


Handelshögskolan
VID GÖTEBORGS UNIVERSITET
Institutionen för informatik

2006-06-06

Exploring a Case of Knowledge Infrastructure in the Pharmaceutical Environment
- A case study at AstraZeneca

Abstract

Knowledge today tends to be an organization's prime tool in order to compete on the business market. For companies like AstraZeneca where research is one of the corner stones there must be an easy way for accessing and sharing knowledge. To be able to meet these demands organizations have to rely intensively on a stable and well-designed knowledge infrastructure. In this masters thesis we study one of AstraZeneca's knowledge sources named AZ Glossary. We investigate how it works as a knowledge infrastructure and how it can be improved to better spread knowledge to the people in the organization. A problem for AstraZeneca today is that the information in the glossary is not machine processable, which means that the information cannot be easily used by other programs and applications. This structure limits the utilization of the glossary and is therefore also affecting the knowledge infrastructure in a less satisfactory way. An ethnographic study based on in-depth interviews with key users and administrators of the glossary was used for mapping AZ Glossary and its organization and usage. An examination of the semantic technologies including languages as XML, RDF (Resource Description Framework) and SKOS (Simple Knowledge Organization System) was carried out in our search to improve the glossary. As a result recommendations are given involving new technical functions tied to the glossary combined with organizational changes to enhance the knowledge infrastructure and knowledge sharing in AstraZeneca.

Keywords: Knowledge Infrastructure, Semantic Web, RDF, SKOS, Knowledge Sharing and Machine Processable

Authours: Johan Engman, Martin Holmberg.
Supervisor: Magnus Bergquist

Masters Thesis, 20 points

Acknowledgement

During our stay at AstraZeneca many people have been most helpful and we would therefore like to express our gratitude. Kerstin Forsberg our organizational supervisor at AstraZeneca has been a tremendous help with her contributions, never ending interest and ideas.

We would also like to express our gratitude to our academic supervisor Magnus Bergquist for his support and many constructive inputs and ideas. Further we would like to thank all the employees at the Information Strategy department, especially Elof Dimenäs and Margareta Mörk, for being very nice and helpful to us during our stay. Last we would like to thank all the selected employees for their cooperation and patience when taking part in our interviews, particularly Thomas Rasmusson and Inger Hägglöf for their continuous support through the whole project.

Gothenburg, 2006-05-29

Johan Engman **Martin Holmberg**

Table of Content

Table of figures

1 Introduction

In the first chapter we introduce the topic and give a short background. The purpose and main question will be presented. Finally we show the disposition of the thesis.

1.1 Background

The role of knowledge management is getting more and more important in today's business environment, for companies like AstraZeneca where research is one of the corner stones there must be an easy way for accessing and sharing the knowledge. A major problem within knowledge sharing for AstraZeneca today is that the information in common information sources such as glossaries and thesaurus are often not structured in a way that makes them machine processable, which means that the information cannot be easily accessed by other programs and applications but only by humans. This limitation holds back the utilization of these information sources. By changing the coding in some of the files the use of the information for different programs and applications all over the organization could be enabled.

One possible solution for this architectural information issue could be with use of the semantic technology. The base in the semantic technology family is languages such as XML, RDF (Resource Description Framework) and different RDF vocabularies such as OWL (Ontology Web Language) and SKOS (Simple Knowledge Organization System), which are used for creating, publishing, searching and structuring semantically rich information. The other dimension is about utilizing the knowledge in AstraZeneca to take advantage of the people in the organization and their common expertise.

1.2 Case – AZ Glossary

AstraZeneca is a large organization with many different departments that are working in different areas. Throughout the years a set of commonly used terms have arisen, and to agree on common definitions AstraZeneca decided to collect all the terms together with their explanations in one place, as an official information source. AZ glossary is a digital glossary that contains terms that are used in the different departments of AstraZeneca. The glossary is only containing terms that are specific for the organization and not the ones that can be found in any other common dictionaries. One purpose with the glossary is to create a standardized terminology in the organization. This is achieved by the help of different communities whose main task is to collect terms and agree upon their definitions in their specific area. Everyone in the company can access the glossary through the organization's intranet and find explanations for the terms that they are looking for. With this service AstraZeneca has established a "standard terminology" that the employees can fall back on and prevent misunderstandings or disagreements of the meanings of the terms.

Today AZ Glossary is built as a normal database with the terms saved in standard text columns. The terms are categorized by their primary owner, which is one of the eight communities that have been created to assess the quality of terms used in the glossary and the organization. In the current situation the terms are not machine processable and therefore the underlying information of the terms cannot be reached

and used by other programs and applications. This limitation holds back the glossary's true potential.

We will look at the possibility to make the terms in the glossary "machine processable" by using the semantic techniques RDF (Resource Description Framework) and SKOS (Simple Knowledge Organization System).

