An evaluation of

Topic Maps

A Master’s Thesis in Computational Linguistics
Göteborg University
May 2002

Anna Carlstedt Mats Nordborg

Supervisors:
Stina Ericsson, Department of Linguistics, Göteborg University
Kerstin Forsberg, AstraZeneca, Information Architecture, R&D IS

Abstract

This Master’s thesis looks at a technology for describing knowledge structures and associating them with information resources, namely Topic Maps. Topic Maps are an ISO/IEC standard that enables the structuring and navigating of large amounts of information. The Topic Map technology is based on the ways of structuring knowledge that can be found in indexes, thesauri, glossaries and semantic networks.

In the thesis we first present the theoretical background to Topic Maps and look at the Topic Map standard itself in some detail. We also present some related technologies and initiatives: Information Retrieval (IR) and the Semantic Web. We then describe our modelling and implementation of a Topic Map over a pharmaceutical research area. The implementation was made in an XML standard called XTM (XML Topic Maps). Also included is the testing of an automated categoriser of documents, Autonomy’s Categorizer, and the possibility of using this in connection with Topic Maps. Lastly we discuss the problems we encountered during our work and put forward some ideas and some possible scenarios for future development of Topic Map usage.

Sammanfattning

Denna magisteruppsats behandlar en teknik för att beskriva kunskapsstrukturer och förbinda dem med informationsresurser, som heter Topic Maps. Topic Maps är en ISO/IEC-standard som gör det möjligt att strukturera och navigera i stora informationsmängder. Tekniken baseras på de sätt att strukturera kunskap som finns i index, thesauri, ordlistor och semantiska nätverk.

I uppsatsen presenterar vi först den teoretiska bakgrunden till Topic Maps och beskriver själva Topic Map-standarden ingående. Vi går också igenom några besläktade tekniker och initiativ: Information Retrieval (IR) och Semantic Web. Sedan beskriver vi vår implementation av en Topic Map över forskningen inom ett farmaceutiskt område. Implementationen är gjord i en XML-standard som heter XTM (XML Topic Maps). Dessutom innehåller uppsatsen en test av en automatisk dokumentkategorisering, Autonomys Categorizer, och möjligheten att använda denna tillsammans med Topic Maps. Slutligen diskuterar vi de problem vi haft under vårt arbete och presenterar några ideér och möjliga scenarion för framtida utveckling av användningen av Topic Map.

Contents

1 Introduction

For the ever-increasing amount of information within a company an easy way for classifying, indexing and searching that information is needed. Within a company there is often a number of different, clear structures such as product families, market segmentation, business processes and different project phases, which all form the basis for structuring information in the operational activity. Other types of structures are less clear, either because they are dynamic or because they are embedded in different professions, vocabularies and local practice.

In AstraZeneca there is a need for tools to share, search, navigate and reuse information from a research area's perspective, that correspond to an exploring research issue, rather than the needs of an operational activity. There is a need for the information to be structured in different ways for different kind of views and sometimes structures within limited research areas need to be made clear.

A potential way of doing this that makes the structures machine processable and possible to navigate is Topic Maps. This thesis will look at the possibility of using Topic Maps as a tool for this and make an implementation where they are used. The implementation will extract information from a pharmaceutical research project to establish a Topic Map covering a limited set of research terms and their relationships, combined with a Topic Map covering a limited business vocabulary. When the knowledge structure is considered complete a number of documents within the research project are categorised, according to the established topics and their associations.

The result is an implemented domain specific Topic Map that is intended to function as a prototype for a component in a portal. It will serve as a map of the knowledge in a community-of-interest. It will investigate possible ways of navigating and making visible the concepts inherent in this community-of-interest.

The first part of this thesis looks at the theoretical background of Topic Maps and reviews related concepts such as taxonomies, thesauri, ontologies and semantic networks. It compares Topic Maps to other technologies in related areas and to similar efforts. Topic Maps themselves are explained in some detail. In the next part of the thesis our implementation of a Topic Map, covering the apolipoprotein research area, is discussed. Finally improvements and future development of Topic Maps are discussed. A few possible scenarios for Topic Map usage are described.

2 Theoretical Background to Topic Maps

In this chapter we will look at the theoretical background of Topic Maps. We will begin by looking at knowledge management as a way of generating, codifying and transferring knowledge. Topic Maps are here relevant as a way of codifying knowledge. After that we will look at the different ways of structuring knowledge that have formed the basis for the Topic Map ideas. We will continue by looking at Topic Maps themselves and give an introduction to the elements included in a Topic Map. We will conclude by looking at a few initiatives and technologies that are related to Topic Maps or to the general ideas behind Topic Maps, namely the Semantic Web and Information Retrieval.

2.1 Knowledge Management

In today's world of ever increasing information flows it becomes more and more important for organisations to find an effective way to handle not only the information excess, but also the inherent knowledge. According to Mack, Ravin and Byrd (2001) an important part of this knowledge comes from "*knowledge work* – solving problems and accomplishing goals by gathering, organizing, analyzing, creating, and synthesizing information and expertise." The people doing this knowledge work are called *knowledge workers*. They accumulate and create knowledge by sharing it with colleagues and communities-of-interest[1]. Further Mack, Ravin and Byrd (2001) says that "*knowledge management* (KM) refers to the methods and tools for capturing, storing, organizing, and making accessible knowledge and expertise within and across communities". Sigel (2000) says that "KM has to ensure strategically that all important knowledge assets and flows are known, utilized and enhanced according to their respective long-term contribution to the business value". Another related area is knowledge organisation (KO) which according to Sigel (2000) "is interested in optimizing the organization (the conceptual access structure) of knowledge repositories to support easier retrieval, creation and sharing of knowledge for user communities".

To be able to effectively access information and get in contact with knowledgeable persons within an organisation, it is important to be clear what the difference really is between information and knowledge, and data. This is so that all three can be taken care of and treated according to their different content and context. According to Davenport (1998) the difference is as follows:

- "Data is a set of objective facts about events." Data in itself contains no meaning; it is merely a description of what has happened. It provides no interpretation or judgement of the event.
- Information is a message and as such has both a sender and a receiver. "Information is meant to change the way the receiver perceives something." It is the receiver, not the sender who decides whether the message is information or merely data, on the basis of whether it changed the receiver's perception of things.
- Knowledge on the other hand is "a fluid mix of framed experience, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates

[1] For a definition of the term *community-of-interest* see *South Bank University Knowledge Library Glossary*: http://litc.sbu.ac.uk/klibrary/glossary.html
Accessed 2002-04-03

> and is applied in the mind of knowers in organizational routines, processes, practices and norms."

Knowledge derives from information just as information derives from data. The biggest difference between the three is that data can be found in records, information in messages and knowledge in individuals (knowers). Knowledge is not easy to pinpoint and it is even harder to translate into digital form.

Once the distinction is understood, the next step for effective KM is to provide a technical solution. It is easy to provide access to data, but much harder to provide access to information or knowledge. How are people within the organisation going to access the information and knowledge that is there? Knowledge management involves connecting people with people, as well as people with information, and the technical solution must take that into account. It is also important to set up the KM system in such a way that it allows for both personal meetings and supports occasions when such meetings are not possible. Davenport (1998) says that ideally "information technology enables the knowledge of an individual or group to be extracted and structured, and then used by other members of the organisation…"

It is also important to realise that it is not enough to provide relevant technology for information to be easily accessed. People are more likely to be happy with local knowledge if that is easy to find, even if it is not as relevant as more distant knowledge, than to go through the effort and uncertainty of trying to find out who in the organisation might know more. In a big organisation this might mean that many things are done again and again just because the information is not easily found or accessed.

Even if distant information is easy to access and it is straightforward to find the expert on the subject within the organisation, there are complicated social and cultural structures behind knowledge sharing. For people to want to share their knowledge with other people they have to have the time to do it and the organisation must have a culture and style that promotes communication and sharing.

Most organisations have informal groups or communities-of-interest consisting of friends or colleagues in which knowledge is shared through informal face to face meetings. These networks of friends or colleagues very often provide good, if not all information on areas of interest to the members. The communities-of-interest have the advantage over digital information that they are always updated. Digital information on the other hand is outdated as soon as it is published. It is very difficult to find appropriate technical support for informal networks, but much knowledge is captured if one finds it. Davenport (1998) says that "when networks of this kind share enough knowledge in common to be able to communicate and collaborate effectively, their ongoing conversation often generates new knowledge…"

Once knowledge structures and networks of knowledge workers are detected in an organisation, the technical solution for storing and accessing it is in place, and a culture and style that promotes communication and sharing is initiated, a codification of knowledge can begin. According to Davenport (1998) "the aim of codification is to put organizational knowledge into a form that makes it accessible to those who need it." Encoded knowledge, i.e. information, exists in many forms in an organisation, both as highly structured (SQL tables and XML messages etc.) and unstructured (web pages, Word documents etc.). In both the technical solution and the codification this has to be taken into consideration. Different methods are needed to provide access to different kinds of knowledge (/information/data).

It is impossible to code all knowledge in an organisation. Therefore a selection as to what is useful and relevant knowledge has to be made. It is important not to choose only knowledge that seems immediately useful however. A careful consideration is necessary. Another important issue is that it is not enough simply making knowledge generally available. There has to be a certain goal to be achieved with the knowledge. Just because it *can* be accessed it doesn't mean that it *will* be, and then the whole codification process has been in vain.

Once knowledge is selected and codified it has to be structured. The key issue in KM is how the knowledge is structured. It is the structuring that decides whether the knowledge will be accessible or not. The structure must make it easy for the users to find what they are looking for, and ideally show them how their area of interest connects to other. An example of this is a knowledge map. It points to knowledge but does not actually contain any documents. The Knowledge Map may point out both people and specific documents. The important thing is that the map points people to where they can find knowledge, not that it points out the actual data. A knowledge map can also function as a map of what resources exist in an organisation. Such knowledge structures can be explored through a portal. It aims to connect users to the information they are looking for, be it documents or human experts.

2.1.1 Portals

The term *portal* is used for a number of different applications. According to Dias (2001) what is now called a portal was in the beginning referred to as a *search engin*e. This was typically a Web-based single point of access to on-line information, with a Boolean search technology applied to HTML-documents. It then evolved to become a *navigation site*. In this stage categorisation was added: popular sites and documents were grouped into different categories based on their content, to enable users to find what they were looking for in an easier and faster way. The third phase of the development is where we are now: the sites are referred to as *portals*. Portals can then be divided into two different categories:

- Corporate portals
- Internet portals

The main difference between Internet portals and corporate portals is that Internet portals are available to anyone, whereas corporate portals are available only to the knowledge workers in an organisation (a comparison can be made to the difference between the Internet and Intranets).

A portal, both corporate and Internet, is an entry point or starting site for either the World Wide Web or for corporate resources. It aims to provide an easy way for its users to find the information they want. It can also provide various other services such as e-mail, chat rooms or message boards, personalised news, or personalised community portal pages[2].
A corporate portal provides a single access point to information found throughout the organisation and often to external information as well. It often provides a browsable

[2] According to *Auburn University Helpdesk Glossary* http://www.auburn.edu/helpdesk/glossary/portal.html
a *portal* is:
"an entry point or starting site for the World-Wide Web, combining a mixture of content and services and attempting to provide a personalized "home base" for its audience with features like customizable start pages to guide users easily through the Web, filterable e-mail, a wide variety of chat rooms and message boards, personalized news and sports headlines options, gaming channels, shopping capabilities, advanced search engines and personal homepage construction kits."
Accessed 25/02/02

topic hierarchy as well as an ordinary key-word search. Another possible feature of the corporate portal is support of user communities and also person-to-person matching. As the user searches for information the user profile is automatically calculated and matched with other users. This enables the user to find other users with similar interests or to find expert users in the same field.

The main aim with the corporate portal is to supply users (knowledge workers) with a way of selecting the information they want, instead of just providing access to *all* information. The corporate portal is one way to try to solve the Infoglut[3] problem. Because of the main aim of corporate portals of supplying information for knowledge workers, they are often called knowledge portals. Corporate portals also "…promote the gathering, sharing, and dissemination of information throughout the enterprise." says Detlor (2000).

The corporate portal is a software application that allows many different components to be presented in a homogeneous style. These different components are called portlets[4]. According to Detlor (2000) a portal's primary function is not to contain information but to provide access to information already available elsewhere in the organisation. Portals provide a way of pulling together all the various computer technologies throughout an organisation.

2.1.2 AstraZeneca Portal Project

An AstraZeneca Portal project, Clinical Informatics Forum, researches the portal concept within Informatics. The vision for the Informatics Forum project is to increase creativity and effectiveness by enabling clinical researchers to share scientific information globally, simultaneously and in real-time. The project focuses on seamless navigation among clinical documents and clinical data across Development R&D while making interaction between researchers easy. It exploits tools to share, search, navigate and reuse information from a research area's perspective, that correspond to an exploring research issue, rather than the needs of an operational activity. Other types of structures than the business driven ones are needed. These seem to be less documented, either because they are dynamic, or because they are embedded in different professions' vocabularies and local practice. The apolipoprotein community-of-interest (see below) is the first initiative to be used in the Clinical Informatics project.

The Topic Map implemented as part of our thesis work is intended to function as a prototype for a portlet in the Clinical Informatics Forum project portal. It will serve as a map of the knowledge structures in the apolipoprotein community-of-interest. It will investigate possible ways of navigating and making visible the concepts inherent in this community-of-interest.

[3] Infoglut is a state of voraciously gathering information, with little or no care for its quality or relevance, closely related to information overload.

[4] According to apache.org (http://cvs.apache.org/viewcvs/jakarta-jetspeed/proposals/portletAPI/) *portlets* are "designed to be aggregatable in the larger context of a portal page. They rely on the portal infrastructure to function, e.g. access to user profile information for the current user, access to the window object that represents the window in which the portlet is displayed, participation in the portal window and action event model, access to web client information, inter-portlet messaging and a standard way of storing and retrieving per-user or per-instance data persistently." Accessed 06/05/02

2.1.3 Apolipoprotein Research Project

The Apolipoprotein Research Project at AstraZeneca examines the risk markers for death and myocardial infarctions (MI) related to abnormal lipids. It is based on AMORIS - Apolipoprotein MOrtality RISk, a study which followed up mortality in 175553 healthy Swedish men (98722) and women (76831) between 1985-1996. Levels of the apolipoproteins apoB, apoA-I, and the ratio of apoB/apoA-I were discovered to be significant predictors of risk of death from acute MI in both sexes.

Cholesterol and triglycerides are transported in blood by proteins called apolipoproteins (apo). ApoB transports the atherogenic lipoproteins (VLDL, IDL, and LDL). There is one apoB per lipoprotein particle. Thus, the apoB value can be used to indicate a number of potentially dangerous, atherogenic particles. ApoA-I transports the protective, anti-atherogenic HDL particles. An imbalance between too many apoB and too few apoA-I particles, determined as the apoB/apoA-I ratio, is a strong cardiovascular risk factor.

2.2 Structuring knowledge

Many different solutions have been found to the problem of how to structure the knowledge of the world and just as many have been rejected. There are however, a few that have prevailed and made a big impression on all areas where knowledge is abundant. In the following chapter we will look at a number of these, namely:

- Ontologies
- Taxonomies
- Thesauri
- Indexes
- Glossaries
- Semantic networks

These form the idea basis on which the Topic Map methodology is built. Another thing all of these have in common is that areas where knowledge structures are important have borrowed all of these expressions and put them to use - sometimes a new and different use. Information Technology is such an area. The word *taxonomy* for example, is often used in a number of ways, some differing quite a lot from the initial. The definitions found below are both the "originals", not typical for the IT world, and other, more IT related ones. Where we felt it necessary, examples of usage have been added.

2.2.1 Ontologies

Ontology[5] is a discipline of Philosophy that deals with what kinds of things exist - what entities there are in the universe. It is a branch of metaphysics, the study of first principles or the essence of things. The word ontology means, according to Russell and Norvig (1995), "a particular theory of the nature of being or existence." It derives from the Greek *onto* (being) and *logia* (written or spoken discourse).

In Information Technology, an ontology is the working model of entities and interactions in some particular domain of knowledge or practices, such as electronic commerce. According to Checkland and Holwell (1998) it is nearly synonymous with conceptual modelling in databases, Highlevel Business Analysis and Soft Systems Methodology. Jacobson, Ericsson and Jacobson (1995) compares it to domain modelling in Object Oriented Design.

In artificial intelligence (AI), an ontology is the specification of conceptualisations, used to help programs and humans share knowledge. In this usage, an ontology is a set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

In topic map terminology, an ontology is a precise description of the kinds of things which are found in the domain covered by the topic map: in other words, the set of topics that are used to define classes of topics, associations, roles, and occurrences.

