities. The digital documents of the lifetime and the activities are therefore separated into life items and digital memories respectively. This leads us to comprehensively describe life items and digital memories, though in the perspective of this thesis, both terms are mentioned alternatively.

Definition 3.1 (Life Item) *A life item is a digital record of personal information. More specifically, by life items we mean digital information objects such as personal notes, digital photographs, web pages from the browsing history, emails, instant messages, and Internet calls. These information objects, in most cases, involve a social context; for example a photo taken in a research conference and an email sent to a colleague about the project status.*

Definition 3.2 (Digital Memories) *Digital memories are record of happening, an experience, or any other kid of information object that a person may perceive and then able to rationalize at a certain point (or period) of time during his/her lifetime. Inherently, the life items, as defined above, are also digital memories but this might not hold in reverse. As a matter of fact, digital memories cover broader sphere of user activities than just life items.*

For example, if a project related document is referred as life item, the records of editing operations on that document by the person are digital memories; explaining 'when' and 'what' was changed. These kind of 'memories' are necessary to answer 'why' and to find patterns from the user activities.

Preserving Digital Memories

The archival of lifetime knowledge raises an obvious problem – long term preservation. The research work presented in (Lee et al., 2002; Ludäscher et al., 2001; Potter, 2002) demonstrate the need for a cross-platform standard for long-term preservation of documents and suggest to use an XML infrastructure for preserving digital archives. It is obvious that a similar strategy could be applied to personal digital documents. Resource Description Format (RDF) has emerged as more elaborated and general purpose solution for information representation and metadata description. The abstract data model of RDF is graph-based (Klyne and Carroll, 2004), but it could also be encoded in XML (Beckett, 2004).

Usage of RDF/XML aggrandizes semantic insight of the structure as well as the content of the life items. For example, the Figure 3.11 depicts the excerpt of a message from 20-newsgroup corpus in three different serialization formats namely RFC-2822, XML, and RDF/XML. It is evident that RDF serialization is in the lead as far as the clarity of the structure and semantics are concerned. *RDFization*[11] of life items allows RDF-aware agents to easily understand the structure of the contents and to interpret the metadata. Furthermore, asserting ontological commitments to the structure enables semantic match-making.

Multifold Semantic Insight

Different manifestation of documents are evident from the previous discussion. Smith (2005) suggested to consider the document as a generalization of the *speech acts* such as statements, and therefrom ontology of documents as a generalization of the ontology of speech acts. In all forms, it is important to manage both the inward and outward focus of the documents. The inward focus is necessary to realize the document contents as a connected graph of facts, activities, and information recorded in the contents. On the other hand, outward focus paves the way to realize the collective document space as connected graph of documents. Increasing the connectedness and semantic understanding of the contents, in-turn, facilitate understanding of the patterns in the information (c.f. Figure 3.12).

[11] Converting original structure to RDF.

From: km@ky3b.pgh.pa.us (Ken Mitchum)
Newsgroups: sci.med
Subject: Re: tuberculosis
Message-ID: <206@ky3b.UUCP>
Date: 3 Apr 93 15:10:01 GMT
References: <1993Mar25.020646.852@news.columbia.edu>
...

(a) Original RFC-822/2822 representation.

```
<email>
  <headers>
    <From>km@ky3b.pgh.pa.us (Ken Mitchum)</From>
    <Newsgroups>sci.med</Newsgroups>
    <Subject>Re: tuberculosis</Subject>
    <Message-ID><206@ky3b.UUCP></Message-ID>
    <Date>3 Apr 93 15:10:01 GMT</Date>
    <References><1993Mar25.020646.852@news.columbia.edu></References>
  ...
</email>
```

(b) XML format as in XMail testbed (Potter, 2002).

```
<rdf:RDF xml:base="http://storm.ifs.tuwien.ac.at/life-items" ... >
<Email rdf:about="mid:206@ky3b.UUCP">
  <from>
    <EmailAgent rdf:about="mailto:km@ky3b.pgh.pa.us">
      <foaf:mbox>km@ky3b.pgh.pa.us</foaf:mbox>
      <display>Ken Mitchum</display>
    </EmailAgent>
  </from>
  <subject rdf:datatype="&xsd;string">Re: tuberculosis</subject>
  <date rdf:datatype="&xsd;dateTime">1993-04-03T15:10:01</date>
  <references>
    <Email rdf:about="mid:1993Mar25.020646.852@news.columbia.edu"/>
  </references>
  ...
</Email>
```

(c) Conversion into RDF increases structuring insight.

Figure 3.11: Excerpt of an Internet Message from 20-newsgroup corpus serialized in different formats.

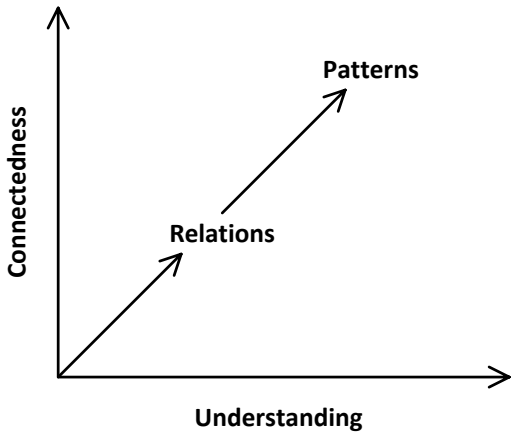


Figure 3.12: Transition from data to knowledge.

Breaking down the target conceptual space is considered as one of the best practices not only in software engineering (Griswold et al., 2006) but also in ontology modeling (Bao and Honavar, 2006; Seidenberg and Rector, 2006). Different approaches for ontology segmentation have been proposed by the researchers. Structural criteria-based segmentation of ontologies is presented in (Schlicht and Stuckenschmidt, 2006; Stuckenschmidt and Klein, 2004). Lee and colleagues (2006) used ontology layers in their ontology architecture based on already established categorization of ontologies such as the meta-layering approach in Model-driven Architecture. A methodology for categorization of ontologies along three dimensions *usage* (community acceptance), *formality* (expressiveness), and *abstraction level* (model scope) is extensively discussed in (Gahleitner et al., 2006; Schaffert et al., 2005). It has to be said that each dimension presents a unique perspective of the ontology contents.

Most of the existing approaches for the ontology classification and ontology segmentation work at the post-development level. Additionally, one aspect not covered in these approaches is the semantic insight offered by the conceptualization present in the ontology. This dimension could be focused to partition the ontology of an information system at the development phase. In the proceeding section we will introduce a novel approach for ontology partitioning into layers based on the semantic insight of life items offered by each layer. The resulting modules are locally complete and thus are easier to maintain separately.

To set off multiple views of the same item we introduce another dimension to the ontology layers. This dimension is based on the information context

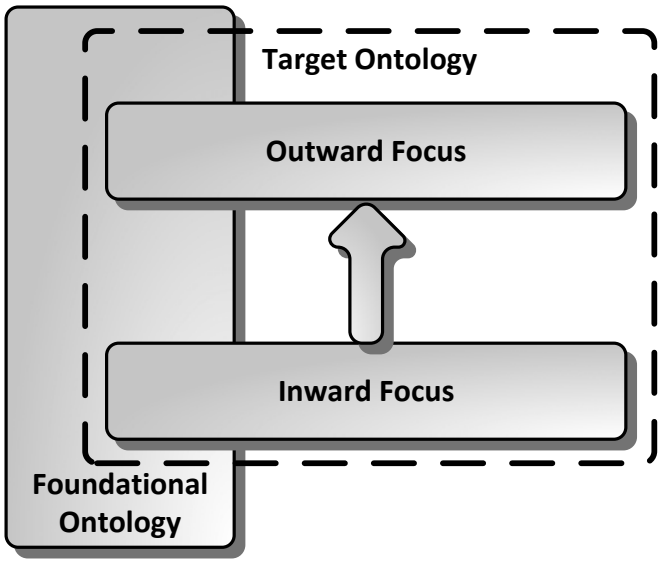


Figure 3.13: Transition from inward to outward focus of an ontology, grounded on the foundational ontology.

and the semantic insight offered by the ontology, such as a layer for semantic enhancements of the underlying resource structure – focusing the inward perspective – and another layer for realizing trails intended to present outward viewpoint (c.f. Figure 3.13). The individual modules in this layering might have same formality level and domain coverage which makes it different both from semantic domain and formality layering. The intrinsic characteristic of these layers is the coverage of the semantic insight for a particular resource in a specific domain. Other than the traditional benefits of ontology segmentation, we report three aspects of these layers.

Separation of concerns: The ontology is divided into layers each spotlighting a specific aspect in relation to the semantic insight of the information objects.

Local Completeness: Each layer is locally complete in its coverage of conceptualization which allows easy maintenance of individual modules.

Dynamicity of Knowledge Model: Some parts of the knowledge model (ontology) might be more stable than others. For instance, it is rare that email structuring would change very frequently where as the contents might include nearly everything from plain text description to compound multimedia objects. Segmenting the ontology in multiple

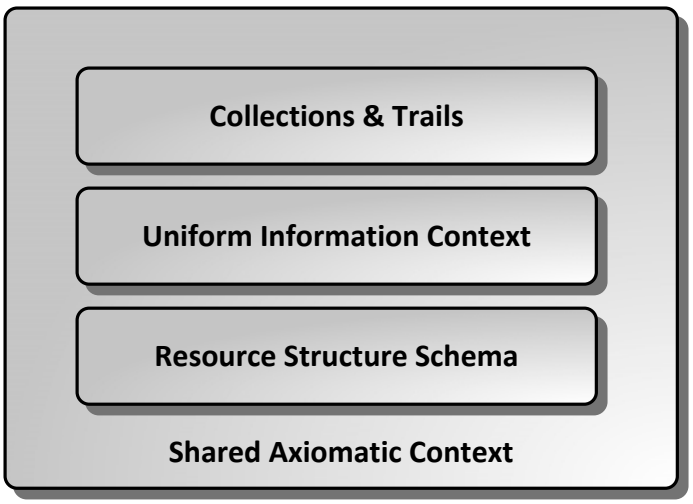


Figure 3.14: Layers of ontology for digital memories following multifold semantic enhancement strategy.

layers allows particular modules to be maintained and evolve separately.

For the personal knowledge box we enrich incoming life items so that the context is elaborated and stored with the fed items. For doing so, we try to avoid ambiguous interpretation of the contents by conforming to the foundational ontology. Figure 3.14 presents an overview of the ontology layers that we have modeled for semantic enhancements of life items in order to realize associative trails with shared axiomatic context.

The bottom most layer focuses on the resource structure and RDFizes the header fields in the case of semi-structured documents such as for emails, web-pages, and address book entries. Further layers are discussed in the next chapters.

3.3.2 Identifying Main Concepts

There exist a number of thesauri built with an effort of thousands of man-months and it would be pity not to reuse those resources. One very prominent example in the ontology community is CYC (Lenat, 1995), and its open source spin off OpenCyc ontology, utilizing a *person-century* effort in a period of more than two decades. As a matter of fact, the success of the Semantic Web is largely connected to efficient methods for allowing reuse, alignment, and mapping of the ontologies.

For modeling the ontology we have borrowed concepts from a number of existing ontologies such as SUMO (Niles and Pease, 2001), OpenCyc, AKT Reference Ontology (Motta and Gibbins, 2003), SWRC (Sure et al., 2005), FOAF, different email models in RDF, and Smith's Document Ontology (Smith, 2005). This satisfies a fundamental feature of conceptual ontology design patterns and DynamOnt methodology about ontology reuse. It is worth mentioning here that the target ontology is not exhaustively complete and the conceptual domain coverage is also not exclusive. For any future extension, we encourage to adopt new types from existing ontological and lexical resources. To start with, we modeled following items from different categories.

Communications: Email, Phone Call, Instant Message, Talk, Dialog, Speech, Meeting

Web: Web-page, Bookmark

Multimedia: Audio/Video Sequence, Music, Song, Photograph, Picture, Image, Figure, Drawing,

Bibliography: Document, Article Journal, Conference, Project, Thesis, Dissertation, Report, Deliverable, Book

General: Task (Todo List), Activity (User Process Monitoring), Note, Project, Event (iCal), Address Book (vCard)

3.3.3 Non-Formal Model

The rationale behind the non-formal model is the RDFization of the structure of the fed item. Very basic information from header fields of semi-structured documents is attached as properties to the life items. These properties are directly concerned with the actual contents such as sender, date, and subject line in email messages and the title of web-pages. These concepts are modeled by following the design principles from DOLCE ontology to achieve shared axiomatic context. For example email is modeled as a sub class of the concept *InformationObject* from DOLCE.

$$
\begin{aligned}
(3.8) \qquad \mathit{Email} \sqsubseteq\ & \mathit{InformationObject} \sqcap \\
& = 1\mathit{subject} \sqcap \exists\, \mathit{orderedBy}.\mathit{RFC}2822 \sqcap \\
& = 1\mathit{sentDate} \sqcap \exists\, \mathit{from}.\mathit{EmailAgent} \sqcap \\
& \exists\, \mathit{recipient}.\mathit{EmailAgent} \sqcap \\
& \exists\, \mathit{expresses}.\mathit{Description}
\end{aligned}
$$

where as

$$\text{(3.9)} \qquad \mathit{sentDate} \sqsubseteq \mathit{temporalLocation}$$

$$\text{(3.10)} \qquad \mathit{recipient} \equiv (\mathit{to} \sqcup \mathit{cc} \sqcup \mathit{bcc})$$

Core and Extended Items

In the Personal Knowledge Box we distinguish between core and extended life items. Core items basically are direct structuring of the digital items present in the current desktops such as emails, web-pages, calendar entries, and files. Initial structuring of items and serialization of semi-structured headers in RDF, on one hand, breaks the ground to construct the knowledge box and, on the other hand, paves the way to realize the Semantic Web in large scale (Bergman, 2007).

The extended items are placed above the core items. They reuse the underlying structure and enhance the semantic insight of the contents. One example is a web-page and an email both expressing a *Call for Papers* (CFP) for the same conference. The web-page and emails are core items with initial RDFization. And the CFP is an extended item which enhances the semantic insight of the contents and also connects the two core items. Consequently we can define email contents (or a web-page) expressing a CFP as depicted in Figure 3.15.

Inward and Outward Focus

In some cases we have modeled the outward focus of the items though resource structuring schema (the informal model) originally intend to cover the inward structuring insight. These outward features are taken, for the most part, from the tools and applications. For instance, we browsers maintain the credentials about the referrer of a page such as stating a page was opened by following the hyperlink from the Google search result page. Another example is the information about the IMAP folders for placing emails in the group. In many situations, such an outward focus increases the understanding of the intent of the user for categorization.

Modeling Decisions

W3C established a Best Practices Working Group[12] with the focus to "provide support for practical issues related to ontology engineering and use for

[12]The working group was closed as of 29 September 2006. See details at http://www.w3.org/2001/sw/BestPractices/

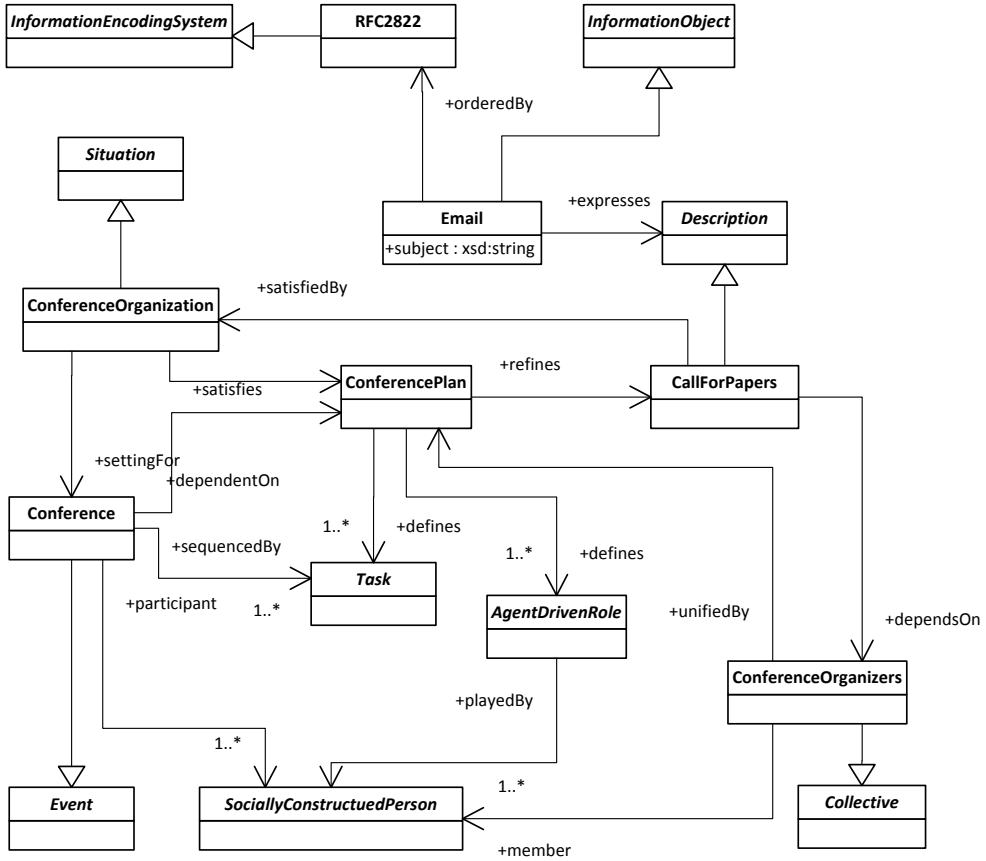


Figure 3.15: Conceptual schema of an email expressing a call for papers for a conference by reusing concepts from DOLCE.

the Semantic Web." Still there are many situations for which no specific best practices are documented.

Number of issues are debated in different semantic web conferences, mailing lists, and other forums with apparently no concrete decision. For example, there had been an intensive discussion about very fundamental debate of Description Logics vs. Datalog paradigm (Patel-Schneider and Horrocks, 2006), open-world vs. close world[13] modeling (Damásio et al., 2006; Horrocks et al., 2005; Mazzocchi, 2005), and proper use of URIs as resource identifiers (Berners-Lee, 2007; Booth, 2007; Cyganiak, 2007; Bouquet et al., 2007). The issue of URIs is easily understood from the following explanation: For a university a valid question might be, is URL of a university's website a unique identifier of the university as a social or physical entity, or the reference to its website? The irresolution in these issues hinder the development of formally consistent ontologies at large scale. Here we will discuss two cases of indecision in modeling Resource Structure Schema specifically and ontology for digital memories in general.

Roles vs. Concepts

Ontologies follow varying strategies to model concepts. A common dichotomy is between roles and classes. For example, a deliverable could be modeled either as a role taken by a report (or software) in a project or as a separate named class. To elaborate the problem and to highlight the issue we present both models in Equation 3.11 and 3.12 respectively:

$$\text{(3.11)} \quad \mathit{Project} \sqsubseteq \mathit{Activity} \sqcap \forall\, \mathit{deliverable}.(\mathit{Report} \sqcup \mathit{Software})$$

$$\text{(3.12)} \quad \mathit{Deliverable} \sqsubseteq (\mathit{Report} \sqcup \mathit{Software}) \sqcap \exists\, \mathit{outcomeOf}\,.\mathit{Project}$$

In most of such situations we have adopted the later model. At the same time, it is worth pointing out that there already are efforts to align both models using concept/role bridges (Ghidini and Serafini, 2006).

Description Logic and Rules

Although we tried our best to restrict the target ontology to only Description Logics, that is the DL version of Web Ontology Language (OWL-DL). In some cases it was inevitable to use rules. A restriction requiring coreference cannot be stated in OWL-DL. Consider the following extended version of

[13] A statement is assumed to be true if its negation can not be proven.

Equation 3.12 where a deliverable is realized as either a software or a report and is modeled as an outcome of a project:

$$\begin{aligned} (3.13) \qquad \mathit{Deliverable} \sqsubseteq\ & (\mathit{Report} \sqcup \mathit{Software}) \sqcap \\ & \exists\, \mathit{outcomeOf}\,.\mathit{Project} \sqcap \\ & \exists\, \mathit{author}.(\exists\, \mathit{memberOf}\,.\mathit{Project}) \end{aligned}$$

The problem in the above statement is that a deliverable could be an outcome of such a project that doesn't have any member at all; as the author of the deliverable might belong to a separate project. Adding a cyclic constraint in OWL-DL is not possible to best of our knowledge. The target system is built for common users who might not be expert in ontologies. For this reason we should expect invalid and inconsistent assertions. To work-out the solution for maintaining the consistency and validity of the asserted statements in such situations, we modeled a set of inference rules. In the case of *Deliverable-Project* problem the following rule is used:

$$\begin{aligned} (3.14) \qquad \forall\, x, y, z\ (\mathit{Person}(x) \wedge \mathit{Project}(y) \wedge \mathit{memberOf}\,(x, y) \wedge \\ \mathit{hasAuthor}(z, x) \wedge \mathit{outcomeOf}\,(z, y)) \\ \Rightarrow \mathit{Deliverable}(z) \end{aligned}$$

The outcome of this rule is not essentially asserted into the knowledge box as we don't want to annotate each and every outcome of a project as a deliverable. The resulting individuals are used only to check if the output of Description Logic reasoner is consistent and valid in the light of above scenario.

So far we have presented a specification of digital memories which mainly focuses on the inward content structuring of the items. RDFization of the items is only the first step in building associations and effectively managing lifetime knowledge box. In the next chapter we will discuss how a formal annotation framework can be used to enhance comprehensibility in the lifetime archive.

Chapter 4

Information Context and Trails

Ever increasing capacity of contemporary storage devices had inspired the continuous archival of information (Gemmel et al., 2003; Fitzgibbon and Reiter, 2004; Ahmed et al., 2004). The enormity of the lifetime information poses a serious challenge in terms of comprehensibility (Muggleton, 2006). An information item is useful only when it is stored and later on being possible to look at it. Now the technology is at such a point that the enormous amount of information can be stored, but is not being exploited effectively and efficiently due to lack of semantics.

As pointed out by (Shirky, 2005), semantic classification without formal categorizations is not suitable for large corpus with unstable and unrestricted entities. A structured semantics enhancement framework is needed to improve the comprehensibility in case of life time capture of personal experiences. We propose to make use of semantic information context to bring together information extracted from diverse media types into an integrated model. The context model provides the semantic insight of life-items related to their spatio-temporal location, involved agents & their activities, and content labels.

First of all, we will present an analysis of existing approaches for personal information organization on the desktop, highlighting the need for a horizontal integration using a uniform information context. In the subsequent section we will introduce the *context ontology* for capturing and enhancing semantics of life-items which also provides a binding for information items. Finally, in Section 4.3, we focus on the issue of collections and semantic associations between information items by exploiting the information context. We discuss the main issues that arise when realizing the vision of associative trails for personal information.

4.1 Personal Information Organization

The choice of how to organize information is not always obvious, since more than one scheme can apply. For that reason applications usually support multiple presentations of their confined contents. For example, emails could be listed based on their sent/receive date, persons involved in the email, and the communication thread.

4.1.1 Horizontal Integration

Modern day desktop applications allow humans to benefit from different approaches for organizing information items. Such applications rarely exploit their semantics and also do not use a common conceptual scheme for information management (Latif and Tjoa, 2006). Thus information items organized by an application following a specific metaphor can not be automatically linked to the items managed by other applications at the horizontal level. Nevertheless, such an inter-relation and inter-linking is important because of the congenital nature of human mind to follow the association of thoughts (Bush, 1945). Otherwise, users have to redundantly re-enter on a new path to find the required information with the associated counterparts managed by different applications in their workspace (Boardman et al., 2003) - users are captured in a "prison of metaphors." Figure 4.1 highlights a scheduled event in Sunbird, the event website as visited by the user, and the photo taken in that event – all managed by different applications with no connections. Lack of integration and interaction support among the personal desktop tools places the information items in different islands. The problem could also be demonstrated by the following scenario:

> Alex is searching an article on *associative trails* knowing the fact that it was saved in a (file system) folder after following the web link forwarded by John in his email from last summer (see Figure 4.2). To add to the complexity, say, Alex has read a lot of articles from the web on the same topic and have saved them in his workspace. Now with traditional information retrieval techniques he can try to search for the article based on the keywords "John" and "associative trails." Such keyword based search will not retrieve the desired article effectively (may rank it too low) because of the fact that no document actually contains both search phrases.

While this trivial example shows the usefulness of portraying involved agent from the context, exploiting more associations emerging from other

Figure 4.1: Part of the website showing program of the event (top left), a picture taken in that event (top right), and scheduled event in Mozilla Sunbird (bottom).

dimensions like location and time can realize even more complex scenarios. And as a result of using structured annotation framework a diverse range of digital memories can be stored, indexed and searched in an integrated and seamless fashion.

4.1.2 Semantics Enhancements with Context

Several definitions of context are present in the literature. In the conceptual modeling and knowledge representation community, the context commonly refers to a particular view of the domain. Context is also widely practiced in pervasive computing where a variety of context models are in use by researchers (Strang and Linnhoff-Popien, 2004).