1.3 Purpose & Question at issue

The purpose of this thesis can be divided into two categories, one organizational and one academic. From the organizational view this thesis is a feasibility study of the possibilities to make data machine processable with semantic technology and how that can contribute to an improved sharing of knowledge in the AstraZeneca organization. The main focus of the work is on the use of the application AZ Glossary and the possibilities to improve its technical architecture, to enable other services to use the content in the database. We will also give recommendations on how to improve the use of AZ Glossary and the organization behind it. A small design proposal will be constructed to show how the semantic technologies SKOS and RDF can be used to improve the dissemination of the content in the glossary.

The academic purpose is to investigate if and how AZ Glossary works as a part of the knowledge infrastructure in AstraZeneca and how it contributes to the spreading and sharing of knowledge in the company.

Based on both the academic and organizational perspective the main question at issue is formulated as follows:

> *"How can AZ Glossary be seen as a part of AstraZeneca's knowledge infrastructure and how can the implementation of semantic technologies improve its function as one?"*

1.4 Delimitation

AstraZeneca is a global company with thousands of employees all over the world, a lot of sources to knowledge are to be found and the knowledge infrastructure in the organization is very complex. The work in this thesis is concentrated on the AZ Glossary and how it works as a knowledge infrastructure, AstraZeneca's general knowledge management and knowledge sharing are not further investigated. Our work will be concentrated on the technical structure and the use of AZ Glossary, the content and its quality will not be closely reviewed.

The technical part of our work is limited to the coding of a base in SKOS/RDF that should serve as a base for further development. No full technical solution will be presented.

1.5 Disposition

The Thesis is structured in the following way:

Chapter 1, Introduction: The first chapter gives the reader an introduction to the thesis. Purpose, background and problems are explained to provide the readers with an understanding to their forthcoming reading.

Chapter 2, Method: The method chapter describes the methods used in our work. A short introduction to ethnographic is also given.

Chapter 3, Theories: The theoretical framework is introduced; the theories are presented and explained.

Chapter 4, Results: All the gathered material in our research is presented.

Chapter 5, Discussion: In this part the results and the theories are tied together and analyzed together with our problem.

Chapter 6, Techniques: This chapter gives an introduction to the techniques that are used in the Design chapter. In this chapter all technical terms used in the thesis will be explained and investigated.

Chapter 7, Design: An example of how the techniques can be used in the presented case.

Chapter 8, Conclusion: The conclusion is a short summary of the thesis and the result and how it responds to the question at issue are discussed and what implications it might lead to.