[5] For a definition of the term *ontology* see *whatis?com IT-specific encyclopedia*
http://whatis.techtarget.com/definition/0,,sid9_gci212702,00.html
Accessed 12/02/02

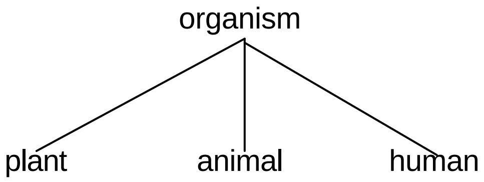
2.2.2 Taxonomies

Taxonomy[6] (from Greek *taxis* meaning arrangement or division and *nomos* meaning law) is the science of classification. A taxonomy is a classification system, organising entities/concepts. It is a way to show the relationships between these concepts. Where the ontology describes a world or domain and the different concepts that make up this world, a taxonomy describes the relationships between the concepts. An ontology completely describes a domain, whereas a taxonomy only indicates class/subclass memberships.

A good taxonomy shows only one dimension or aspect of things. The categories in the taxonomy should be mutually exclusive. One individual concept should be found in one place only in the taxonomy. The taxonomy should be exhaustive; all possibilities should be included.

One of the best-known taxonomies is the one devised by the Swedish scientist, Carolus Linnaeus classifying living organisms. His books are considered the beginning of modern botanical and zoological nomenclature.

The work involved in creating an ontology is essentially a classification, according to Blackburn and Bos (1999). One has to divide the concepts that make up a world or domain into different classes. One has to look at the relationships that hold between the different concepts that make up the world or domain. These relationships say a lot about the meaning of the individual concepts. For making an ontology also making a taxonomy therefore is necessary. One cannot look at what concepts make up a world without looking at the relationships between those concepts. An example of an ontology with a taxonomy could look as follows:



Blackburn and Bos (1999)

Some of the classes will be disjoint, i.e. they are mutually exclusive (an example of this is plant/animal). The classes will also inherit properties from other classes above them. For example any animal in the above tree-structure will also be an organism (but cannot at the same time be both plant and animal). All daughters of a mother node are disjoint. All the daughter nodes inherit information from the mother nodes. A daughter node can have more properties than its mother node, but it will at least have all the properties of its mother node.

In short, an ontology describes all concepts that exist within a world/domain. A taxonomy describes the relationships between the concepts. An ontology and a taxonomy taken together is a hierarchical structure of concepts that represent a world or a domain. The nodes of the structure denote concepts and the arches between the nodes denote the relationships between the concepts.

[6] For a definition of the term *taxonomy* see *whatis?com IT-specific encyclopedia*
http://whatis.techtarget.com/definition/0,,sid9_gci331416,00.html
Accessed 12/02/02

2.2.3 Thesauri

A thesaurus is a network of interrelated terms within a particular domain. The network contains cross-references and gives the associations between terms. Thesauri depend upon the concept of a controlled vocabulary in order to describe the preferred term when many synonyms are available. Given a certain term the thesaurus indicates which terms mean the same, which denote a broader category and which terms denote a narrower. It also indicates which are related in some other way. It may contain other features such as definitions and examples of usage, but the main feature is the interrelation between terms. According to the ISO 2788, 1986:2 standard (1986) "a thesaurus is the vocabulary of a controlled indexing language formally organized so that the a priori relationships between concepts (for example as "broader" and "narrower") are made explicit." Thesauri are designed to help users find a particular word when they have a concept in mind, whereas dictionaries are designed to give users information about unfamiliar concepts.

2.2.4 Indexes

An index[7] is an alphabetised list of names, places, and subjects treated in a printed work, giving the page or pages on which each item is mentioned. It lists the topics covered in the printed work with, ideally, all possible names for each topic. It shows the occurrences of a topic, both by giving the actual page where it can be found and by giving *see* and *see also* references, where the first one allows multiple entrances to the same topic and the latter points out related topics. Another possible feature is showing what type of occurrence is indicated, by differences in font (e.g. bold font meaning main entrance for a specific topic) or through the use of explanatory labels (e.g. Tosca (opera) and Tosca (character)).

2.2.5 Glossaries

A glossary[8] is a list of often difficult or specialised words with their definitions, often placed at the back of a book. Instead of pointing to an occurrence of a topic like an index it just gives the definition of the topic. It may contain additional information such as *see* and *see also* references or give guidance about language use or pronunciation.

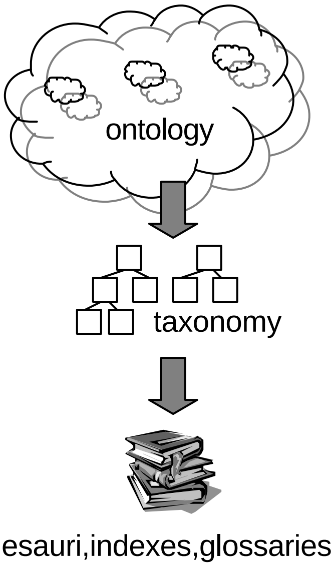
2.2.6 Bringing it all together

There is a strong connection between all the above ways of expressing knowledge and how it is structured. An ontology is what describes what concepts exist in a world. In it all concepts relevant for a certain domain are defined. The taxonomy goes on to clarify the relationships between these concepts. What are the relationship between them? How are they grouped together? The concepts and their relationships are described and defined in different ways in thesauri, glossaries and indexes. Different aspects of the knowledge structures are implemented in each of

[7] For a definition of the term *index* see *The American Heritage Dictionary of the English Language: Fourth Edition*
http://www.bartleby.com/61/7/I0100700.html
Accessed 25/02/02

[8] For a definition of the term *glossary* see *The American Heritage Dictionary of the English Language: Fourth Edition*
http://www.bartleby.com/61/66/G0156600.html
Accessed 25/02/02

them. Thesauri and indexes show the knowledge structure, i.e. the ontology and the taxonomy, whereas a glossary gives the definitions of the concepts.



The relationship between ontologies, taxonomies and thesauri, indexes and glossaries.

2.2.7 Relevance for Topic Maps

A Topic Map is a map over the knowledge that can be found in a document base. It shows the relevant concepts and the relationships between them, in a way similar to that of a thesaurus or an index. It also gives the definition of concepts like a glossary. It aims to arrange the concepts in an ontology and a taxonomy.

Topic Maps aim to take the structures found in all the above and use the different techniques in them to make the structures machine processable and possible to navigate. TMs also provide advanced techniques for linking and addressing the knowledge structure and the document base.

2.2.8 Semantic Networks

In semantic networks, objects are represented as nodes in a graph, with relations between objects being represented by named arcs. The nodes are organised in a taxonomic structure and the arcs represent binary structures. Everything that can be expressed in first order logic can also be expressed as a semantic network. An example of a semantic network, WordNet[9], which is relevant for Topic Maps, is discussed below.

WordNet

WordNet[10] is a semantic dictionary that is built as a network. Its design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organised into four separate semantic nets which are

[9] For more information on *WordNet* see http://www.cogsci.princeton.edu/~wn/index.shtml
[10] There is a Swedish version of WordNet called *SwordNet*. For more information see http://www.ling.lu.se/projects/Swordnet/

then organised into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

According to Fellbaum (1999), WordNet contains compounds, phrasal verbs, collocations, and idiomatic phrases, but the word is the base unit. The words are not decomposed into smaller meaningful units. It also contains short phrases, such as *bad person*, that are not paraphrasable by a single word. These short phrases reflect lexical gaps in English, but may well exist as single words in other languages. The gaps come from WordNet's relational structure, which sometimes connects two concepts via a third that is not a word in English.

WordNet organises words in groups according to their meaning (semantic properties) rather than according to their spelling. Since the relationship between words and meanings is not always a one-to-one relationship, words are represented as sets of synonyms or *synsets* in WordNet. This is to disambiguate their meaning. A word can mean more than one thing (polysemy). The word *design* can mean both *a sketch* and *to create something*. In WordNet the synsets show this difference through pairing the word *design* with synonyms:

{design, sketch}
{design, create}

A concept or meaning can also have more than one word that describes it (synonymity). For example the words *phone* and *telephone* describe the same thing.

The semantic relationships between synsets in WordNet are among other antonymy, hyponymy, hypernymy, and meronymy. Antonyms are words that have opposite meanings. For example, *black* and *white* are antonyms of one another. Usually there is only one antonym for a given meaning of a word. Hyponyms are words which are "contained" by another word. For example, *car* is a hyponym of *vehicle*; all cars are vehicles, but not all vehicles are cars. Hypernyms are words which "contain" another word. In the above example *vehicle* is the hypernym of *car*. Meronyms are words which are parts (or mem bers) of a given word. For example, *engine* is a meronym of *car*.

WordNet as a Thesaurus

WordNet's design is similar to the design of a thesaurus in the respect that both organises words according to what concepts they belong to rather than according to their initial letter. Unlike a thesaurus, though, WordNet makes the relationships between words and concepts explicit, by labelling them. WordNet also shows concepts that are not yet lexicalised, through its use of short phrases, which a thesaurus does not.

WordNet as a Dictionary

WordNet in some ways resembles a traditional dictionary. It gives definitions of words and examples of usage for most synsets.

WordNet and Topic Maps

There are many similarities between WordNet and Topic Maps. Both consist of nodes and arcs, where the nodes denote concepts and the arcs denote the relationships between the concepts (this is true for any semantic network). The main difference is that in a TM a node (topic) can be anything one chooses to define as a topic, whereas in WordNet the nodes consist of synsets. Similarly the arcs in a TM can be anything that associates topics, but in WordNet, the arcs denote one of a number of semantic relationships. The concept behind both WordNet and TMs is the same, though. The aim is to show the meaning of different concepts through the contexts they can be found in. The relationships between different concepts identify their meaning. TMs have taken this a step further and allows *occurrences* (see page 16), i.e. pointing out information resources that contain data on the different concepts.

Topic Maps are much more general than WordNet, but then the expected usage is totally different. WordNet is a semantic dictionary built as a semantic network, whereas a Topic Map can be used to describe any domain with anything described as topics and the associations between them. Topic Maps use the ideas of WordNet (or any semantic network) just as they use the ideas behind thesauri and glossaries. The basic model of semantic networks, with concepts represented as nodes and the relationships between the concepts represented as arcs, is very similar to that of the topics and associations found in Topic Maps.

2.3 Topic Maps

Topic Maps (TM) are an ISO standard (ISO/IEC 13250:2000) which provides a standardised notation for representing information about the structure of information resources used to define topics, and the relationships between topics. A set of one or more interrelated documents that employs the notation and grammar defined by the ISO/IEC 13250 International Standard is called a 'topic map'. In general, the structural information conveyed by topic maps includes:

- groupings of addressable information objects around topics (occurrences)
- relationships between topics (associations).

According to ISO/IEC 13250 (2000) a topic map defines a multidimensional topic space, a space in which the locations are topics, and in which the distances between topics are measurable in terms of the number of intervening topics which must be visited in order to get from one topic to another, and the kinds of relationships that define the path from one topic to another, if any, through the intervening topics, if any.

Put in another way the standard can encode knowledge structures and associate them with information resources. As such they constitute an enabling technology for knowledge management, according to Pepper (2000). Biezunski (1999) also points out that they are tools for organising information in a way that is optimised for navigation in the information resources. A topic map organises large sets of information and builds a structured semantic linked network over the resources.

The Standard was set in early 2000, but the origins of the TM paradigm date itself back to 1993 when it was first expressed in the context of the Davenport Group[11]. The paradigm was more fully developed thereafter in the context of the GCA Research Institute (now known as IDEAlliance[12]) and after that it was independently developed, implemented and promulgated. Early in 2000, after several years of continuous efforts by an international group of individuals the paradigm became the ISO/IEC 13250 standard that today is known as Topic Maps.

So what can the TM standard be used for? As Steve Newcomb (a pioneer of TM) points out, in an article by Pepper (1999), the original motivation was to see to the need of merging indexes of different sets of documentation. The insight Newcomb arrived at was that:

> “ Indexes if they have any self-consistency at all, conform to models of the structure of the knowledge available in the material that they index. But such models are implicit, and nowhere to be found. If such models could be captured formally, then they could guide and facilitate the process of merging indexes. “

The standard was extended to cover more navigational aids such as tables of content, glossaries, thesauri, cross references, etc. Common to all these organising theories is the attempt to provide access to information based on the model of the knowledge they contain, and in the middle lies the concept of the topic (see below). Today the topic paradigm appears to have a broader applicability. It seems, according to Pepper (1999), that in many contexts it is the fundamental organising principle for the creation and maintenance of information besides the basis for

[11] The Davenport Group has merged with OASIS. OASIS, Organization for the Advancement of Structured Information Standards, creates interoperable industry specifications based on public standards. http://www.oasis-open.org/committees/docbook/

[12] IDEAlliance, developer of standards. http://www.idealliance.org/

navigational aid. This is definitely the case in reference works publishing and in legal publishing, but it seems that the principle can be equally useful in all commercial branches.

2.3.1 The TAO of Topic Maps

While it is possible to build immensely complex structures with TMs, the basic concepts of the TM model, Topics, Associations, Occurrences (TAO), are easily grasped and will be explained in the next sections. Other concepts of the model are those of Scope, Public Subject and Facet, which will be explained as well.

Topics (T)

Topics are clearly the most fundamental concept in a TM. A topic, as expressed in an article by Pepper (2000), in its most generic sense, can be any “thing”, regardless of whether it exists or have any other characteristics, about which anything whatsoever may be asserted by any means whatsoever. For example a person, an entity, a concept - anything. This is how the ISO/IEC 13250 (2000) defines the term Subject, the term that is used for the real world thing. The link between the subject and the topic is what the author had in mind when it was created. One could say that a topic reifies[13] a subject, that is, makes it real for the system. So to reify a topic is to create it in for example XML or Java. If one for example would like to represent the members of an organisation one would probably like to have titles and the individual persons as topics.

Topic types

Topics can be categorised according to their kind. In a TM, any given topic is an instance of zero or more topic types. A topic type is one of the classes a topic can belong to. Topic types represent a typical class-instance relationship. Pepper (1999) says that what one chooses to regard as topics in an application will vary according to the needs of the application, the information and the uses to which the TM will be put. Topic types are themselves defined as topics by the standard. Continuing the example from above, one might say that a topic type would be ‘title’ and topics of this type could be ‘manager’, ‘secretary’ and ‘president’.

Topic names

Topics have explicit names, since that gives us a way to talk about them. The TM standard does not try to enumerate all different types of names (e.g. formal names, nicknames etc.). Instead it recognises the need for some forms of names to be defined in a standardised way, in order for applications to be able to do something meaningful with them, and for complete freedom to define application-specific name types, according to Pepper (2000). Each name may exist in multiple forms. A name always has exactly one base form, and it may, in addition, have one or more variants for use in specific processing contexts, e.g. one name for the processing context of English and another for Swedish, according to ISO/IEC 13250 (2000). Therefore the standard provides an element form for topic names, which includes the following kinds: base name (required), display name (optional) and sort name (optional).

[13] *Reify* is to regard (something abstract) as a material or concrete thing. Synonym: *Entify*, from *Entity*.

	Context – English	Context – Swedish
Base name	secretary	sekreterare
Display name	Secretary	Sekreterare
Sort name	position3	position3

Example of name types

Occurrences (O)

An occurrence is any information that is specified as being relevant to a given subject. Either it could be a reference to an article about the topic, or a picture that describes the topic, or a simple mention of a topic in the context of something else. Occurrences of this type are mostly outside of the TM, but could be inside. They are pointed at using any pointing-mechanism that the system supports, says Pepper (2000). One advantage here is that the documents themselves do not have to be touched, as opposed to the systems used today to create similar structures. This means that the TM is separated from its occurrences into two layers. This is one of the clues to the power of TMs, which makes them portable and possible to apply on different information resources. According to Pepper (1999) other systems often use some mark-up in the documents to be indexed (i.e. a bottom-up approach instead of TMs top-down approach).

Occurrence roles

As mentioned above, occurrences may be of different types (reference to an article or a mention of a topic). These distinctions are supported in the standard by the concepts of occurrence roles and occurrence role types. In general terms the distinction between the two is small but subtle. Pepper (2000) says that both are about the same thing, namely the way in which the occurrence contributes information to the subject in question, though the role is simply a mnemonic and the type is a reference to a topic in the TM which further characterises the relevance of the role.

Associations (A)

The previously discussed concepts of topic, topic type, topic name, occurrence and occurrence role allow us to organise our information resource according to topics, but not much more. But the standard gives us a way to describe relations between topics through associations in a construct called topic association. As described in ISO/IEC 13250 (2000) a topic association specifies a relationship among specific topics (e.g. that Person *is founder of* Organisation or Organisation *is located in* Place). A topic association is a link between topics, each of which plays a role as a member of that association. This is important because a relationship that holds between topics is probably interesting even without the given context that the topic and its associations were created for (e.g. City *is located in* Country, that a city is located in a country relation is valid in most contexts). They are completely independent of whatever information resource that may or may not exist or is considered as occurrences of those topics, says Pepper (1999).