We use the term *information context* in the following sense: *The context of a life-item is the semantic insights of its contents and relation with other items.* Such an information context could be used for personal knowledge management either in the information capture and archiving stage or during

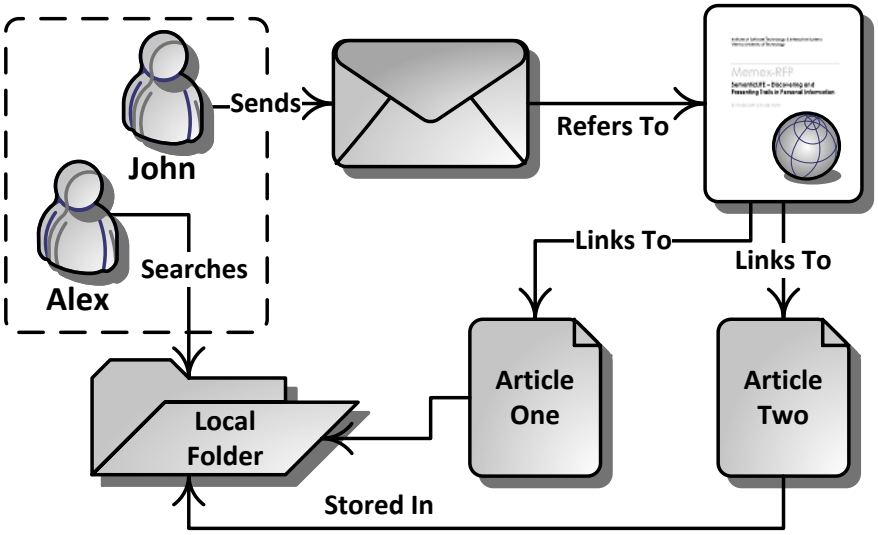


Figure 4.2: Interrelation of different information items.

the retrieval time.

The diversity of context models raises the question of what aspects should be captured for personal information as part of the context. Context principles, such as time, although used somehow in different applications mostly implicate that their inter-relation is missing in many ways. Semantics of the information items are not modeled explicitly in personal information management softwares and there is no binding of properties possessed by one life-item with the others. Providing a unified view of the information space consisting of such objects becomes substantially difficult and hence implies the absence of morphism from one contextual organization to any other context in personal information space. Time and location are the widely used metaphors for organizing personal information (Aris et al., 2004; Buchanan et al., 2004; Freeman and Gelernter, 1996; Rekimoto, 1999; Ringel et al., 2003). Different studies have revealed that although important but, time should not be the only principle to organize personal information (Dumais et al., 2003; Teevan, 2004). A little effort has been put to identify other generic dimensions for information organization.

We propose to model aspects which may well be used to organize information items. Thus the information context will symbolize the personal information space in which each aspect represents one dimension or view of the information.

4.1.3 Analysis of Information Organization Models

One of the prominent information organization model is facet-based classification. Faceted classification plays an integral part in many information retrieval methods (Broughton, 2006), but it is more useful when supported by a structured framework. For example, (Ranganathan, 1963) suggested to structure the facets using the following basic dimensions: Personality (primary facet), Matter (physical materials), Energy (Action), and Space & Time. Information about an item can be added within these slots.

Wurman and colleagues (2000) identified that organization of information is finite and there are only five principles: Location, Alphabet, Time, Category, and Hierarchy, known as LATCH. Our study of existing desktop and personal information management systems has revealed that most of them use one or more LATCH principles for information organization (Latif and Tjoa, 2006). But, all of its principles may not be taken as input to context model for digital memories due to their non-contextual nature. Alphabet, for example, is more an ordering principle and could be applied to any other context dimension such as lexical ordering of locations.

Similarly hierarchy could be taken as a mean to organize categories and locations. This is because the hierarchy typically implies arrangement of items in a tree structure. On the other hand, category is also a subjective dimension and could be used to organize other principles such as a category of locations (all islands).

Another aspect missing in the LATCH, if considered as context model, is the *agent*. Most of the activities such as personal communication (e.g. emails, instant messages, phone calls) and collaboration, as in research projects or in office work, embody other persons. Photos also encircle agents mostly human agents. Yet there are life-items which depict non-human agents such as correspondence with some research funding agency or university. Thus replacing alphabets with agents in the LATCH make it suitable for modeling as context metaphor in personal information management.

Interestingly spatial location, time, and agent correlate with where, when, and who respectively. Additionally we introduce semantic labels to highlight the content semantics, referring to 'what' (see Figure 4.3). This dimension is analogous to *Personality* facet in Ranganathan's structural framework. Modeling these dimensions in personal information as context not only explicitly amplifies content insight but also provide a foundation for bindings between the information items. Having described the drive behind the uniform information context framework, we proceed toward its development.

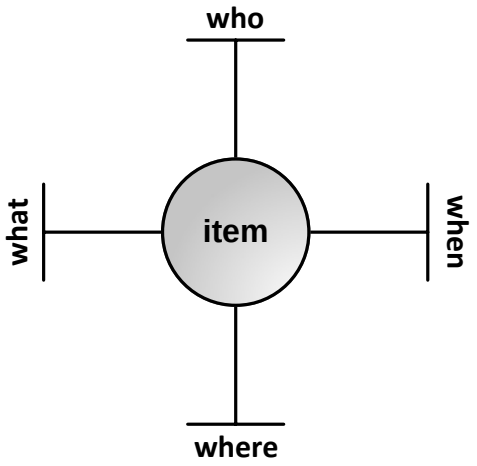


Figure 4.3: Overview of STeAL model.

4.2 Information Context Ontology

In order to retrieve precise and semantically correct information, when dealing with context in personal knowledge box, it is necessary to organize metadata of the life-items in effective and comprehensive fashion. The context framework provides the semantic insight of life-items related to their *Spatio-Temporal location, involved Agents and their activities, and content Labels* (STeAL for short). Life-items that are close to each other, for example in space or time, become connected. The individual dimensions in the context model are explained in the subsequent sections.

A central question for any ontology is how properties (*qualities*) and property values should be modeled. We draw from the advantages of conceptual spaces (Gärdenfors, 2000) to build the conceptual schema for spatio-temporal location (first two dimensions) in the context framework. An important aspect of conceptual spaces is that the property values can be structured into *quality domains*; spatial concepts belong to one domain, concepts for color values to a different domain, kinship relations to a third, and so on. These quality domains make up the dimensions $D_1, \ldots, D_n$ of the conceptual space. Each dimension is endowed with a certain geometrical or topological structure. It should be noted that some dimensions have only a discrete structure, that is, they merely divide objects into disjoint classes. A *point* in the space is represented by a vector $\vec{v} = \langle d_1, \ldots, d_n \rangle$ with one index for each dimension. Consequently, a property reflects a *region* of the conceptual space S. Now the objects could be represented as points in the conceptual space. In this way, the similarity of two objects can be defined via the distance between their representing points in the space; the smaller distance between the rep-

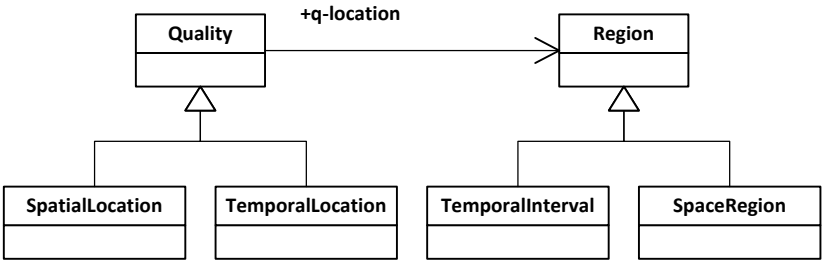


Figure 4.4: Quality and regions adopted from DOLCE.

resentations of two objects entails more similarity between them and vice versa.

4.2.1 Spatial Location

Location describes a point or extent of a life-item in the geographic space. DOLCE follows the approach of conceptual spaces for modeling spatial location of objects (Gangemi et al., 2002). An object can be declared to have spatial property, value of which is located in the space region within the geographical coordinates (c.f. Figure 4.5). Two objects are near each other on the location axis provided their spatial regions are closely located.

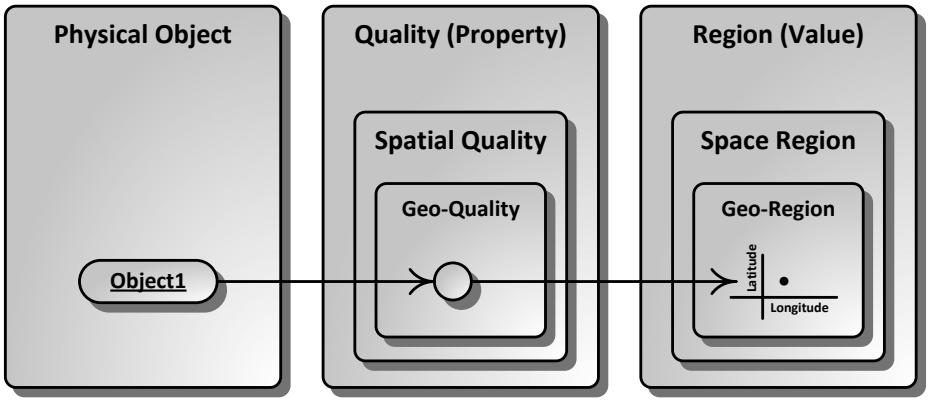


Figure 4.5: Presence of an object in geographic space region.

OpenCyc describes a comprehensive vocabulary of geographic concepts including *Continent*, *Country*, and so forth. Including these concepts in the context framework is necessary as we can't expect the user to attach precise geographic coordinates with each and every life item. In contrast, these concepts are aligned with *GeoPoliticalEntity* in DOLCE, which in turns refers

to *GeographicalObject* (a *PhysicalObject* that can have spatial properties). The model is depicted in the figure below.

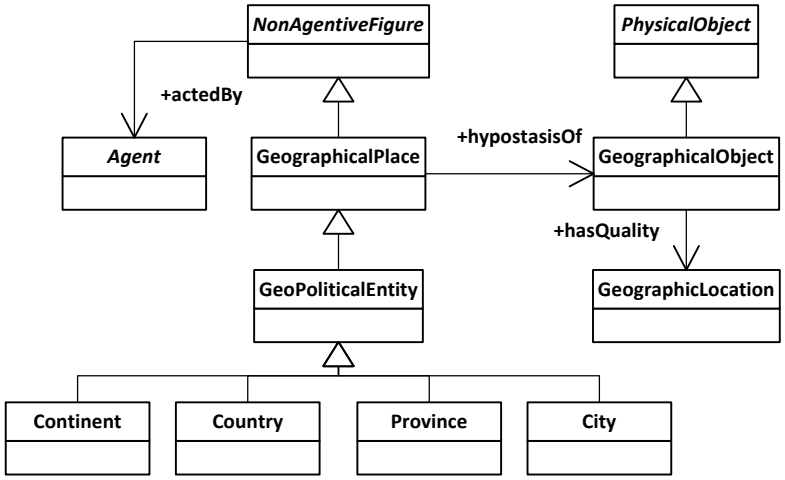


Figure 4.6: Geographical and political concepts adopted from OpenCyc and DOLCE.

Interestingly OpenCyc includes comprehensive vocabulary on proximity, containing the concepts such as *near*, *adjacentTo*, and *onPath* which we found very useful for supporting continuum organization and building associated trails. For instance, the user can declare that a photo was taken *near* the geographic location of a city.

4.2.2 Temporal Location

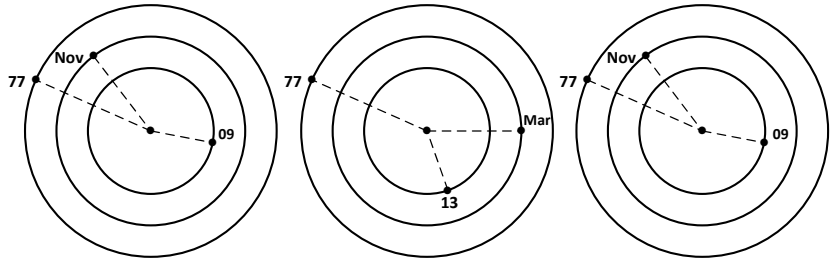
Time could be modeled in two unique ways, either as a one-dimensional line of real numbers or as a circular structure. In the first case, if we assume *present* time as the zero point on the line, the future corresponds to the infinite positive real line and the past to the infinite negative line. One dimensional view of the time has certain limitations. Imprecise, relative, and recurring time values are difficult to model. People in different cultures might have different time dimensions as a part of their cognitive structures. In some cultural contexts, time is viewed as a circular structure (Gärdenfors, 2000, p.6-7). For instance, every *Easter Monday* is similar though occurring on different exact date every changing year. This circular nature happens at different granularity levels from seconds to centuries[1].

[1]Seconds and centuries are mentioned as examples otherwise there exist other leves of granularity.

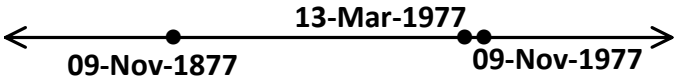
Similarity could be modeled as an exponentially decaying function of the distance (Nosofsky, 1986, Eq.4b). For time intervals, the similarity could be measured along different levels (dimensions) of the intervals. If $\sigma_T(t_1, t_2)$ expresses the similarity between two temporal values t_1 and t_2, and d_i their distance along ith dimension then the following formula expresses the relation between the two measures:

$$\sigma_T(t_1, t_2) = e^{-c(\sum_{i=1}^{k} w_i d_i)^2} \tag{4.1}$$

where c is a general "sensitivity" parameter and is assumed to be mute by default (i.e. $c = 1$), and weights w_i are context dependent variables that represent the relative degree of salience assigned to different dimensions. Large value of w_i stretch the temporal space along the ith dimension, while small values of w_i will shrink the temporal space along that dimension (Gärdenfors, 2000, p.20). A weight of zero would make the dimension carry no effect on the distance and hence has no effect on similarity. Over time, the knowledge and interests of the user can influence the salience weights (Gärdenfors, 2000, p.104).



(a) Time as circular structure.



(b) One dimensional view of time values.

Figure 4.7: Different representations of time values.

Consider three date values: 09-Nov-1877 (t_1), 13-Mar-1977 (t_2), and 09-Nov-1977 (t_3). On the linear scale, as depicted in the Figure 4.7, the first and second values are closer than the first and third date values. In a particular context, say birthday and assuming the first date value signifies the actual date of birth, the first and the third date values can be inferred to be closely similar which otherwise are very far away on the linear scale. This could be calculated by measuring the similarity along four dimensions: date in month

(d_1), month (d_2), year in century (d_3), and century (d_4). We set the weights along these dimensions to 0.0025, 0.0757, 0.9217, and 0.0001 respectively. Now using City Block distance we can measure, that $\sigma_T(t_1, t_2) = 0.6841$, $\sigma_T(t_2, t_3) = 0.6841$, and $\sigma_T(t_1, t_3) = 0.9999$ which proves that t_1 and t_3 are semantically closer. Although triangular inequality[2] holds in this example as $(0.6841 \times 0.6841) < 0.9999$, but it can't be ensured in general for semantic similarity matching (Rodríguez and Egenhofer, 2003, p.446).

We also modeled many relationship predicates to measure temporal similarity in events (Allen, 1983) such as the following:

$$\begin{aligned}(4.2) \qquad \forall\, t_1, t_2, ep_{t_1}(&\mathit{TemporalThing}(t_1) \wedge \mathit{TemporalThing}(t_2) \wedge \\ &\mathit{EndPoint}(t_1, ep_{t_1}) \wedge \mathit{TemporalBoundsContain}(t_2, ep_{t_1})) \\ &\Rightarrow \mathit{EndsDuring}(t_1, t_2)\end{aligned}$$

OWL-Time (Hobbs and Pan, 2004) was suggested by Semantic Web Best Practices Working Group (SWBPD) for modeling the time in OWL. In future, we will examine how the time model we adopted from Cyc and DOLCE could be aligned with OWL-Time.

4.2.3 Agents and Activities

This dimension covers two aspects. First of all, it describes the agents involved in the life-item and secondly the *actions* that agents are performing such as a person eating sushi, playing football, and delivering a speech.

Agent is a generic notion, an individual or an organization, which can take a role for carrying out operations. This dimension does away with the ambiguities in classification of agents, actors, roles, and actions. This distinction is important because most user scenarios are not clear about that. Typically the agent is clearly referred in the domain description and the actual role he/she plays can be captured from the situation.

- Agents refer to the real world entity such as *Person* and *Organization*. Figure 4.8 describes the agent model of context framework. A special case is $\mathit{GeoPoliticalEntity} \sqsubseteq \mathit{GeographicalPlace}$ (such as a country) which is a *NonAgentiveFigure*. Figures have strong corelation with agents in DOLCE based on the *actedBy* predicate.

[2]Triangle inequality $\sigma(a, b).\sigma(b, c) \leq \sigma(a, c)$ is one of the tenets of similarity in metric spaces. Other two essential properties are *symmetry* $\sigma(a, b) = \sigma(b, a)$ and *maximality* $\sigma(a, b) \leq \sigma(a, a) = 1$.

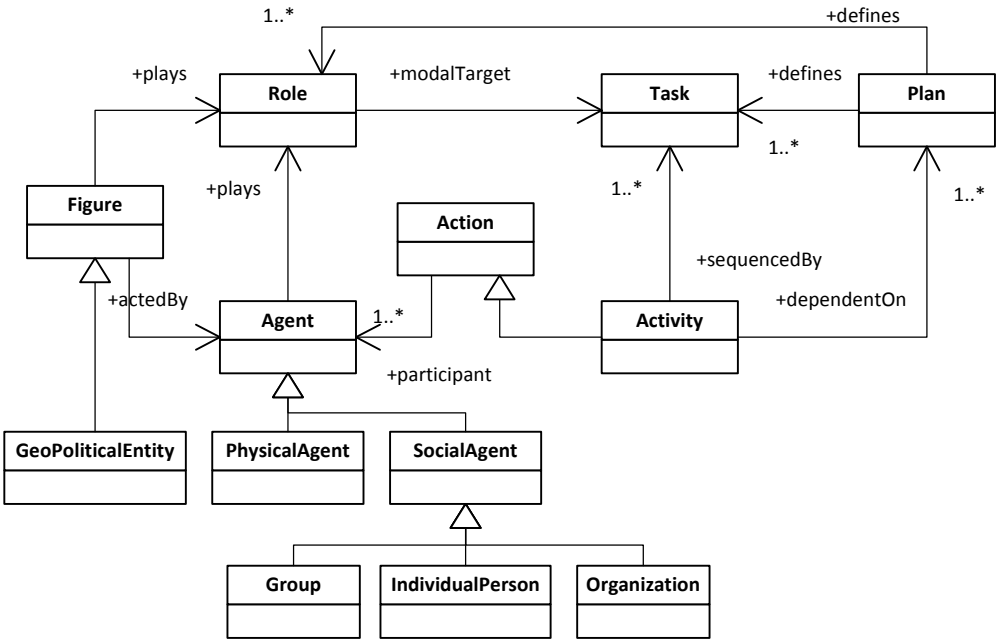


Figure 4.8: Agents and activity model as a specialization of *Role-Task* ontology pattern.

- Role refers to the actual function an agent has in the process or its part such as Author and Publisher. A role may be used as an abstraction for a specific actor.

Moreover, activities typically include an actor, an agent who plays a role, to perform the activity. Occurrences of actor-action tuple in the natural language text is common and in many cases it could be discretely categorized into finite classes (Schank, 1973). The defining attribute in these occurrences is the type of the action such as "physical" and "mental" acts. Additionally, WordNet allows to look at the *troponyms* (further types) of predicates. Troponyms could be used to extract and classify reasonable amount of action predicates from the natural language text. The verb *communicate*, for example, has the troponym *speak* which is derivationally related to *speech*. On the other hand, **MTrans** – one primitive action defined by Schank – represents a change in the mental control of a conceptualization including the communicate action. So a mention of speech situation in the text could be mapped to MTrans act – leading to *action-based associations* of digital memories.

Example

Agent and activity model can be best explained with an example of a keynote speech in a conference. First of all we have modeled different speech and communication settings:

$$\text{(4.3)} \quad \mathit{SpeechAct} \sqsubseteq \mathit{Act\_HumanAction} \sqcap \exists\, \mathit{performedBy}.\mathit{Agent}$$
$$\text{(4.4)} \quad \mathit{Communication} \sqsubseteq \mathit{Act\_HumanAction}$$
$$\text{(4.5)} \quad \mathit{SpeechCommunication} \sqsubseteq \mathit{AuditoryCommunication}$$
$$\text{(4.6)} \quad \mathit{Address\_Speech} \sqsubseteq \mathit{SpeechAct}$$
$$\text{(4.7)} \quad \mathit{Colloquium} \sqsubseteq \mathit{Address\_Speech}$$
$$\text{(4.8)} \quad \mathit{Lecture} \sqsubseteq \mathit{Address\_Speech}$$
$$\text{(4.9)} \quad \mathit{Discussion} \sqsubseteq \mathit{SpeechCommunication}$$

This lead us to model a keynote speech as follows:

$$\text{(4.10)} \quad \mathit{KeynoteSpeech} \sqsubseteq \mathit{Address\_Speech}$$

Now consider processing a transcript[3] of *Tom Gruber*'s keynote speech in the *International Semantic Web Conference (ISWC)*. Clearly the transcript should be annotated as *about* the keynote speech.

```
:iswc-keynote-transcript
   rdf:type     :Text ;
   :title       "Social Web meets Semantic Web" ;
   :about       :iswc-keynote-speech .
:iswc-keynote-speech
   rdf:type     :KeynoteSpeech ;
   :partOf      :int-semweb-conf ;
   :deliveredBy  :TomGruber .
:TomGruber
   rdf:type     :Person .
```

where *deliveredBy* $\sqsubseteq$ *performedBy* is a functional participation relation between agents and actions. Additionally in DOLCE's terminology, aboutness of an entity is perceived within a certain context. At the same time, an information object (transcript of the keynote speech in this case) expresses

[3]Or even an MP3 file (*AudioSequence* $\sqsubseteq$ *InformationObject*) of the recording.

a conceptualization which is further satisfied by the context. For example, look at the following modified version of the Equation 4.10.

$$\text{KeynoteSpeech} \equiv \text{Address\_Speech} \sqcap \exists\, \text{setting}.\text{KeynoteSpeechSituation} \tag{4.11}$$

where as

$$\begin{aligned} \text{SpeechSituation} \equiv\ & \text{Situation} \sqcap (\\ & \exists\, \text{settingFor}.\text{Address\_Speech} \sqcup \\ & \exists\, \text{satisfies}.(\text{Description} \sqcap \\ & \exists\, \text{defines}.\text{SpeechRole})) \end{aligned} \tag{4.12}$$

$$\begin{aligned} \text{KeynoteSpeechSituation} \equiv\ & \text{SpeechSituation} \sqcap (\\ & \exists\, \text{settingFor}.\text{KeynoteSpeech} \sqcup \\ & \exists\, \text{satisfies}.(\text{Description} \sqcap \\ & \exists\, \text{defines}.\text{KeynoteSpeechRole})) \end{aligned} \tag{4.13}$$

and

$$\text{SpeechRole} \sqsubseteq \text{AgentDrivenRole} \sqcap \forall\, \text{playedBy}.\text{Person} \tag{4.14}$$

$$\text{KeynoteSpeechRole} \sqsubseteq \text{SpeechRole} \tag{4.15}$$

$$\text{Person} \sqsubseteq \text{Agent} \tag{4.16}$$

So, feeding the text file in the knowledge box shall produce the following RDF statements:

```
:iswc-keynote-speech
   rdf:type      :KeynoteSpeech ;
   :partOf       :int-semweb-conf ;
   :deliveredBy   :TomGruber ;
   :setting      :iswc06-keynote-speech-context .
:iswc-keynote-transcript
   rdf:type      :Text ;
   :title        "Social Web meets Semantic Web" ;
   :about        :iswc-keynote-speech ;
   :expresses    :iswc-keynote-description .
:iswc-keynote-description
```

```
   rdf:type     :Description ;
   :defines     :iswc06-keynote-speech-role .
:iswc06-keynote-speech-context
   rdf:type     :Situation ;
   :satisfies   :iswc06-keynote-description ;
   :settingFor  :iswc-keynote-speech .
:iswc06-keynote-speech-role
   rdf:type     :KeynoteSpeechRole ;
   :playedBy    :TomGruber .
```

Now the query to search for "Speeches of Tom Gruber" could be modeled in SPARQL as follows:

```
SELECT ?speech WHERE {
  ?speech rdf:type :Speech ;
      :deliveredBy :TomGruber .
}
```

And the query to search for "Keynote speeches in ISWC'06" would unfold as the following:

```
SELECT ?speech WHERE {
  ?speech rdf:type :KeynoteSpeech ;
      :partOf   :int-semweb-conf .
}
```

Digital photos are categorized as personal, professional photos, and art work (van Ossenbruggen et al., 2006). Personal photo collections may include a photo of an art object. Such photos have a special actor, the *creator*, annotated as the original author of the object depicted in the photo. One such example is a painting depicted in a photo being annotated with the *artist* (where artist *is-a* creator). The user might ask for *the photos depicting Picasso*. Inference rules play a certain role in these situations. In the current context a rule can state that actually its the painting depicting Picasso being depicted in the photo.

$$\forall x, y, z \; (\mathit{Depicts}(x, y) \wedge \mathit{Depicts}(y, z)) \Rightarrow \mathit{Depicts}(x, z) \tag{4.17}$$

Figure 4.9: Snapshot of popular tags from del.icio.us from August 09, 2007.

4.2.4 Semantic Labels

The collaborative social tagging has resulted in many interesting applications for the next generation of the Web. A tag, the basic building block of social tagging, is an informal keyword or term which is assigned to a web resource. There are many-to-many mappings between terms and concepts. A single concept can be expressed by synonymous terms, variations, abbreviations, and acronyms (Economist, 2007b). Conversely, the same term can represent different concepts. This is a well-known problem in information retrieval (Agirre and Edmonds, 2006). Traditional text indexing and retrieval methods do not effectively comprehend different senses of a term.