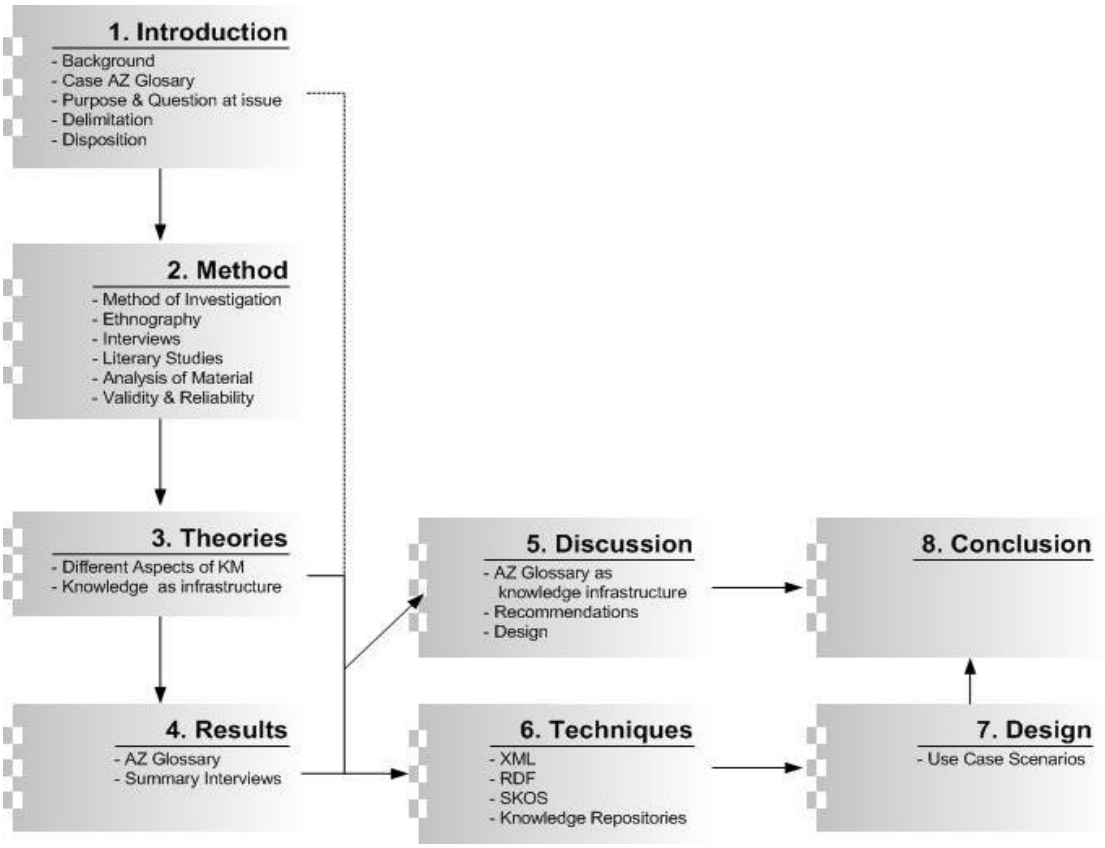


Figure 1: The picture shows the disposition of our thesis and the connections between the different chapters.

The picture (Figure 1) shows the connections between the different chapters in our thesis. The questions and case, presented in the first chapter, are together with the theories in chapter three and the results from chapter four, the base in our discussion chapter. In the chapter Techniques we describe the Semantic Technology, which is the base in the design chapter together with the Results. In the conclusion everything is summarized.

2 Method

In the second chapter we present the way we have worked. We give a brief introduction to the essential parts of ethnography and explain the methods used in our work and discuss why we chose to work as we did.

2.1 Method of investigation

The studies in this thesis are done with an ethnographic method, which is based on in-depth interviews with people in the AstraZeneca organization that are linked to the work with or use of AZ Glossary.

The objective with the investigation was to find out how the application AZ Glossary was used throughout the organization and if it is applicable on the theories about Knowledge Infrastructure and Knowledge as Infrastructure. Therefore we needed to know exactly what the application is used for, who are the users and why does it look the way it does. To get the right answers it is important with probing and the underlying causes are necessary to find out. This would not be possible with a method of the quantitative kind, which is the reason for choosing the qualitative approach.

2.2 Ethnography

Ethnography is a research method used for understanding which effect certain activities have on the people performing these activities. It means to reach for an understanding of the circumstances in which the activities can occur – the circumstances that give the meaning to these activities. (Harper, 2000)

Ethnography involves a range of different methods for finding information, for example participant observation and interviewing. Interviews can be conducted in several ways; often a separation is made between structured and unstructured interviews. However, according to Hammersley and Atkinson (1983) there are no such things as structured or unstructured interviews. All interviews are structured, just like any other social interaction, by both the interviewer and the interviewee. The distinction should lie between standardized and reflexive interviewing. In the standardized form not much space is given to dialog and follow up questions while the reflexive kind is very flexible. Hammersley and Atkinson (1983) claim that ethnographers do not decide beforehand what questions should be asked in an interview, they neither ask exactly the same questions in every interview. The most common way of doing qualitative interviews is to bring an interview guide with areas to be covered, the interviewer can then control the interview with the use of follow up questions and discussion and make sure that nothing is left out. This form of interview is also known as focused interview. (Langemar, 2005) How reflexive or standardized the interview should be is connected to the problem of investigation, the more open problem the more reflexive it should be. The interview guide should cover the whole area of the investigation and nothing else and it should also be able to give an answer to the problem of the research.

2.3 Interviews

To get an understanding of how the AZ Glossary really was used in the organization, interviews were conducted. We chose people that were related to the glossary in one

way or another. Mainly the interviewees were chosen from three different groups: end-users of the glossary, administrators and those with interest to use the content of the glossary in other services.

Most of the interviews were carried out as phone interviews due to great distances, but in a few cases we had the possibility to talk face-to-face. We decided to use the reflexive interview technique in our investigation. Interview guides were put together, one for each of the three groups. We stated the basic questions and left space for free talking and follow up questions. All the interviewees received a copy of the interview guide before the interview to prepare themselves.

In total 10 people were interviewed. Four of them represented the administration side, three were people with interests to use AZ Glossary in other contents and the last three were normal end-users. All of the interviews were conducted and transcribed in Swedish and then they were summarized in English. The interviews are not presented in their full version in this thesis, the relevant material from all of the interviews are put together and presented in the result part. All the quotes presented in the interviews are translated to English from the original interview. To keep the identity of the interviewees concealed no names are given in the text. Instead all quotes are followed by the interviewee's role in AZ Glossary. See appendix 1, 2 and 3 for the full versions of the three different interview guides.

When performing an ethnographic research in an organization with purpose of making profit, it can be difficult to get full and totally objective answers. Sometimes people are afraid of exposing details about sensitive information and how they work. We had that in mind when performing our interviews and were ready to review the material extra carefully. However, in our research in the AstraZeneca organization we got the feeling that everybody was really helpful, showed interest in our work and gave us open and trustworthy answers on all our questions. Therefore we do not think there is any reason in that matter to doubt the reliability of our gathered material in the interview part.