Association types

As topics and occurrences can be groped according to type, so too can associations be groped according to their type. From the example above we would get the following association types: is_founder_of and is_located_in. Association types like these are themselves defined as topics. In the above example is_founder_of and is_located_in would be topics in the topic map. The typing ability makes it possible to group together the sets of topics that have the same relationship to a topic, which is, according to Pepper (2000) of great importance when graphically navigating vast pools of information. It is also important for the separation of the information resource and the TM. This means, as said by Pepper (2000) that the same TM can be overlaid on different information resources, just as different TMs can be overlaid on the same pool of information to provide different views of the information to different users.

Association roles

Each topic that participates in an association plays a role in that association, the association role. For the relationship "Paris *is located in* France", the roles might be 'city' and 'country'. The association role is defined as a topic, that is 'city' and 'country' are specified as topics in the map. According to Pepper (2000) all associations are inherently multidirectional, unlike associations in mathematics. In a TM it does not make sense to say that A is related to B, but B is not related to A. If A is related to B then, by definition, B must be related to A. For example if we have an influenced-by association, we need to know who was influenced by whom (i.e. influencer, influencee). The labelling of roles is the act of naming, not direction. So if the association is a founder-relationship it means that if we look at it from the founder-role perspective the relation could express the role *founder_of* and from the other direction *was_founded_by.* This means that one can easily navigate from one role in an association to the other and back again.

A topic's names, occurrences, and roles played in associations are collectively known as its characteristics.

Public subjects

Sometimes the same subject is represented by more than one topic. It is then necessary to have some way of establishing the identity between those topics. The case could be when two topic maps are being merged to establish one topic map of the two. For example there may be two topics, 'English' and 'engelska' (Swedish for 'English'), then there is a need to be able to assert that the two refer to the same subject. This is enabled by the concept of public subject and it uses an attribute called identity. The attribute is specified in the topic and addresses a resource that as unambiguously as possible identifies the subject of discourse. The resource could be a publicly available document or a definitional description within or outside the topic map. According to Pepper (2000), any two topics that reference the same subject by means of their identity, are considered semantically equivalent to a single topic, that has the union of the topics characteristics. In the map a single topic results from combining the characteristics of the two topics.

Facets

Facets contribute a filtering mechanism based on properties of the information resource to the standard. Basically, facets are property-value pairs over the resource that otherwise would have been provided by SGML or XML attributes, says Pepper (2000). This could include properties such as language, security, user level etc. By applying facets to the information resource one can filter out information that is redundant for a specific user of the TM, i.e. language=swe, user level=beginner will produce documents in Swedish where the user level equals beginner.

Scopes

The TM model allows three things to be said about any given topic: what names it has, what associations it partakes in and what occurrences it has (the topic characteristics). Pepper and Grønmo (2001) say that when one makes assignments of such characteristics to a topic, one is essentially making an assertion about that topic. However, not all assertions are universally valid. For example a name may be appropriate in some contexts but not in others, an occurrence relevant in some situations, but not in others and an association might state an opinion that is not shared by others. To deal with problems of this type the standard offers the concept of scope.

According to ISO/IEC 13250 (2000) scope is said to specify the limit of the validity of a topic characteristic. It establishes the context in which a name or an occurrence is assigned to a given topic, and the context in which topics are related to each other through associations. Every characteristic has a scope, which may be specified either explicitly, as a set of topics, or implicitly, in which case it is known as the *unconstrained scope*. Assignments made in the unconstrained scope are always valid.

2.3.2 The Thesis Topic Map

In our implementation of a Topic Map we will use the concepts available in the Topic Map standard to express the concepts inherent in the apolipoprotein community-of-interest. We will examine the ways that a Topic Map can mirror the knowledge structures and the information contained in an organisation.

2.4 Related Ideas and Technologies

2.4.1 Semantic Web

The Semantic Web is Tim Berners-Lee's, Director of the World Wide Web Consortium (W3C) and inventor of the Web, future vision of the Internet. It is not a separate web, but an extension of the existing one. The goal is to create a technology to enable machines to make more sense of the web, and through this make it more useful for humans. Computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically. "Semantic" in this context means machine-processable, and has nothing to do with the sense of natural language semantics. Berners-Lee (1998) says that:

> "The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. One of the major obstacles to this has been the fact that most information on the Web is designed for human consumption, and even if it was derived from a database with well defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the web. Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine processable form."

An example of how the Semantic web could be used is as follows, according to Truog (2001): A new camera is developed that can send photos with data about the photos via the Internet. This data is represented as tags. Every tag will have two definitions: one that is human-readable with technical explanations and one that is computer processable. Printer makers and image processing software companies will use these tags to better print the photos and to make them easier to manipulate in programs like Photoshop. The different products can use the tags to communicate over the Internet. This will result in better and easier-to-use products.

According to Swartz (2001), the Semantic Web is built on syntaxes using Uniform Resource Identifiers (URIs) to display data. These syntaxes are called Resource Description Framework (RDF) syntaxes. The W3C has developed an XML serialisation of RDF, called RDF/XML, which is considered the standard interchange format for RDF on the Semantic Web.

A URI is a Web identifier, e.g. the strings identifying a specific web page beginning with *http://*. It can identify anything. It is simply a name for a resource, that may or may not be accessible on the Internet. URIs identify not only information resources such as web pages, but also indirectly refer to physical resources such as people, places and things. RDF uses triples of URIs. These triples consist of a subject, a predicate and an object, much like an ordinary natural language sentence.

The first layer of the semantic Web is the syntax described above. It is very simple and what one can do with it is make assertions about assertions. The next layer is the schema layer. It was designed to be a data typing model for RDF. The schema layer allows one to create different concepts such as resources, classes, properties, ranges and domains. In short this layer helps describe the meanings and relationships of terms. A related system that is also used is the Darpa Agent Markup Language with Ontology Inference Layer (DAML+OIL).

Above this comes the ontology layer. An ontology is capable of describing relationships between types of things. It explains what things are, but it does not say what can be done with them. The Semantic Web depends on ontologies being published on the Internet accessible for all, so that one can define the concepts one uses in a way similar to others.

The next layer is the logical layer. This layer adds the means to use rules to make inferences. New knowledge can be derived. This layer has not yet been fully developed. Another layer which is undeveloped is the proof-layer. Once the information on the web is accessible a program could in theory use a list of items to derive a new item or fact. To do this a reasoning engine is needed. As of yet, no such thing is included in the RDF model.

A third layer which is undeveloped is the trust-layer. The Semantic Web will depend on whether people trust the data included and therefore this layer is very important. It will work as a reasoning engine with a digital signature checker built into it. Digital signatures provides proof that a certain person wrote (or agrees with) a document or a statement. The result of this will be a system that allows one to decide what or whom to trust on the Semantic Web.

The Semantic Web and Topic Maps

The idea behind both the Semantic Web and Topic Maps is the same. Both aims to make large document bases machine processable. Both aims to make it easier for the user to find exactly the information he or she is looking for. But whereas the Semantic Web is a bottom-up approach that requires every document in the document base to be tagged in a specific way, TMs is a top-down approach that does not require the individual documents to be changed at all. A TM is a map placed on top of the document base, but the Semantic Web means that the whole document base is labelled in a specific way.

Both Topic Maps and the Semantic Web requires published ontologies to function in the intended way. There needs to be a way to make sure that the concept one is talking about is the same, that is both human and machine readable. In the Semantic Web this is done through the use of RDF triples of URIs. These URIs can then be published as an ontology for all to use and refer to. The TM version of this is the use of Published Subject Indicators (PSI), which allows a TM to reference its topics to public directories of PSI.

Underlying both the Semantic Web and Topic Maps is the philosophy of a navigable space, with a mapping from concepts to resources. The concepts are in this context identifiers for a resource, not ways to retrieve it. Both technologies depend on external resources for explaining what the concepts used mean.

2.4.2 Information Retrieval

The main goal of Information Retrieval (IR) is to retrieve information which might be useful or relevant to the user from a large information collection. The technology is designed so that the user will not have to scan through the entire information collection, but only look at the documents the IR system return as relevant. This is normally done by the user specifying a query that expresses the user information need. This query is then translated into a set of keywords, which summarises the user information need.

The user of the IR system can use it in two different ways:

- for information browsing
- for information retrieval

The first type of usage occurs when the user does not know for sure what document or information resource he or she is looking for. He or she might start with a general query and then simply look around the document collection that is the output from the IR system, following whatever link that seems interesting. The user is still retrieving information, but the purpose is not well defined from the beginning and may change during the interaction.

The second type of usage is the opposite. The user knows what he or she is looking for and starts with a specific query. The purpose is well defined from the start and is not likely to change during the interaction.

In IR, the information collection that relevant documents are extracted from, is traditionally seen as a text collection. In order to return a relevant document collection to the user, the IR system must interpret the content of the documents and compare this to the content of the user query. To do this syntactic and semantic information in the documents are used.

The main aspect of IR is *relevance*. An IR system's effectiveness can be measured in terms of how relevant the returned documents are and how many of the relevant documents in the entire collection are returned. In short, IR is about returning the relevant documents to a user, or put in another way; filtering out the irrelevant ones.

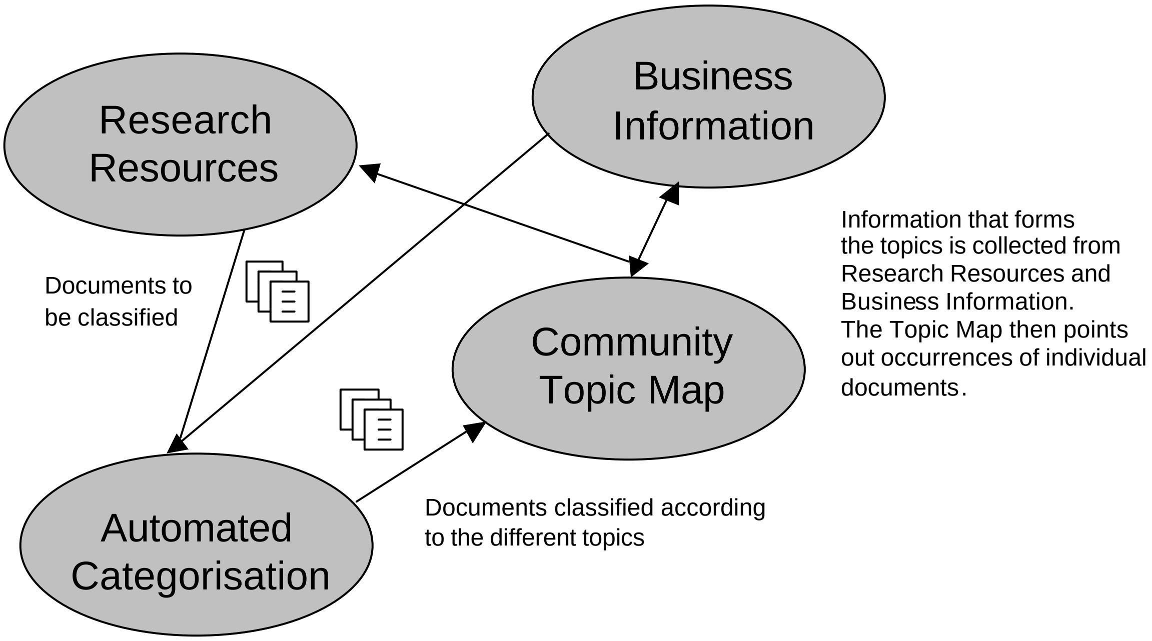
Information Retrieval and Topic Maps

The general idea behind both IR and TMs is more or less the same. Both technologies aim to provide the user with easy access to the information he or she is looking for, and only to that information. Both technologies also allow for some browsing, although in different ways. In IR entirely unknown documents can be searched for relevant information, whereas in TMs all documents are "known", since they have to be linked to a specific topic. In IR the user specifies a query which the system then uses to decide what documents to return as relevant information. In TMs the user does not have to specify a query, but instead he or she has to navigate through a knowledge structure to find the relevant information. IR systems do much of the work that TMs leave to the user, but TMs, on the other hand, give a better view as to how the different documents and the concepts in them are connected. A TM does more than an IR system. It wants to give the user an overview of what the knowledge structure that contains the wanted information looks like, not just return the relevant documents.

3 Topic Map implementation

In this chapter we will discuss our implementation of a Topic Map over the apolipoprotein research area. We will begin by describing how we modelled our domain, continue by looking at the technologies we used to implement our TM, and finish the chapter by explaining how we categorised the documents pointed to by the occurrences in the TM.

The Topic Map is made up of concepts from Research Resources and Business Information, combined to a Community Topic Map for the apolipoprotein community-of-interest. Documents are collected from these areas. The documents are categorised according to the various concepts of the Topic Map and then added to the Topic Map.


Research Resources
Business Information
Community Topic Map
Automated Categorisation
Documents to be classified
Documents classified according to the different topics
Information that forms the topics is collected from Research Resources and Business Information.
The Topic Map then points out occurrences of individual documents.

Overview of the Thesis Topic Map and the resources used to create it

We have chosen to implement our Topic Map in an XML (eXtensible Markup Language) standard called XTM (XML Topic Maps, see section 3.2.1) that is specifically created for implementing Topic Maps. For visualising the Topic Map we have looked at a number of different programs, but in the end decided to use the Omnigator, which is a developing tool for Topic Maps from Ontopia[14]. The reason we decided to use the Omnigator is not because we think it is a very good visualisation, but because there is no better one. The Omnigator shows the Topic Map as a list of topics and associations and allows a user to navigate in this list by clicking on them as links. Ideally we would have liked the visualisation to show the TM as a graph with the topics as nodes and the associations as arcs between them, but we were unable to find such a visualisation tool, using XTM encoded knowledge structures.

To model the apolipoprotein domain which we wanted to express as a Topic Map a number of different approaches were necessary. We tried three main approaches. The first step was to look through the AstraZeneca thesaurus, glossary and term collection to extract terms and relationships used in the organisation. This gave us very general terms, but not many relevant for the apolipoprotein area. Our next step

[14] For more information on *the Omnigator* see www.ontopia.net

was to extract 200 documents in the apolipoprotein area from a database. From these documents we then extracted keywords. These keywords also gave us very general terms. Our last and the most important step was the regular meetings we had with Göran Walldius, who is a professor at Karolinska Institutet and a researcher at AstraZeneca Research & Development (R&D) Mölndal, and whose area of research is also the area we wanted to model.

Once the domain was in place the next step was to connect it to documents. In a Topic Map this is done by defining *occurrences* for the different topics. To connect the relevant documents to the topics we tried to use a categoriser from Autonomy[15]. It allows you to create categories and to give example documents of what kind of documents each category should contain. After doing that the Categorizer automatically sorts the documents you feed it, into the right category. This means that a large number of documents can be categorised with minimal human effort and time.

[15] For more information on the *Autonomy* suite see www.autonomy.com For more information on the Categorizer see www.autonomy.com/Content/IDOL/APPOLS/Categorizer/

3.1 Domain modelling

When we set out to define our domain (the apolipoprotein research area) and to model the different concepts and relationships that it comprises, we thought that our main work would consist of using terms and structures already existing in different forms in AstraZeneca's different resources. From these resources we would then extract relevant concepts and combine and implement them as a Topic Map. The idea from the beginning was to take concepts from the research area and combine them with business concepts, thus reflecting the entire organisation involved in the creation of a new product.

3.1.1 Resources

We expected to use many of the resources available at the AstraZeneca library such as the thesauri called *PL@net* and *Amiracle*. Other material we expected to use was the different glossaries (*AZ Glossary*[16], *AZURE*[17] *Business Glossary*) and terminology initiatives (*Clinical Terminology Team*) which all aim to standardise the usage of terms and concepts throughout the organisation.

Amiracle is an information management system containing several sets of information. The largest database is the product literature database PL@net. In Amiracle one can search for documents in a normal keyword search system and also look up concepts in the thesaurus.

The AZ Glossary is an application that provides information on terms and acronyms, commonly used within AstraZeneca. For each term, a definition is given together with additional information about its ownership and use. The AZURE Business Glossary is a subset of the AZ Glossary, which contains business related terms only. It also provides instances of key business entities, such as sites, companies, countries, currencies etc.

The Clinical Terminology Team aims to harmonise language in order to improve understanding and to make content searchable and reusable. It consists of a list of terms with definitions and also information on ownership and use.

3.1.2 The modelling

The first step in modelling the domain was to look through the resources mentioned above, to try to extract relevant concepts to use as topics in the TM. This did not result in very many topics, since the apolipoprotein area is highly specialised, while the resources cover all research and business areas in a general way. It is possible to find information on all broader categories that the apolipoproteins belong to, but it is not possible to find specific information on the apolipoproteins. It is also possible to find general business concepts such as *Clinical Study* explained but nothing on individual studies.