A similar challenge is posed in social tagging where users select natural language terms to describe web resources such as documents. The same concept can be expressed by different terms or words by the users. Figure 4.9 depicts a snapshot of popular tags from the famous social tagging website del.icio.us[4]. Variants of these tags which were also present among popular tags are highlighted in Table 4.1. Furnas and colleagues (1987) observed that the probability of two persons choosing the same word to describe the same concept is less than 0.2. A case study in (Bar-Ilan et al., 2006) concluded that different interpretations of the meaning of the tags may worsen the retrieval and recommended to experiment with a system where the users can provide both the tags and their context.

[4]http://del.icio.us/tag/

Tag	Variants or related tags
game	games, gaming
blog	blogs
apple	mac, osx
article	articles
free	freeware, opensource
operating system	linux, ubuntu, osx, windows
technology	tech
social	web2.0
finance	money
fun	funny, humor
recipe	recipes, cooking
music	audio, mp3
media	tv, movies, films
tutorial	tutorials, howto
rubyonrails	rails, ruby

Table 4.1: Variants of popular tags from del.icio.us

Tags are bottom up labels without any context semantics. Knowledge Organization Systems (KOS) are known to provide a conceptual and representational foundation for the context. Combining tags with context semantics modeled in KOS creates a structured data framework which could be used to build retrieval-efficient knowledge management system. The success of social tagging and its wide spread adoption has attracted researchers from the Semantic Web community to work around the problem of variable terminology. (Gregorowicz and Kramer, 2006) addressed the problem of generating a domain independent map of keywords to concepts from Wikipedia. Their work does not address the problem of mapping domain specific resources to the concept space. In contrast, (Mika, 2005) exploited the potential of enhancing social tagging with semantics and by following concept-based schemata.

Our work for semantic labels is focused in this direction; that is, extending the tagging mechanism with semantics by aligning the natural language tags to the concepts from shared conceptualizations. In Section 3.2.3, we have explained how concepts in the emergent KOS are aligned with OntoWordNet – an enhanced version of WordNet being aligned with the DOLCE foundational ontology. Additionally, the $\langle \mathit{Label}, \mathit{LifeItem} \rangle$ tuple is given a relevance score to further amputate uncertainty in categorization. Users are allowed to annotate the items with the weighted labels activated in their mind. These labels, in the OntoWordNet aligned form, constitute a conceptual index of

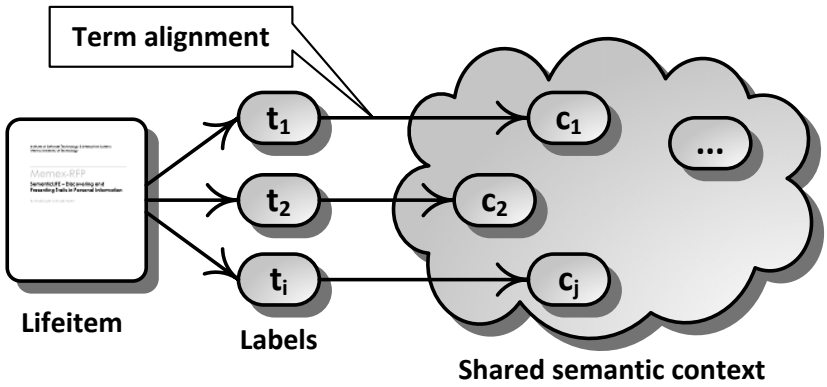


Figure 4.10: Achieving shared specification of labels through alignments.

life-items which characterizes the important point of one's life.

Labels and Categorization

In cognitive psychology, there is a general agreement about activation of certain concepts in users mind during the categorization process (Sinha, 2005), and such neural activity could even be correlate to haemodynamic response of the brain (Logothetis et al., 2001). Selection of ultimate categories is driven by the cultural knowledge (shared specification of conceptualization) about the item under scrutiny (Boster, 2005). The process is similar for the digital documents such as for placing files in different folders. The categorization in the digital world has another dimension to it; the aspect of retrieval afterwards. Our hypothesis is that achieving shared specifications of the labels, essentially aligning them to foundational ontologies (c.f. Figure 4.10), would also guarantee easier (semi-) automatic categorization and efficient retrieval. So the labels are first step towards categorization and building trails in the knowledge box.

Label Representation

A term t within a conceptual context κ is assumed to refer to a concept c i.e. $(t, \kappa) \Rightarrow c$. The term-concept map represents a bridge from the natural language domain to the concept domain. A number of formal languages exist to model the connectedness and shared semantic understanding of the terms and concepts. Simple Knowledge Organization System (SKOS), developed under the W3C framework (Miles and Brickley, 2005), is designed for representation of taxonomies and concept schemes. SKOS provides a single class for representing concepts, `skos:concept`. Term to concept relationships are

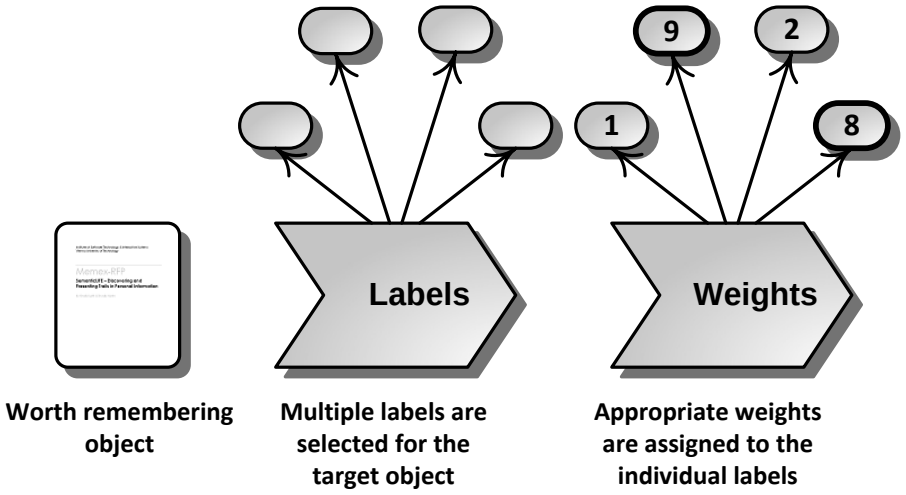


Figure 4.11: The process of assigning weighted labels to life-items.

defined by three mutually-exclusive properties of concepts:

prefLabel is the preferred term for a concept.

altLabel represent alternative terms for the concept including acronyms, abbreviations, spelling variants, and irregular plural/singular forms.

hiddenLabel are terms that should not appear in the user interface, but may be used in free text search operations; typically used for common misspellings.

Inter-concept relations are materialized in SKOS by thesaurus-like notions such as *broader*, *narrower*, and *related*. More complex concept-concept relationship can be established by using SKOS extension and mapping vocabulary. In addition, SKOS provides *subject* and *primarySubject* predicates to relate resources with concepts.

SKOS concepts could be used directly to annotate digital memories with labels with the only exception of ranking the labels. In certain scenarios a label might be preferred as more important over other labels. This is usually the case when user assigns multiple labels to a resource and wants to specify importance of some of the labels (c.f. Figure 4.11). Label weights are also useful when displaying metadata about an item. The weights can influence the sizes of the labels as a mean to portray their importance.

Influenced by the Newman's[5] and Gruber's[6] tagging ontologies, we modeled a weighted labeling scheme. RDF serialization of the model is listed below:

```
:LabelContext a owl:Class ;
  rdfs:label "Label context" ;
  rdfs:comment "Label context is used to annotate life item" .

:label a owl:ObjectProperty ;
  rdfs:subPropertyOf skos:subject ;
  rdfs:range :LabelContext ;
  rdfs:label "has label context" ;
  rdfs:comment "Indicates a life item tagged with a label" .

:weight a owl:DatatypeProperty ;
  rdfs:domain :LabelContext ;
  rdfs:range xsd:positiveInteger ;
  rdfs:label "Label weight" ;
  rdfs:comment "Provides the weight of the label subject" .

:subject a owl:ObjectProperty ;
  rdfs:subPropertyOf skos:subject ;
  rdfs:domain :LabelContext ;
  rdfs:range skos:Concept ;
  rdfs:label "Label subject" ;
  rdfs:comment "Subject of the attached label" .
```

The subject of the labels are SKOS concepts and are aligned with the OntoWordNet concepts. Additionally, the subject mentioned in the label context is inferred to be associated with the life item using the following rule.

$$\forall\, \phi, \kappa, t\ (\mathit{LifeItem}(\phi) \wedge \mathit{LabelContext}(\phi, \kappa) \wedge \mathit{Subject}(\kappa, t)) \Rightarrow \mathit{Subject}(\phi, t) \tag{4.18}$$

where $\mathit{LabelContext}(\phi, \kappa)$ is read as ϕ has a context κ for the label, and $\mathit{Subject}(\kappa, t)$ as κ is annotated with subject t – a SKOS concept/term.

[5] http://www.holygoat.co.uk/owl/redwood/0.1/tags/

[6] http://tomgruber.org/writing/tagontology.htm

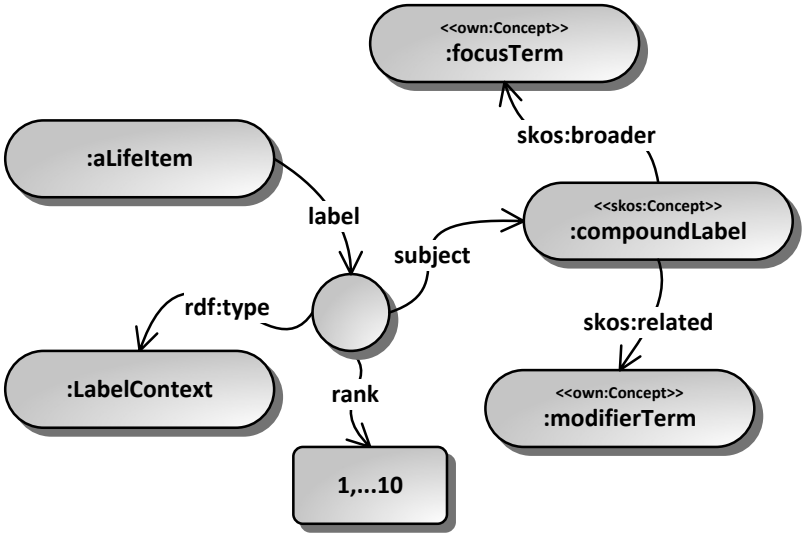


Figure 4.12: Abstract model of weighted semantic labels.

Compound Labels

Traditionally keyword based text indexing methodologies reckon on single terms for representing document vectors. The statistical analysis usually overlook the semantics behind the keywords associations (Schank et al., 1981). The use of compound terms can solve different issues involving semantic similarity between the adjacent terms. For example, with the famous *copper coating on lead pipes* problem, the search on *copper pipes* and *lead coating* could be handled properly.

Compound terms have two parts: the head noun (*focus*) and the difference (*modifier*). The focus identifies the broader class of concepts to which the term as a whole refers (ANSI-NISO-Z39-19, 2005), for example coating in lead-coating. The modifier part refers to a characteristic which narrows the focus by specifying a subclass of the broader concept represented by the focus, for instance lead in lead coating. In our semantic label scheme the compound term is annotated as the narrower concept of the focus and being semantically related to the modifier. The model is depicted in Figure 4.12.

4.2.5 Semantic Similarity

The similarity among life items is measures along all four dimensions in a holistic manner. If ϕ_1 and ϕ_2 represent two life items and $\sigma_S(\phi_1, \phi_2)$, $\sigma_T(\phi_1, \phi_2)$, $\sigma_A(\phi_1, \phi_2)$, and $\sigma_L(\phi_1, \phi_2)$ their similarity along spatial-location, temporal-location, agent and activity, and content labels respectively, then

the aggregated similarity of the items is measured as follows:

$$\begin{aligned}(4.19)\qquad \sigma_{STeAL}(\phi_1,\phi_2) &= \lambda_S\,\sigma_S(\phi_1,\phi_2) + \\ &\quad \lambda_T\,\sigma_T(\phi_1,\phi_2) + \\ &\quad \lambda_A\,\sigma_A(\phi_1,\phi_2) + \\ &\quad \lambda_L\,\sigma_L(\phi_1,\phi_2)\end{aligned}$$

where $\lambda_S + \lambda_T + \lambda_A + \lambda_L = 1$ are salience weights for each context dimension. The choice of weights for a particular context dimension depends on the story you want to tell. In a specific application setting certain characteristics of the items may be considered more critical than others, such as *time sensitivity* in health care and the *communication agent* in personal and business communications. Each dimension permits a unique understanding of the personal experiences.

4.2.6 Analogy of STeAL Model to Named Entities

Most research endeavors in information retrieval are focused to ascertain meanings out of unstructured resources such as text documents. The Semantic Web promises to overcome some of the issues by explicitly (either manually or automatically) annotating resources with meta-data. Many practitioners have questioned the wide-spread adoption of semantic web initiative due to complexity and expressiveness of current triple-based semantic web languages. For example Rob McCool, in his series of articles in IEEE Internet Computing (McCool, 2005; McCool, 2006), criticized by saying that RDF and OWL are leading towards a "shadow web" in which semantic annotations are maintained in documents separate from the original resource mainly because translation from natural language to triples and vice versa is difficult.

In contrast, researchers are pushing for a "named entity web" (see Figure 4.13 for an example) where *entities* should be annotated within the original resource such as using Microformats (Khare, 2006). W3C has also indulged in a similar effort[7]. Emergence of RDFa (Adida and Birbeck, 2007) and GRDDL (Connolly, 2007) under W3C's umbrella is an evidence of the need of a consistent representation formalism. Standardization of the representation languages aside, automation of the named entity annotation process is very important. That is to say, discovering and annotating named entities on the fly.

[7]http://www.w3.org/News/2006#x20060714a

```
<div class="vevent">
  <span class="summary">Internet Semantic Web Conference</span>
  will be held from <abbr class="dtstart" title="2006-11-05">
  November 5</abbr> to <abbr class="dtend" title="2006-11-09">9
  </abbr> at the <span class="location">GA Center, Athens, GA</span>
</div>
```

```
<p><span class="ical:Vevent">
 <span class="ical:summary">Internet Semantic Web Conference</span>
 will be held from <span property="ical:dtstart" content="2006-11-05">
 November 5</span> to <span property="ical:dtend" content="2006-11-09">9
 </span> at the at the <span property="ical:location">GA Center, Athens, GA
</span>.</p>
```

Figure 4.13: Example of annotations in the named entity web using Microformat syntax (above) and RDFa (below).

Dimensions in STeAL model have certain similarities with generic named entities. The generic named entities are those common in most of the domains such as person, organization, location, or date and time. Some of the types of named entities are listed in the Table 4.2. Recognizing these named entities allow more complex text-mining tasks to be addressed (Cohen and Hersh, 2005) and lays the foundation for further extraction of relationships and other semantic information by identifying the key concepts of interest (Zhang et al., 2004). As a next step those concepts can be represented in some consistent formalism.

The need of recognizing named entities from personal information is evident as demonstrated by (Dumais et al., 2003). The work reports that "the most common query types in our logs were People/places/things... Their importance is highlighted by the fact that 25% of the queries involved people's names, suggesting that people are a powerful memory cue for personal content. In contrast, general information queries are less prevalent." Survey of semantic annotation solutions using named entity recognition are performed by (Reeve and Han, 2005; Sazedj and Pinto, 2005). A comparative analysis of applying different NER tools to the personal information items was carried out in our work and the results are discussed in Section 5.2.2. There we have also demonstrated how much automation of STeAL annotations is possible using current breed of generic named entity recognizers.

4.3 Collections and Associative Trails

Partitioning the knowledge box into collections is an essential feature for managing digital memories of lifetime. One benefit of this approach is that

Named Entities	Example(s)
Places	Country and city names
Persons	Person names, titles etc
Organizations	Company names, Institutes and other organizations
Date & Time	Different date/time labels
Identifiers	URIs, email address, file name mentions

Table 4.2: Generic named entities.

every item in the knowledge box can be archived and then located in relation with other items in the same collection. Further benefits include focused archival, efficient retrieval based on associations, and better handling of the lifetime repository. Collections are useful for the users to organize life-items while doing a specific task. And as demonstrated by (Gemmel et al., 2003) collections are a valuable tool for building trails.

On the one hand, individual items can live in multiple collections, and on the other hand, items with varying topics can be located within the same trail. For example, an email from the semantic web mailing list expressing a call for papers could be filed in both "semantic web" and "conference calls" collections. On the other hand, both an article on sight seeing tours in Beijing and an email for registration in a conference held in Beijing can be positioned close to each other on the location axis. But their topic is different from each other, characterized by annotating with appropriate semantic labels. Still, both artifacts can belong to the same collection (say conference participation).

4.3.1 Collections in Personal Desktops

Categories, hierarchies, or other kinds of classifications are used in existing desktop applications. The inter-relation among applications for sharing the members of the categories is missing for the most part (Ravasio et al., 2004). For example, the bookmarks items, emails, working documents, and presentations live in separate homes (see Figure 4.14). Thus an information item present in a category in one application has no explicit associations with its counter parts in other applications in the similar category, which is necessary for building trails and allowing humans to follow the association of thoughts to locate an information object. One possible solution to this problem is (1) using a shared conceptualization of information items, (2) exploiting the semantics using a context framework, and finally (3) annotating and linking the information items based on that ontology. In the previous sections we have explained the structured framework to capture information context

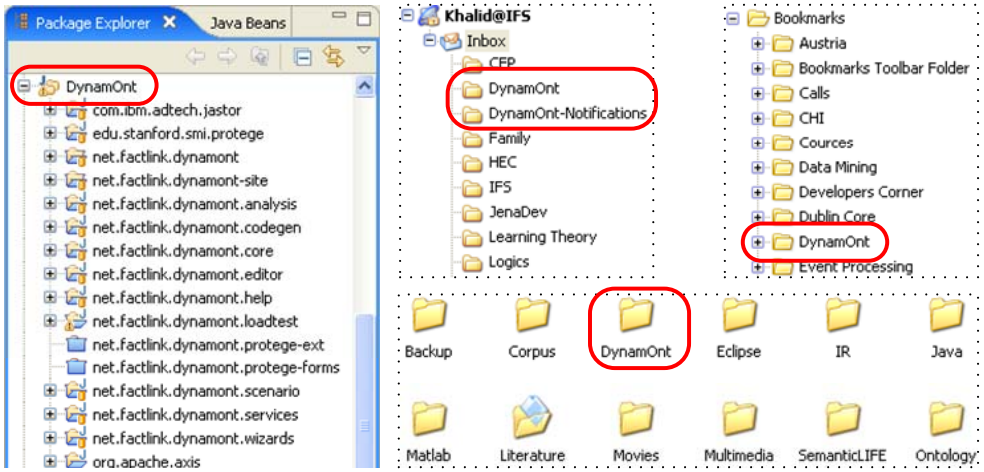


Figure 4.14: Isolated collections of information as depicted in development workbench (left), IMAP folders (top center), bookmarks (top right) and file-system directories (bottom); similar collections from different applications are highlighted.

which provides the foundation for integrating the items. In the subsequent section we will present the model for building collections and trails to link items sharing semantic similarity which, otherwise, were captured in isolated collections.

4.3.2 Modeling Trails

A trail Θ consists of

- an explicit or implicit membership criteria that is the characterization of the context. The explicit criteria is asserted by following the STeAL model and implicit (tacit) criteria may be extracted from the context information of member items as a common denominator.

- a forward navigation function η_f. Given an item ϕ_a the function η_f can map to the next item ϕ_b in the navigation queue, i.e. $\eta_f : \phi_a \rightarrow \phi_b$.

- a backward navigation function η_b. Given an item ϕ_b the function η_b can map to the previous item ϕ_a in the navigation queue. The navigation functions, in a way, control the ordering of the items and are an important aspect of the storytelling and building trails. From the pragmatic point of view, the navigation function could be envisioned as the *semantic* linked lists where having one item in hand it is possible

to navigate back and forth depending on linear or on-linear nature of the ordering adopted for a particular trail.

Trails are divided into following types based on the membership criteria:

Silent trails are essentially arbitrary *user collections* with tacit membership criteria. Each item in the silent trail is explicitly listed in the collection. User collections are distinct from other trails because they can contain variety of life-items without following any restriction.

Live trails, on the other hand, have explicit membership criteria. Such a trail could be annotated to optionally index the members in which case items fed to the system are automatically filed into the trail if they fulfill the criteria. In contrast, non-indexing live trails are similar to database views. An example of the later is a temporally ordered trail of all items labeled with "semantic web." Member items are retrieved from the knowledge box based on the user query and the membership criteria.

4.3.3 Dynamic Semantic Links

In the physical world, entities are usually interconnected, either by physical or by semantic means; in the latter case, the semantic meaning is added by human interaction (in an abstract sense) with the physical world. Digital memories can be understood as digital information entities and, in most cases, they are representations of such physical entities. They are connected to other life-items according to their semantic meaning. There are a number of questions one can ask about how to construct such conceptual associations.

In the section about semantic labels we explained that abstract labels can be associated with the life-items. Such labels are mapped to the axiomatic space of the foundational ontology. Furthermore, the other three dimensions of the STeAL model expose lightweight semantics about the item. Now the associations can be built by making comparisons between semantic similarity among the items.

For example, consider that *Alex* fed an event information from his personal calendar about a workshop to be held on 12-Nov-2006 in Salzburg. Now if Alex fed an image taken on 12th November 2006, most probably (even if not annotated directly) both items have an association – a weighted connection between them. The weight describes the strength and quality of the association and is calculated using previously described semantic similarity measures. It is important to mention here that weak links (with low weights)

may get stronger either by manual annotations or through new items which eventually confirms the relationship established between the previous items. Thus personal knowledge box is, in a way, *dynamic*, as it develops and modifies permanently during the system and user lifetime.

4.3.4 Effective Organization

Living systems have different characteristics such as *reproduction*, *growth*, and *self-regulation of processes* (Nicolau, 1995). For a personal knowledge box, it is impossible beforehand to anticipate all categories. Analogous to living systems, the personal knowledge box should be able to effectively reorganize its categories; creating sub-categories if one gets larger. These characteristics could be envisioned in a semantic way. The categorization of ones digital memories such as documents, persons, places, organizations, events and tasks, could be partitioned into collections which grow and reproduce new classifications over time.

Self similarity of a collection is the average pair-wise similarity between its members (Chakrabarti, 2003, p.85). If Θ is the trail of items and $\mid \Theta \mid$ denotes the count of its members then its self similarity is calculated as follows:

$$\sigma(\Theta) = \frac{2}{\mid \Theta \mid (\mid \Theta \mid -1)} \sum_{\phi_i, \phi_j \in \Theta, i \neq j} \sigma(\phi_i, \phi_j) \tag{4.20}$$

Initiated by the user, the trail is divided into two new collections iff $\sigma(\Theta) \leq w_r$ where w_r is the user defined re-classification factor.

4.3.5 Information Landmarks

Humans make use of variety of practices for recollection. *Method of loci* (also known as *mnemonics*) is one example of such practices originated with the ancient Greeks. The idea is to relate parts of the information to well-known landmarks. Recent example of its use is in rescue operations after earthquake, during 2005, in northern areas of south Asia where American pilots had difficulty to remember South Asian city names. For efficient communication they virtually named the affected cities (Balakot, Bagh etc.) after city names in USA. Use of landmark events is also investigated for personal information space by (Ringel et al., 2003).

The notion of landmarks is also used in graph drawing of co-citation networks (Chen, 2004, p.285-287) for highlighting the importance of a node,

such as a highly cited article. In hypertext systems, the opening web page is considered a landmark and every other web page in that particular web application is linked with it (Sorrows, 2004). While the first two examples show random associations the later are more consistent and logical. For our work it is not important if a landmark is used as mnemonic or as a reference point, more crucial aspect is building the trails by linking together different items. Though we focus on modeling the context and landmarks for building logical associations, users are not impeded in manually constructing random ones.

Continuum Organization

In continuum organization, characteristics of an information item are identified relative to a significant event, or state of the same or other item. Declaring that a picture, for example, was taken after few days of a momentous event entails reconciliation of subjective and objective views. A story from the selected pictures could easily be created provided the information space is organized in such a way. For the case of personal information any object or its significant state can be associated with other objects and thus creating a cognitive map of the life-items. Items that share meaning or physical similarity with the landmark become associated with it and selection of the landmark activates other linked items, and vice versa. The degree of activation depends on the strength of association.

In contrast to public and personal landmarks on time axis as proposed in (Ringel et al., 2003) we argue that: landmarks could be located in various axis not only in time, and the significance of landmarks is better remembrance of items so they should be all personal. By stating so we do not negate that a momentous news story can be a landmark. The point is that a landmark should be induced by the user and not by the system. It should emerge from the user's life-items and shouldn't be an external item which might be of least interest to the user, so it is personal in that sense.

Landmarks Model

For declaring a life-item as landmark user simply selects it and assigns a non-negative weight value w_L. The landmark weight follows a certain scale characterized by the maximum allowed value w_m, by default set to 10. The weight w_L provides the attraction force and determines the strength[8] of the landmark in terms of semantic depth. The landmark weight also contributes

[8]The weight of a landmark photo is also used in determining the size of thumbnail in photo collection view and also in search results view.

in determining the nearness of one landmark with other items. Based on the value of weight w it is decided if an item is linked and could be followed through the landmark even if it is not directly linked with it.