2.4 Literature Survey

The information that describes the technology parts was received by searching in literature. To get deeper in to the subject the project started with a lot of research of information about RDF, SKOS and the Semantic Web. This gave us a greater understanding of the subject and an overview of the area, it also became a backbone to fall back on in the continuous work. Most of the information in this area is found in articles of different kinds and on websites of the organizations behind the development.

2.5 Analysis of material

All our interviews were recorded on tape and then transcribed to files on our computers. After each interview we analysed what had been said and how the information could be tied together with our theories. Even if we did not start to compose our analysis chapter until after the interview session, the work was simultaneously going on in our minds.

The interviews gave us a wider view of AZ Glossary; we saw how the organization behind it works and how it is used throughout the company. This information was

used to analyse how the AZ Glossary works as a Knowledge Infrastructure in the AstraZeneca organization.

Easterby-Smith et al. 1991 discusses two main methods when analysing qualitative data, "Content analysis" and "Grounded theory". The differences between the two methods are that the grounded theory is more holistic, inductive and goes closer to the data, while content analysis is more bitty and deductive. *"The grounded theory provides a more open approach to data analysis which is particularly good for dealing with transcripts".* (Easterby-Smith et al., 1991) (P.108) We found the grounded theory appropriate for our material and followed the seven steps model presented in the book:

1. **Familiarize with the subject**
   By re-reading the transcripts and re-listening to the tapes it is possible to find out more about the attitude and the level of confidence of the interviewee. This tells a lot about the reliability of the interview.
   Comment: When transcribing our interviews we listened extra carefully and searched for signs that could affect the reliability of the interviewee. As we mentioned before we do not think there is any reason to question the reliability of the interviews.

2. **Reflect**
   The gathered information should be categorized and non-relevant information taken out. Evaluation of data is done, is the information enough? Have any new questions arisen?
   Comment: The first step was to transcribe all the interviews, after that we started to work with the material. Sift out all the non-relevant information and focus on important aspects that were coming back several times.

3. **Conception of concepts**
   In this phase there are normally some concepts or variables that are vital for understanding what is going on. However, it is too early to see the true meaning of these concepts, are they valid and reliable? Are they related in a consistent way to how the individual really sees the picture or has there been a misinterpretation?
   Comment: Keywords were taken out from the material and the different interviews were compared to each to insure that everybody was speaking the same language.

4. **Cataloguing of Concepts**
   When it is assured that the identified concepts are indeed what the interviewees meant during the interviews they can be categorized and written down for further analyse.
   Comment: After the found keywords were confirmed, we wrote them down as headlines for the summary of the material. A few of these headlines were also used in the discussion chapters.

5. **Recoding**
   When all the concepts have been gathered they have to be compared to each other in the situations they occurred. One person may have a different meaning with a term than other people.
   Comment: This step was done simultaneously with step three.

6. **Linkage**
   The analytic framework and explanations are coming clearer. Emerging patterns are found and concepts should be fit together. The identified concepts and variables can now be linked to a more holistic view. This phase often results in a first draft. This draft is produced in a quite early stage of the project and should therefore be reviewed by others.
   Comment: In this phase we started to analyze the material. This analyze later became the base in our discussion chapter.

7. **Re-evaluation**
   Given feedback from others, the work with the first draft goes on. Some areas might not be complete and the maybe a change in the structure is needed.
   Comment: We had close contact with both our supervisors throughout the whole project and they assisted us regularly during our work and gave us useful criticism.

2.6 Validity and Reliability

In quantitative research the terms validity and reliability refer to if the research measures what it is supposed to measure and how trustworthy the result is. When it comes to more qualitative research, some people do not even use the terms, for instance, reliability is sometimes used for telling the reliability and trustworthiness in surveys, and surveys are not used in qualitative research. In quantitative interviews it is important for the reliability that the questions are always asked in the same sequence in every interview, which is not the same in qualitative interviews where it is more important that all relevant areas are covered. Those areas can be different for different people. This makes it more alike content validity in quantitative methods. (Langemar, 2005) However, when used in qualitative researches the term reliability can be defined as*: "Will similar observations be made by other researchers on different occasions".* (Easterby-Smith et al., 1991) P.41 Seen to our study in the AstraZeneca organization we think that our result is reliable. If other researchers would do the same investigation the result would probably be about the same. The interviews were performed with a wide range of people from different parts of the organization. The common opinion about the glossary was pretty much the same regardless of position in the organization, even if people of course had some different thoughts of the details.