Once we realised this we decided on a different course of action. 200 documents on apolipoproteins and related subjects were extracted from a database called Medline, which is a bibliographic database produced by the US National Library of Medicine. The documents in Medline are tagged with keywords specifying their content. We

[16] AZ Glossary = AstraZeneca Glossary
[17] AZURE = AstraZeneca Uniform Reference Environment

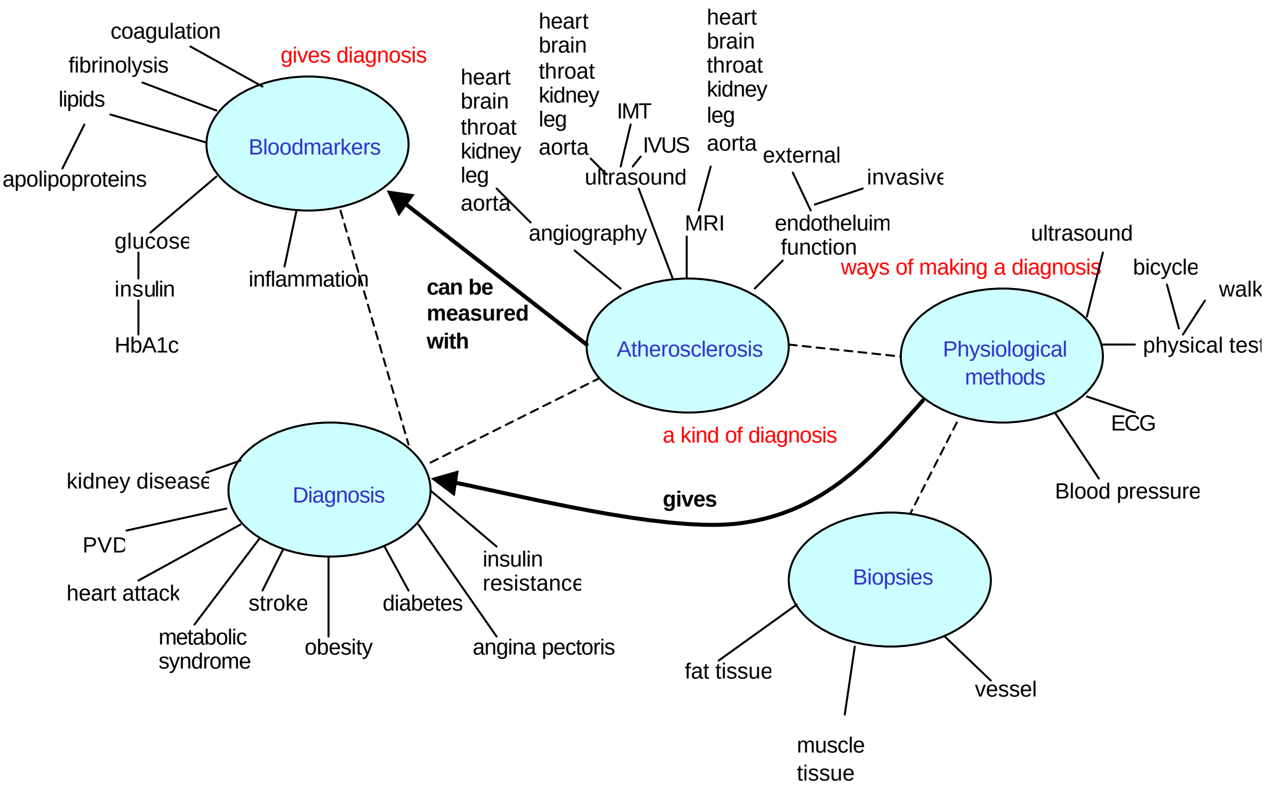
extracted these keywords in the hope that they would form relevant topics for our domain. Once again we made the same discovery that we had made looking through the above resources: the keywords were far too general and did not contain material specific enough for the apolipoprotein area. The fifteen most common keywords we extracted were:

1. *Human*
2. Apolipoprotein A-I/bl [Blood]
3. *Male*
4. *Female*
5. *Middle Age*
6. Apolipoprotein A-I/me [Metabolism]
7. Lipoproteins, HDL Cholesterol/bl [Blood]
8. *Adult*
9. Apolipoprotein A-I/ge [Genetics]
10. Triglycerides/bl [Blood]
11. *Animal*
12. Apolipoproteins B/bl [Blood]
13. Cholesterol/bl [Blood]
14. *Aged*
15. Lipoproteins, LDL Cholesterol/bl [Blood]

As shown above in italics, seven (no.s 1, 3, 4, 5, 8, 11, 14) of the fifteen most common words are not related to the apolipoprotein area, but are general research terms.

At this point we realised that maybe we would not be able to use the existing resources like we had expected. From them we could get general information to place our specific concepts in a wider context, but we could not do the opposite, i.e. extract specific concepts and then expand them into a wider context. Instead of using the existing resources we would have to do a lot of modelling by hand. We would have to decide what concepts where relevant and the relationships between them without the help of the existing resources. Once we had the concepts we could use the resources to place them in a wider context. It also meant that we had to talk to someone who knew the domain and extract relevant concepts and relationships from this person's knowledge. Our lack of domain knowledge made it impossible for us to do it on our own. A meeting was set up with Göran Walldius, who is a professor at Karolinska and also a researcher in the apolipoprotein area at AstraZeneca, R&D Mölndal. During this meeting he drew us a map over the concepts relevant for the apolipoprotein research area and the relationships between them. He also gave us some general background information on the area. With this information the modelling of the domain could begin. Since most of our concepts come from a researcher in the apolipoprotein area there is a distinct lack of business terms among them. Also relevant business terms have proved harder to find in the existing resources, than have research terms.

The modelling was a complex task, that included drawing diagrams over the concepts, their subclasses/subcategories and the relationships between them. Most of the drawing was done in Microsoft PowerPoint, since Omnigator (the visualisation tool) does not display the TM as a graph with the topics and associations shown as nodes and arcs, but as a list of topics and associations that allows a user to navigate in the list by clicking them as links. This is not suitable for the modelling stage of the development, because it makes it difficult to have an opinion on the extent of the Topic Map and on the accuracy of the relationships between topics.

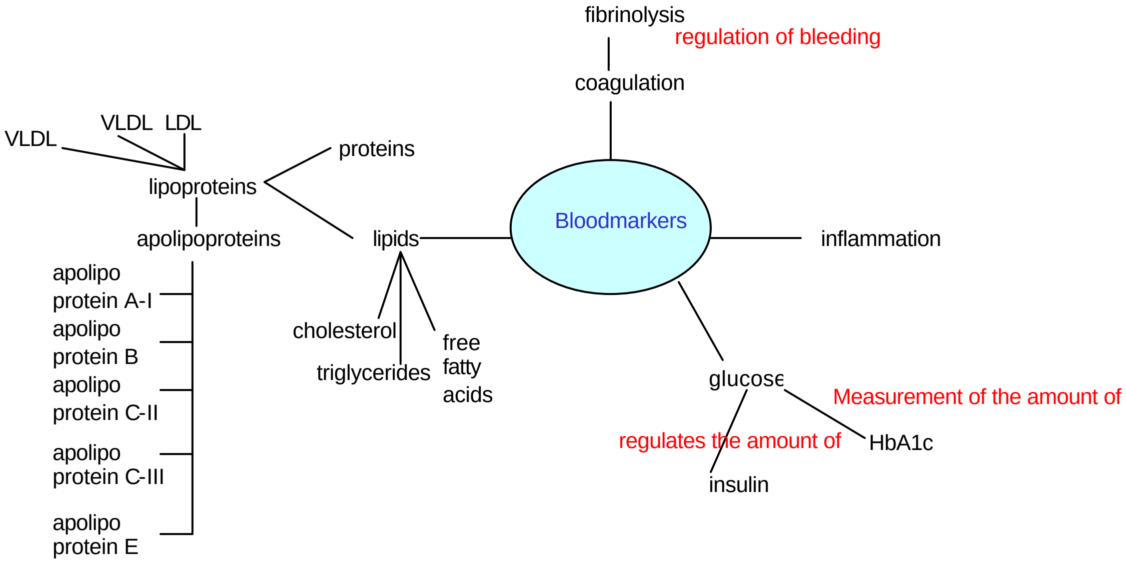


Picture 1 The first overview of the Topic Map with all concepts and some relationships added

Once a model that covered the apolipoprotein research area was established (see Picture 1) we started to look up the topics in it in the existing resources. The thesauri proved most useful at this stage. In them the superclass/subclass relationships can be clearly seen and what broader categories a term belong to is easy to find.

At this point we started to implement our model into a Topic Map (henceforth referred to as the thesis TM). For this we used the XML standard called XTM, described below. The XTM files were then visualised in a visualisation tool called Omnigator (see section 3.3). Once the first version of the TM was implemented and possible to navigate, the refining work began. At this stage most of the relationships in the TM were of the type *instance of*, which is the way to express a subclass/superclass relation in XTM. This was a simplification. The concepts were related, but the relationships between them were of many types, most of them not of the *instance of* type, but of a more complex nature. Using mainly the thesauri, but also the glossaries, we tried to establish these complex relationships, such as *indicates/is measured through*. We also added some new concepts to get more of a business view on the TM. These concepts were of the type clinical study (AMORIS) and pharmaceutical product (Crestor). This resulted in a slightly different model and another meeting with Göran Walldius was set up, to make sure that the relationships were accurate.

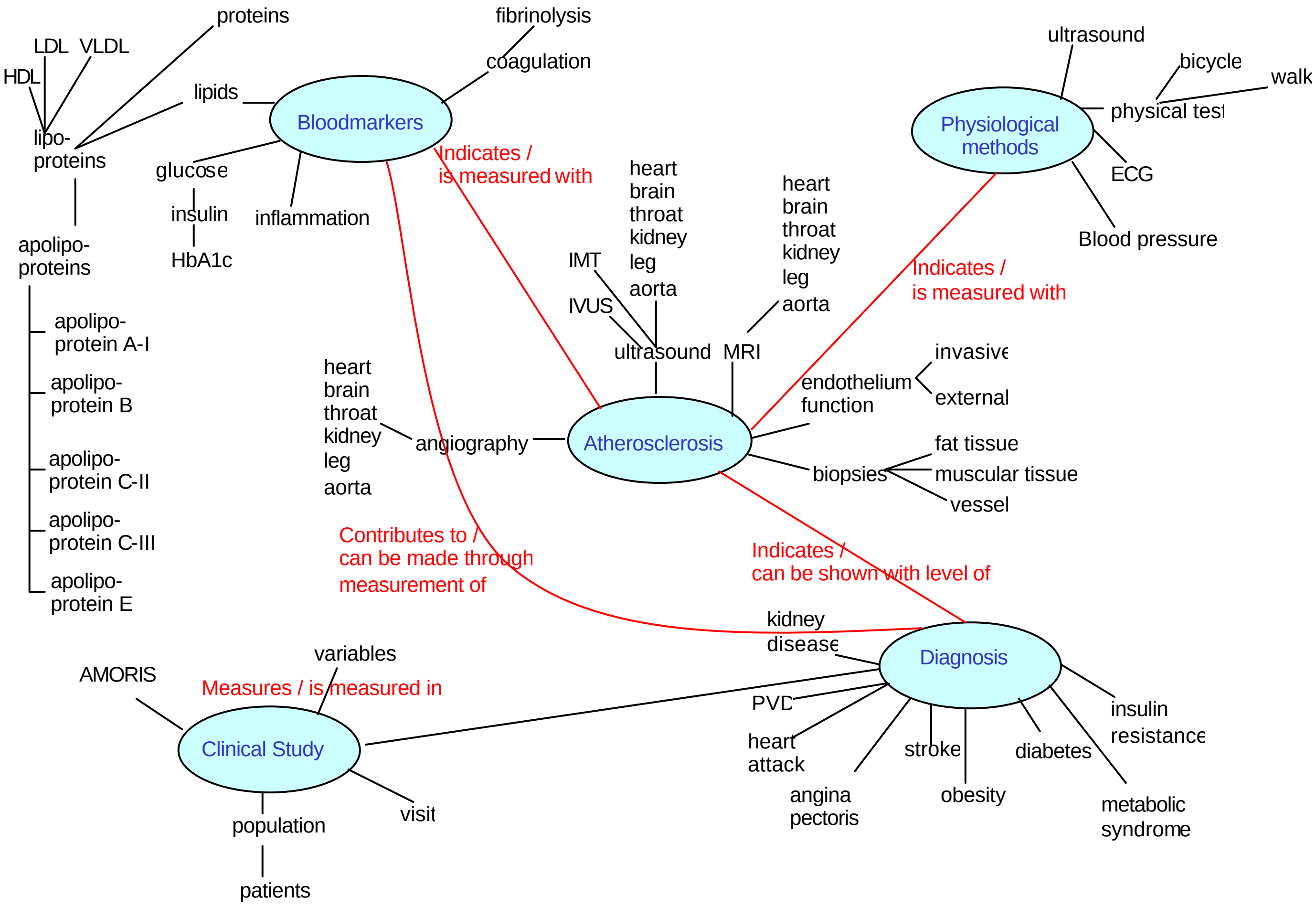
During this meeting the relationships we had established were approved. A number of new concepts were also added in the bloodmarker area of the TM (see Picture 2), which is the most important area from the apolipoprotein view, since this is where the apolipoproteins and their related concepts fit in.



Picture 2 The bloodmarker area of the thesis Topic Map, with all concepts shown and the relationships visualised as lines

A whole new aspect of the apolipoprotein research area was also discussed, namely *the risk area*. This includes concepts like *prevalence* - the occurrence of a diagnosis in society, *incidence* - how many with this diagnosis will fall ill and how many will die from it, *absolute* and *relative risk*, and *primary* and *secondary prevention.* The risk area gives a whole new viewpoint to the TM. It looks at the apolipoprotein research area from a quality of life perspective, which is to be seen as a supplementary research area, partly overlapping and extending the TM, and also referring to a new or broader set of occurrences. After the meeting we continued the discussion and decided not to add the risk area to the thesis TM. This was mainly because of the difficulty of finding relevant concepts in the resources. The relationships and the new concepts in the bloodmarker area were implemented. We also added term definitions and descriptions that explained the different concepts and relationships in the TM. The definitions and descriptions together form TM *internal information*. At this stage the thesis TM was almost finished. Only some refinements were needed. The thesis TM is a bilingual TM. Everything is implemented in both English and Swedish. This is because we wanted to try how well this worked with Topic Maps and also because we wanted everyone to be able to use the TM.

We had a third meeting with Göran Walldius to further clarify the relationships between the concepts in the bloodmarker part of the TM. A number of new concepts were once again added, all of them in the bloodmarker area (see Picture 3).The concept biopsies was also moved to become an undercategory to *Atherosclerosis*. These changes were implemented and finally occurrences were added to the different concepts. An occurrence is a link that points to a document that contains information relevant to a concept. Our occurrences consist of 27 documents that can be found in the apolipoprotein community-of-interest on the Informatics Forum portal. They are all relevant for the apolipoprotein research at AstraZeneca. The occurrences form the *external information* in the TM. With the addition of the occurrences our implementation of the TM was finished.



Picture 3 This is a model of the final Topic Map
(for readability only some of the relationships are shown).

All through our modelling of the domain we have tried to bring together similar concepts under a single parent concept, e.g. method of measurement as a joint concept for among other things all concepts around a*therosclerosis*: *angiography*, *ultrasound*, *MRI*, *endothelium function* and *biopsies*. Doing this means that one only has to describe the parent concept's relationships as roles in association types, instead of specifying one role for each child that plays a role in the association. In the above example this means that one only has to describe the relationship between *atherosclerosis* and *method of measurement*, instead of describing different roles between *atherosclerosis* and *angiography*, *atherosclerosis* and *ultrasound*, *atherosclerosis* and *MRI* etc.

This way of describing the concepts in the Topic Map also gives an abstract layer to the TM from which the individual concepts can be instantiated. It makes it easier to talk about relationships between concepts since one can do it on both an abstract level and on a concept specific level. However, it has proved difficult to do this throughout the TM modelling, since not all concepts have been possible to provide an abstraction to. Therefore it is provided where possible and where not possible, the instantiated forms are kept. For a quick overview of the relations implemented in thesis TM see table 1 below. From left to right in the first line it is read like *Blood markers* **contributes to** *Diagnosis* in the association *contributes to/can be made through measurement of* and from right to left *Diagnosis* **can be made through measurement of** *Blood Markers*.