An item ϕ is said to be semantically associated and with a landmark ϕ_L having weight w_L if ϕ is in its neighborhood measured as following[9]:

$$N(\phi_L) = \{\phi_i \mid \forall\, i\ \sigma_{STeAL}(\phi_L, \phi_i) \geq w_n\} \tag{4.21}$$

where w_n is normalized weight of the landmark calculated as $w_n = \frac{w_L}{w_m}$. We have implemented different inference rules and employ different semantic distance measures that effect the semantic similarity of the items. For example, the following conditions contribute to the semantic nearness of a landmark item ϕ_L having weight w_L with item ϕ. The similarity increases if any or all of the following hold true:

- ϕ_L and ϕ are in the same collection.
- ϕ_L and ϕ are directly connected as related items through manual linking.
- Both ϕ_L and ϕ are annotated with the same concept C from the ontology.
- ϕ_L has an annotation of concept type c_1; ϕ has an annotation of concept type c_2, and $SemanticDistance(c_1, c_2) < w_n$. The semantic distance between concepts is computed in several ways such as the manual associations, property-entity associations (Aleman-Meza et al., 2005), topic similarity (Equation 2.3), and hierarchical concept distance (Equation 2.1). More detail of calculating semantic distance in nearness discovery of landmarks is presented in our previous work (Latif and Tjoa, 2006).

Additionally, every rule gets a weight and the accumulated score is used to rank similar photos for getting the k most relevant items. The photo viewer uses these rules to find landmark photos near the currently selected photo. User can set a threshold value (default to 4) for the number of relevant photos to show in the photo viewer. The priority is given to the photos with higher landmark weight. While viewing one photo from a collection the user is provided with photos which are semantically near the photo from other collections. Thus the whole photo collection turned out to be a web of trails.

[9]The neighborhood measure is adopted from (Rodríguez and Egenhofer, 2003).

For manually associating one or more information items with a landmark user drags and drop those in an item list widget (see Section 5.3 for more details on interface issues). The desired landmark is later on selected and finally the user commits to establish the association. Comments, both free text or using a category hierarchy, could be attached with the association. In principle, the process of associating items with landmark could be applied to connect an information item with any other information item.

Perspective Customization

Landmarks guarantee efficient retrieval in large information space by exploiting associations and by guiding the user in exploring the large information space. Driven by the vision about trails, the Knowledge Box allows the user to select any path. For example, the user can select *agent* axis as starting point, and Pakistan as the figure. This will get him/her information items related with Pakistan such as the news story "Austria helped Pakistan with water processing plant in earthquake rescue operations" as the most recent landmark item associated with the concept Pakistan in the context of agent. Noticeably Pakistan and Austria both are instance of *GeoPoliticalEntity* which in this scenario means a *NonAgentiveFigure* acted by an *Agent*. Selecting the news story will present the user with all information items associated with the landmark such as the news stories of the earthquake, photos of the scenes and the fact that a fund raising lunch was arranged in the United Nations headquarter in Vienna.

Now the user can look for the life-items on the location axis by zooming in to Vienna. Items with fine-grained locations of Vienna will also be presented to the user such as the collection "Talks at institute IFS." The only limit remains imagination as the user can choose the time axis to view items close to a specific talk on timeline axis.

The enormity of the lifetime information poses a serious challenge in terms of comprehensibility. The structured context framework for semantic annotations proposed here not only brings together information extracted from diverse media types into an integrated model but also improves the comprehensibility by enhancing semantic content insight of the life items. The context framework is further augmented with collections, trails, and landmarks for achieving associative information exploration. In the next chapter we have elaborated the software implementation issues for realizing this vision.

Chapter 5

Implementation and Results

Web services provide a systematic and extensible framework for application-to-application interaction. Services allow automatic and dynamic interoperability between software systems. However, the implementation and effective use of Web services is not yet fully explored for the personal information management and desktop applications. The process of assembling "pieces of functionality" into complex processes is often thinkable just for big enterprises, and more recently for news syndication. For ordinary computer users, there is no easy way to interact with the Web service ecosystem.

In this chapter we present a service-oriented architecture for the personal knowledge box. We also introduce a software architecture pattern for semantic enhancements and give a detailed account of using the services and the architecture pattern in the reference implementation. The chapter is structured as follows: First of all we discuss the semantics enhancement architecture pattern and then present service oriented pipelines which are lightweight implementation of service-oriented architecture and service orchestration. The core functionality of the framework is based upon semantic data enrichment and subsequent ontological storage. Finally, we describe a novel technique for the visualization of lifetime information using *Night Sky* metaphor. The night sky visualization facilitates better understanding of the underlying data by exploiting the *overview & details-on-demand* interaction technique.

5.1 Semantics Enhancement Architecture

Design patterns are common solutions to recurring problem (Gamma et al., 1994, p.2-4). They represent best practices exercised by the community in a certain situation to solve a particular problem. Architectural software

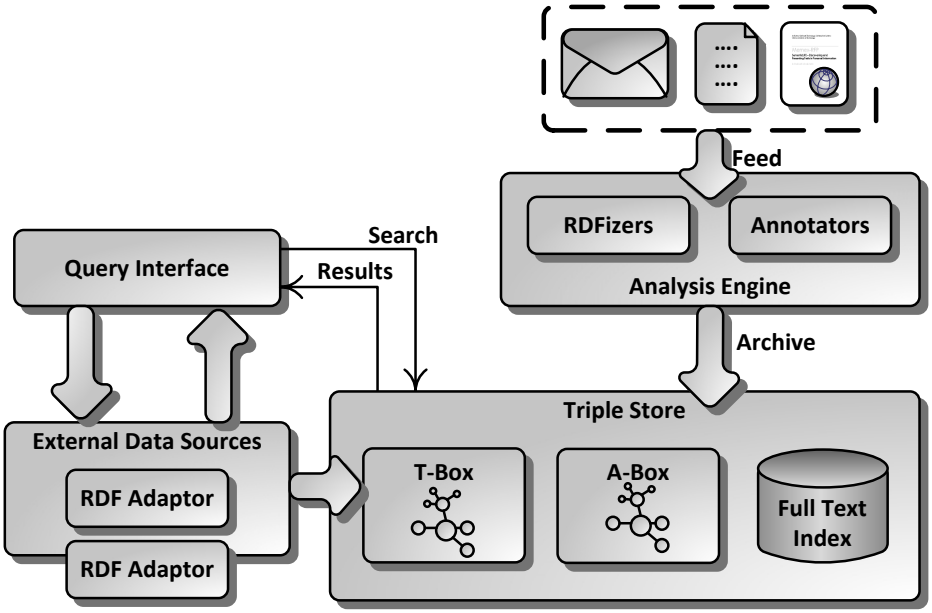


Figure 5.1: Semantics enhancement architecture for the knowledge box

design patterns are special kind of design patterns, and their reuse allows easy refactoring and customization of the software systems (Booch, 2007).

The primary goal of this research is to build a personal information management system by exploiting the semantics of the information and thus realizing the vision of trails and association of thoughts. For the Personal Knowledge Box we envision five principle functionalities: *Capture*, *Process*, *Archive*, *Adapt*, and *Interact*. These aspects are combined in the Semantics Enhancement Architecture Pattern (see Figure 5.1) using a service-oriented strategy.

The architecture of the knowledge box has a highly modular structure. It relies on plug-in mechanism in order to guarantee flexibility and extensibility. Communication within the system is based on a message-oriented design. This has the advantage of loose coupling, i.e. various modules (as described below) can be easily connected and controlled using a central message queue.

5.1.1 Services and Pipelines

Service-oriented architectures have three prominent components including providers, consumers, and registries (Huhns and Singh, 2005). Providers expose pieces of functionalities as services. The services could be combined using service composition standards (such as BPEL) to perform a bigger task involving multiple operations (Pasley, 2005). Consequently, services and

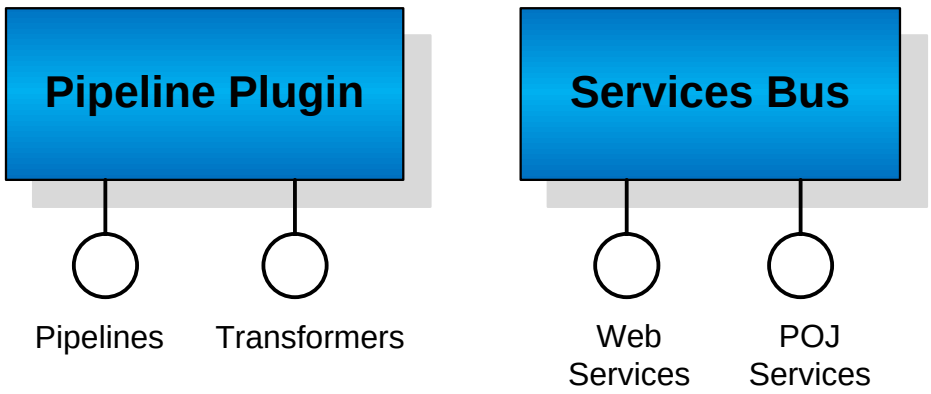


Figure 5.2: Components of service-oriented pipeline architecture.

pipelines (orchestration of services) are two basic elements of the knowledge box[1]. The services S could be GUI services S_v, internal analytic processors S_p, or external web-services S_w. GUI services extend user interaction and confirmation support through extension points where service provider can also provide a visual service for visual analytics.

Pipelines P_i orchestrate different services and apply transformation T_i to render the results back to the user or to another pipeline i.e.

$$P = \{S_0 \dots S_n \in S, T\} \tag{5.1}$$

Services and pipelines are managed by two plug-ins: *Service Bus* and *Pipeline Manager*. With extension point mechanism of Eclipse (Bolour, 2003), these plug-ins can expose extension points where other services and pipelines can be connected (see Figure 5.2).

Services Bus

The Service Bus can be seen as the door to the knowledge box. It is responsible for routing and monitoring message traffic, adding time-stamps to messages, and logging system states. This allows analyzing the behavior of the system in case of problems. Moreover, the usage of a message oriented

[1]Service-Oriented Pipeline Architecture is a lightweight implementation of Service-Oriented Framework for combining Rich Client components and business processing components for richer desktop integration. The earlier version of the service-oriented pipeline architecture was presented at the JAX Innovation Award 2006 and was selected among top 10 nominations. Other notable nominations were Spring Framework and Rich Ajax Platform: http://jax-award.de/jax_award06/nominierungen_en.php

design provides means for future enhancements to guarantee scalability and flexibility.

Standard Java classes can be published as Web services by utilizing the services extension point offered by the Service Bus. Automatic deployment is done using embedded Jetty and Apache AXIS. Additionally, Service Bus is responsible for routing the service call requests to the actual connected service. Thus it provides a uniform access layer and transparency to internal and external services. The integration with other applications on the desktop, and external feeding modules is also supported through the Web service interface. Binary contents of the items are encoded to Base64 for keeping them safe from validating and parsing errors.

Pipeline Manager

Pipelines are uniquely named set of service-calls and intermediate transformations. The idea of pipelines in our framework is inspired by Apache Cocoon[2] which is built around the concept of *separation of concerns* and component based web development. Pipelines provide a unique view over the available local and remote services and are useful for realizing scenarios by combining services.

Pipelines have three core elements and are serialized in XML structure. The `pipeline` element describes the basic pipeline meta-data such as its name and required parameters. A series of `call` elements are used to call services. And finally `transform` element is used to mention the required XSLT transformation which has to be applied on the results to prepare it as an input to a visualization.

The workbench supports interaction with the existing tools specifically tailored for different knowledge visualizations. The interoperability measures such as transforming output of a pipeline to be used as input in a visualization were carried out based on the supported data models of the tools. Thus harmonizing the tools and philosophy behind them to deal with the complexity of personal knowledge management is one of the crucial job of the transformers.

A concrete example of a pipeline is presented below which combines results of two service calls and transforms the results to prepare a timeline based data model.

```
<pipeline name="at.slife.search.label-dt">
  <parameters>
    <parameter name="label" type="string"/>
```

[2]http://cocoon.apache.org

```
    <parameter name="dtstart" type="dateTime"/>
    <parameter name="dtend" type="dateTime"/>
  </parameters>
  <intersection>
    <call service="at.slife.search" operation="labelled">
      <parameter>{label}</parameter>
    </call>
    <call service="at.slife.search" operation="dtimed">
      <parameter>{dtstart}</parameter>
      <parameter>{dtend}</parameter>
    </call>
  </intersection>
  <transform xsl="timeline.xsl"/>
</pipeline>
```

Conditional Calls The results of the service call are maintained in an XML element identified by the `id` of the call. Any subsequent operation can reuse the results as parameters through XPath statements. Additionally, the structure of pipelines allows to call services based on some defined condition. For this purpose the following XSL commands are permitted within a pipeline:

- xsl:for-each – Loops through each node for repeated invocation
- xsl:value-of – Extracts the value of a selected node which could makeup a parameter for a service call
- xsl:if – Conditional invocation based on XPath boolean expression testing
- xsl:choose, xsl:when and xsl:otherwise

The following pipeline structure depicts the use of `xpath` expression and `xsl:for-each` statement to iteratively process the result of a previous operation.

```
<xsl:for-each select="/results/item">
  <call service="at.slife.profiler" operation="rank">
    <parameter>{query}</parameter>
    <parameter>{xpath:/item/title}</parameter>
    <parameter>{xpath:/item/dtstart}</parameter>
    <parameter>{xpath:/item/dtend}</parameter>
    ...
```

Call Embedding It is also possible to make nested calls to the services; i.e. the services may be chained together to exchange the parameters and results. The following example depicts such situation:

```
...
  <call service="com.example.currency"
        operation="convert">
    <parameter name="amount">{amount_dollar}</parameter>
    <parameter name="rate">
      <call service="com.myforex"
            operation="exchangeRate" returns="double"/>
        <parameter name="from">USD</parameter>
        <parameter name="to">EUR</parameter>
      </call>
    </parameter>
  </call>
  <!-- now do the transfer -->
  ...
```

Distributed Calls There are two aspects of the heterogeneity in personal information. Firstly, users tend to keep information at multiple places such as personal laptop, office desktop, and mobile phone. Secondly, the information may be distributed across the peers such as the colleagues working on the same project. For that reason, the pipelines allow to mention calls to distributed services identified by the host address in the service name. Accordingly the request is transferred to the Service Bus of the target machine and after due authentication the results are transmitted back to the originating pipeline.

```
...
<call service="at.slife.search@168.168.168.10"
...
```

5.1.2 Desktop Integration

The workbench supports integration with the existing applications on the desktop. The coherency between the parts of the tools being used in the workbench is achieved by transforming their input/output to fit the needs of the others. The workbench sits in between the tools and provide new context for already existing concepts and methods by dictating its service-oriented

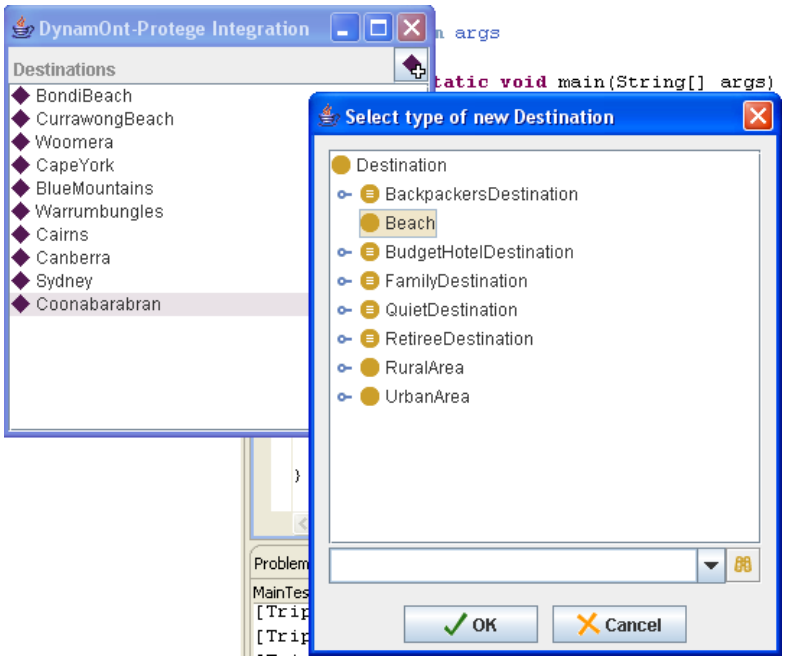


Figure 5.3: A custom Protégé instance form.

methodology. Thus the workbench is a mean to automate, speed-up, and reduce the cost of prior personal information management.

Tools like MS Word and Internet Explorer were embedded into the workbench using ActiveX controls following the multi-tool plug-in model. Such contribution is mainly on the user interface for better user interaction. Protégé is a prominent ontology editor. It can be reused in different levels and settings. It could be used as an ontology editor, an ontology management engine, or even as a collection of pre-fabricated ontology visualization widgets (c.f. Figure 5.3). We have integrated Protégé with the workbench and benefit from it as being an ontology editor and also by reusing its forms.

5.2 Implementation Details

The core of the system is based on its analysis and metadata extraction capabilities. New data sources emerge by the time and need to be treated by the system. It would become a time consuming task to add support for newer data sources if the system had tight coupling with its components. A light weight messaging based solution is required where new modules can be plugged into the system. Furthermore, openness is an extremely important issue considering systems that are designed for a lifetime acquisition of

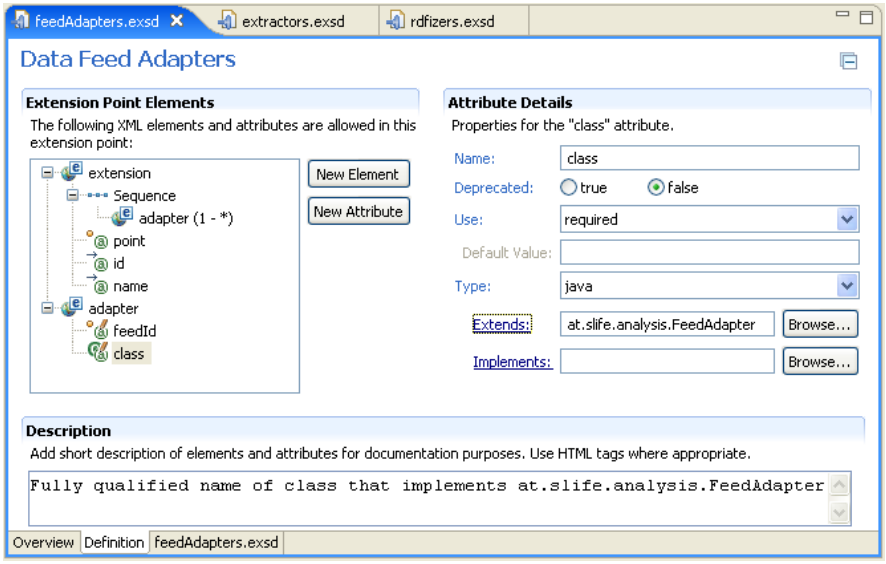


Figure 5.4: Extension point schema for feed adaptors.

data and metadata (McBride, 2002a). Hence the reference implementation is made public with an Open Source license which will allow various groups of developers and users to enhance and maintain the system.

The current implementation uses extension-point mechanism supported by Eclipse Rich Client Platform (RCP). It exposes following core extension points for 1) Triple Store, 2) Similarity Search Plug-in, 3) Feed Adaptors, 4) Content Analyzers, 5) External Data Source Adaptors, and 6) Item Editors for visualization. Figure 5.4 represents extension point schema for feed adaptor plug-ins. Details of individual modules is presented in subsequent sections.

5.2.1 Continuous Acquisition and Archival

Data that are fed into the knowledge box come from various sources and in various formats. So the acquisition module must be able to handle various types of data. Instead of constructing a "central" acquisition module with multi-source support, we developed several independent modules, each handling a specific data source. The range of data sources starts from communication data (emails, phone calls, chat sessions) to personal documents, pictures, web-browsing sessions and calendar data, and may include a whole range of additional sources up to sensory data (e.g. temperature, geographic location, blood pressure). The capture-store process consists of the following steps:

```
/**
 * Asynchronously stores the item contents and given metadata.
 *
 * @param feedType Type id of the feed such as <b>at.slife.feed.email</b>
 * @param contents The original item contents encoded in Base64
 * @param contentType Optional MIME type of the contents such as text/html.
 * @param metadata (Semi-)structured metadata about the contents such as
 * referrer URL for web-pages and IMAP folder name for emails. The metadata
 * contents should be a valid XML document and its root element should be
 * <b>metadata</b>. For example metadata for a web page might be:<br>
 * <pre>
 * <metadata>
 *   <url>http://www.domain.com/example.html</url>
 *   <referrer>http://www.referrer.com/link.html</referrer>
 *   <visited>2005-05-05 05:05:05</visited>
 * </metadata>
 * </pre>
 */
public void feed(String feedType, String contents, String contentType, String metadata) {
    // Create unique item ID
    final String itemId = "sl-"+System.currentTimeMillis();
    // Archive original contents
    archiveItem(itemId, contents);
    // Invoke analysis plug-in to extract text contents
    String extracted = extractText(contentType, contents);
    // Index text contents using lucene
    indexItem(itemId, extracted);
    // RDFize item contents
    OntModel model = rdfizeItem(feedType, contents, contentType);
    // Combine metadata and extracted contents
    combineMetadata(model, metadata, feedType);
    // Finally store generated triples in the triple store
    storeItem(model);
}
```

Figure 5.5: Code listing of feed service excluding the details of authentication and multi-threading.

1. Capture digital memories and associated information.
2. Extract metadata from the contents and allow the user to enrich the item manually with semantic annotations based on STeAL model.
3. Store the item including semantic context in the triple store.
4. Allow the user to query the data with context-sensitive information.

The above steps are performed by `ItemFeedThread` which is initiated by the web-service interface of the Service Bus. Figure 5.5 lists the Java code of the routine excluding the exception handling details.

Two types of data acquisition can be distinguished, namely *automatic (scheduled) feeding* and *manual feeding*. Retrieving emails, monitoring user's application processes and web-browsing sessions are examples of the former, while manual upload of documents, audio or video sequences, and synchronizing calendar data with the system falls into the second category. Special focus is directed toward time information originating primarily from calendar data. The connections between calendar entries and other information items provide huge potential for conceptual exploration.

Privacy concerns become an issue as the system tries to capture as much information as possible in an automatic manner. To support a large degree of user control over the feeding process, a range of filtering mechanisms were implemented that allow to specify which data items should be forwarded into the system. Examples are time or domain based exclusion of certain web-browsing activities; feeding of email based on sender address or subject fields/keywords; or the differentiation between public and private calendar entries. Figure 5.6 shows the configuration settings for document change monitoring and user activity monitoring feeds.

5.2.2 Information Analysis

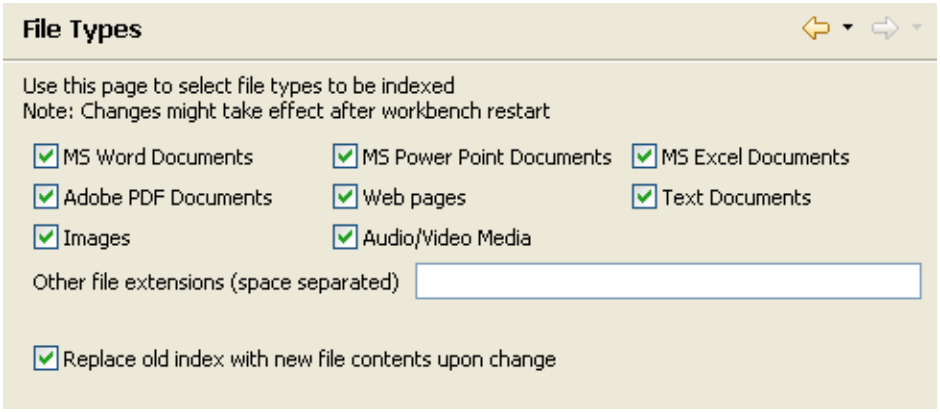
The analysis engine contains a number of specific analysis plug-ins providing semantic mark-up by applying a range of feature extraction and indexing techniques.

Cascaded Plug-ins

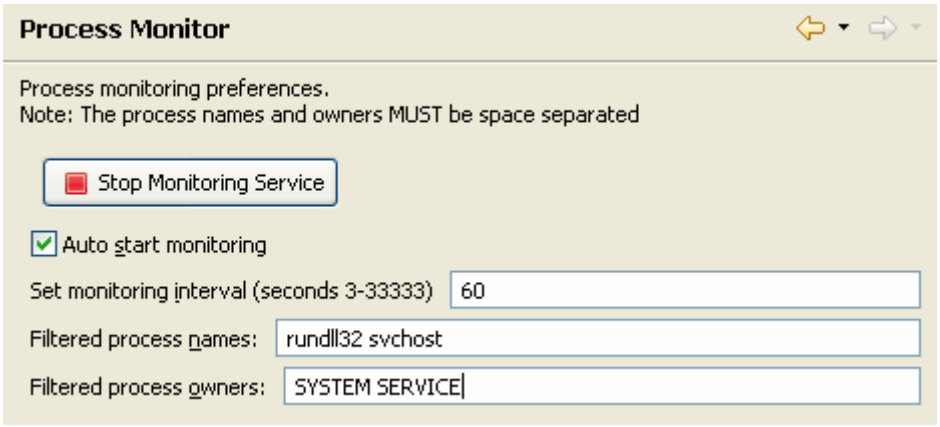
The incoming items may contain information in a nested manner e.g., ZIP archives comprising office documents, which in turn may include image or audio objects. To effectively deal with nested structuring of information a cascaded design is followed. This is basically a plug-in mechanism that allows adding various analysis modules to process items of certain types. All the items are analyzed in a nested way for extraction of metadata available in their content. Depending on the item type more than one analysis plug-in may be invoked for processing.

Content Analysis

The analysis modules primarily extract metadata and adds them to the life-items without changing the original contents. Parts of the functionality of text analysis modules in the current implementation is attributed to existing



(a) Automatic file system monitoring and indexing.