The term validity is more applicable in qualitative methods where it is usually referring to things such as quality, trustworthiness, and meaningfulness. Important aspects are the amount of interviews and in what degree the interviewees can be representative for the organization. Looking at our qualitative research in the AstraZeneca organization, we have not conducted that many interviews but it should be more than enough for the purpose of this investigation. The people with the most important knowledge about AZ Glossary, such as owner and designer, were interviewed to get a picture of the area. To get a broader view we also complemented the investigation with the perspective from end-users and other people with interests for the application.

3 Theoretical Framework

In this chapter we describe theories from different aspects of knowledge management that are applicable on our case with AZ Glossary. The presented theories will serve as one of the cornerstones in the Discussion chapter.

The chapter starts with some general aspects of knowledge and knowledge management and in the second part we go deeper into Ole Hanseth's theory "Knowledge as infrastructure".

3.1 Knowledge Management

Knowledge management is crucial for today's organizations due to that knowledge itself has become the primary strategic resource. Therefore, the interest around knowledge management has increased and it is used as an important strategy tool for improving organizations' competitiveness and performances. (Wong and Aspinwall, 2006)

An organization's knowledge base and the ideas and the insight that lie in the heads of the people working there, are the things that set the value to the organization. It can often be difficult to define what knowledge really is because of differences between data, information and knowledge. Most often data is described as pure raw facts like a set of numbers, but it cannot be used for anything with out being organized and structured by a human, which in that case it becomes information, something that can be interpreted and useful for a human or machine. Information that is meaningful and has a value added to it because it has been filtered through a human mind fits the meaning of knowledge, which has the highest value compared to data and information. (Wong and Aspinwall, 2006)

Further knowledge can be classified as either tacit or explicit, where the first is primarily knowledge that is stored in the minds of humans and is often hard to transfer to other humans or to documents. Explicit is the opposite side of knowledge than tacit, which has been transformed or expressed to other humans, documents or code in either physical or electronic form.

Blackler speaks of five categories of knowledge, "embodied" and "embedded" that symbolize knowledge located in one's body and routines. Knowledge located in the human brain, dialogues and symbols categorizes in "enbrained", "encultured" and "encoded". These five categories could belong to tacit or explicit, or a little bit of both. (Blackler, 1995)

In general terms, knowledge, when viewed as an object, can be perceived to be any piece of idea, insight, know-what, know-how or meaningful information that can be used to achieve an objective, (Wong and Aspinwall, 2006).

According to Gupta et al. (2000), knowledge management can be defined as "*a process that helps organizations find, select, organize, disseminate and transfer important information and expertise necessary for activities such as problem solving, dynamic learning, strategic planning and decision making. (Wong and Aspinwall, 2006)*

Another view from Liebowitz (2003) is to describe knowledge management as *"dealing with capturing, sharing, applying and creating knowledge in an organization to best leverage this resource internally and externally. (Wong and Aspinwall, 2006)*

A knowledge organization practises knowledge management, which includes sharing, storing, organizing and transferring the knowledge. However, in order to make knowledge management function properly it requires to be supported by a knowledge infrastructure. This is because an organization has to manage its knowledge technically through an infrastructure in order to make it reachable and accessible.

"In fully fledged form, a knowledge infrastructure is one of the organization's core tools and, like the nervous system, it links the other tools. A powerful knowledge infrastructure strengthens the capabilities of the organization; without one, an organization functions at diminished capacity." (Sivan, 2001)

Further, knowledge itself can be seen as infrastructure, it does not need to be all about techniques. Ole Hanseth's theory about knowledge as infrastructure places the interest around the concept of relationships, inter-dependencies and standards of knowledge. This is also very important to become aware of when trying to obtain a rigid infrastructure for managing knowledge.

3.2 Knowledge as Infrastructure

Technology in general signifies and is expected to enable more efficient ways for people and organizations to perform different working tasks. Radical paradigm changes in technology, especially with Information Communication Technology (ICT), have enabled communication from locally within an organization to communication all over the world. New technology most certainly offers better ways for organizations to carry out their type of business, but it is not done in a flash it takes time. In fact, ICT is more often designed to rationalize, assist and maintain activities in the current way they are already performed. Therefore, new knowledge is required to better understand the new design that comes with this paradigm change to make organizations' business processes better.