Players	Name	Association	Name	Players
Blood markers	contributes to	contributes to / can be made through measurement of	can be made through measurement of	Diagnosis
IMT IVUS	gives value to measurement method	gives value to / is measured in	is measured with measurement	Ultrasound
Atherosclerosis	Indicates	Indicates / can be shown with level of	can be shown with level of	Diagnosis
Blood markers Physiological methods	Indicates	Indicates / is measured with	is measured with	Atherosclerosis
Glucose	is measured with test method	measures / is measured by	measures	HbA1c IVETT POGTT
Atherosclerosis	is measured with measurement method		measures	Angiography Biopsies Endothelium function MRI Ultrasound
AMORIS	measures variables	measures / is measured in	is measured in clinical study	Apolipoprotein A -1 Apolipoprotein B HDL LDL VLDL Cholesterol Triglycerides
External Invasive	is measure procedure for	is measured with / is measure procedure for	is measured with procedure	Endothelium function
Brain Fat tissue Heart Kidney Leg Muscular tissue Throat Vessel	is measure location for	is measured on / is measure location for	is measured on	Angiography Biopsies MRI Ultrasound
Coagulation	adjusts	adjusts	adjusts	Fibrinolysis
Clinical study	consists of	is part of / consists of	is part of	Diagnosis Population Variable Visit
Lipoproteins				Proteins Lipids

Table 1 – All associations in thesis TM

The relationships between the concepts in the domain were difficult to find in the existing resources, since these mostly contain terms, and say nothing on the connections between these terms. This meant that for modelling most of them, we used Göran Walldius' pictures over the apolipoprotein area. Therefore the relationships in the thesis TM mirrors his view on what relationships are needed and also on what these relationships are. The hierarchical relationships of the superclass/subclass type can be found in the thesauri, and therefore these are also extracts from the resources.

3.2 XML – eXtensible Markup Language

The eXtensible Markup Language, abbreviated XML, is a subset of Standard Generalized Markup Language (SGML). Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with Hyper Text Markup Language (HTML). XML has been designed for ease of implementation and for interoperability with both SGML and HTML.

Almost all documents have some structure. A mark-up language is a mechanism to identify structures in a document. The XML specification[18] defines a standard way to add mark-up to documents. XML is a mark-up language for documents containing structured information with a linear syntax. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.).

XML describes a class of data objects called XML documents. XML is an application profile or restricted form of SGML. By construction, XML documents are conforming to SGML documents.

The fundamental idea of XML is to get the semantics out of documents as opposed to HTML where marking is based on how text is to be displayed. For example, if you are to describe an e-mail message the structure could look like this:

```
 <MESSAGE>
       <FROM>mark.johnson@company.com</FROM>
       <TO>john.markson@company.com</TO>
       <SUBJECT>Lunch-meeting?</SUBJECT>
       <DATE>2002-04-11</DATE>
       <BODY>Hello John! Don’t forget our lunch-meeting
             today 1200 hours. Mark</BODY>
</MESSAGE>
```

E-mail message in XML

3.2.1 XTM – XML Topic Maps

XML Topic Maps (XTM) 1.0[19] is the specification that provides a model and grammar for representing the structure of information resources used to define topics, and the associations (relationships) between topics in XML (see section about TM). Names, resources, and relationships are said to be *characteristics* of abstract subjects, which are called *topics*. Topics have their characteristics within *scopes,* i.e. the limited contexts within which the names and resources are regarded as their name, resource, and relationship characteristics. One or more interrelated documents employing this grammar is called a *topic map*.

The specification was developed by topicmaps.org which is an independent consortium of parties developing the applicability of the topic map paradigm (ISO/IEC 13250:2000) to the World Wide Web by leveraging the XML family of specifications.

[18] *Extensible Markup Language (XML) 1.0*, Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen, 10 February 1998. Available at http://www.w3.org/TR/REC-xml
[19] *XTM – XML Topic Maps specification*. Available at http://www.topicmaps.org/xtm/index.html

The specification describes an abstract model and XML grammar for interchanging web-based topic maps.
The design goals for XTM were that:

1. XTM shall be straightforwardly usable over the Internet.
2. XTM shall support a wide variety of applications.
3. XTM shall be compatible with XML, XLink, and ISO 13250.
4. It shall be easy to write programs that process XTM documents.
5. The number of optional features in XTM is to be kept to the absolute minimum, ideally zero.
6. XTM documents should be human-legible and reasonably clear.
7. The XTM design should be prepared quickly.
8. The design of XTM shall be formal and concise.
9. XTM documents shall be easy to create.
10. Terseness in XTM mark-up is of minimal importance.

XTM Elements

The element-set that comes with XTM is reasonably clear and is described in the specification. It includes the elements showed in table 2.

Element	Description
`<association>`	Topic Association
`<baseName>`	Base Name of a Topic
`<baseNameString>`	Base Name String container
`<instanceOf>`	Points to a Topic representing a class
`<member>`	Member in Topic Association
`<mergeMap>`	Merge with another Topic Map
`<occurrence>`	Resources regarded as an Occurrence
`<parameters>`	Processing context for Variant
`<resourceData>`	Container for Resource data
`<resourceRef>`	Reference to a Resource
`<roleSpec>`	Points to a Topic serving as an Association
`<scope>`	Reference to Topic(s) that comprise the Scope
`<subjectIdentity>`	Subject reified by Topic
`<subjectIndicatorRef>`	Reference to a Subject Indicator
`<topic>`	Topic element
`<topicMap>`	Topic Map document element
`<topicRef>`	Reference to a Topic element
`<variant>`	Alternate forms of Base Name
`<variantName>`	Container for Variant Name

Table 2 - Element-set for XTM

The following section gives a simple introduction to the different elements in the specification. It is referenced to the XTM specification (http://www.topicmaps.org/xtm/index.html).The interested reader should consult the specification for a deeper explanation.

`<association>`
Element that asserts a relationship among topics that play roles as members of the association. The class to which an **`<association>`** belongs is specified with an **`<instanceOf>`** element. Since **`<association>`** is a characteristic you may apply it within a **`<scope>`** element.

<baseName>
Element that specifies a topic name. A topic name is represented with one string specified in the **<baseNameString>** element. A base name is a characteristic so it is governed by **<scope>**.

<baseNameString>
Element that contains a string representation of a topic base name.

<instanceOf>
Element that specifies via a **<topicRef>** or a **<subjectIndicatorRef>** the classes that the topic is an instance of.

<member>
Element that specifies all topics that play a given role in an association.

<mergeMap>
Element that references an external topic map which the topic map, that **<mergeMap>** is specified in, is to be merged with.

<occurrence>
Element that specifies a resource supplying information relevant to a topic. The context within which the occurrence is valid may be specified via a **<scope>** element, since an occurrence is considered a characteristic of a topic.

<parameters>
Element which consists of one or more **<subjectIndicatorRef>** or **<topicRef>** elements, which specifies additional processing contexts.

<subjectIdentity>
Element that specifies the subject that is reified by a topic, via a **<rescourceRef>**, **<subjectIndicatorRef>** and/or a **<topicRef>** element.

<subjectIndicatorRef>
Element which provides a URI reference to a resource that acts as a subject indicator.

<topic>
Element that specifies the name and occurrence characteristics of a single topic. It has a single unique identifier, and the ability to state the class(es) of which it is an instance, with **<instanceOf>**, and the identity of the subject with **<subjectIdentity>**.

<topicMap>
Root element of a topic map document and the parent of all **<topic>**, **<association>** and **<mergeMap>** elements. The element may also be a subtree inside an XML document containing other information than the topic map itself.

<topicRef>
Element which provides a URI reference to a topic. **<topicRef>**s are identical to **<subjectIndicatorRef>**s except for the additional constraint that they must point to a **<topic>** element.

<variant>
Element which specifies an alternate form of a topic's base name appropriate for a specific processing context.

`<variantName>`
Element that provides the resource to be used as a variant of a base name.

3.2.2 Available Technology and Explorations

Beside this pure XML notation for topic maps there exists a notation called Linear Topic Map (LTM)[20] where one can express topic, associations and occurrences in just a few lines. The reason for us to choose the XTM standard was in fact just that it is a standard and thereby supported by most applications. The LTM notation is just a proposal for a simple textual format for topic maps and the number of applications that supports it, is low.

Topic Maps Query Language (TMQL)[21] is a guide that sets down the requirements for a query language for topic maps. An implementation of these requirements was made through the creation of Tolog[22]. Tolog is a Prolog inspired query language for Topic Maps. Central to both languages is the concept of a fact database containing statements about the universe of discourse, i.e. the subject area of the database. The only things that exist in this universe is what is encoded.

TM4J[23] (Topic Map Engine for Java) is an open-source project which develops topic map processing tools and applications. The current focus of the TM4J project is on the development of a topic map engine which processes files conforming to the XML Topic Maps (XTM) specification and stores them either in memory or in a more persistent store using an object-oriented database. Future development plans is to include a framework for developing web applications using topic maps and a navigation/editing application for topic maps.

Many other aspects of the topic map technology are under development. Most of them are fairly new and not yet standardised. For example there are working documents for a schema language (constraint language for, among other things inference and consistent topic maps) for topic maps and directives for Published Subject Indicators. There is also an ongoing discussion between RDF and XTM developers as to the ultimate compatibility of their specification and the possibilities of using RDF to represent the basic concepts of Topic Maps.

[20] Non-formal description available at http://www.ontopia.net/download/ltm.html
[21] A draft user requirements document for the new ISO standard available at http://www.y12.doe.gov/sgml/sc34/document/0227.htm
[22] Available at http://www.ontopia.net/download/tolog.htm
[23] TM4J Project website available at http://tm4j.org/

3.2.3 Examples

Next follows some examples from the topic map we constructed in our thesis with comments on different constructions. For brevity references have been cut down in length and parts have been left out where they do not contribute to the example in question. Some examples or parts thereof are just made up to establish an understanding for certain constructions.

```
1 <topic id="apolipoproteins">
2   <instanceOf>
3     <topicRef xlink:href="#lipoproteins"/>
4   </instanceOf>
5     <baseName>
6       <scope><topicRef xlink:href="#swedish"/></scope>
7       <baseNameString>Apolipoproteiner</baseNameString>
8     <variant>
9       <parameters><topicRef xlink:href="#sort"/></parameters>
10       <variantName><resourceData>lipoprotein</resourceData></variantName>
11    </variant>
12    </baseName>
13</topic>
```

Example 1 - topic

Example 1 describes a topic with the unique id *apolipoproteins* (line 1). It is of type *lipoproteins* which follows from lines 2 - 4. Its name 'Apolipoproteiner' in the scope of Swedish is specified with the name elements and scope elements in lines 5 to 12. Lines 8 -11 within the base name declaration is for the specific processing context of sorting, which means that the topic is sorted according to this name but displays the name specified in the **`<baseNameString>`** element at line 7. For a fully consistent topic map the topics *lipoproteins*, *swedish* and *sort* must be created in the topic map or referenced to another topic map outside of the one where this topic is defined.

```
1 <topic id="lipoproteins">
2   <subjectIdentity>
3     <subjectIndicatorRef xlink:href="http://www.lipoproteins.com"/>
4   </subjectIdentity>
5      . . .
6     <!-- Instances, base names and occurrences go here -->
7      . . .
8 </topic>
```

Example 2 – subject identity

The interesting point to notice with this example is between lines 2 – 4, where a **`<subjectIdentity>`** element is being used to establish the topic identity. The **`<subjectIndicatorRef>`** element points to the resource that as unambiguously as possibly identifies the subject of discourse. Subject identity is what makes the topic real for humans. The identity is what the author of the topic map had in mind when the topic was created. The address specified in line 3 is a nonsense address, it does not exist. It is there to clarify this example, nothing else.

```
1 <topic id="apolipoproteins">
2   . . .
3  <!-- Instances and base names go here -->
4   . . .
5  <occurrence>
6    <instanceOf>
7      <topicRef xlink:href="#definition"/>
8    </instanceOf>
9      <scope><topicRef xlink:href="#swedish"/></scope>
10       <resourceData>Proteinkomponenten i en lipoprotein.</resourceData>
11 </occurrence>
12 <occurrence>
```

```
13   <instanceOf>
14     <topicRef xlink:href="#article"/>
15   </instanceOf>
16     <scope><topicRef xlink:href="#swedish"/></scope>
17       <resourceRef xlink:href="file:///C|/APO/values.pdf"/>
18 </occurrence>
19</topic>
```

Example 3 - occurrence

The *apolipoproteins* topic has two occurrences. One (lines 5 to 11) is internal, which means that the resource is specified within the topic map via the **`<resourceData>`** element (line 10). The other (lines 12 to 18) is external and point through a **`<resourceRef>`** element (line 17) out of the map to specify the resource that contribute to the topic *apolipoprotein*. Lines 6 – 8 and 13 – 15 specifies which type the occurrence is, in this case a definition and an article respectively.

```
1<association>
2  <instanceOf>
3    <topicRef xlink:href="#measures/is_measured_in"/>
4  </instanceOf>
5  <member>
6    <roleSpec>
7      <topicRef xlink:href="#clinical_study"/>
8    </roleSpec>
9    <topicRef xlink:href="#amoris"/>
10 </member>
11 <member>
12   <roleSpec>
13     <topicRef xlink:href="#variable"/>
14   </roleSpec>
15   <topicRef xlink:href="#apolipo_a-1"/>
16 </member>
17</association>
```

Example 4 - association

As mentioned above an association asserts relationships among topics that play roles as members of the association. An association is expressed in XTM notation as shown in example 4. It is an instance of a topic called *measured/is_measured_in* (lines 2- 4) which is its type. Lines 5 to 10 and 11 to 16 specifies the members of the association where lines 6 – 8 and 12 – 14 gives the role types with the **`<roleSpec>`** element and lines 9 and 15 reference the roles via a **`<topicRef>`** element (the role players). Both the role and role type are declared as topics. The association should be read that the topic *amoris* plays the role of *clinical_study* and that the topic *apolipo_a-1* plays the role of *variable* in the association type called *measures/is_measured_in*.

```
1<topic id="measures/is_measured_in">
2  <baseName>
3    <scope><topicRef xlink:href="#swedish"/></scope>
4      <baseNameString>mäter/mäts i</baseNameString>
5  </baseName>
6  <baseName>
7    <scope><topicRef xlink:href="#swedish"/>
8           <topicRef xlink:href="#clinical_study"/></scope>
9      <baseNameString>measures variables</baseNameString>
10 </baseName>
11 <baseName>
12   <scope><topicRef xlink:href="#swedish"/>
13          <topicRef xlink:href="#variable"/></scope>
14     <baseNameString>is measured in clinical study</baseNameString>
15 </baseName>
16</topic>
```

Example 5 – association type

An association type as in example 5 gives names to the roles in the association. It is declared as a topic. Between lines 2 – 5 the association's name is declared. Lines 6 – 10 specifies the name for a role in the scope of *swedish* and *clinical_study* and in this scope the name expresses the string 'measures variables'. A second role is specified between lines 11 and 15. It has as scope *swedish* and *variable* and expresses 'is measured in clinical study'. It is this that gives the multi-directional aspect of the association in example 4. When seen from the role player *apolipo_a-1* the relationship is described as 'is measured in clinical study' and from the other end when seen from the *amoris* topic it expresses 'measures variables'. If one has other topics that have a similar relationship between each other one just has to add other base names in other scopes that express these roles.

3.3 Topic Map visualisation

The Omnigator is an application that can be used to navigate any topic map using a standard web browser. It is developed by Ontopia[24]. The name the Omnigator is a contraction of "omnivorous navigator", which underlines the application's principle design goal - to be able to make sense of any conforming topic map. The Omnigator is intended as a teaching and developing aid, not as an application to be used by end users.

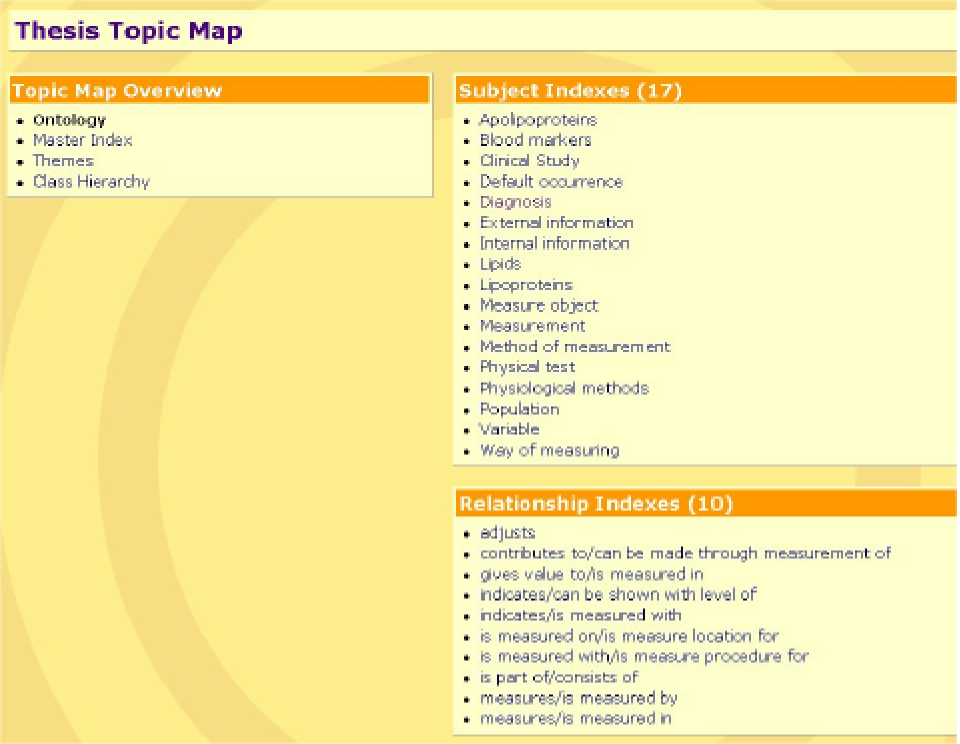
According to Ontopia's description, the Omnigator uses a simple client-server architecture based on a standard http protocol. On the server side there is a Java 2, Enterprise Edition (J2EE) web application built using the Ontopia Topic Map Engine and Navigator Framework, that runs in the Tomcat web server. Tomcat is the servlet container that is used in the official Reference Implementation for the Java Servlet and JavaServer Pages technologies. Tomcat is developed in an open and participatory environment and released under the Apache Software License. Tomcat is intended to be a collaboration of the best-of-breed developers from around the world. This application reads (and writes) topic maps and generates HTML pages. On the client, a standard web browser receives these HTML pages and displays a view of some portion of the topic map. This view is rich in links, built from the data structures that constitute the topic map. Each time the user clicks on a link, a request is sent to the server application, resulting in a new set of information extracted from the topic map.