(b) Settings for activity monitoring.

Figure 5.6: Configuration for different feeding modules.

APIs including Apache Lucene, POI, and different named entity recognition components such as GATE (Cunningham et al., 2002; Maynard et al., 2002). Term extraction plug-in, for example, uses Lucene and different term weighting measures to obtain important terms from the textual documents (see Figure 5.7). Most of the text analysis components deal with syntactic analysis of the contents with no involvement of ontologies. However, during the analysis phase metadata is generated by mapping the results of syntactic analysis and named entity recognition to concepts in the ontology. For images, the current implementation analyzes EXIF headers and may be further extended in future to process, for example, color histograms, texture and contour and thus enabling automatic shape annotations.

Some life-items, such as email, involve originating network IP address as part of their structure. The IP address or the host name could be used

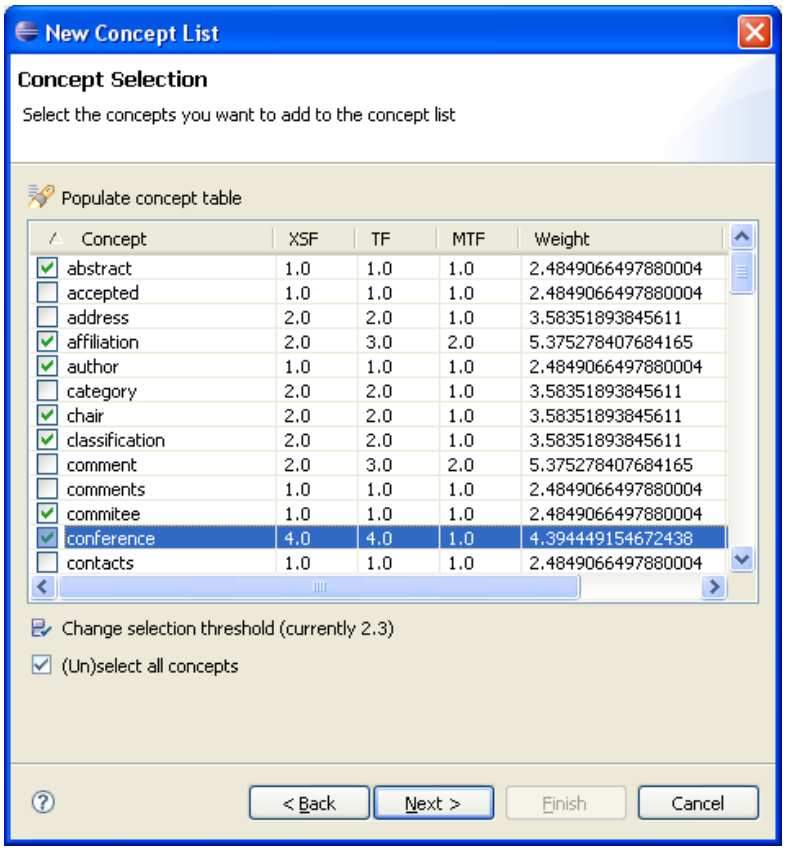


Figure 5.7: Result of term extraction component.

to lookup the associated geographical location. Different Geo-IP mapping services are available such as GeoIPTool[3], WebsiteGoodies[4], and HostIP[5]. The later, HostIP, was used intensively to help the user in locating the geographic address of the sender of email. It is important to highlight that such as lookup may reveal wrong location and therefore our algorithm solely depend on the target user to correct the location. For example, the following lookup[6] for the IP address `12.215.42.19` is not 100% correct in terms of longitude and latitude values but it provides a useful hint about the originating country and city.

[3]http://www.geoiptool.com
[4]http://www.websitegoodies.com/tools/geoip.php
[5]http://www.hostip.info
[6]An HTTP GET command was used for the lookup:
http://api.hostip.info/get_html.php?ip=12.215.42.19& position=true

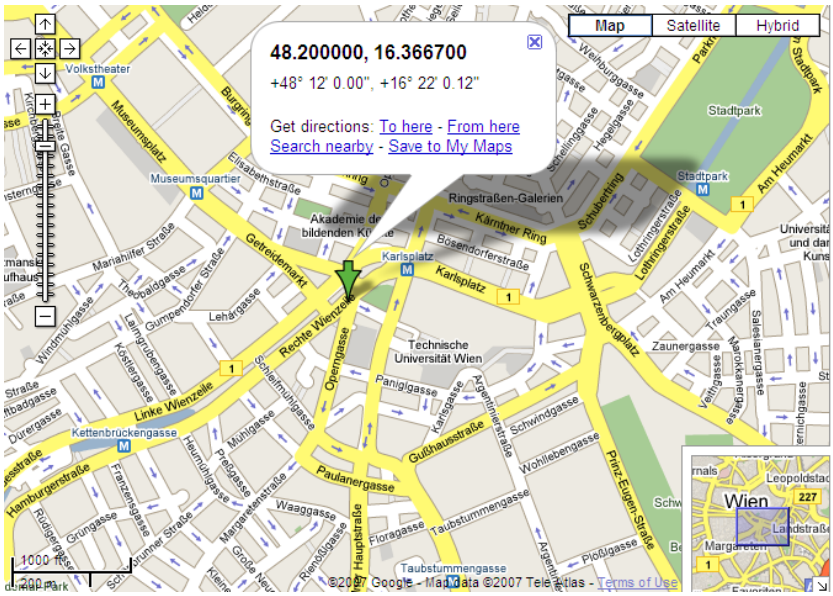


Figure 5.8: Result of the address lookup as shown by Google Maps.

Automatic Annotations

The goal of named entity recognition (NER) is to identify the instances of a name for a specific type of thing within a collection of text. NER could be categorized as domain specific or generic (domain independent). In first case, such as for bio-medical text corpus, NER could be regarded as recognition of drug names, chemical and biological symbols, or the gene names. The generic NER tools extract mentions of date, time, organization, places, etc (c.f. Figure 5.9).

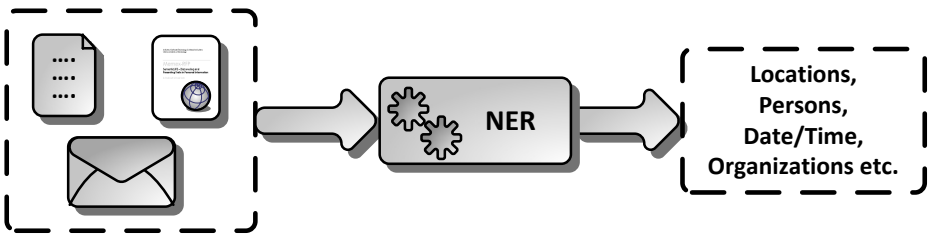


Figure 5.9: Named entities extraction process.

We used different NER solutions for our work. The tools were selected based on certain criteria explained comprehensively in (Latif and Rauber, 2006). The short listed tools are also presented in the table below.

Conventionally, effectiveness of information extraction systems is mea-

Tool	Thesaurus	Custom Rules	Relation Mining
ANNIE	Gazetteer lists	Yes	No
LingPipe	NA	-	Yes
OpenNLP	NA	-	No
MinorThird	NA	Yes	No
KIM	Uses gazetteer lists of ANNIE		Yes
UIMA	NA	Yes	Yes
Callisto	NA	No	Yes
ESpotter	Lexicon	Yes	-
Ellogon	Yes (Greek)	-	-

Table 5.1: Selected named entity recognition tools and overview of their features.

sured as precision and recall. The combination of precision and recall for NER is influenced by the following outcomes:

- *True positive*: There was mentioned of a named entity and was recognized successfully
- *True negative*: There was no mentioned of a named entity and was recognized so.
- *False positive*: There was no mentioned of a named entity but the system recognized one.
- *False negative*: There was mentioned of a named entity but the system failed to recognize.

The precision and recall measures show the performance of an NER tools and are computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

where TP is the count of *true positives*, FP is *false positives*, and FN is *false negatives*. The tradeoff between precision and recall can be adjusted by applying a parameter α in F-measure to weight either recall or precision.

$$F = \frac{(\alpha^2 + 1)P \times R}{\alpha^2 P + R} \tag{5.4}$$

Conventional precision and recall measures not necessarily are true representative of potential of the entity recognition algorithm. We observed that these four situations not necessarily cover all situations in named entity recognition. At least two other situations were identified in our experimentation with NER tools, known as labeling and boundary errors[7]:

- *False Label*: There was mentioned of a named entity and was recognized successfully but give it a wrong label.
- *False Boundary*: There was mentioned of a named entity and was recognized successfully but gets its boundary wrong.

By exploiting the boundary and labeling errors, we have introduced new precision measures. These are:

$$P_a = \frac{TP_a}{TP + FP} \tag{5.5}$$

where TP_a counts only those true positives not having boundary and labeling errors. This is most strict measure of precision where NER tool should not only get the boundaries right but also have to guess its label correctly. The next measure is:

$$P_b = \frac{TP_b}{TP + FP} \tag{5.6}$$

where TP_b counts only those true positives with *correct boundary*. This measure represent how accurately the NER tool can extract the entity regardless of its type which might be incorrect and ignored in this case. Another similar measure is for *correct labeling* regardless of boundary mistakes and is computed as follows:

$$P_c = \frac{TP_c}{TP + FP} \tag{5.7}$$

[7]Christopher Manning has also discussed the boundary and labeling error in detail in this weblog: http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

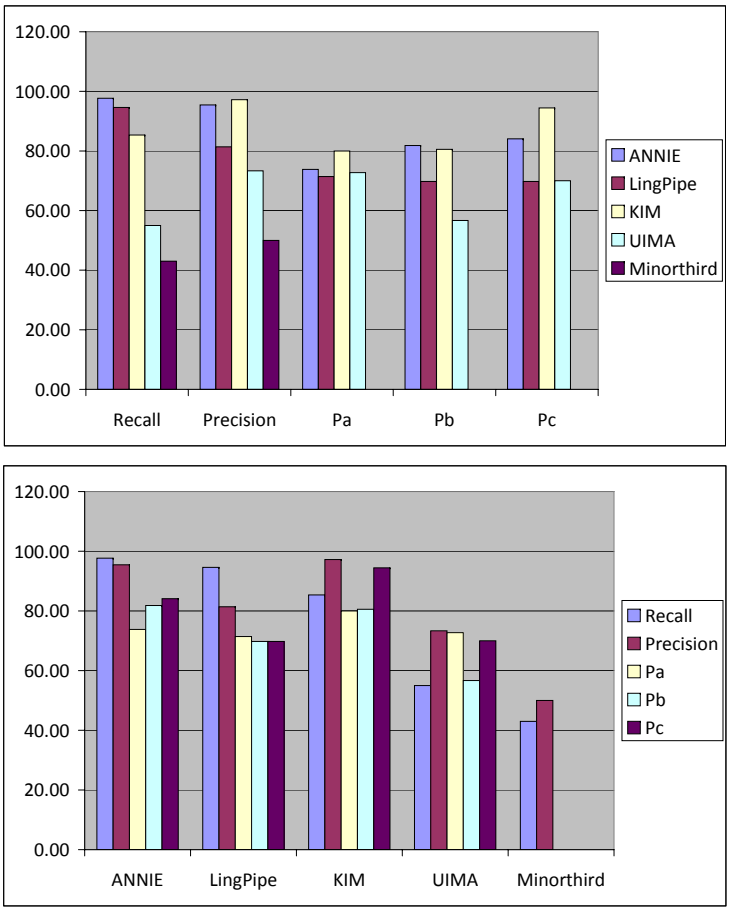


Figure 5.10: Precision and recall in named entity recognition.

Finally the short listed NER tools were evaluated on a small corpus. Figure 5.10 compares the precision and recall of different tools. The corpus was generated from different news stories, blogs, emails, chat logs, web pages and other text documents of varying length which is very typical of a researcher's daily log of life items. Almost half of the documents were in German and other half in English language. The documents were first manually annotated with 381 distinct entities by external subjects. In many cases a single entity occurred multiple times in the document and in different variants. For example, the person name "Ray Nagin" occurred multiple times in just one document. In some occurrences it was only mentioned as "Nagin." In another document the entity "IBM" occurred thrice as "IBM", "IBM Research", and "IBM Consulting" whereas it was only mentioned as "IBM" in the manual annotations.

The knowledge box of lifetime grows gradually with not more than tens

of items fed at a time. So we are not required to measure the efficiency for many thousands of documents. Initially, performance for each tool was monitored when applied to only one document. This gives the setup time for the tool needed to prepare its analytic processes for named entity recognition. Afterwards the process was repeated for 10, 40, and finally 80 selected documents which gives an approximation of its efficiency degradation when applied to a bigger corpus in other scenarios. Figure 5.11 presents the efficiency trend of the selected NER solutions. Details of evaluation setup are described in (Latif and Rauber, 2006).

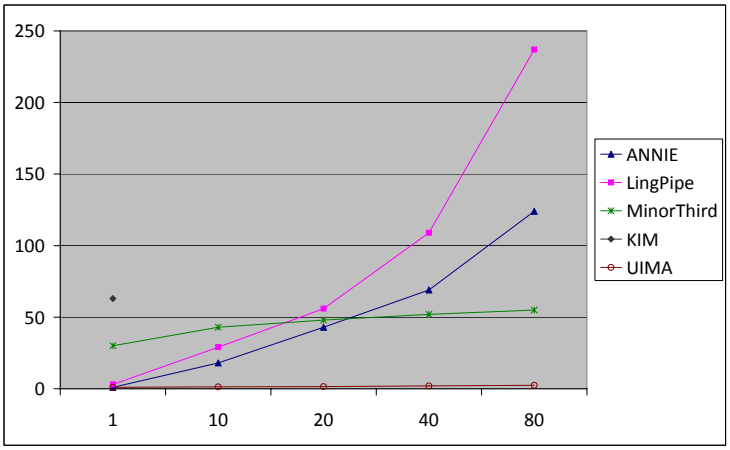


Figure 5.11: Efficiency trend of named entity recognition process.

Some NER tools involve supervised training of a statistical model. The training data must be labeled with all of the entities of interest and their types. Furthermore, it is important that the training data match the target data on which NER is applied. If a system is trained on a news-corpus and is then run on weblogs and emails, as is the case for our work, performance may degrade quite significantly as it will be tuned to the clues provided by typical news stories and will miss the clues provided in emails. Additionally, some NER solutions do not come with a default thesaurus. As the evaluation was performed on their baseline capabilities so the results might be different from their capabilities. For example, IBM's UIMA[8] was tested against built-in name annotator which is implemented using regular expressions. Even with this configuration it can achieve more than 73% precision and 55% recall which (without using gazetteers) is convincing.

Evaluating NER tools on a small bilingual (German & English) and multi-genre corpus, containing news stories, blogs, emails, chat logs, and web pages,

[8]http://www.research.ibm.com/UIMA/

proved to be a difficult task as some tools were trained for entity recognition against a specific genre and language. Evaluating their baseline capabilities against such a diverse corpus obviously resulted in varying statistics which were quite different from original claims of the tools. Among those, UIMA and LingPipe[9] were very convincing. For instance, LingPipe in most cases recognized the *topic-words* as entities but failed to assign the right label. UIMA, on the other hand, was good in recognizing person names from English documents but failed to do any good with German documents. Its regular expression based person annotator identified all the occurrences of persons for documents in German but with very bad precision. We believe that these tools, if trained for multilingual & multi-genre entity recognition, can perform much better than the current findings. The tools which used dictionary/gazetteer lookup performed better than others in general. Nonetheless the combined coverage of different tools (c.f. Table 5.2) provides the evidence to prove our hypothesis that named entity recognition can be used to automatically annotate life-items with STeAL to a reasonable extent.

From all of the foregoing, it is clear that named entity recognition contributes in realization of the personal knowledge box in particular through automatic annotations.

Manual Annotations

Currently, many research groups are working on automatic information extraction. Still, it is extremely difficult to bridge the so-called semantic gap, i.e. to extract (structured) descriptions of the content of a picture or a movie. Though we explored different strategies for ontology guided information extraction, manual annotations are also supported to enrich the contents. Human annotations to data items being fed into the system will ultimately improve its quality (Handschuh et al., 2002; Kahan and Koivunen, 2001). They act as a complement to the content analysis and automatic metadata extraction described earlier.

5.2.3 Inference and Metadata Management

Most of the life-items, in their raw form, lack explicit semantic annotations. The analysis engine enriches them with metadata extracted by the analysis plug-ins. The storage module goes one step further by putting the data into context using the available metadata. The items are assigned to ontological concepts and relationships between those items are established.

[9] www.aliasi.com/lingpipe

Manual Annotations		**Automatic Extraction**	
Entity	**Label**	**Entity**	**Label**
New Orleans	Location	New Orleans	City
Louisiana	Location	Louisiana	Province
CNN	Organization	CNN	Company
Ray Nagin	Person	*Mayor* Ray Nagin	Male
Terry Ebbert	Person	Terry Ebbert	Person
Hurricane	Topic	Hurricane	Unknown
		Katrina	Person
		Amtrak	Company
Louisiana Department of Health and Hospitals	Organization	Louisiana Department of Health	Organization
Mitch Landrieu	Person	*Lt.* Mitch Landrieu	Male
20-May	Time	May 20	Date
		May 23	Date
Homeland Security Department	Organization	Homeland Security Department	Organization
evacuation	Topic	evacuation	Name
airline	Topic	airlines	Organization
		Transport Security Administration	Organization
		Convention Center	Organization
		St. Rita Nursing Home	Organization

Table 5.2: Comparison of manually annotated entities and those recognized by NER solutions from one document.

Such relationships can be built automatically or with human intervention leading to the concept of weak and strong links, respectively. The weak link creation is carried out periodically or triggered by certain events. For example, email objects can be related based on the sender and (groups of) recipient addresses, or by the subjects. On the one hand names, project acronyms or locations can be used to connect data items from different sources on a semantic level (Lei et al., 2006), and on the othe hand various types of analysis provide a time-line based context to each life item. Sometimes, however, relations based on some criteria given by the user are also needed to establish explicit connections with a much higher confidence than those automatically created by the system.

The final storage of the life items takes place in three steps: The metadata extracted by the analysis modules is stored in the triple store built on top of Jena and MySQL. The triple store is divided into A-Box (data) and T-Box (ontologies). Separating data from ontologies allows to provide different perspectives on the data. Inference rules are separately stored in rule files and are loaded into the inference model of the triple store during the initialization

phase. Description Logics reasoning is also achieved through generic rule-based reasoner of Jena. The full text index of the contents is stored within the same instance of MySQL but in a separate catalog. Finally, the original contents are stored in a table as blob. These are necessary step to create a system capable of handling the huge amounts of data accumulating over time, and specifically to capture, represent and process the myriad of semantic relationships evolving in it.

5.2.4 Search and Retrieval

In this section we will demonstrate how can queries that require sophisticated interpretation be handled by the knowledge box.

Free Text Queries

As the data is already stored semantically enriched and with the possibility to invoke external data sources on the fly, it is possible to provide more powerful "imprecise searches", that go far beyond "simple" full-text indices. Here, the term "imprecise" has two meanings: firstly, the generated queries are to satisfy fuzzily defined information needs. Secondly, the target of the query is specified but there is ambiguity in the query. Therefore, the system has to solve these problems during query generation, by exploring the system's database and ontology repository and then generating SPARQL queries.

We have constructed an index of all the class names from OntoWordNet, user labels, named entities, and also from concepts in the ontology of digital memories. The lookup operation is performed against every query term q_{t_i} to find if its mapping either with a concept or some mention of an instance exists. Failure in finding the mapping renders a message for the user. Otherwise, the query term q_{t_i} is mapped to its relevant concept c_i and the items related to the concept c_i are returned, ranked based on their semantic similarity with c_i. For instance, if the user is searching for "Keynote Gruber", the lookup operation reveales that "Gruber" denotes a person and "Keynote", on the other hand, refers to $KeynoteSpeech$. Consequently the search results are augmented with appropriate metada about speeches involving Tom Gruber. The semantic search process may also retrieve an item which necessarily doesn't mentions c_i but is somehow associated with it. For instance, any search for *conference* related items during 2006-10-05 to 2006-1105 might also return mentions of workshop and symposium, though ranked lower than those mentioning conference. The results are combined with the full-text index results for improved recall.

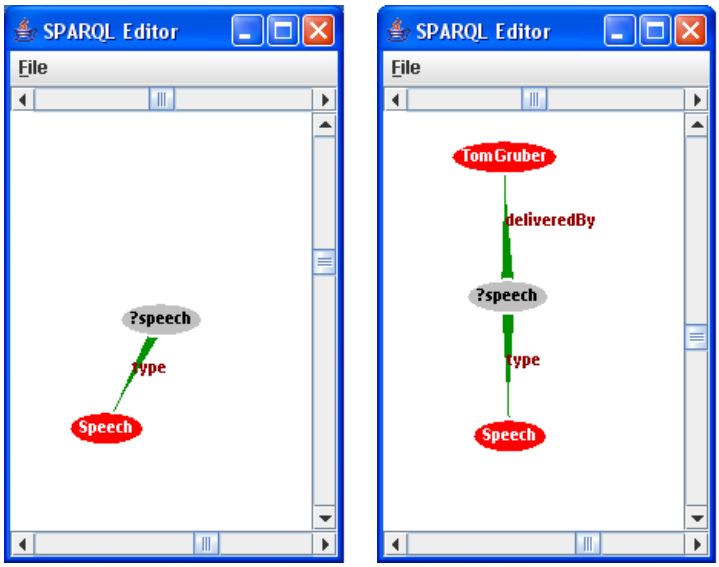


Figure 5.12: A session with graph based SPARQL query editor.

Structured Queries

Structured database systems like relational or object oriented databases usually provide query mechanisms that allow powerful queries on highly structured data. It is difficult for the common users to define highly structured SPARQL queries. Consequently, for ontology-based information seeking (Garcia and Sicilia, 2003), interactive retrieval sessions are a necessity which could be further enhanced by the structural organization offered by the STeAL model. We have developed a graph-based SPARQL query editor (c.f. Figure 5.12) where user can model the query in an interactive manner.

5.3 User Interaction and Navigation

Digital memories and the associated knowledge could be archived, semantically enhanced, and effectively retrieved but the next big challenge is how should people interact with them and the question of visualizing trails of the lifetime. The lifetime capture of digital memories results into a stockpile of information with hundred of thousands of items even after filtering. The semantics overload should also be reduced at the interface level, not only at the time of archiving. The *Towards 2020 Science* roadmap (Towards2020, 2006) also emphasizes on developing interfaces that ensure the usability and friendly navigation support for users (esp. scientists).

5.3.1 Managing Photos of Lifetime

Photo collections are one of the promising sources to tell story of life in this digital era. Organizing a huge photo collection of a lifetime requires effective use of the photo metadata. This metadata can be separated into two categories: the general photo characteristics and the photo contents. The first category provides information about photo resolution, format, size, etc. Such information is present in the EXIF header of digital photos and is easily extracted. Currently available personal photo management tools mostly exploit first type of metadata with support for unstructured labels and comments (Girgensohn et al., 2003; Schneiderman and Kang, 2000). The second category describes what is depicted by the photo. The contents of personal digital photo vary largely, and may include a wide range of domains such as sports, entertainment, and sightseeing. The content semantics although can be extracted for low level feature description, are manually annotated to lower the "semantic gap" (von Ahn, 2006; van Ossenbruggen et al., 2006).

Utility of semantic annotations for describing such diverse photo contents is well established (Bloehdorn et al., 2005; Hollink et al., 2003). But, improving comprehensibility in lifetime photo space and usability of Semantic Web technologies from the user interaction point of view is an ongoing endeavor. Common users mostly want easy access to their photo collections for viewing, using in their homepage, creating presentations, or sharing with other people. It is difficult to provide a unified way for annotating personal photos with arbitrary RDF. Even simple and otherwise trivial annotations are complex and hard to grasp for non-experts, regardless of any simplification in the visualization. On the other hand, a simplified annotation model can lead to pragmatic interfaces.

The hypothesis is that a structuring of annotation template, such as the STeAL model, on one hand provides adequate semantics to organize personal photo collections and on the other hand is easily comprehended by the user. The values for slots are filled from concepts in the existing ontologies. Compared to keyword search such semantic annotations allow concept-based searching where users can, among other features, specialize or generalize a query based on the concept hierarchy.