The characteristics of knowledge can be seen from various perspectives from different communities that are interested in issues in organizational learning, innovation and knowledge management. Some communities describe knowledge as something built up as different elements that we can store in our brain or in a computer. Other communities describe knowledge as a cognitive material, with knowledge such as explicit or tacit. There are more ways knowledge can be looked upon, as in a phenomenological perspective, meaning knowledge that is personified and rooted into one's physical body and practices. Knowledge can also be seen as embedded into institutions and material structures like buildings and information systems. Ole Hanseth uses all these views on knowledge merged together but adds one aspect, which is that individual pieces of knowledge are not independent of each other. This is because knowledge is very systemic, meaning that these individual pieces of knowledge are linked together in different ways into complex structures. Further, these structural features play an important function in the way knowledge can be processed, when it comes to construction, distribution and implementation of knowledge. (Hanseth, 2004)

To understand the theory knowledge as infrastructure first one must go back to basic knowledge concepts. Looking at knowledge as a network including actors in the network, two fundamental assumptions are made. Step one; individual pieces of knowledge are related and mutually dependent of each other. Step two; various individuals adopt the same piece of knowledge, and that piece of knowledge is embedded into routines and practices, these routines and practises are linked together and become interdependent. (Hanseth, 2004)

Internet is at its current state an example of a paradigm in the information communication technology category. Looking at it as a knowledge network, Internet is made up by a huge number of computers connected to each other and to make the communication possible, standards are crucial. Compatibility standards as TCP/IP protocols are necessary to enable computers around the worlds to connect and communicate. This standard is defined as "horizontal". Other important standards are "vertical" standards that relate more to in depth technology as software modules including file systems, operating systems etc. In a vertical standard the structure between applications and operating system can also be seen as a paradigm, because the relationship between them has become a standardized structure and thereby the have become interdependent.

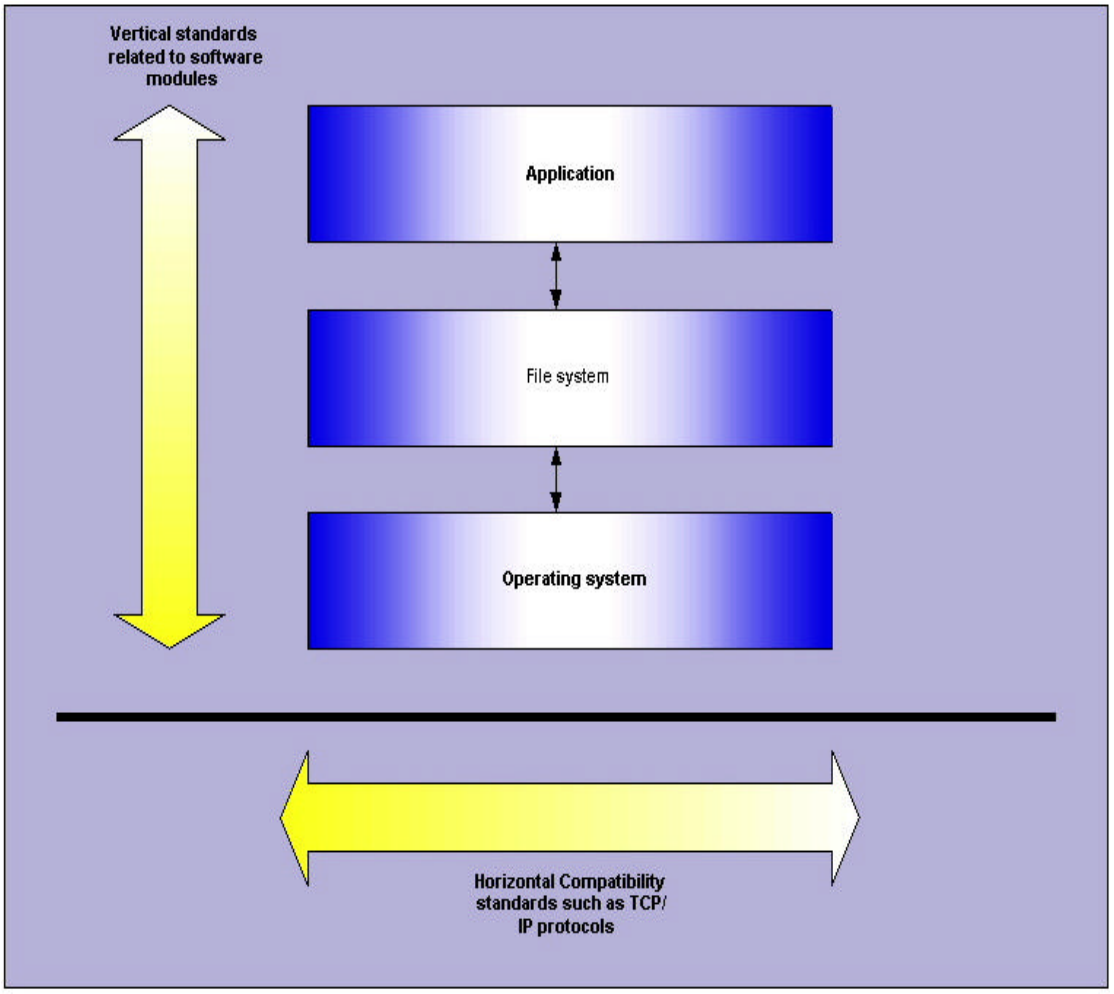


Figure 2: The picture illustrates horizontal and vertical standards.