The Omnigator visualisation tool is part of a bigger toolkit, under development at Ontopia, called the Ontopia knowledge suite (OKS)[25]. In the toolkit is a data model (engine) for topic maps which includes interfaces, utilities and readers and writers of the data. On top of the engine is a navigator framework on which the Omnigator is built. The backend consists of In-Memory and Relational Data Base Management storage. Under development are administration tools for editing, creating and maintaining topic maps, as well as an autogeneration toolkit for processing of other data into topic maps. Future development of the OKS includes a client editor framework for desktop editors, a topic map server for distributed applications and some sort of virtuality for dynamic topic maps.

[24] See www.ontopia.net

[25] For more information on *the Ontopia Knowledge Suite* see http://www.ontopia.net

In the Omnigator a Topic Map is visualised as a list of links that lets the user navigate around in the list by clicking the links. It does not display the TM as a graph with the topics and associations shown as nodes and links, and so offers no easy way to get an overview of how topics are interrelated and how complex the TM is.



Thesis Topic Map

Topic Map Overview

- Ontology
- Master Index
- Themes
- Class Hierarchy

Subject Indexes (17)

- Apolipoproteins
- Blood markers
- Clinical Study
- Default occurrence
- Diagnosis
- External information
- Internal information
- Lipids
- Lipoproteins
- Measure object
- Measurement
- Method of measurement
- Physical test
- Physiological methods
- Population
- Variable
- Way of measuring

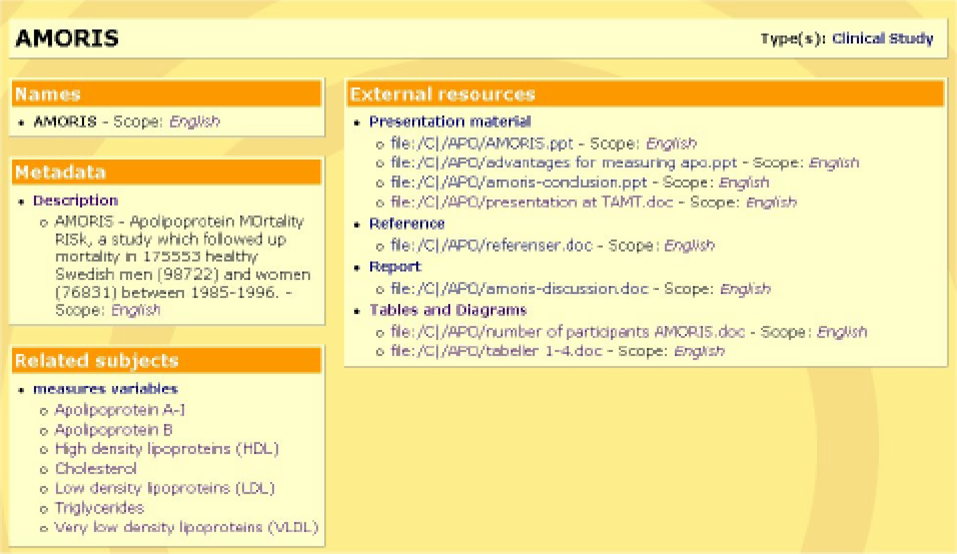
Relationship Indexes (10)

- adjusts
- contributes to/can be made through measurement of
- gives value to/is measured in
- indicates/can be shown with level of
- indicates/is measured with
- is measured on/is measure location for
- is measured with/is measure procedure for
- is part of/consists of
- measures/is measured by
- measures/is measured in

Example 1

In example 1 part of the main view of the Thesis Topic Map is shown. The parent concepts and the different relationships can be seen to the right. Every item in the lists is also a link that leads to further information.

When a link for a topic is clicked on, the topic is displayed with all the information about the topic, which could be associations, occurrences etc.



AMORIS

Type(s): Clinical Study

Names

- AMORIS - Scope: *English*

Metadata

- Description
  - AMORIS - Apolipoprotein MOrtality RISk, a study which followed up mortality in 175553 healthy Swedish men (98722) and women (76831) between 1985-1996. - Scope: *English*

Related subjects

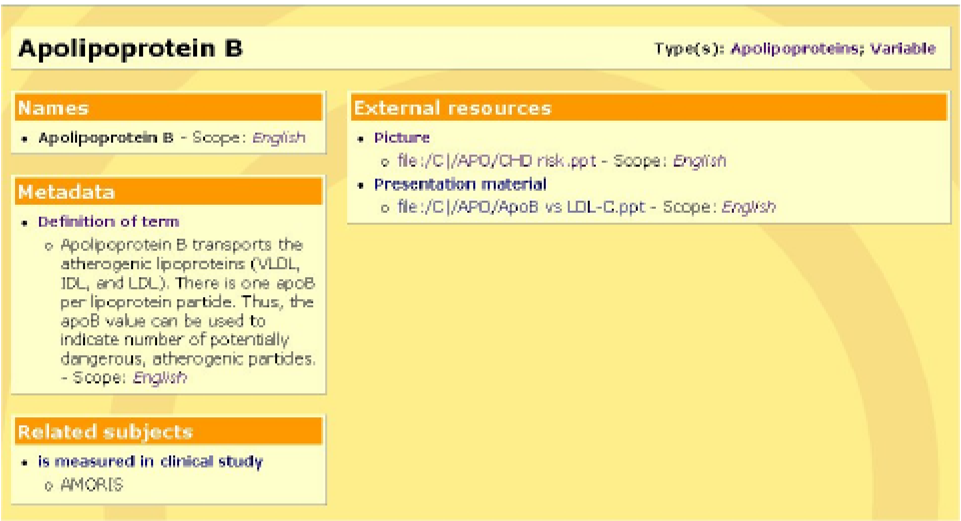
- measures variables
  - Apolipoprotein A-I
  - Apolipoprotein B
  - High density lipoproteins (HDL)
  - Cholesterol
  - Low density lipoproteins (LDL)
  - Triglycerides
  - Very low density lipoproteins (VLDL)

External resources

- Presentation material
  - file:/C|/APO/AMORIS.ppt - Scope: *English*
  - file:/C|/APO/advantages for measuring apo.ppt - Scope: *English*
  - file:/C|/APO/amoris-conclusion.ppt - Scope: *English*
  - file:/C|/APO/presentation at TAMT.doc - Scope: *English*
- Reference
  - file:/C|/APO/referenser.doc - Scope: *English*
- Report
  - file:/C|/APO/amoris-discussion.doc - Scope: *English*
- Tables and Diagrams
  - file:/C|/APO/number of participants AMORIS.doc - Scope: *English*
  - file:/C|/APO/tabeller 1-4.doc - Scope: *English*

Example 2

In example 2 the link for the topic *AMORIS* has been clicked. All information on the topic is shown, for example the internal information (*description*) and the external information (*presentation material, reference, report*, and *tables and diagrams*). Under the headline *Related subjects* the relationships the topic is part of can be seen. In this case it is *measures variables.* The items in the list under *measures variables* are topics that are also part of the relationship, but play another role.



Apolipoprotein B

Type(s): Apolipoproteins; Variable

Names

- Apolipoprotein B - Scope: *English*

Metadata

- Definition of term
  - Apolipoprotein B transports the atherogenic lipoproteins (VLDL, IDL, and LDL). There is one apoB per lipoprotein particle. Thus, the apoB value can be used to indicate number of potentially dangerous, atherogenic particles. - Scope: *English*

Related subjects

- is measured in clinical study
  - AMORIS

External resources

- Picture
  - file:/C|/APO/CHD risk.ppt - Scope: *English*
- Presentation material
  - file:/C|/APO/ApoB vs LDL-C.ppt - Scope: *English*

Example 3

In example 3 the relationship *measures variables* from example 2 is seen from the end, i.e. from the view of the topic *Apolipoprotein B*. Here the relationship is *is measured in clinical study*.

As mentioned in the introduction, the main reasons for us to choose the Omnigator for visualising the thesis Topic Map are two:

- There are not that many visualising tools freely available.
- Of the ones available, the Omnigator is the one easiest to use and the one that displayed the Topic Map most intuitively to our mind.

Ideally we would have wanted a visualising tool that displayed the TM as a graph, with the topics as nodes and the relationships between them as arcs. This would have made the modelling easier and facilitated modelling the TM in the visualising tool immediately without using drawing tools. It would also have presented a better view of the TM for the user.

3.4 Document categorisation

In a Topic Map a large number of documents can be pointed out as relevant for specific topics through the use of *occurrences*. An occurrence functions as a link that points to a document that contains information relevant to a topic. This document can have any form – web pages, presentation material, sound clips, videos etc. In order to classify these documents in the thesis Topic Map we wanted to try an automated categoriser, namely Autonomy's Categorizer.

3.4.1 Autonomy

Autonomy[26] software uses advanced pattern matching techniques utilising statistical methods such as Bayesian Inference and Shannon's Information Theory. The software identifies the patterns that naturally occur in a text, through usage or frequency of words. Based on this Autonomy extracts the digital fingerprint of a text (or other information form). Through this, documents with both structured and unstructured information can be categorised as being similar or not. This enables the software to perform various operations on the document, where the one relevant for our purposes is the automated categorisation, which is done with the Categorizer.

The Categorizer allows you to look through a number of documents manually and sort them into relevant categories. These documents are called the training material. The categories can be created like a folder structure, where the folder is the category and the relevant documents for each category are placed in the folder corresponding to the correct category. Any one document can be placed in more than one folder, if the document is judged to belong to more than one category. It can also be discarded if it is thought not to fit into any category. The Categorizer then uses these manually sorted documents to create a digital fingerprint of every category. Once this digital fingerprint is made, documents can be automatically categorised as belonging or not belonging to the existing categories.

3.4.2 Planned use of Autonomy

The categories used by the Categorizer can be seen as corresponding to part of, or all, topics in a Topic Map, and the manually sorted documents that the Categorizer needs can be seen as the occurrences in the TM. Our thought was that since we already had the topics and the occurrences in the thesis TM, it would be easy to use this structure for categorisation. We would only need to translate it to a folder structure and then any number of documents could be automatically categorised and added to the TM as occurrences. This would make the TM as good as any keyword search when it came to finding relevant documents, and it would be uncomplicated to add a large number of documents to it. This could also be used to keep the TM updated – new documents could be categorised and added with minimal human effort. The only thing that needed to be done was to add the code for new occurrences.

Because of the problems we encountered in using the Categorizer we decided not to include automated categorisation in the final thesis TM. We had too few documents to get a good and reliable result. Furthermore our lack of domain knowledge made it impossible for us to categorise the required training documents well enough for the

[26] See www.autonomy.com

result to be reliable even if we had had a sufficient amount of documents. For further discussion on these problems see section 4.3.3.

3.4.3 Summary

The domain for our model of the apolipoprotein research area became broader and more general than expected. It covers many concepts not specific to the apolipoprotein research area that are still relevant to it. While the research part of the map is firmly established in the existing resources and in the research community, the business terms were difficult to find and hard to connect to the apolipoprotein area. Therefore only a few business related terms have been included, such as *Clinical Study*. Also much more modelling by hand was needed than we thought from the beginning. This was mainly because the apolipoprotein area is too specific to be included in any detail in existing resources, but also due to the design of the visualisation tool we used in our implementation.

The visualisation tools available proved not as good as we had hoped. None of them displayed the TM as a graph with the topics shown as nodes and the associations between them shown as arcs. The visualisation tool that we used, the Omnigator, displays the TM as a list of links that allows the user to navigate in the list by clicking on the links. It is not a tool meant for end users, but a development tool.

Due to the problems we encountered in using the Categorizer to categorise the documents for the thesis TM, we decided not to use any automated categorisation. We had too few documents to get a reliable result and too little domain knowledge to be able to categorise the training data required by the Categorizer.

4 Discussion and conclusions

In this chapter we will discuss the problems we have had during our thesis work and present the conclusions we have come to. We will begin by going through the problems we have encountered, both during our domain modelling and during our Topic Map implementation. This discussion follows, like the Topic Map Implementation chapter, a chronological order. The chapter ends with a presentation of our conclusions and some scenarios of possible TM usage.

4.1 Problems with the domain modelling

In using Topic Maps we have come to the conclusion that it is not difficult to model the concepts that are implemented as topics. Difficulties in modelling arises with relationships and occurrences. To model those there is a strong need for domain knowledge. Relationships are also more difficult to find in the existing resources, since these mostly contain terms and do not describe how the terms are related to each other.

4.1.1 Topics

The reason why it is so easy to model the concepts implemented as topics is that one only needs to know the name one wants to represent them by – everything else is optional. Providing one knows the domain well enough to know the relevant concepts, or that one has good resources to extract these concepts from, it is not difficult to build the basic topic structure, implementing simply the names.

It is tempting to add a large number of topics, just because they do not present any trouble. It can be compared to a programmer's wish to add every feature possible to a program just because he can – not because they will be useful for the user. This should be avoided in the TM, for the TM to be of maximum help and service to the user. Only the topics that are really relevant for a domain should be included in the TM.

Even though we have had some difficulties in finding the apolipoprotein specific concepts we needed for our domain, this was problems due to the resources available to us, rather than to the actual modelling. The resources we used (see section 3.1.1) did not contain enough apolipoprotein specific terms, but only gave the general terms associated with them. From the thesauri, for example, it was possible to extract all broader terms to *apolipoproteins*, but no narrower terms.

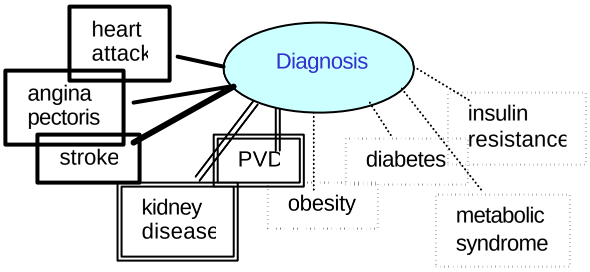
4.1.2 Associations

When it comes to the relationships between topics (called associations in Topic Maps), it is almost impossible to model them without domain knowledge. In the existing resources most relationships are either implicit, and therefore not possible to extract at all, or of the type found in thesauri, i.e. hierarchical. Hierarchical relationships present a restricted and simplified way of expressing the relationships compared to the possibilities given in a TM. Some hierarchical relationships are certainly desirable, to give a subclass/superclass structure, but not as the only relationships in the TM. We do have quite a few hierarchical relationships in the thesis TM, for example *apolipoproteins* are *instance_of lipoproteins*, but we also have a large number of other relationships, of other, more complex kinds. Some examples are *indicates/is measured with*, *contributes to/can be made through measurement of*, and *indicates/is shown with level of* (for all relationships implemented see table 1, section 3.1.2).

Another problem with modelling the relationships, that requires domain knowledge, is that all associations are given in two directions. One must be able to name them from both sides of the relationship. This means that to model a relationship between for example *bloodmarkers* and *atherosclerosis* it is not enough to know that *bloodmarkers indicates atherosclerosis*, one also has to know in what way *atherosclerosis* affects *bloodmarkers*. Sometimes one side of a relationship is available in the resources, but not very often two. When both sides of the relationship are available, they may not be shown together and it takes domain knowledge to see that they are two sides of the same relationship.

Also, to model a relationship so that is mirrors reality, often it has to be between three concepts or more, rather than between just two concepts. The TM does not put any restrictions on how many topics can be part of a relationship, but the modelling is made more difficult when more concepts are involved. The kinds of relationships that involve more than two concepts are even more difficult to find in the resources.

TMs do not support all desirable features that are needed to mirror the reality of a modelled area. One example of this is that different topics have different relevance for the focus and perspective taken in the research community. They may all be necessary to include in the modelling, but they may be more or less relevant for the domain. They may also belong to different subgroups, with internal similarities. These groups may not be possible to express as specific topics, but one still wants to group the topics belonging to these subgroups together. This is not possible to do in a TM. Instead one has to trust the user’s domain knowledge to implicitly add this grouping. In our domain this can be seen in the diagnosis area: *heart attack*, *angina pectoris* and *stroke* are the most relevant diagnoses in the apolipoprotein research (see example below).