Annotations through Linking Items

The repository of knowledge box is fed with different desktop information such as calendar entries/appointments, web browsing cache, emails, and address book. The photo annotation is an integral part of the system, so it utilizes and reuses the existing information by far. Most of the personal

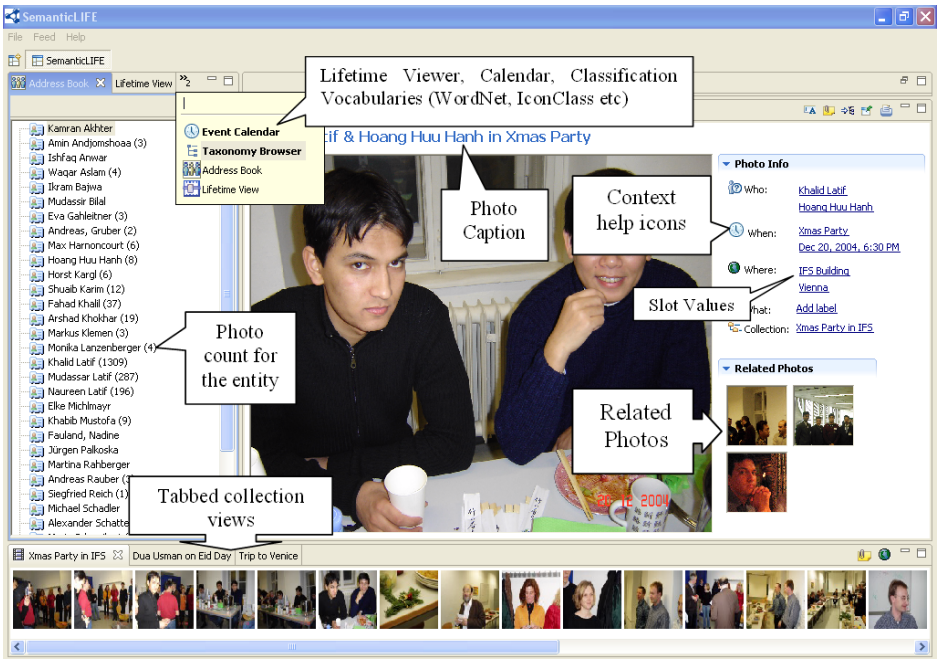


Figure 5.13: Overview of photo annotation interface.

photos come from planned events, such as birthday party or a conference. Information about such events (if present) is fed by Outlook and Sunbird adaptors, and is stored in the repository after appropriate transformation to RDF. Such existing items are an added help to the user in photo annotations. Items in the same date/time range are suggested to the user for their possible reuse during photo annotation. As explained in Section 5.2.2, we also apply named entity recognition to the recently fed items (such as web pages and emails) and clicking the *context help icon* in photo view (c.f. Figure 5.13) displays a list of relevant extracted entities. Any of these items/entities could be dragged and dropped on the photo or whole collection. Depending on the item type and its meta-data the appropriate slot is filled, thus users do not have to re-type.

Region Annotation

A specific region of the photo could be annotated by first selecting a rectangular area within a photo through the annotation marker and then associating it with the target concept from existing vocabulary(Latif et al., 2006). This annotates the image region with currently selected concept. The taxonomy browser loads the ontology vocabularies in a tree structure for this purpose. An abridged RDF listing of an annotation showing a rectangle within a photo

which depicts a concept *Gondola* from OntoWordNet is presented below:

```
@prefix reg: <http://www.w3.org/2004/02/image-regions#> .
@prefix own: <http://www.loa-cnr.it/ontologies/WordNet/OWN#> .

:s1004285
  a  foaf:Image ;
  reg:hasRegion :s1004285-r1 .

:s1004285-r1
  a  reg:Rectangle ;
  reg:coord  ... ;
  reg:regionDepicts :s1004285-g1 .

:s1004285-g1
  a  own:GONDOLA ;
  rdfs:label "Gondola" .
```

where as

$$hasRegion \sqsubseteq hasPart \tag{5.8}$$

$$regionDepicts \sqsubseteq depicts \sqsubseteq about \tag{5.9}$$

(5.10)

and

$$\forall X, Y, Z\ (hasPart(X, Y) \wedge depicts(Y, Z)) \Rightarrow depicts(X, Z) \tag{5.11}$$

The rectangular region is hidden in the photo view unless the target concept is selected from the photo information which highlights the region. Figure 5.14 shows a photo with region annotation.

Photo Collections

Collections can be created either manually by dragging and dropping the selecting photos or suggested by the system for un-sorted photos. The later task examines the EXIF header for possible match in date/time and other available characteristics. Similar to photo annotations, whole collections can also be tagged with ontology concepts or linked to other personal information

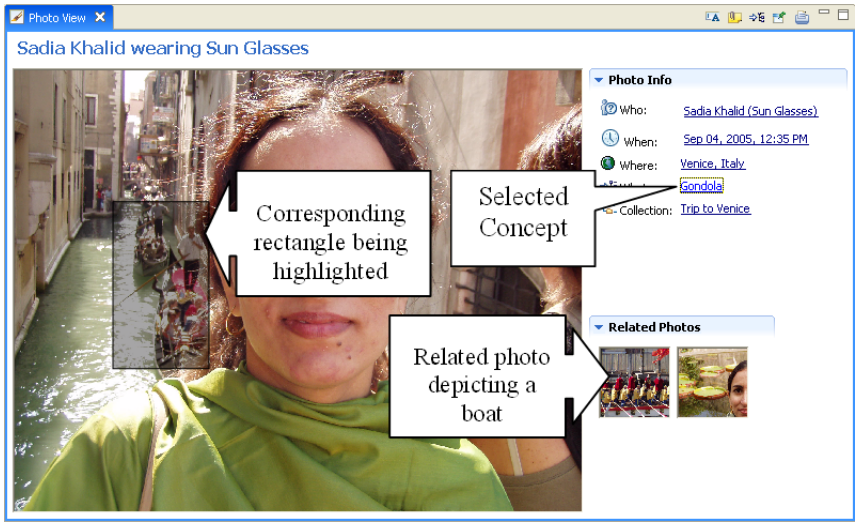


Figure 5.14: The photo viewer with concept and region highlighting support. The selection of concept *Gondola* has highlighted the associated region.

items such as an event from the calendar data. Associating the metadata with the collection replicates the semantics to all member photos. Moreover a collection can also become a part of another collection.

Rich Client Interfaces

Other than the taxonomy browser (classification view) which is an integral part of workbench, three views are provided to the user for navigation and annotation of photos: (1) lifetime photos view, (2) collection view, and (3) the photo view. In the lifetime view representative thumbnails of all categories are displayed along with their titles and event information. The thumbnails are generated from combination of the landmark photos in the collection. The collections in this view are sorted based on the timeline.

The collection view by default uses date/time for sorting the photos. Photos can also be sorted and filtered based on the concepts in the taxonomy hierarchy. A scattered plot mode with a background location map is also supported (see Figure 5.15). Users can freely place the photos on the map. The settings are preserved and could be seen anytime by selecting the location map mode in the collection view toolbar. All mentioned visualizations are implemented as JFace Views within Eclipse *Rich Client Platform* (RCP) that allows the user to arrange the views at the position of their choice.

Figure 5.15: Photos arranged on a map by the user. Magnification of a photo thumbnail depends on its landmark weight.

5.3.2 Sky of Lifetime Knowledge

The Self-Organizing Map (SOM) is a prominent tool for data mining and knowledge management. It is an unsupervised neural network model that provides a mapping from a high-dimensional input space to a lower, often two-dimensional, output space (Kohonen, 1995). An important property of this mapping is that it is topology preserving – elements which are located close to each other in the input space will also be closely located in the output space. Part of SOMs popularity can be attributed to the various visualization methods. SOM visualizations can utilize the output space as a platform (Vesanto, 1999), where quantitative information is most commonly depicted as color values or markers of different sizes. More advanced approaches use e.g. the analogy to geography (Skupin, 2004). These visualizations allow an easier interpretation of the cluster structures and correlations in the content by highlighting cluster boundaries and cumulations in the map (Pölzlbauer et al., 2005b; Pölzlbauer et al., 2005a; Pölzlbauer et al., 2006; Ultsch and Siemon, 1990; Ultsch, 1999). Thus the SOM, coupled with an effective visualization, summarize the characteristics of the data set and

Feature	Sky Metaphor
Individual Item	Star
Landmark (Prominent Item)	Guiding Star
Collection	Star Cluster and Galaxy
Associated Trail	Constellation

Table 5.3: Comparison between features of life-items in knowledge box and sky metaphor.

help the user in understanding and analyzing the underlying structure in the input data. The location of the input objects on the map allows the user to quickly identify similar and different objects.

However, the mapping of an input onto single map units is coarse and inaccurate to some extent in current visualizations. Depending on the resolution of the map, i.e. the number of units, inputs mapped onto the same unit might bear significant differences, which are not easy to transmit or visualize. Therefore, we propose a novel visualization technique that takes into account not only the best-matching unit of an input object, but also the *input's distances to the neighboring units*. As a result, the objects will not be placed at the center of the map unit, but drift toward some of the neighboring units. This helps the user on one hand to more easily distinguish between the items in the same unit, and on the other hand to grasp the similarities between data objects across unit boundaries.

In this section, we present the implementation details of Night-Sky visualization technique, which was first introduced in (Latif and Mayer, 2007). The map uses a black background to resemble the night sky. Individual objects from the input space are represented as stars, which together with other similar objects may form star clusters. This effect can be enhanced by using Smoothed Data Histograms (Pampalk et al., 2002) visualization on top of the background, resembling galaxies. Units that do not contain any inputs remain black and will resemble dark nebulae. Further, we make use of the notion of guiding stars and constellations.

The sky metaphor has been used, in parts, in other tools. Cloud of tags is a common visualization in the Web2.0 community (Hassan-Montero and Herrero-Solana, 2006). Such clouds usually depict the popularity in terms of the usage of the tags. A similar label cloud was also developed for the personal knowledge box where the size of the label represent the item frequency (IF), the number of life-items to which a label has been applied – analogous to document frequency (DF) in term frequency measure. The IN-SPIRE (Wong et al., 2004) tool builds on galaxy visualization by mak-

ing use of the metaphor of star clusters. Each star represents an individual document, and clusters around center points represent themes. The galaxy metaphor is also investigated in a prior work (Hetzler et al., 1998) to visualize document similarity. InfoSky also uses the notion of sky for visualizing documents. It is contingent on the assumption that documents are already organized in a hierarchy of collections (Andrews et al., 2002). The collections are rendered as Voronoi cells, and hierarchically related collections are placed alongside each other. In contrast, we use SOM for organizing the collections, and the night sky metaphor as a novel visualization above the output space of the SOM.

Different interaction strategies such as zooming & panning, individual document and area selection are not specific to the sky metaphor, but are supported by our SOM toolkit (Neumayer et al., 2005).

Initial Processing

The input space consists of collections of life items which are represented in the numerical form - e.g. a vector space bag-of-words representation of text documents, features extracted from audio or images, and user labels on the life items. The output space is organized as a rectangular grid of units, a representation that is easily understandable for users due to its analogy to 2-D maps. Each of the units on the map is assigned a *weight vector* $\mathbf{m}_i$, which is of the same dimensionality as the vectors $\mathbf{x}_i$ in the input space. During the training process, the vectors $\mathbf{x}_i$ are presented to the Self-Organizing Map, and the unit with the most similar weight vector to this input vector, the *best-matching unit*, is determined. The weight vector of this unit, and, to a lesser extent, of the neighboring units, are adapted towards the input vector, i.e. their distance in the input space is reduced – the output space 'folds' as closely as possible into the input space. After the training is finished, the inputs are mapped onto their ultimate best-matching unit. Some units might accumulate a lot of inputs, while others, probably located between clusters, may be left empty.

Star Clusters

Traditionally, the SOM algorithm assigns input objects only to a discrete map unit. We however want to reveal more details about the relations between the objects that are mapped onto the same unit, and also the similarities of the objects to other objects in neighboring units. Therefore, we propose to place the input objects not in the center of a unit cell, but spread them across the cell.

Neighborhood Forces

We calculate the exact location of an input $\mathbf{x}$ which is mapped onto its best-matching unit U. The distance from the input to the weight vector of U is denoted as d. Our assumption is that the location of the input $\mathbf{x}$ in unit U is driven by the position of the next closest units, with the distance of $\mathbf{x}$ to these units acting as a pull force to the input. More specifically, the pull force ($\mathbf{F}$) of a unit is *inversely proportional to the distance of the input from the unit* and is relative to the distance of the input to its best matching unit. This relationship is given in Equation 5.12.

$$\mathbf{F}_i \propto \frac{d(x, U_1)}{d(x, U_i)} \quad \text{for} \quad i > 1 \tag{5.12}$$

where d denotes the metric measuring the distance from the input to the weight vector of a unit.

As the second-best matching unit is nearer to the input than third-best matching unit, its pull force is higher in magnitude. For this reason, the displacement effect is insignificant for farther units. In most of the cases the second and third closest units, denoted as U_2 and U_3, are sufficient for calculating the displacement of the input $\mathbf{x_i}$ from the center of the unit U_1. Their pull forces make up a virtual triangle, as illustrated in Figure 5.16. There is one rare exception to this assumption, namely in cases where both the second and third best matching unit are found to be on one axis with U_1. This implies that the input $\mathbf{x_i}$ would drift along only one dimension as a triangle effect can not be realized. In those cases, the fourth closest unit U_4 is taken into account (c.f. Figure 5.16(d)).

Additionally, if the next closer unit is on the same axis, its force – and hence the displacement along that axis – is not computed to reduce the computation overhead. In other words, the displacement is zero if the units are not pulling in different directions. Finally, the x and y coordinates of the exact position $\mathbf{p}$ of input $\mathbf{x}$ on unit U can be defined as:

$$\mathbf{p}_{<x,y>} = < \lambda * \sum_{i=2}^{k} \mathbf{F}_i * \frac{1}{U_{i<x>} - U_{1<x>}},\ \lambda * \sum_{i=2}^{k} \mathbf{F}_i * \frac{1}{U_{i<y>} - U_{1<y>}} >$$

where k is an index over the two or three nearer units U_2, U_3 and U_4 respectively, i.e. $k = 3$ or $k = 4$. A grid-constant λ is used to reconcile the displacement according to the display co-ordinates and is initially set to approximately a quarter of the unit's pixel size. In some cases two or more

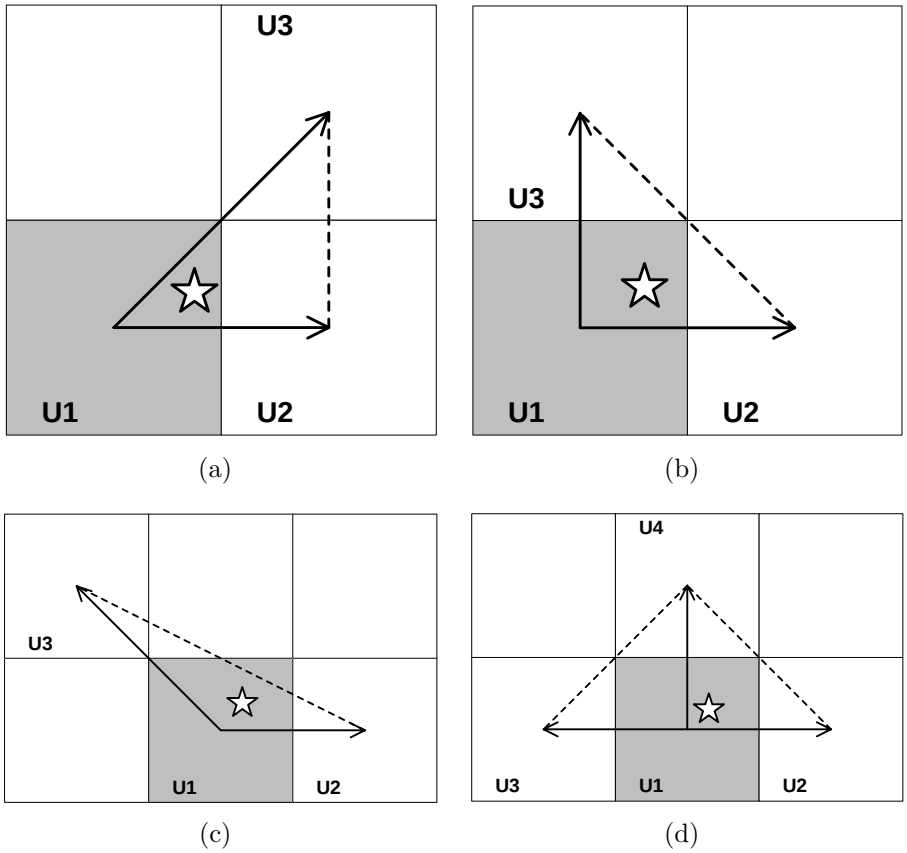


Figure 5.16: Neighbouring forces on an item.

inputs may overlap each other too much due to very high similarity in their winner units and weights. In such a situation we marginally shift the inputs apart by applying a force of repulsion, where overlapping units push each other in opposite direction.

Constellations as Interconnection Trails

In the physical world, entities are usually interconnected either by physical or by semantic means. In the proposed night-sky visualization, the interconnections are realized by exploiting the notion of constellations. Closely related stars form a pattern and highlight the relationship between the inputs, which may otherwise be mapped to different units. We allow both user defined and automatic trails (such as based on meta-data) to illustrate usefulness of constellations by drawing connection lines between the stars.

Guiding Stars and Semantic Zoom

The night sky has other characteristics which make it an interesting metaphor for information visualization and exploration – landmarks and details-on-demand effect. Some of the stars are more prominent than others and in a way guide the exploration activity. Additionally not all stars are visible to the naked (bare) eye, and we need to zoom in using telescope to view the details.

These traits are realized by making use of input-unit distance metric. The size of the document representative stars depends on the relative distance and relevance in the unit based on the user preferences. The star for most relevant documents is larger in size to other documents in the unit. And initially, only some of the more prominent documents are displayed. Detail levels are defined before hand (c.f. Table 5.4) and more stars are made visible as the user zooms in to a particular level. On the other hand, this option could be switched off to display all stars and thus clusters would be visible by many stars next to each other.

Zoom Level	0	1	2	3	4
Stars' Visibility	10%	20%	40%	70%	100%

Table 5.4: Zoom level and visibility threshold of stars.

5.3.3 Experiments and Results

In contrast to text retrieval, the research in personal knowledge management lacks benchmark corpuses and metrics for a reliable evaluation of the developed tools (Kelly, 2006). For the experiments described in this section, we used two synthetic data sets to demonstrate the visualization, and one benchmark text corpus to test its applicability to a large real-life corpus.

Figure 5.17 shows experiments with the 'chain-link' data set, i.e. two intertwining rings in three-dimensional space. This data-set cannot be projected to two dimensions while preserving the ring-structures, the normal behavior is for the rings to 'break'. The visualization resembles the structure of the two rings well, with the points stretching over the cell space in such a way that an almost continuous line is formed. This is very similar to the original data, which also doesn't form the ring as a continuous data chain, but rather as several small clusters of data points.

Figure 5.18 depicts a plot of the two principal components of a ten-dimensional data set, generated using several Gaussian distributions with different centers and kernel widths. By not placing the data items in the

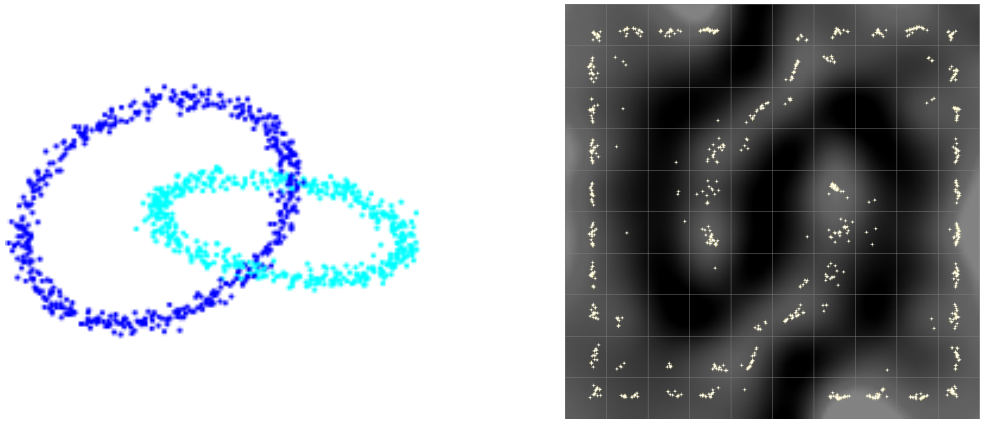


Figure 5.17: Chain-link data set and a trained map.

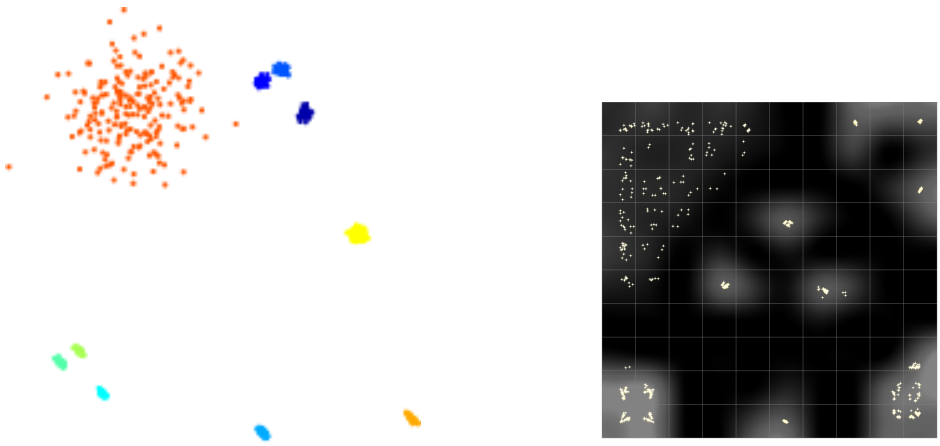


Figure 5.18: A data set of several different Gaussian clusters and a trained map.

center of the units, the Night Sky visualization shows the concentration of inputs more effectively and also provides clear cluster boundaries.

The text corpus we used for our last experiment is the *20 newsgroups* data set[10], which has become very popular benchmark corpus. It consists of 1000 newsgroup postings for each of its 20 different newsgroups, such as *alt.atheism* and *comp.sys.mac.hardware*. From each newsgroup, 1,000 articles from the year 1993 have been taken. We considered only the subject, references, and the message body, but omitted other header lines. A standard *bag-of-words* indexing approach was used, applying a manually created stop-word list and document frequency threshold to reduce the dimensional-

[10]http://people.csail.mit.edu/jrennie/20Newsgroups

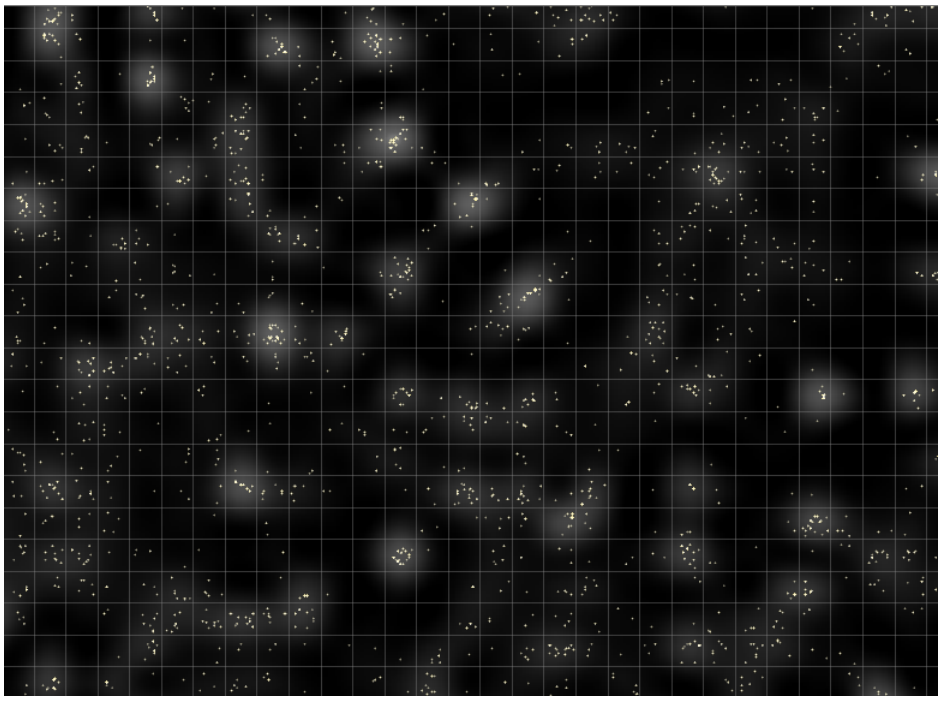


Figure 5.19: Overview of sky visualization for 20 newsgroups data set.

ity. A $tf \times idf$ weighting scheme was employed to obtain the vector values for the 2896 remaining terms. Finally, we trained a SOM of the size of 50×40 units.

Figure 5.19 depicts the overview of the trained map. Due to spreading the inputs over the SOM cells and the tendency to the inputs being moved towards the cluster centers, these become more compact and dense, while the areas between two clusters become larger – it becomes easier to identify groups of similar inputs.

Figure 5.20 depicts two sections of the map. The left image in the fig-

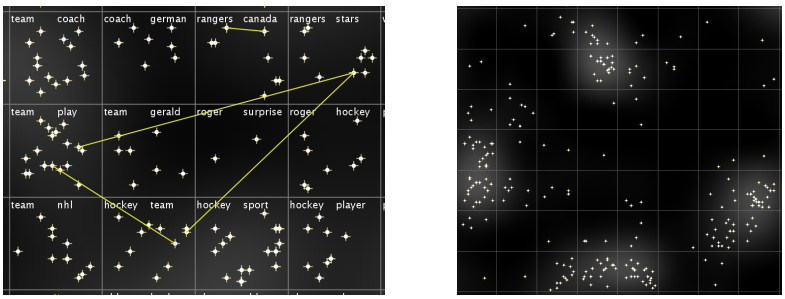


Figure 5.20: Detailed view of the 20 newsgroups map.

ure illustrates the concept of constellations: postings that are in relation to each other, here direct replies to other postings, are linked. Such associative referencing allows instantly recalling other linked items in the data set. The linkage information was automatically extracted from the headers of the postings. And, the labels attached to the units have been automatically extracted as the most important terms describing the unit's content using the LabelSOM method. The right image shows a detailed view of cluster boundaries between two *sci.med* clusters in the upper-left and upper-middle area, and two *rec.motorcycles* and *rec.autos* clusters located in the lower-middle and right-middle area. Even though there are only few or no empty units between the cluster centers, the inputs on the units between those centers have been placed closer towards the centers, and therefore the cluster boundaries become easily visible.