The concepts of relationships, interdependencies and various kinds of standards in a knowledge network are not limited to operation systems, Internet or technology in general; it can be applied to many areas. Like common areas or organizations that have actors collaborating, sharing and communicating knowledge through a network in the organization. For example, doctors with different specialties in hospitals share knowledge between them to better understand how to cure patients, but in addition to the pieces of knowledge upon such collaboration a certain level of compatibility and standardization is necessary. (Hanseth, 2004)

Considering knowledge as a network also implies the theories known as "information economics". The key concepts from these disciplines are network externalities, increasing returns, path-dependencies and lock-ins. (Shapiro, 1999) in (Hanseth, 2004).

3.2.1 Network Externalities

"Network externalities denote the fact that an economic transaction may have effects for a network of actors external to the transaction itself: those actors not involved in the transaction itself". (Hanseth, 2004)

Trying to explain this, parallels can be drawn to when a person buys an Internet connection that includes an email address. The transaction will have an effect on those people already having an email, because the existing users will have one more person that they are able to send emails to. It is not the technical functions that come with the email system but the number of people that are using email as a standard that determines the value of the system.

The fact is, when a user adopts an email system, that specific user is implementing a specific standard. Further the value of the email systems for the pre-existing users, and for the whole network, increases due to the fact that there is one more user to with whom they can communicate. The same principle goes for knowledge. (Hanseth, 2004)

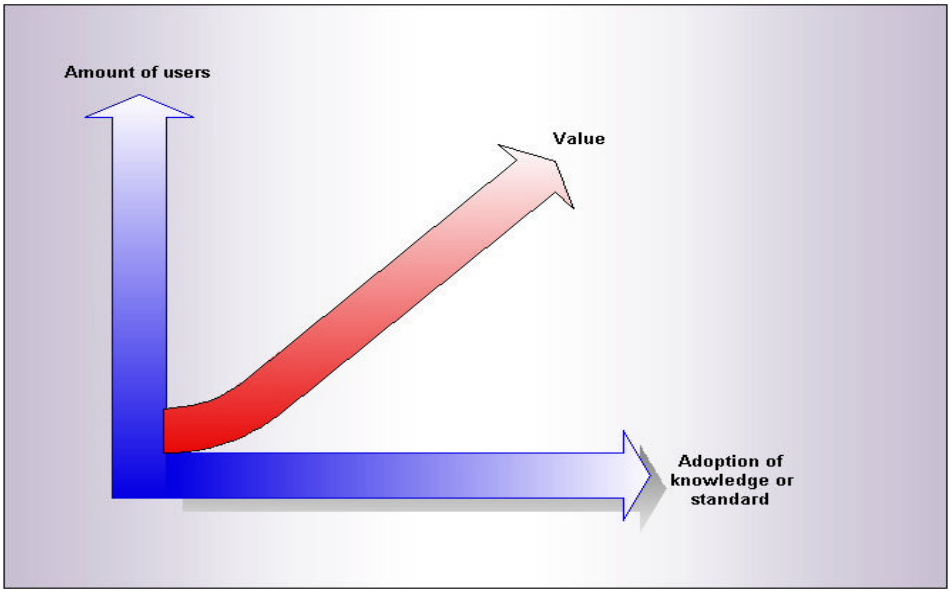


Figure 3: The graph shows how the value of knowledge or a standard increases the more users adopting it.

3.2.2 Increasing Returns

For the users, network externalities like the example with the email leads to "increasing returns", meaning that one user's adoption of a specific technology increases the value of that technology for those who already have adopted it, so the value of the technology as such increases as more users adopting it (Arthur 1994; Shapiro and Varian 1999). When this occur *"standards are important because the value of a specific technology depends on its numbers of users, and the technology will have its highest value when all potentials users adopt a version of this technology following the same standards (Hanseth, 2004)* Parallels can be drawn with TCP/IP protocols for Internet.

When network externalities and increasing returns seem to contribute to knowledge, the standardization of knowledge becomes valuable. The number of externalities will set the primary value of a standard, but a standard cannot be used for anything within it self, because it only enables communication with the adopters of the same standard. So the core value of a standard is the sheer number of current users who have adopted the standard.

3.2.3 Infrastructure

Taking all these concepts described about knowledge as a network one can understand how it can be taken to the next level, namely knowledge as infrastructure. The knowledge infrastructure is a standardized network due to certain differentials. One aspect of knowledge as infrastructure is that it contains various numbers of shared resources that are used to support the whole organization or different activities in a community.