An example of a structure in the apolipoprotein research area not possible to implement in the TM

They are also closely related to each other. *PVD* and *kidney disease* are both on the same relevance level, but not closely related to each other. Next comes *diabetes*, *metabolic syndrome*, *insulin resistance* and *obesity*, which are not as relevant, but all closely related. In the TM they are all simply expressed as types of diagnoses, without any internal differences or relations.

4.1.3 Occurrences

For occurrences the main difficulties come with trying to separate documents as belonging to different topics. This is difficult since the topics are both very similar and very close in meaning to each other, whereas the documents are most often general with aspects concerning many of the topics. At the same time it is necessary for the topics to be similar and close, since the apolipoprotein research area is small and the different topics have very diverse significance and consequence.

Another aspect of this is that it is not possible to let a document function as an occurrence for too many topics, as the user will get frustrated by finding the same document representing almost every topic.

A third issue that arises when the occurrences are decided is a question of what the usage of the TM is planned to be. Is it just a map showing the knowledge structures of a specific area? In that case too many occurrences are not good, as they will blur the knowledge map and make the structures in it more difficult to see. If the usage, on the other hand, is to help the user find documents on certain areas, like a normal keyword search or information retrieval system, then many documents are necessary, or the user will not have to use the TM for long before he or she has seen all documents included. For this to work some sort of automated update of the occurrences is also necessary, or they will soon be outdated.

4.1.4 Conclusions

To model a TM a lot of domain knowledge is necessary. Both to choose what concepts to model as topics and how to describe the relationships between them, one needs to have an acquaintance with the domain to fully express it. It is also difficult to find the relevant concepts both for topics and for relationships between them in the existing resources, mostly because the resources are very general, while the apolipoprotein area contains highly specialised terms and relationships. The ideal for the TM domain modelling seems to be a language technologist working together with a researcher with extensive domain knowledge. This way, both the domain knowledge and the knowledge necessary for ontology and taxonomy modelling are represented.

4.2 Problems with the Topic Map implementation

In the next sections we will discuss the implementation of our topic map and some of the problems that surfaced along the way.

4.2.1 Associations

When modelling associations, we encountered difficulties in abstraction when it came to specifying the association roles. What we wanted to do was to have an abstract layer of topics for our top topics. These parent topics would consist the world ontology (see below), and always, if necessary, provide an abstraction to allow using these topics as role players. The abstractions that we could not make were for the topics *blood markers*, *diagnosis*, *atherosclerosis* etc. Therefore we had to specify these both as role players and as role types in the associations (see example 1 below). Doing this meant that some topics were implemented as playing the roles of themselves. It also contributed to an inefficient and exhaustive way of establishing associations, as we had to specify a role for almost every topic we wanted to use in the association.

```
<association>
  <member>
    <roleSpec>
      <topicRef xlink:href="#blood_markers"/>
    </roleSpec>
    <topicRef xlink:href="#blood_markers"/>
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="#atherosclerosis"/>
    </roleSpec>
    <topicRef xlink:href="#atherosclerosis"/>
  </member>
</association>
```

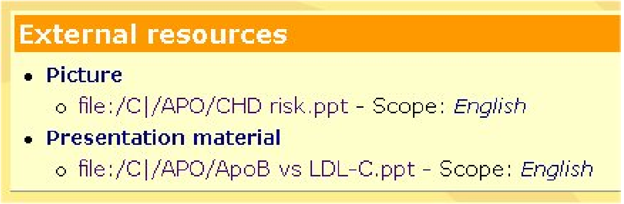
Example 1 – topics as both roles and players in an association

4.2.2 Scope

Because of our chosen bilinguality for the topic map, the characteristics (names, occurrences and roles played in associations) became a difficult part to implement. Since the scope used to get a bilingual topic map could not be used in a satisfactory way, one had to specify every characteristic twice, once for each scope. If the characteristics were not specified in both scopes they were only visible in the unconstrained scope, and not when one of the two scopes was chosen. This, as with the associations, gave an inefficient and exhaustive way of establishing characteristics.

4.2.3 Occurrences

An aspect of the occurrences, that was not satisfactory, was that the reference to them had to be specified showing the URL to the resource (see picture 1).

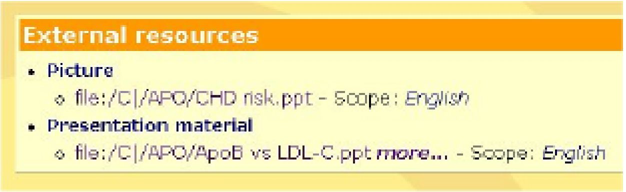


External resources

- Picture
  - file:/C|/APO/CHD risk.ppt - Scope: *English*
- Presentation material
  - file:/C|/APO/ApoB vs LDL-C.ppt - Scope: *English*

Picture 1 - Occurrences

Ideally, there should have been a functionality similar to that which one gets with the element **<a>** in HTML where one can specify a descriptional string for the URL. This is not possible in XTM, however. What one can do, is to give the occurrence an identity and reify a topic that has this identity as its subject identity. One can then, in the Omnigator, see a *'more'* reference to that topic after the URL specified as the occurrence (see picture 2).



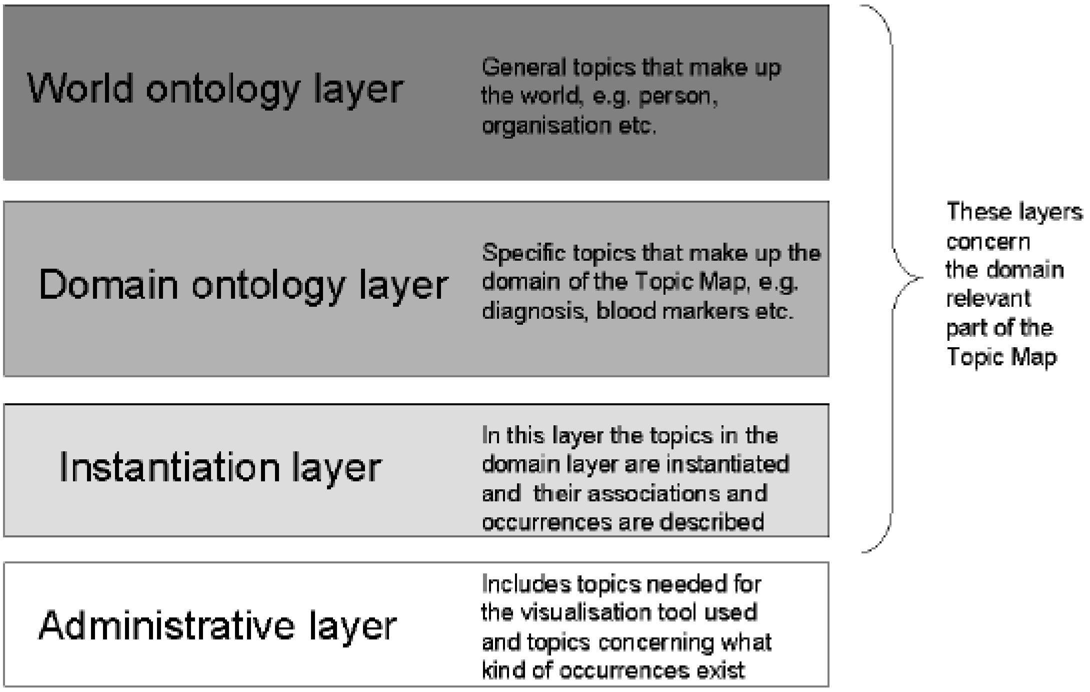
External resources

- Picture
  - file:/C|/APO/CHD risk.ppt - Scope: *English*
- Presentation material
  - file:/C|/APO/ApoB vs LDL-C.ppt *more...* - Scope: *English*

Picture 2 – Occurrences, with 'more' reference

4.2.4 Possible structure for TM

When we implemented the thesis topic map we came across problems like the one described in the association part of this chapter. These problems mostly occurred because of our lack of domain knowledge. Because of this we could not do all the abstractions needed. To avoid this in a future use of the topic map standard, we have outlined a possible structure for the organisation of topic maps in different layers containing different aspects of a domain. Below we will discuss this structure (visualised in picture 3) and see how the thesis topic map differs from it.


World ontology layer
General topics that make up the world, e.g. person, organisation etc.
Domain ontology layer
Specific topics that make up the domain of the Topic Map, e.g. diagnosis, blood markers etc.
Instantiation layer
In this layer the topics in the domain layer are instantiated and their associations and occurrences are described
Administrative layer
Includes topics needed for the visualisation tool used and topics concerning what kind of occurrences exist
These layers concern the domain relevant part of the Topic Map

Picture 3 – the possible structure, topic map layers

The *world ontology layer* consists of knowledge of the world in general. That is knowledge that is applicable on most domains (e.g. person, organisation, organism etc.). From the world ontology, one simply selects topics relevant for the domain one is modelling. This is where the topics needed for abstraction of the top topics of the domain ontology are found. In our topic map we do not have a very well reasoned world ontology layer, even though we are now arguing for this structure. Using it would give a possibility for standardised world ontologies which would make merging of different topic maps easier.

The *domain ontology layer* is where one specifies the upper ontology for the domain one is modelling, the Topic Map Ontology. These topics might or might not be instantiated by the world ontology layer. Topics in this layer also constitute the role types and association types used in the relationship among topics in the domain. Without knowing it when modelling, some domain ontology layer topics can still be found in our implementation (e.g. *measurement*, *method of measurement* etc.).

The *instantiation layer* provides the individuals of the domain, i.e. the leaves of the instantiation hierarchy. These are also the topics that play roles in associations and constitutes the occurrences. This layer is where most of our topics exist (e.g. *proteins*, *lipids*, *lipoproteins*, *apolipoproteins* etc.), which is often contrary to the layered structure we now argue for.

The *administrative layer* consists of topics which are application specific and it contributes information to display, handle, sort etc. topics accordingly. It also consists of topics concerning what kinds of occurrences exists.

The topics used for scopes, one could argue, could belong to any layer since what one might want to use as a scope may exist in any of them. For example one might want to use the topic 'male' as scope (existing in the world ontology layer) as well as the individual topic 'Charles' (existing in the instantiation layer).

Administrative layer

When we created our topic map and used the Omnigator as the visualisation tool we used certain topics which had a specific subject identity. The identity pointed to Published Subject Indicators at the Ontopia website. When one uses these, the TM achieves a certain look in the Omnigator. One example is that the Omnigator makes a distinction between the occurrences that are subclasses to a topic with the identity of the Published Subject. It displays these occurrences in a field called Metadata, (see example 2 and 3, section 3.3), while it displays occurrences which are not subclasses to that topic under a category in the Omnigator called external resources. (see example 2 and 3, section 3.3)

This also gives an administrative set of topics to the map one creates, if one chooses to use the full expressiveness of the Omnigator. Since the Omnigator is a development tool for Topic Map authors it shows all topics available in the map. This makes the administrative set of topics accessible to navigate for the user. When one is about to make an application based on topic maps, one needs to consider carefully what is to be visualised to the user. The involvement of an administrative set of topics can present knowledge in the map that (probably) is not relevant to the domain one is modelling, and which might confuse the users of the Topic Map.

4.2.5 Conclusions

The problems we had during our implementation of the thesis TM where mainly due to two things:

- The bilingualism of our TM
- The chosen visualisation tool

Our TM being bilingual led to an inefficient and exhaustive way of coding. We basically had to repeat everything twice, once for each language. Therefore our opinion is that the TM standard does not yet support the implementation of a bilingual TM in a satisfactory way. The chosen visualisation tool, the Omnigator, also gave us some problems, since it required using certain topics which had a specific subject identity. This was to give the TM a certain look in the Omnigator and it also gave an administrative set of topics to the TM. This in turn led us to the thought that a division of the TM into different layers was to be preferred. The development of different layers of the TM also gave us a solution to another problem, namely that some topics had to be implemented as playing the role of themselves.

4.3 Ideal Topic Map

Ideally the thesis Topic Map would have been modelled on extracts from existing resources. It would display the apolipoprotein area in a way that showed the knowledge structures in the apolipoprotein community-of-interest both from a business and from a research perspective. The documents shown as occurrences for each topic would have been automatically categorised using Autonomy's Categorizer. The visualisation would display the TM as a graph, where the topics were represented as nodes and the associations between them as arcs. It would afford clicking on both the nodes and the arcs and would then display the information available about them, next to the graph representation of the whole TM.

The above scenario has proved impossible to implement in real life, mainly due to:

- The limitations of available resources
- Our lack of domain knowledge
- Autonomy's Categorizer
- The visualisation tool

Below follows a discussion on our problems in implementing the thesis TM concerning these points.

4.3.1 Resources for the ideal TM

In order to model the TM on extracts from existing resources, these resources should be more similar in using the same terms for the same type of knowledge. They should be coded in some structured document format (ideally XML) for easy transfer to XTM. To really benefit from using the resources, these have to contain more instantiated and specific information, not only general. Another alternative is to use them only when modelling *the World Ontology layer* of the TM.

According to the TM standard the resources used as occurrences can have any form. Both structured and unstructured material can be pointed to. Because of this, we do not see it as necessary to point out any specific restrictions or desires on these resources.

4.3.2 The need for domain knowledge

We have discovered, mainly in our modelling of the thesis TM, that the work requires a large amount of domain knowledge. What topics to include, how to express the relationships between them and to which topic to assign which occurrence are all tasks that present a great difficulty unless one knows the domain well.

Also, the representation of the domain shows someone's view on what is important and how to represent that. One important point with a TM is to show a person's or a group's special view of a domain (and to give the possibility of combining this view with others by merging the different TMs). To do this also requires a lot of domain knowledge, since one is not likely to find a certain person's or group's view of a domain in the general resources. It is only the person or group itself who can give that.

4.3.3 Autonomy's Categorizer

In our implementation of a Topic Map we tried to use Autonomy's Categorizer to categorise documents automatically. Our theory was that since we already had the topics and the occurrences in the thesis TM, it would be easy to use this structure for categorisation. We would only need to translate it to a folder structure and then any number of documents could be automatically categorised and added to the TM as occurrences.

When trying this we came upon two problems:

- There were too few documents available.
- The topics were too similar

To get a good and reliable result from the Categorizer at least a few hundred documents are needed, preferably even more. The digital fingerprint extracted from one or two documents per category is not reliable and will not give a good result when used for automatic categorisation. The digital fingerprints will be too similar for all categories to make a consistent categorisation possible. The occurrences in the thesis TM that we wanted to use as the manually sorted training data for the Categorizer, consisted of only 27 documents in all, which meant that no category would have had more than at most around 10 documents.

The other problem was that the topics were too similar to make good categories. It is difficult to separate documents as belonging to different topics, or to different categories, when the topics or categories are very close in meaning. At the same time it was necessary in our modelling for the topics to be close. This is because the apolipoprotein area is small, with many closely related concepts. To be able to model it properly, closely related concepts need to be expressed as different topics. This is in contrast to the requirements for creating good categories. For this the concepts expressed by the categories need to be disparate and clearly defined. If they are not, more documents are needed. This problem is made more difficult to solve by the fact that the documents are not created with a specific category in mind, but rather to explain relevant concepts for the apolipoprotein area as well as possible, which often means that a number of concepts (or categories from the Categorizer point of view) are mentioned and explained in the same document. This problem also means that Autonomy's Categorizer might never work well with TMs, since most TMs will have topics too similar to make good categories for the Categorizer. One can imagine a scenario, though, where the rough sorting at the top level is made by the Categorizer, and the domain expert then does the exact sorting at lower levels.

As always there is also a problem with our lack of domain knowledge. It is virtually impossible without domain knowledge to decide which documents are representative of a certain category and which documents belong to more than one category to make up the training material. To get a good result from the Categorizer, a domain expert is needed, to look through what documents to put in what category. Only if this manual sorting is well-made will the automatic categorisation give a reliable result.

Another problem with an automated categorisation is that it will result in hundreds or thousands of categorised documents. These have to be added as occurrences in the topic map. This have to be done manually or a way of automatically adding them as XTM code will have to be developed. A topic map with hundreds of documents on each topic may also be too large to let the user get a good overview of each topic.

4.3.4 Omnigator

The visualisation tool we used, the Omnigator, is meant to be used as a development tool, not as a user interface for end users. This explains why it sometimes shows topics not relevant for a user of the TM. We would have wanted to use a visualisation tool more aimed at end users, that displayed the TM as a graph, like the models of the domain we have made (see section 3.1.2), but unfortunately we have been unable to find one that meets these requirements. All visualisation tools we have looked at displays the TMs as lists of links. The layout varies, but the basic visualisation idea is the same. This entails a difficulty in testing the thesis TM on users. They find it difficult to see the TM "through" the interface and imagine how it could be used, if it had a different visualisation. The type of visualisation the Omnigator uses, which is very similar to all visualisations we have seen, is a clumsy way of displaying the possibilities of a TM. It does not show any of the map-qualities of the Topic Map idea. It is good neither for end users, nor for the developer, especially not during the modelling stage. The Omnigator is, however easier to use than the other visualisation tools we have tried. It also supports more of the TM standard features, such as the merging of different TMs, which made us choose it.