We presented a novel method for visualizing lifetime information using Self-Organizing Maps. The night-sky metaphor is used to represent and interactively explore the digital memories. The relationship of similarity between the items was depicted through star clusters and other complex interconnections by constellations. Our experiments with different data sets show that even a large stockpile of data could be turned into very useful knowledge map with effective visualizations.

Chapter 6

Outlook

Human life is blended with multiple entanglements ranging from the workplace to family issues. Coverage of this diversity in a single software system although could be interesting for rationalizing trails but is a difficult task. Moreover it is very difficult to draw a boundary between the pubic and private life. For example, over-head cameras can record everything a person sees which means a knowledge box might contain others' life history too. Confronting the social implications and privacy issues was out of scope of our work. In the current implementation we restricted the data feed modules to cover only digital items already available from the desktop such as emails, documents, photos, contacts, calendar entries, and browsing history. This data is archived on users own computer and he/she is solely responsible for sharing this repository with others.

The personal knowledge box is designed as a generic framework. A very convincing case study is a diary of a *researcher* handling a wide range of information accumulated over a lifetime. Continuous archival is supplemented by associating metadata with contents using ontologies. The possibility of adding annotations to all stored objects enriches the potential use of such a diary.

6.1 Summary and Discussion

The research presented in this thesis has two principal contributions: Layered ontology for modeling semantics of digital memories and the implemented framework. Both contributions cover multiple facets of the knowledge box and are summarized subsequently.

Multifold Semantics Enhancements

First contribution of this thesis is the multifold ontology for modeling the semantics of a diverse range of digital memories and also their associations. Web ontologies are rarely static. The changes are influenced from different directions such as correcting errors, adding new axioms, or even by improving the domain model. Consequently, ontology of digital memories is partitioned into layers based on the information context and the semantic insight, such as a layer for semantic enhancements of the underlying resource structure – focusing the inward perspective – and another layer for realizing trails intended to present outward viewpoint. Each layer is locally complete in its coverage of conceptualization which allows easy maintenance of growing individual modules.

Dynamic Knowledge Model

There is a strong relationship between human thoughts and the social context; personal information is never an individual product and that it emerges through connections, dialog and social interaction. For the ontology of digital memories, such a shared context is achieved by reusing an existing foundational ontology as a base. The reuse of foundational ontologies can facilitate mutual understanding and interoperability among varying terminologies. The guided questions derived from the foundational ontology provide a formal but transparent basis for the users' negotiation to disambiguate any confusion in interpretation of their terminology.

Structured Annotation Framework

The enormity of the lifetime information poses a serious challenge in terms of comprehensibility. To improve the comprehensibility in case of life time capture of personal experiences, we used a structured context annotation framework. The annotation framework brought together information extracted from diverse media types into an integrated model. The annotation framework provided the semantic insight of life-items related to their spatio-temporal location, involved agents & their activities, and semantic content labels. The process of annotation was further automated using named entity recognition and term extraction.

Collections and Trails

Comprehensibility was further improved by partitioning the knowledge space into collections. These collections lay the foundation for constructing associative trails – an essential feature for managing digital memories of lifetime. Two types of trails were used in our work which are characterized by their membership criteria and the navigation function. The later is an important aspect of the storytelling. Silent trails are essentially arbitrary user collections with tacit membership criteria. On the other hand, live trails have explicit membership criteria. Trails were also constructed in our work by means of landmarks. Significance of landmarks is better remembrance of personal experiences. They emerge from the user's lifetime knowledge space. And items that share meaning or physical similarity with the landmark become associated with it and selection of the landmark activates linked items, and vice versa.

Knowledge Space Navigation

Finally, a software architecture pattern for semantic enrichment was introduced. This pattern was used to develop an extensible service-oriented framework for demonstrating capabilities of the lifetime store. A number of dedicated data acquisition and analysis modules were implemented to enrichment the life items with semantics and for subsequent ontological storage. Innovative visualization techniques were introduced to effectively navigate in the lifetime archive. The proposed night sky visualization facilitated better understanding of the underlying data by exploiting the overview and details-on-demand interaction technique. Other interaction strategies such as zooming, panning, and document and area selection were also supported. Individual items from the input space were represented as stars, which together with other similar items formed star clusters. Further, we made use of the notion of guiding stars and constellations. From the results of our experiments with different data sets we conclude that sky visualization, on top of concept-focused associative search, can make it easy to locate, link, and learn from even a huge repository.

6.2 Research Questions Revisited

In this section we will re-examine the questions stipulated in the beginning of this thesis about realizing a system for managing digital memories, in the light of the current findings.

- *Is it possible to accurately model the semantics of a diverse range of digital memories and also their associations? How Semantic Web technologies can help in this regard?*
  Modeling whole life of a person might had required an effort similar to Cyc ontology in scale. We adopted a modular approach for modeling the semantics of digital memories in multifolds. The layers are based on the information context and the semantic insight. Each layer is locally complete in its coverage of conceptualization which allows easy maintenance of growing individual modules. The guided questions derived from the foundational ontology provide a formal but transparent basis for the users' negotiation for future extensions to the ontology.

- *Manual building of trails for thousands, or even hundreds, of items is a diligent task. How can we realize an efficient and productive system of construction of trails and how much automation in constructing associative trails is possible by exploiting the semantic insights of the contents?*
  Foremost, we admit that it is difficult to avoid the manual effort in building trails and explicitly annotating semantics of good quality with the digital memories. Automation is possible, though, in different aspect of trails construction. For example, the concept based associations formed a conceptual index of life-items which characterized the important point of one's life. These associations were exploited to suggest related items to the user. Secondly, we made use of the notion of landmarks in the lifetime space of memories. Similar to highly cited articles in co-citation networks, the landmarks are important items (persons, events, or documents) that emerge from ones life and guide to other related items.

- *Which contextual information and other annotations should be stored along with the actual memories and can these be acquired automatically or do they need to be manually entered?*
  Our analysis of the personal information organization on the desktop and faceted knowledge organization revealed that, for the most part, digital memories have following core dimensions: spatial & temporal location, involved agents & activities, and content descriptors. These dimensions were modeled in a structured annotation framework – STeAL. Different text retrieval techniques, such as named entity recognition and term extraction, were applied to textual contents and the results were suggested to the user during the annotation process.

- *Personal memories have a strong relationship with the social interac-*

tions. Can we guarantee shared semantic context for items in the networked environment?
Depending upon the coverage of concepts and the model scope, ontologies are categorized as foundational and domain ontologies. Foundational ontologies reflect shared specifications by using conceptual design patterns and are marked by agreement inside the Community of Practice, driving the development of the foundational ontology. Foundational ontologies, thus provide a strong foundation for mediation in heterogeneous environments & knowledge sharing and their reuse provides taxonomic and axiomatic context. The reuse of foundational ontologies can further facilitate mutual understanding and interoperability in a networked environment. For the ontology of digital memories we adopted a high level view from DOLCE foundational ontology. Any subsequent extension is also subjected to alignment with the foundational ontology through a question-driven approach.

- *How can we easily and effectively retrieve useful information from the stockpile of digital objects spanning a human lifetime? Additionally, how can we overcome the comprehensibility problem and semantics overload in the lifetime knowledge box, presumably containing millions of items.*
We used a hybrid strategy to persist digital memories. We built full text index of the contents to guarantee efficient retrieval & improved recall and combined it with the metadata stored in the triple store for retrieving precise results of semantic searches on the data. A graph based query editor was implemented for interactively writing structured queries on the metadata. Finally, the Nigh-Sky visualization was developed to facilitate better understanding of the underlying data by exploiting the overview and details-on-demand interaction technique.

6.3 Future Work and Conclusion

Visualization is an important issue and needs to be addressed in detail. We will focus on the visualization of search results, ontologies, and the means to incorporate the changes made by user into the system. It is also necessary to assist the user in refining queries in an iterative manner influenced by the previous search results.

Only term aligning component of our implemented framework was subject to external evaluation. In an informal setting, the alignment algorithm was demonstrated to three linguistic scientists. It is worthwhile to do an

empirical user study of the whole framework. We have planned to perform a comprehensive study involving all facets of the system from visualization to effectiveness of semantic search.

Personal desktops are gradually disappearing and the use of hand held devices is increasing by every coming day. Most parts of our digital lives are now available on mobile phones than on the desktop. Such pervasiveness presents different challenges such as how to keep every bit of life synchronized at heterogeneous devices. We strongly feel that research in personal information management would eventually lean toward this area. Current breed of Semantic Web tools and APIs are not very memory efficient. By the time smart and efficient API's emerge, we will definitely see mushrooming of pervasive semantic Webs on mobile devices, handling our life bits with much more clarity of results. We conclude this thesis by quoting Vannevar Bush from his landmark paper "As We May Think."

> "Man cannot hope fully to duplicate this mental process [of association of thoughts] artificially, but he certainly ought to be able to learn from it. In minor ways he may even improve, for his records have relative permanency... One cannot hope thus to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage."

Bibliography

Adar, E., Karger, D., and Stein, L. A. (1999). Haystack: Per-user information environments. In *Proceedings of 8th International Conference on Information and Knowledge Management (CIKM)*, pages 413–422, Kansas City, Missouri. ACM Press.

Adida, B. and Birbeck, M. (2007). RDFa Primer 1.0: Embedding RDF in XHTML. Working draft, W3C. http://www.w3.org/TR/xhtml-rdfa-primer.

Agirre, E. and Edmonds, P., editors (2006). *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer.

Ahmed, M., Hoang, H. H., Karim, S., Khusro, S., Lanzenberger, M., Latif, K., Michlmayr, E., Mustofa, K., Nguyen, H. T., Rauber, A., Schatten, A., Tho, M. N., and Tjoa, A. M. (2004). SemanticLIFE - A framework for managing information of a human lifetime. In Bressan, S., Taniar, D., Kotsis, G., and Ibrahim, I. K., editors, *Proceedings of 6th International Conference on Information Integration and Web-based Applications & Services (iiWAS)*, volume 183 of *books@ocg.at*, pages 687–696, Jakarta, Indonesia. Austrian Computer Society.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., and Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21.

Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, B., Ramakrishnan, C., and Sheth, A. (2005). Ranking complex relationships on the semantic web. *IEEE Internet Computing*, 9(3):37–44.

Alexaki, S., Christophides, V., Karvounarakis, G., Plexousakis, D., and Tolle, K. (2001). On storing voluminous RDF descriptions: The case of web portal catalogs. In Mecca, G. and Siméon, J., editors, *Proceedings of 4th International Workshop on the Web and Databases (WebDB)*, pages 43–48, Santa Barbara, CA.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., and Tochtermann, K. (2002). The InfoSky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3/4):166–181.

ANSI-NISO-Z39-19 (2005). ANSI/NISO Z39.19-2005, Guidelines for the construction, format, and management of monolingual controlled vocabularies. National Information Standards Organization (NISO) Press, Bethesda, Maryland. ISSN: 1041-5653.

Anyanwu, K. and Sheth, A. (2003). **p**-queries: Enabling querying for semantic annotations on the semantic web. In (WWW, 2003).

Aris, A., Gemmell, J., , and Lueder, R. (2004). Exploiting location and time for photo search and storytelling in MyLifeBits. Technical Report MSR-TR-2004-102, Microsoft Research.

Aroyo, L., Denaux, R., Dimitrova, V., and Pye, M. (2006). Interactive ontology-based user knowledge acquisition: A case study. In (Sure and Domingue, 2006), pages 560–574.

Auer, S. and Lehmann, J. (2007). What have Innsbruck and Leipzig in common? extracting semantics from wiki content. In (Franconi et al., 2007), pages 503–517.

Bao, J. and Honavar, V. (2006). Divide and conquer semantic web with modular ontologies - a brief review of modular ontology language formalisms. In (Haase et al., 2006).

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., and Shachak, A. (2006). Structured vs. unstructured tagging - a case study. In *Collaborative Web Tagging Workshop at WWW'06*, Edinburgh. http://www.rawsugar.com/www2006/12.pdf.

Beckett, D. (2002). Scalability and storage: Survey of free software / open source RDF storage systems. Deliverable 10.1, W3C Semantic Web Advanced Development (SWAD) for Europe. http://www.w3.org/2001/sw/Europe/reports/pdf/10.2.pdf.

Beckett, D. (2004). RDF/XML syntax specification (revised). Recommendation, W3C. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210.

Beckett, D. and Grant, J. (2002). Mapping semantic web data with RDBMSes. Deliverable 10.2, W3C Semantic Web Advanced Development (SWAD) for Europe. http://www.w3.org/2001/sw/Europe/reports/pdf/10.1.pdf.

Behrendt, W., Gahleitner, E., Latif, K., Gruber, A., Weippl, E., Schaffert, S., and Kargl, H. (2005). Upper ontologies with specific consideration of DOLCE, SUMO and Sowa's upper level ontology. Deliverable D121, DynamOnt Project. http://dynamont.factlink.net/216753.1.

Bell, G. and Gemmell, J. (2007). A digital life. *Scientific American*, March.

Bergman, M. K. (2007). Structure paves the way to the semantic web. *IEEE Intelligent Systems*, 22(3):84–86.

Berners-Lee, T. (2007). Re: What if an URI also is a URL. semantic-web@w3.org Mail Archives. http://lists.w3.org/Archives/Public/semantic-web/2007Jun/0066.html.

Berners-Lee, T., Fielding, R. T., and Masinter, L. (2005). Uniform Resource Identifier (URI): Generic Syntax. http://www.ietf.org/rfc/rfc3986.txt.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 279:34–43.

Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., and Strintzis, M. (2005). Semantic annotation of images and videos for multimedia analysis. In Gómez-Pérez, A. and Euzenat, J., editors, *The Semantic Web: Research and Applications (Proceedings of 2nd European Semantic Web Conference)*, volume 3532 of *Lecture Notes in Computer Science*, pages 592–607, Heraklion, Greece. Springer.

Boardman, R. (2004). *Improving Tool Support for Personal Information Management*. PhD thesis, Department of Electrical and Electronic Engineering, Imperial College London, Intelligent and Interactive Systems Group, SW7 2TE, UK.

Boardman, R., Spence, R., and Sasse, M. A. (2003). Too many hierarchies? the daily struggle for control of the workspace. In Jacko, J. A. and Stephanidis, C., editors, *Proceedings of 10th International Conference on Human-Computer Interaction (HCII)*, pages 616–620, Crete, Greece. Lawrence Erlbaum Associates.

Bolour, A. (2003). Notes on the Eclipse plug-in architecture. Eclipse Corner Article. http://www.eclipse.org/articles/Article-Plug-in-architecture/.

Booch, G. (2007). The well-tempered architecture. *IEEE Software*, 24(4):24–25.

Booth, D. (2007). URI declaration versus use (was: Terminology question). semantic-web@w3.org Mail Archives. http://lists.w3.org/Archives/Public/semantic-web/2007Jul/0435.html.

Boster, J. (2005). Categories and cognitive anthropology. In Lefebvre, C. and Cohen, H., editors, *Handbook of Categorization in Cognitive Science*, pages 92–117. Elsevier.

Bouquet, P., Stoermer, H., Tummarello, G., and Halpin, H., editors (2007). *Proceedings of the WWW2007 Workshop I3: Identity, Identifiers, Identification*, volume 249 of *CEUR Workshop Proceedings*, Banff, Canada. http://ceur-ws.org/Vol-249.

Bressan, S., Küng, J., and Wagner, R., editors (2006). *Proceedings of 17th*

International Conference on Database and Expert Systems Applications, volume 4080 of *Lecture Notes in Computer Science*, Krakow, Poland. Springer.

Brickley, D. and Guha, R. (2004). RDF vocabulary description language 1.0: RDF Schema. Recommendation, W3C. http://www.w3.org/TR/2004/REC-rdf-schema-20040210.

Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: a generic architecture for storing and querying RDF and RDF schema. In (Horrocks and Hendler, 2002), pages 54–68.

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings*, 58(1/2):49–72.

Buchanan, G., Blandford, A., Thimbleby, H., and Jones, M. (2004). Integrating information seeking and structuring: exploring the role of spatial hypertext in a digital library. In *Proceedings of the 15th ACM conference on Hypertext & Hypermedia (HYPERTEXT)*, pages 225–234, Santa Cruz, CA. ACM Press.

Budin, G. (2003). Ontology-driven translation management. In *Knowledge Systems in Text and Translation*, EU High Level Scientific Conference Series, Aarhus, Denmark.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1):101–108.

Celjuska, D. and Vargas-Vera, M. (2004). Semi-automatic population of ontologies from text. In Paralic, J., Pölzlbauer, G., and Rauber, A., editors, *5th Workshop on Data Analysis (WDA)*, Tatranska Polianka, Slovakia.

Chakrabarti, S. (2003). *Mining the Web: Discovering Knowledge from Hypertext Data*. Data Management Systems. Morgan Kaufman Publishers.

Chandrasekaran, B., Josephson, J., and Benjamins, R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20–26.

Chen, C. (2004). *Information Visualization: Beyond the Horizon*. Springer-Verlag London, 2nd edition.

Cheyer, A., Park, J., and Giuli, R. (2005). IRIS: Integrate. Relate. Infer. Share. In Decker, S., Park, J., Quan, D., and Sauermann, L., editors, *Next Generation Information Management & Collaboration Infrastructure (Proceedings of ISWC Workshop on The Semantic Desktop)*, volume 175 of *CEUR Workshop Proceedings*, pages 64–78, Galway, Ireland.

Cimiano, P., Staab, S., and Tane, J. (2003). Deriving concept hierarchies from text by smooth formal concept analysis. In *Proceedings of the GI Workshops on Lehren - Lernen - Wissen - Adaptivität (LLWA), Fachgruppe Maschinelles Lernen, Wissenentdeckung, Data Mining*, pages 72–79, Karlsruhe, Germany.

Ciravegna, F. (2001a). Adaptive information extraction from text by rule

induction and generalisation. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, Seattle, Washington.

Ciravegna, F. (2001b). $(LP)^2$ an adaptive algorithm for information extraction from web-related texts. In *Workshop on Adaptive Text Extraction and Mining in conjuction with 17th Internation Joint Conference on Artificial Intelligence (IJCAI)*, Seattle, USA.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in BioInformatics*, 6(1):57–71.

Connolly, D. (2007). Gleaning resource descriptions from dialects of languages (GRDDL). Proposed recommendation, W3C. http://www.w3.org/TR/grddl/.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. (1998). Learning to extract symbolic knowledge from the world wide web. In *15th National Conference on Artificial Intelligence*, pages 509–516, Madison, USA. AAAI Press.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust nlp tools and applications. In *40th Anniversary Meeting, Association for Computational Linguistics (ACL)*, Philadelphia, USA.

Cutrell, E., Robbins, D., Dumais, S., and Sarin, R. (2006). Fast, flexible filtering with phlat. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 261–270, Montreal, Canada. ACM Press.

Cyganiak, R. (2007). Re: Cool URIs for the semantic web. semantic-web@w3.org Mail Archives. http://lists.w3.org/Archives/Public/semantic-web/2007Mar/0055.html.

Czerwinski, M., Gage, D. W., Gemmell, J., Marshall, C. C., Manuel A. Pérez-Qui n., Skeels, M. M., and Catarci, T. (2006). Digital memories in an era of ubiquitous computing and abundant storage. *Communications of the ACM*, 49(1):44–50.

Damásio, C. V., Analyti, A., Antoniou, G., and Wagner, G. (2006). Supporting open and closed world reasoning on the web. In Alferes, J. J., Bailey, J., May, W., and Schwertel, U., editors, *Principles and Practice of Semantic Web Reasoning (Revised selected papers from 4th International Workshop PPSWR'06)*, volume 4187 of *Lecture Notes in Computer Science*, pages 149–163, Budva, Montenegro. Springer.

Damjanović, V., Behrendt, W., Plössnig, M., and Holzapfel, M. (2007). Developing ontologies for collaborative engineering in mechatronics. In (Franconi et al., 2007), pages 190–204.

Davies, J., Fensel, D., and van Harmelen, F., editors (2003). *Towards The Semantic Web: Ontology-driven Knowledge Management*. John Wiley &

Sons.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., and Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 652–659, Washington, D.C. ACM Press.

Dittrich, J.-P. and Salles, M. A. V. (2006). iDM: A unified and versatile data model for personal dataspace management. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pages 367–378, Seoul, Korea.

Dittrich, J.-P., Salles, M. A. V., Kossmann, D., and Blunschi, L. (2005). iMeMex: Escapes from the personal information jungle. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB)*, pages 1306–1309, Trondheim, Norway.

Dong, X. and Halevy, A. (2005). A platform for personal information management and integration. In Stonebraker, M., Weikum, G., and DeWitt, D., editors, *Proceedings of the 2nd Conference on Innovative Data Systems Research (CIDR)*, pages 119–130, Asilomar, CA.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C. (2003). Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th SIGIR Conference on Research and Development in Information Retrieval*, pages 72–79, Toronto, Canada. ACM Press.

Economist (2007a). Watching the web grow up. Technology Quarterly.

Economist (2007b). What's in a name? Technology Quarterly.

Edgington, T., Choi, B., Henson, K., Raghu, T., and Vinze, A. (2004). Adopting ontology to facilitate knowledge sharing. *Communications of ACM*, 47(11):85–90.

Embley, D. (2004). Toward semantic understanding - an approach based on information extraction ontologies. In Schewe, K.-D. and Williams, H., editors, *Proceedings of 15th Australian Database Conference (ADC)*, volume 27 of *Conferences in Research and Practice in Information Technology*, pages 3–12, Dunedin, New Zealand. Australian Computer Society.

Euzenat, J., Bouquet, P., Dieng, R., Ehrig, M., Hauswirth, M., Jarrar, M., Stuckenschmidt, H., and Shvaiko, P. (2004). State of the art on ontology alignment. Deliverable 2.2.3 (v1.2), Knowledge Web Project.

Falconer, S. M., Noy, N. F., and Storey, M.-A. (2006). Towards understanding the needs of cognitive support for ontology mapping. In *Proceedings of International Workshop on Ontology Matching (OM-2006)*, Georgia, USA.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT

Press, Cambridge, MA.

Fensel, D., Bussler, C., and Maedche, A. (2002). Semantic web enabled web services. In (Horrocks and Hendler, 2002).

Fitzgibbon, A. and Reiter, E. (2004). Memories for life: managing information over a human lifetime. In Hoare, T. and Milner, R., editors, *Grand Challenges in Computing Research*, pages 13–16. The British Computer Society.

Fluit, C., Sabou, M., and van Harmelen, F. (2003). Ontology-based information visualization. In Geroimenko, V. and Chen, C., editors, *Visualizing the Semantic Web*, pages 36–48. Springer-Verlag.

Franconi, E., Kifer, M., and May, W., editors (2007). *Proceedings of 4th European Semantic Web Conference*, volume 4519 of *Lecture Notes in Computer Science*, Innsbruck, Austria. Springer.

Freeman, E. and Gelernter, D. (1996). Lifestreams: A storage model for personal data. *SIGMOD Record*, 25(1):80–86.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Gahleitner, E., Behrendt, W., Palkoska, J., and Weippl, E. (2005). On cooperatively creating dynamic ontologies. In *Proceedings of the 16th ACM Conference on Hypertext and Hypermedia*, pages 208–210, Salzburg, Austria. ACM Press.

Gahleitner, E., Latif, K., Gruber, A., and Westenthaler, R. (2006). Specification of methodology and workbench for dynamic ontology creation. Deliverable D201, DynamOnt Project. http://dynamont.factlink.net/258501.0.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison Wesley.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In (Gil et al., 2005), pages 262–276.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In Gómez-Pérez, A. and Benjamins, V. R., editors, *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume 2473 of *Lecture Notes In Computer Science*, pages 166–181, Siguenza, Spain. Springer-Verlag.

Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet project: extension and axiomatization of conceptual relations in WordNet. In *On The Move to Meaningful Internet Systems: CoopIS, DOA, and ODBASE*, volume 2888 of *LNCS*, pages 820–838, Catania, Italy. Springer.

Garcia, E. and Sicilia, M.-A. (2003). User interface tactics in ontology-based information seeking. *PsychNology Journal*, 1(3):242–255.

Gärdenfors, P. (2000). *Conceptual Spaces (The geometry of thoughts)*. MIT Press.

Gemmel, J., Bell, G., and Lueder, R. (2003). Living with a lifetime store. In *Proceedings of ATR Workshop on Ubiquitous Experience Media*, Kyoto, Japan.

Gemmel, J., Bell, G., Lueder, R., Drucker, S., and Wong, C. (2002). MyLifeBits: Fulfilling the Memex Vision. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 235–238, Juan Les Pins, France. ACM Press.

Ghidini, C. and Serafini, L. (2006). Reconciling concepts and relations in heterogeneous ontologies. In (Sure and Domingue, 2006), pages 50–64.

Gil, Y., Motta, E., Benjamins, R., and Musen, M., editors (2005). *Proceedings of 4th International Semantic Web Conference (ISWC)*, volume 3729 of *Lecture Notes in Computer Science*, Galway, Ireland. Springer.

Girgensohn, A., Adcock, J., Cooper, M., Foote, J., and Wilcox, L. (2003). Simplifying the management of large photo collections. In (Rauterberg et al., 2003), pages 196–203.

Goldin, D. and Wegner, P. (2006). Principles of interactive computation. In Goldin, D., Smolka, S. A., and Wegner, P., editors, *Interactive Computation - The New Paradigm*, pages 25–37. Springer.

Gomez, F. (2001). An algorithm for aspects of semantic interpretation using an enhanced WordNet. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, CMU, Pittsburgh.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2003). *Ontological Engineering*. Springer.

Gregorowicz, A. and Kramer, M. (2006). Mining a large scale term-concept network from Wikipedia. Technical paper, MITRE Corporation.

Griswold, W. G., Shonle, M., Sullivan, K., Song, Y., Tewari, N., Cai, Y., and Rajan, H. (2006). Modular software design with crosscutting interfaces. *IEEE Software*, 23(1):51–60.

Grosz, B. and Sidner, C. (1986). Attention, intention and the structure of discourse. *Computational Linguistic*, 12(3):175–204.

Gruber, A., Latif, K., Westenthaler, R., and Thalbauer, S. (2007). Specification of methodology and workbench for dynamic ontology creation. Deliverable D202, DynamOnt Project.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Human-Computer Studies*, 43(1):907–928.

Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2-3):293–310.

Guarino, N. (1998). Formal ontology and information systems. In *Inter-*

national Conference on Formal Ontology in Information Systems, pages 3–15, Trento, Italy. IOS Press.

Guo, Y., Pan, Z., and Heflin, J. (2004). An evaluation of knowledge base systems for large OWL datasets. In (McIlraith et al., 2004), pages 274–288.

Haase, P., Honava, V., Kutz, O., Sure, Y., and Tamilin, A., editors (2006). *Proceedings of the ISWC Workshop on Modular Ontologies (WoMO)*, volume 232 of *CEUR Workshop Proceedings*, Athens, GA.

Hammerton, J., Osborne, M., Armstrong, S., and Daelemans, W. (2002). Introduction to special issue on machine learning approaches shallow parsing. *Machine Learning Research*, 2(1):551–558.

Handschuh, S., Staab, S., and Ciravegna, F. (2002). S-CREAM - semi-automatic creation of metadata. In *Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, volume 2473 of *Lecture Notes in Computer Science*, pages 358–372, Siguenza, Spain. Springer-Verlag.

Harel, D. and Rumpe, B. (2004). Meaningful modeling: What's the semantics of "semantics". *IEEE Computer*, 37(10):64–72.

Harth, A. and Decker, S. (2005). Optimized index structures for querying RDF from the web. In *Proceedings of 3rd Latin American Web Congress (LA-Web)*, Buenos Aires, Argentina. IEEE Computer Society.

Hassan-Montero, Y. and Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In Guerrero-Bote, V. P., editor, *Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies (InSciT).*, Mérida, Spain.

Hayes, P. (2004). RDF Semantics. Recommendation, W3C. http://www.w3.org/TR/2004/REC-rdf-mt-20040210.

Heflin, J. and Hendler, J. (2000). Dynamic ontologies on the web. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 443–449, Menlo Park, CA. AAAI/MIT Press.

Herman, I. (2006). Semantic Web @ W3C : Activities, recommendations and state of adoption. In *Industry Track of 5th International Semantic Web Conference*, Athens, GA. http://www.w3.org/2006/Talks/1109-Athens-IH.

Herman, I. (2007). Introduction to the semantic web (tutorial). In *International Conference on Semantic Web & Digital Libraries*, Bangalore, India. http://www.w3.org/2007/Talks/0221-Bangalore-IH/.

Hetzler, B., Harris, M., Havre, S., and Whitney, P. (1998). Visualizing the full spectrum of document relationships. In *Proceedings of 5th International Conference on Structures and Relations in Knowledge Organization*, pages 168–175, Lille, France. ERGON Verlag.

Hobbs, J. R. and Pan, F. (2004). An ontology of time for the semantic web. *ACM Transactions on Asian Language Processing (Special issue on Temporal Information Processing)*, 3(1):66–85.

Hollink, L., Schreiber, G., Wielemaker, J., and Wielinga, B. (2003). Semantic annotation of image collections. In Handschuh, S., Koivunen, M.-R., Dieng-Kuntz, R., and Staab, S., editors, *Proceedings of the KCAP Workshop on Knowledge Markup and Semantic Annotation*, Sanibel, Florida.

Horrocks, I. and Hendler, J., editors (2002). *Proceedings of 1st International Semantic Web Conference (ISWC)*, volume 2342 of *Lecture Notes in Computer Science*, Sardinia, Italy. Springer.

Horrocks, I., Parsia, B., Patel-Schneider, P., and Hendler, J. (2005). Semantic web architecture: Stack or two towers? In Fages, F. and Soliman, S., editors, *Principles and Practice of Semantic Web Reasoning (Proceedings of Third International Workshop PPSWR'05)*, volume 3703 of *Lecture Notes in Computer Science*, pages 37–41, Dagstuhl Castle, Germany. Springer.

Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26.

Huhns, M. N. and Singh, M. P. (2005). Service-oriented computing: key concepts and principles. *IEEE Internet Computing*, 9(1):75–81.

Jeh, G. and Widom, J. (2002). Simrank: A measure of structural-context similarity. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, Edmonton, Alberta. ACM Press.

Jones, D., Bench-Capon, T., and Visser, P. (1998). Methodologies for ontology development. In Cuena, J., editor, *Proceedings of IFIP XV IT & KNOWS*, pages 62–75, Budapest, Hungary.

Kahan, J. and Koivunen, M.-R. (2001). Annotea: An open rdf infrastrucrue for shared web annotations. In *Proceedings of 10th Iinternational Conference on World Wide Web (WWW)*, pages 623–632, Hong Kong. ACM Press.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology mapping: The state of the art. *The Knowledge Engineering Review*, 18.

Karim, S., Latif, K., and Tjoa, A. M. (2007). Providing universal accessibility using connecting ontologies: A holistic approach. In *HCII'07*, volume 4556 of *Lecture Notes in Computer Science*, Beijing, China. Springer.

Kelly, D. (2006). Evaluating personal information management behaviors and tools. *Communications of the ACM*, 49(1):84–86.

Kettler, D. (1967). Sociology of knowledge and moral philosophy: The place of traditional problems in the formation of Mannheim's thought. *Political Science Quarterly*, 82(3):399–426.

Khan, L., McLeod, D., and Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 13:71–85.

Khare, R. (2006). Microformats: the next (small) thing on the semantic web? *IEEE Internet Computing*, 10(1):68–75.

Klyne, G. and Carroll, J. J. (2004). Resource description framework (RDF): Concepts and abstract syntax. Recommendation, W3C. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210.

Kohonen, T. (1995). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer.

Krogstie, J. (1995). *Conceptual Modeling for Computerized Information Systems Support in Organizations*. PhD thesis, Faculty of Electrical Engineering and Computer Science, Norwegian Institute of Technology, Trondheim, Norway.

Latif, K. and Mayer, R. (2007). Sky-metaphor visualisation for self-organising maps. In Tochtermann, K. and Maurer, H., editors, *Journal of Universal Computer Science (Proceedings of 7th International Conference on Knowledge Management)*, pages 400–407, Graz, Austria.

Latif, K., Mustofa, K., and Tjoa, A. M. (2006). An approach for a personal information management system for photos of a lifetime by exploiting semantics. In (Bressan et al., 2006), pages 467–477.

Latif, K. and Rauber, A. (2006). Named entity recognition tools - a comparative analysis. Technical report, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria.

Latif, K. and Tjoa, A. M. (2006). Combining context ontology and landmarks for personal information management. In *Proceedings of IEEE International Conference on Computing & Informatics (ICOCI)*, Kuala Lumpur, Malaysia. IEEE Computer Society.

Latif, K., Weippl, E., and Tjoa, A. M. (2007). Question driven semantics interpretation for collaborative knowledge engineering and ontology reuse. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 170–176, Las Vegas, NV.

Lee, J., Chae, H., Kim, K., and Kim, C.-H. (2006). An ontology architecture for integration of ontologies. In Mizoguchi, R., Shi, Z., and Giunchiglia, F., editors, *Proceedings of Asian Semantic Web Conference (ASWC)*, volume 4185 of *LNCS*, pages 205–211, Beijing, China. Springer.

Lee, K.-H., Slattery, O., Lu, R., Tang, X., and McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology*, 107(1):93106.

Lee, R. (2004). Scalability report on triple store applications. Technical report, The SIMILE Project. http://simile.mit.edu/reports/stores/.

Lei, Y., Sabou, M., Lopez, V., Zhu, J., Uren, V., and Motta, E. (2006). An infrastructure for acquiring high quality semantic metadata. In (Sure and Domingue, 2006), pages 230–244.

Lenat, D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157.

Lopez, V., Motta, E., and Uren, V. (2006). Poweraqua: Fishing the semantic web. In (Sure and Domingue, 2006), pages 393–410.

Ludäscher, B., Marciano, R., and Moore, R. (2001). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *ACM SIGMOD Record*, 30(3):54–63.

Lyttleton, O., Sinclair, D., and Tracey, D. (2005). Mediating between heterogeneous ontologies using schema matching techniques. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 247–252, Las Vegas, Nevada. IEEE SMC Society.

Ma, L., Su, Z., Pan, Y., Zhang, L., and Liu, T. (2004). RStar: An RDF storage and query system for enterprise resource management. In *Proceedings of the 13th International Conference on Information and Knowledge Management(CIKM)*, pages 484–491, Washington D.C. ACM Press.

Maedche, A. (2002). *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers.

Maedche, A. and Zacharias, V. (2002). Clustering ontology-based metadata in the semantic web. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Proceedings of 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, volume 2431 of *Lecture Notes in Computer Science*, pages 348–360, Helsinki, Finland. Springer.

Markram, H. (2006). The Blue Brain project. *Nature Reviews Neuroscience*, 7(2):153–160.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2003). The WonderWeb library of foundational ontologies. Deliverable 18, WonderWeb Project.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., and Wilks, Y. (2002). Architectural elements of language engineering robustness. *Natural Language Engineering (Robust Methods in Analysis of Natural Language Data)*, 8(2-3):257–274.

Mazzocchi, S. (2005). Closed world vs. open world: the first semantic web battle. http://www.betaversion.org/ stefano/linotype/news/91/.

McBride, B. (2002a). Four steps towards the widespread adoption of a semantic web. In (Horrocks and Hendler, 2002), pages 419–422.

McBride, B. (2002b). Jena: a semantic web toolkit. *IEEE Internet Computing*, 6(6):55–59.

McCool, R. (2005). Rethinking the semantic web, part 1. *IEEE Internet Computing*, 9(6):86–88.

McCool, R. (2006). Rethinking the semantic web, part 2. *IEEE Internet Computing*, 10(1):93–96.

McIlraith, S. A., Plexousakis, D., and van Harmelen, F., editors (2004). *Proceedings of 3rd International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, Hiroshima, Japan. Springer.

McIlraith, S. A., Son, T. C., and Zeng, H. (2001). Semantic web services. *IEEE Intelligent Systems*, March/April:46–53.

Menczer, F. (2005). Mapping the semantics of web text and links. *IEEE Internet Computing*, 9(3):27–36.

Meziane, F. and Rezgui, Y. (2004). A document management methodology based on similarity contents. *Information Sciences*, 158:15–36.

Mika, P. (2005). Ontologies are us: A unifed model of social networks and semantics. In (Gil et al., 2005), pages 522–536.

Miles, A. and Brickley, D. (2005). SKOS core vocabulary specification. Working Draft 2nd, W3C. http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102.

Morris, P. (1994). *Introduction to Game Theory*. Universitext. Springer-Verlag, New York.

Motta, E. and Gibbins, N. (2003). AKT Reference Ontology. v2.0. http://www.aktors.org/publications/ontology/.

Muggleton, S. H. (2006). Exceeding human limits. *Nature*, 440(7083):409–410.

Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., and Risch, T. (2002). EDUTELLA: A P2P networking infrastructure based on RDF. In *Proceedings of the 11th International Conference on World Wide Web (WWW)*, pages 604–615, Honolulu, Hawaii. ACM Press.

Neumayer, R., Dittenbach, M., and Rauber, A. (2005). PlaySOM and PocketSOMPlayer: Alternative interfaces to large music collections. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 618–623, London, UK.

Nicolau, J. M. (1995). On thoughts about the brain. In Moreno-Díaz, R. and Mira-Mira, J., editors, *Brain Processes, Theories and Models*, pages 71–77. The MIT Press, Cambridge, Massachusetts.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of 2nd International Conference on Formal Ontology in Information Systems (FOIS)*, pages 2–9, Ogunquit, Maine. ACM Special

Interest Group on Artificial Intelligence (SIGART), ACM Press.

Niles, I. and Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Arabnia, H. R., editor, *Proceedings of International Conference on Information and Knowledge Engineering*, volume 2, pages 412–416, Las Vegas, Nevada. CSREA Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

O'Hara, K., Morris, R., Shadbolt, N., Hitch, G. J., Hall, W., and Beagrie, N. (2006). Memories for life: a review of the science and technology. *Journal of The Royal Society Interface*, 3(8):351–365.

Oren, E., Völkel, M., Breslin, J. G., and Decker, S. (2006). Semantic wikis for personal knowledge management. In (Bressan et al., 2006), pages 509–518.

Pampalk, E., Rauber, A., and Merkl, D. (2002). Using Smoothed Data Histograms for cluster visualization in Self-Organizing Maps. In *Proceedings of the International Conference on Artifical Neural Networks (ICANN)*, pages 871–876, Madrid, Spain. Springer.

Park, Y. (2004). Glossont: A concept-focused ontology building tool. In Dubois, D., Welty, C. A., and Williams, M.-A., editors, *Proceedings of the 9th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 498–506, Whistler, Canada. AAAI Press.

Pasley, J. (2005). How bpel and soa are changing web services development. *IEEE Internet Computing*, 9(3):60–67.

Patel, C., Supekar, K., and Lee, Y. (2003a). OntoGenie: Extracting ontology instances from WWW. In *Proceedings of ISWC Workshop on Human Language Technology for the Semantic Web and Web Services*, Florida.

Patel, C., Supekar, K., Lee, Y., and Park, E. K. (2003b). Ontokhoj: A semantic web prtal for ontology searching, ranking and classification. In *Proceedings of 5th CIKM International Workshop on Web Information and Data Management (WIDM)*, pages 58–61, New Orleans, Louisiana. ACM Press.

Patel-Schneider, P. F., Hayes, P., and Horrocks, I. (2004). OWL web ontology language: Semantics and abstract syntax. Recommendation, W3C. http://www.w3.org/TR/2004/REC-owl-semantics-20040210.

Patel-Schneider, P. F. and Horrocks, I. (2006). A comparison of two modelling paradigms in the semantic web. In (WWW, 2006), pages 3–12.

Payne, J. (2007). Thai king's archives to go online. Associated Press Writer. http://hosted.ap.org/dynamic/stories/T/THAILAND_ROYALTY_ONLINE (Accessed: June 12, 2007).

PIM-SIGIR (2006). *Proceedings of the SIGIR Workshop on Personal Information Management*, Seattle, Washington.

Pölzlbauer, G., Dittenbach, M., and Rauber, A. (2005a). A visualization technique for self-organizing maps with vector fields to obtain the cluster structure at desired levels of detail. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1558–1563, Montreal, Canada. IEEE Computer Society.

Pölzlbauer, G., Dittenbach, M., and Rauber, A. (2006). Advanced visualization of self-organizing maps with vector fields. *Neural Networks*, 19(6-7):911–922.

Pölzlbauer, G., Rauber, A., and Dittenbach, M. (2005b). Advanced visualization techniques for self-organizing maps with graph-based methods. In *Proceedings of the 2nd International Symposium on Neural Networks*, pages 75–80, Chongqing, China. Springer.

Potter, M. (2002). XML for digital preservation: XML implementation options for e-mails. In *Erpanet workshop on XML and Digital Preservation*, Urbino, Italy. http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=299.

Ranganathan, S. R. (1963). *Colon Classification: Basic Classification*. Asia Publishing House, Bombay, India, 6th edition.

Rauterberg, M., Menozzi, M., and Wesson, J., editors (2003). *Proceedings of 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, Zurich, Switzerland. IOS Press.

Ravasio, P., Schär, S. G., and Krueger, H. (2004). In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction*, 11(2):156–180.

Reeve, L. and Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of ACM Symposium on Applied Computing (SAC)*, pages 1634–1638, Santa Fe, New Mexico. ACM Press.

Rekimoto, J. (1999). Time-machine computing: a time-centric approach for the information environment. In *Proceedings of the 12th annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 45–54, Asheville, North Carolina. ACM Press.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworths.

Ringel, M., Cutrell, E., Dumais, S., and Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. In (Rauterberg et al., 2003), pages 184–191.

Rodríguez, A. and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456.

Sauermann, L., Kiesel, M., Fluit, C., Maus, H., Heim, D., Nadeem, D.,

Horak, B., and Dengel, A. (2006). Semantic desktop 2.0: The gnowsis experience. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *Proceedings of 5th International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 887–900, Athens, GA. Springer.

Sazedj, P. and Pinto, S. (2005). Time to evaluate: Targeting annotation tools. In Handschuh, S., Declerck, T., and Koivunen, M.-R., editors, *5th International Workshop on Knowledge Markup and Semantic Annotation*, volume 185, Galway, Ireland. CEUR-WS.

Schaffert, S., Gruber, A., and Westenthaler, R. (2005). A semantic wiki for collaborative knowledge formation. In *Proceedings of Semantics*, Vienna, Austria.

Schank, R. C. (1973). The fourteen primitive actions and their inferences. MEMO AIM-183 CS-TR-73-344, Stanford University, Department of Computer Science.

Schank, R. C., Kolodner, J. L., and DeJong, G. (1981). Conceptual information retrieval. In Oddy, R. N., Robertson, S. E., van Rijsbergen, C. J., and Williams, P. W., editors, *Proceedings of the 3rd ACM Conference on Research and Development in Information Retrieval (SIGIR'80)*, pages 94–116, Cambridge, UK. Butterworth & Co.

Schlicht, A. and Stuckenschmidt, H. (2006). Towards structural criteria for ontology modularization. In (Haase et al., 2006).

Schneiderman, B. and Kang, H. (2000). Direct annotation: A drag-and-drop strategy for labeling photos. In Banissi, E., Bannatyne, M., Chen, C., Khosrowshahi, F., Sarfraz, M., and Ursyn, A., editors, *Proceedings International Conference on Information Visualisation*, pages 88–95, London, England. IEEE Computer Society.

Seidenberg, J. and Rector, A. (2006). Web ontology segmentation: Analysis, classification and use. In (WWW, 2006), pages 13–22.

Sheth, A., Arpinar, B., and Kashyap, V. (2003). Relationships at the heart of semantic web: Modeling, discoverying, and exploiting complex semantic relationships. In Nikravesh, M., Azvin, B., Yager, R., and Zadeh, L., editors, *Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing*. Springer-Verlag.

Sheth, A. and Ramakrishnan, C. (2003). Semantic (web) technology in action: Ontology driven information systems for search, integration and analysis. *IEEE Data Engineering Bulletin (Making the Semantic Web Real)*, 26(4):40–48.

Shirky, C. (2005). Ontology is overrated: Links, tags, and post-hoc metadata. In *O'Reilly Emerging Technology Conference*, San Diego, CA.

Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching

approaches. *Journal on Data Semantics*, IV:146–171.

Sinha, R. (2005). A cognitive analysis of tagging (how the lower cognitive cost of tagging makes it popular). Thoughts on technology, design & cognition. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.htm.

Skupin, A. (2004). A picture from a thousand words. *Computing in Science and Engineering*, 6(5):84–88.

Smith, B. (2003). *Blackwell Guide to the Philosophy of Computing and Information*, chapter Ontology, pages 155–166. Blackwell Philosophy Guides. Blackwell Publishing.

Smith, B. (2005). How to do things with paper: The ontology of documents and the technologies of identification. Ontolog Forum. http://ontology.buffalo.edu/document_ontology/.

Smith, B. and Rosse, C. (2004). The role of foundational relations in the alignment of biomedical ontologies. In *Proceedings of 11th World Congress on Medical Informatics (MedInfo)*, volume 107 of *Studies in Health Technology and Informatics*, pages 444–448, San Francisco, CA. IOS Press.

Sorrows, M. E. (2004). *Recall of Landmarks in Information Space*. Phd dissertation, School of Information Sciences, University of Pittsburgh.

Storey, M.-A., Musen, M., Silva, J., Best, C., Ernst, N., Fergerson, R., and Noy, N. (2001). Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in protégé. In *Proceedings of the Workshop on Interactive Tools for Knowledge Capture*, Victoria, B.C.

Strang, T. and Linnhoff-Popien, C. (2004). A context modeling survey. In *Proceedings of the UbiComp Workshop on Advanced Context Modelling, Reasoning and Management*, Nottingham, England.

Strube, M. and Hahn, U. (1999). Functional centring - grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Stuckenschmidt, H. and Klein, M. C. A. (2004). Structure-based partitioning of large concept hierarchies. In (McIlraith et al., 2004), pages 289–303.

Stuckenschmidt, H. and van Harmelen, F. (2001). Ontology-based metadata generation from semi-structured information. In *Proceedings of International Conference on Knowledge Capture (K-CAP)*, pages 163–170, Victoria, British Columbia. ACM Press.

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., and Oberle, D. (2005). The SWRC ontology semantic web for research communities. In Bento, C., Cardoso, A., and Dias, G., editors, *Proceedings of 12th Portuguese Conference on Artificial Intelligence*, volume 3808 of *Lecture Notes in Artificial Intelligence*, pages 218–231, Covilha, Portugal. Springer.

Sure, Y. and Domingue, J., editors (2006). *Proceedings of 3rd European Semantic Web Conference (ESWC)*, volume 4011 of *Lecture Notes in Com-*

puter Science, Budva, Montenegro. Springer.

Teevan, J. (2004). How people re-find information when the web changes. MIT-CSAIL AI Memo 2004-012, Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA.

Tjoa, A. M., Latif, K., and Karim, S. (2006). Exploiting semantic web for open source software development: Opportunities and challenges. In *International Conference on Open Source Technologies (ICOST)*, Lahore, Pakistan. Al-Khwarizmi Institute of Computer Science (KICS), University of Engineering and Technology.

Towards2020 (2006). Roadmap: Towards 2020 science. Microsoft Corporation.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

Tzitzikas, Y. and Hainaut, J.-L. (2006). On the visualization of large-sized ontologies. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI)*, pages 99–102, Venezia, Italy. ACM Press.

Ultsch, A. (1999). Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 33–45. Elsevier, Amsterdam.

Ultsch, A. and Siemon, H. P. (1990). Kohonen's self-organizing feature maps for exploratory data analysis. In *Proceedings of the International Neural Network Conference*, pages 305–308, Paris, France. Kluwer.

Uschold, M. (1996). Building ontologies: Towards a unified methodology. In *Proceedings of 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*, Cambridge, UK.

Uschold, M. and Jasper, R. (1999). A framework for understanding and classifying ontology applications. In Benjamins, V. R., editor, *Proceedings of IJCAI Workshop on Ontologies and Problem-Solving Methods*, volume 18 of *CEUR Workshop Proceedings*, pages 11–1–11–12, Stockholm, Sweden.

Valarakos, A., Paliouras, G., Karkaletsis, V., and Vouros, G. (2004). Enhancing ontological knowledge through ontology population and enrichment. In Motta, E., Stutt, A., Shadbolt, N., and Gibbins, N., editors, *Proceedings of 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume 3257 of *Lecture Notes in Computer Science*, pages 144–156, Whittlebury Hall, UK. Springer.

van Ossenbruggen, J., Troncy, R., Stamou, G., and Pan, J. (2006). Image annotation on the semantic web. Technical report, W3C Semantic Web Best Practices Working Group. http://www.w3.org/2001/sw/BestPractices/MM/image_annotation.html.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126.

Volz, R., Oberle, D., Staab, S., and Motik, B. (2003). KAON SERVER - a semantic web management system. In (WWW, 2003).

von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, 39(6):92–94.

Walker, A. (2003). Lightweight english heavyweight inference and a semantic distance measure. In *NIST/NSF International Workshop on Semantic Distance*.

Wang, B., McKay, B., Abbass, H., and Barlow, M. (2003). A comparitive study for domain ontology guided feature extraction. In Oudshoorn, M., editor, *Proceedings of 25th Australian Computer Science Conference (ACSC)*, volume 16 of *Conferences in Research and Practice in Information Technology*, Adelaide, Australia. Australian Computer Society.

Weinstein, P. and Alloway, G. (1997). Seed ontologies: growing digital libraries as distributed, intelligent systems. In *Proceedings of the 2nd International Conference on Digital Libraries (DL)*, pages 83–91, Philadelphia, Pennsylvania. ACM Press.

Wolfe, S. and Keller, R. (2005). Workspaces in the semantic web. In *Workshop on Contexts and Ontologies: Theory, Practice and Applications (AAAI-05)*, Pittsburgh, PA. http://sciencedesk.arc.nasa.gov.

Wong, P. C., Hetzler, B., Posse, C., Whiting, M., Havre, S., Cramer, N., Shah, A., Singhal, M., Turner, A., and Thomas, J. (2004). IN-SPIRE InfoVis 2004 Contest Entry. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 216–217, Washington, DC. IEEE Computer Society.

Wood, D., Gearon, P., and Adam, T. (2005). Kowari: A platform for semantic web storage and analysis. In *Proceedings of XTech Conference*, Amsterdam, Netherlands. http://www.idealliance.org/proceedings/xtech05/papers/04-02-04/.

Wurman, R. S., Sume, D., and Leifer, L. (2000). *Information Anxiety 2*. Que.

WWW (2003). *Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary. ACM Press.

WWW (2006). *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland. ACM Press.

Yang, M. (1993). *COMIS - A Conceptual Model for Information Systems*. PhD thesis, Faculty of Electrical Engineering and Computer Science, Norwegian Institute of Technology, Trondheim, Norway.

Zhang, L., Pan, Y., and Zhang, T. (2004). Focused named entity recognition using machine learning. In *Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 281–288, Sheffield, UK. ACM Press.