4.3.5 Conclusions

For the ideal TM to be possible a number of thing are required. Most important is to have well structured resources to use when establishing the topics. If the resources contain all the concepts used as topics, this means that there is an independent description of the concepts available, which is approved across the organisation. This could also function as Published Subject Indicators (PSI) for the TM and thus make it easier to merge different TMs. The modelling of the TM would then be a question of extracting the relevant concepts from the resources rather than self inventing them. To be able to do this, domain knowledge is needed, but when using the resources, rather then inventing oneself, it is not required to the same extent. It is also necessary to have a good visualisation tool, both for displaying the TM and for modelling it. The last thing preferable for the ideal TM is an automated categorisation, to use to categorise the occurrences. This allows hundreds of occurrences to be added, without manual labour. During our modelling and implementation most of the above were not available or possible to use in the intended way.

4.4 Future development and scenarios for TM usage

Topic Maps are good in a number of contexts. They can be used for searching vast amounts of information, structuring and illustrating the knowledge of a domain, and navigating through the knowledge structures and/or information amounts. During our work with Topic Maps a lot of questions and ideas on the usage of them have arisen. The technologies and products around the Topic Map standard are not yet fully developed and therefore leaves a lot to do in the future. It also opens a lot of exciting new possibilities. In this part we will begin by presenting some possible scenarios for TM usage, then discuss the question of keyword search in TMs and go on to discuss possibilities for future development.

4.4.1 Scenarios for Topic Map usage

How can then Topic Maps be used? Based on the experiences we have gained during our modelling and implementation of a TM, we below present some possible scenarios for topic maps usage.

Scenario One

“You are newly employed in a company (or new to an area in general) and do not understand the domain or the routines in the company very well.”

Here topic maps could be used as a navigation tool to improve the understanding of a domain and to help illustrate what is related to what in the domain. If a domain expert has modelled the map, you will get all relevant information. This could also be used for educational needs, both in distance and ordinary teaching. The teacher has one map, which is a model of the course from an administrative perspective, as well as one for the concepts of the course and the relations between them. If it is within a program of courses, the model of the course could contain the relevant relations to other courses.

Scenario Two

“You are about to structure a pool of information and looking for a way to do it.”

A topic map could here be used as an implemented underlying structure for modelling, in an application. Here all types of editor applications for creating and maintaining topic maps and generation of some type of graph representation or other kind of visualisation lie.

Scenario Three

“You are about to search for information on an Intranet or the Internet. You get many hits, and since you do not know the area very well, you do not know what is relevant information.”

In this scenario a topic map could be used as an alternative to the search, where one can navigate a topic map based on the search result and see how the hits in the search are related to each other. A topic map is dynamically built over the results one gets from the search. This will probably involve other types of software for sorting and extracting information from the results to build up the topic map.

Scenario Four

"You are about to add a new source of information to a community page in a corporate portal and want to have a community specific way of organising the categories of information."

A topic map could here be used as the community specific way of organising the categories of information. The topic map would then constitute the new information source. This would also enable navigating in the information source.

4.4.2 Keyword search

When using the thesis TM we have discovered a need for finding information one already knows exists in it. How does one for example find a particular document if one already knows that it is an occurrence in the TM, but one does not know to what topic the occurrence belongs? How is the TM technology related to normal keyword search? For a keyword search to work the addition of synonyms is necessary, or the search will give limited results. This is not a problem in a TM, though, since the standard allows every topic to have any number of names, which means that the synonyms can be added as variant names. There also exists a requirements guide for a query language for Topic Maps, Topic Maps Query Language (TMQL). These requirements are implemented through Tolog (see section 3.2.2), but are still under development. One of the ideas with the TM technology, though, is that in a TM, search is navigating through the TM. To find what one is looking for one needs to navigate through the knowledge structure that constitutes the TM. TMs require a new way of thinking about searching for information. This works against the addition of a keyword search functionality, but for users to fully appreciate a TM, we feel that some kind of search function that allows searching within the TM is necessary.

4.4.3 Future development

The future development of the Topic Map technology gives a lot of possibilities. It can be divided into two categories:

- Development of the TM technology alone.
- Development of technologies and products where TMs interact with other technologies and products.

Development of the TM technology alone

The main development of the TM technology lies in the visualisation part. There already exists standards for how to code TMs in XML and work is being done on a query language. The available visualisation tools, however, leave a lot to be wished for. The ones we have looked at all present the TM very similarly to the one we used, the Omnigator, namely like a list of links, which , when clicked on show the information available on that specific topic. We have found none that represent the TM as a graph, which would be ideal. There are tools for expressing knowledge structures as graphs, but these do not support XTM or any TM standard, but requires adaptation to a special way of implementing the TMs and are therefore not generally applicable.

One of the most important additions that should be made to the Topic Map implementation is the possibility of showing the way one has moved through the map, *the active context*. This could be shown as a graph in a part of the display window. It should allow the user to see the whole route to the current topic, both the topics one has visited and the associations between them one has followed. It would make it easier to understand and remember the connections between topics. This would also make it easier to see the links between topics when these are not of the *instance_of* type. It would make also the non-hierarchical relationships clear. The occurrences one has chosen to visit should also be shown. It should be possible to save all this, as the user could then see exactly where in the map he or she was last time using it, and what documents where looked at. From the thesis TM an example could be something like: *bloodmarkers – lipids – lipoproteins – apolipoproteins – apolipoprotein B:* picture CHD risk.ppt.

It should also be possible for the user to make his or her own occurrence list containing for example relevant research documents. To this could be added all the occurrences that the user finds especially important at the time. It should be possible for the user to have a number of different, personal occurrence lists. This could be implemented as an extra layer of the TM – *the personal layer*. It should also be possible to choose parts of documents to add to the personal occurrence list.

Another possibility with pointing out parts of documents is that a whole document could be pointed to from a superclass topic and different parts of the same document could be pointed to from the different subclass topics. This would solve the problem of documents often being very general, containing information on many topics, while the different topics are highly specific and therefore many topics are covered by one document.

There are two different kinds of potential TM users. The one just browsing and using the TM as a way to find information and the one administrating the TM. The browser needs an interface where the TM is displayed as a graph, while the administrator needs a graphical user interface (GUI) that allows him or her to add and update topics, associations and occurrences without having to code it. This way the administrator can be anyone with domain knowledge, but does not have to have any programming skills. A possibility to add at least occurrences is needed for the browser as well, if the opportunity to add personal occurrences is to be implemented.

Development of TM interaction

The TM could be connected to a keyword search function that allows the user to search for documents that are not occurrences, in resources not indexed in the TM. The potential keywords could be the topics the user has visited on his or her way through the TM. It should also be possible to choose some of the visited topics as keywords if not all of them are of interest to the user. The important thing in this is to make sure that the documents searched for in the keyword search are related to the whole combination of topics and to the relationships between the topics, rather than just the mentioning of one or two of them. There should also be a possibility to save the result list from the keyword search as a personal occurrence. A keyword search function for searching within the TM should also be added.

One could also imagine that one of the occurrences for each topic is an agent that looks through the available information resources (e.g. databases, the Internet, shared document resources etc.) and when clicked presents a list of the latest documents or database entrances on the topic. This agent should be pre-trained. Autonomy offers this kind of agent function, so the possibility to use other parts of Autonomy, apart from the Categorizer is interesting.

4.4.4 Conclusions

During our implementation of a topic map we have come to the conclusions that to model a TM, a lot of domain knowledge is necessary. Both to choose what concepts to model as topics and how to describe the relationships between them, one needs to have an acquaintance with the domain to fully express it. Well structured resources to use when establishing the topics are also necessary (and can in some ways reduce the need for domain knowledge). If the resources contain all the concepts used as topics, this means that there is an independent description of the concepts available, which is approved across the organisation. We would also have liked to have an automated categorisation, to use to categorise the occurrrences. This would allow hundreds of documents to be added without manual labour.

Most of the limiting factors of TM usage today lie in the software. There exists no good visualisation tool for TMs. We have found none that displays the TM as a graph, which to us is the ideal visualisation. Neither have we found any software that provides an easy GUI for administration of a TM. TMs are not supported by many other applications. All this is due to the fact that the TM technology and standard are fairly new. Many things are still under development. With time many new applications will evolve, and then the TM technology opens many new and exciting ideas for navigating and structuring large amounts of information.

The basis of the Topic Map idea, *topics*, *associations*, *occurrences*, are simple to model and to understand. The full expressiveness of a Topic Map is achieved through the concepts of *scope* and *subject identity*. *Scope* allows multiple viewpoints and customised views of a topic map. However, it does not support the implementation of a bilingual TM in a satisfactory way *Subject Identity* gives a one-to-one relationship between a topic and a real world object and thus provides a way of identification and disambiguation of topics.

It is important to spend most of the time and energy on the ontology modelling. A good ontology means a good Topic Map. Building a Topic Map one can start small – the Topic Map can then grow ”organically”, and also be merged with other Topic Maps.

Bibliography

Berners-Lee, T. (1998). *Semantic Web Road Map*. The World Wide Web Consortium (W3C), http://www.w3.org/DesignIssues/Semantic
Accessed 14/02/02

Biezunski, M. (1999). *Topic Maps at a Glance*. XML Europe 99
http://www.infoloom.com/tmsample/bie0.htm
Accessed 03/02/02

Blackburn, P. and Bos, J. (2000). *Presentation and Inference for Natural Language. A First Course in Computational Semantics*. University of the Saarland, Computational Linguistics. Draft available at http://www.comsem.org

Checkland, P. and Holwell, S.E. (1998). *Information, Systems and Information Systems*, Chichester: John Wiley & Sons.

Davenport , T. H. and Prusak, L. (1998). *Working Knowledge: How Organizations Manage What They Know*. Boston: Harvard Business School Press.

Detlor, B. (2000). The Corporate Portal As Information Infrastructure: Towards a Framework for Portal Design. *International Journal of Information Management* 20, 91-101

Dias, C. (2001). Corporate Portals: A Literature Review of a New Concept in Information Management. *International Journal of Information Management* 21, 269-287

Fellbaum, C. (1999). Introduction. In Fellbaum, Christiane (ed). *WordNet: An Electronic Lexical Database.* Cambridge: The MIT Press.

International Organisation for Standardization (1986). *Guidelines for the Establishment and Development of Monolingual Thesauri.*
http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=7776
Accessed 25/02/02

Organisation for Standardization (2000). *ISO/IEC 13250 Information Technology – SGML - Topic Maps.* International
http://www.y12.doe.gov/sgml/sc34/document/0129.pdf
Accessed 31/01/02

Jacobson, I., Ericsson, M., and Jacobson, A. (1995). *The Object Advantage: Business Process Re-Engineering with Object Technology*. Reading: Addison-Wesley.

Mack, R., Ravin, Y. and Byrd, R.J. (2001). Knowledge Portals and the Emerging Digital Workplace*. IBM Systems Journal*, vol. 40, no 4, 925-955

Pepper, S. (1999). *Euler, Topic Maps, and Revolution*
http://www.ontopia.net/topicmaps/materials/euler.pdf
Accessed 31/01/02

Pepper, S. (2000). *The TAO of Topic Maps: Finding the Way in the Age of Infoglut*
http://www.ontopia.net/topicmaps/materials/tao.pdf
Accessed 31/01/02

Pepper, S. and Grønmo, G. (2001). *Towards a General Theory of Scope*
http://www.ontopia.net/topicmaps/materials/scope.htm
Accessed 26/02/02

Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* New Jersey: Prentice-Hall.

Sigel, A. (2000) *Towards Knowledge Organization with Topic Maps*. XML Europe 2000, http://www.gca.org/papers/xmleurope2000/papers/s22-02.html.
Accessed 18/12/01

Swartz, A. (2001). *The Semantic Web in Breadth*.
http://www.logicerror.com/semanticWeb-long
Accessed 14/02/02

Truog, D. (2001).*How the X Internet Will Communicate.* The Forrester Report
http://www.invention-machine.com/DOWNLOAD/0,2254,20774,00.pdf
Accessed 11/03/02

Topicmaps.org. (2000). *XML Topic Maps (XTM) 1.0 Specification*
http://www.topicmaps.org/xtm/1.0/xtm1-20010806.html
Accessed 31/01/02

These appendices show code examples from our implementation in XTM code. Only a fraction is shown, because of the similar structure of all the code. To include all code would have covered about 30 pages.

Appendix I

In this appendix a topic map is shown. For space saving reasons only one topic is shown to its full extent. The rest of the topics in the topic map are shown only with the **`<topic>`** element. This is because the rest of the topics' structures are the same as the one shown.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<topicMap xmlns=http://www.topicmaps.org/xtm/1.0/
          xmlns:xlink="http://www.w3.org/1999/xlink">

  <topic id="lipids">
    <instanceOf>
      <topicRef xlink:href="psi.xtm#blood_markers"/>
    </instanceOf>
    <baseName>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
      <baseNameString>Lipider</baseNameString>
    </baseName>
    <baseName>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <baseNameString>Lipids</baseNameString>
    </baseName>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#def"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
      <resourceData>Fettliknande substans som existerar i människans vävnad
      och utgör en viktig del av människans kost.</resourceData>
    </occurrence>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#def"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <resourceData>Fatlike substance which exists in human tissue and forms
      an important part of the human diet.</resourceData>
    </occurrence>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#report"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
      <resourceRef xlink:href="file:///C|/APO/guidelines.pdf"/>
    </occurrence>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#report"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <resourceRef xlink:href="file:///C|/APO/guidelines.pdf"/>
    </occurrence>
    <occurrence>
      <instanceOf>

        <topicRef xlink:href="psi.xtm#pres"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
      <resourceRef
            xlink:href="file:///C|/APO/new trends in the lipid field.ppt"/>
    </occurrence>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#pres" />
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <resourceRef
            xlink:href="file:///C|/APO/new trends in the lipid field.ppt"/>
    </occurrence>
  </topic>
  <topic id="cholesterol">
  <topic id="triglycerides">
  <topic id="ffa">
  <topic id="hdl">
  <topic id="ldl">
  <topic id="vldl">
  <topic id="proteins">
  <topic id="lipoproteins">
  <topic id="apolipoproteins">
  <topic id="apolipo_a-1">
  <topic id="apolipo_b">
  <topic id="apolipo_cII">
  <topic id="apolipo_cIII">
  <topic id="apolipo_e">
  <topic id="glucose">
  <topic id="insulin">
  <topic id="hba1c">
  <topic id="pogtt">
  <topic id="ivett">
  <topic id="inflammation">
  <topic id="coagulation">
  <topic id="fibrinolysis">
</topicMap>
```

Appendix II

This appendix shows how one can name one’s topic map by reifying a topic whose subject identity is the id in the **`<topicMap>`** element. This also means that one can assign characteristics to the topic map itself. It also shows all merge directives we used for merging our separate topic maps into one.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<topicMap id="t-t-m" xmlns=http://www.topicmaps.org/xtm/1.0/
                     xmlns:xlink="http://www.w3.org/1999/xlink">

  <!-- Naming of the topic map   -->
  <topic id="thesis_topic_map">
    <subjectIdentity>
      <subjectIndicatorRef xlink:href="#t-t-m"/>
    </subjectIdentity>
    <baseName>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
      <baseNameString>Examensarbete Topic Map</baseNameString>
    </baseName>
    <baseName>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <baseNameString>Thesis Topic Map</baseNameString>
    </baseName>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#desc"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#swedish"/>
      </scope>
        <resourceData>Examensarbete Topic Map extraherar information från
        apolipoprotein forskningsprojektet för att upprätta en Topic Map som
       täcker en begränsad uppsättning forskningstermer och deras
       relationer. Detta kombineras med en Topic Map som behandlar en
       begränsad uppsättning verksamhetstermer.</resourceData>
    </occurrence>
    <occurrence>
      <instanceOf>
        <topicRef xlink:href="psi.xtm#desc"/>
      </instanceOf>
      <scope>
        <topicRef xlink:href="needed.xtm#english"/>
      </scope>
      <resourceData>The Thesis Topic Map will extract information from the
      apolipoprotein research project to establish a Topic Map covering a
      limited set of research terms and their relationships combined with a
      Topic Map covering a limited business vocabulary.</resourceData>
    </occurrence>
  </topic>

  <!-- All merge directives   -->
  <mergeMap xlink:href="blood_markers.xtm"/>
  <mergeMap xlink:href="diagnosis.xtm"/>
  <mergeMap xlink:href="atherosclerosis.xtm"/>
  <mergeMap xlink:href="physiological_methods.xtm"/>
  <mergeMap xlink:href="psi.xtm"/>
  <mergeMap xlink:href="needed.xtm"/>
  <mergeMap xlink:href="associations_and_role_types.xtm"/>
  <mergeMap xlink:href="clinical_study.xtm"/>
</topicMap>
```