In a more abstract view the concept of an infrastructure versus a network, the first is more solid, thought through, wealthier and planed with a purpose compared to a common network. The infrastructure is something that acts like a framework for an organizations knowledge environment and its way to carry out daily routines. The common employee is constantly working accordingly to this infrastructure most often without knowing it, this is because an infrastructure works as a hidden layer, to huge and unyielding to grasp. Knowledge does not have the same physical touch but it has some common features with infrastructure. *(Hanseth, 2004)*

Chapter Summary

Knowledge in general and knowledge management in particular is getting a more important role in the modern organizations. Today it acts as a strategic tool in the competition on the business market. Knowledge can be divided into tacit and explicit knowledge. Tacit knowledge is a form of knowledge that is stored in our brains and bodies while explicit is possible to save and document so others can take part of it. This sharing of knowledge is a major part of the knowledge management today, to share the knowledge in a big organization a well-developed infrastructure is needed. This infrastructure is constituted by all documents, applications and other sources that help spreading the knowledge. Ole Hanseth adds another perspective and claims that the knowledge itself also can be seen as an infrastructure. It has the same stable and systemic characteristics and it is a shared resource that is used to support the different activities in an organization.

4 Results

This chapter starts with a closer description of AZ Glossary and its organization and structure. In the second half of the chapter we will present a summary of the interviews from our research and give the users' points of view.

4.1 AZ Glossary

AZ Glossary is a web-based source that contains information resources describing conceptual resources that are used in the different departments in AstraZeneca. It has been built up as a webpage, accessible from the organization's intranet and with the intention to be easy to use for the employees. The glossary's initiative is that it should function as a single global reference resource for terms and acronyms that are within interest throughout the entire organization.

The main purpose with the glossary is to create a standardized terminology to help employees to orient themselves through the many terms in the organization. This standard terminology also helps preventing misunderstandings and disagreements that might occur of the meanings of the terms. The glossary also aims to improve the understanding between the different sites and functions in the organization in order to make information sharing more efficient, and help new employees getting familiarised with the company's terminology.

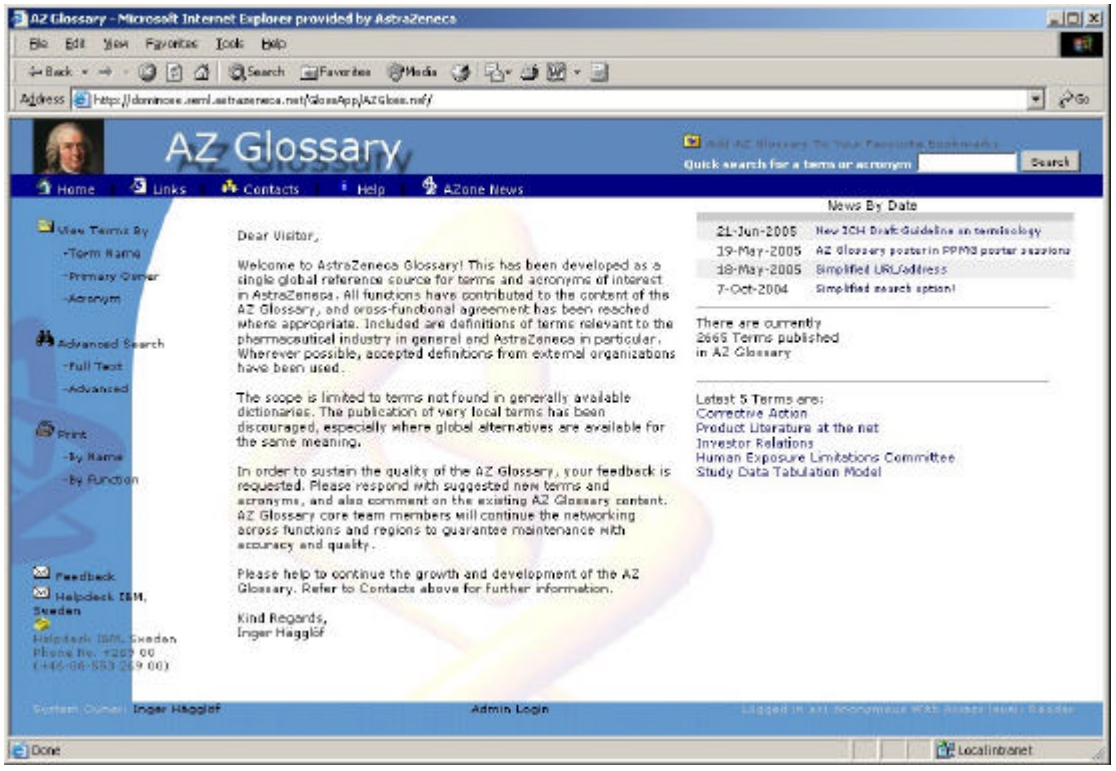


Figure 4: The welcoming page of the AZ Glossary application.

4.1.1 Interface

This is a print screen picture from the AZ Glossary representing the term “Area under the curve”.

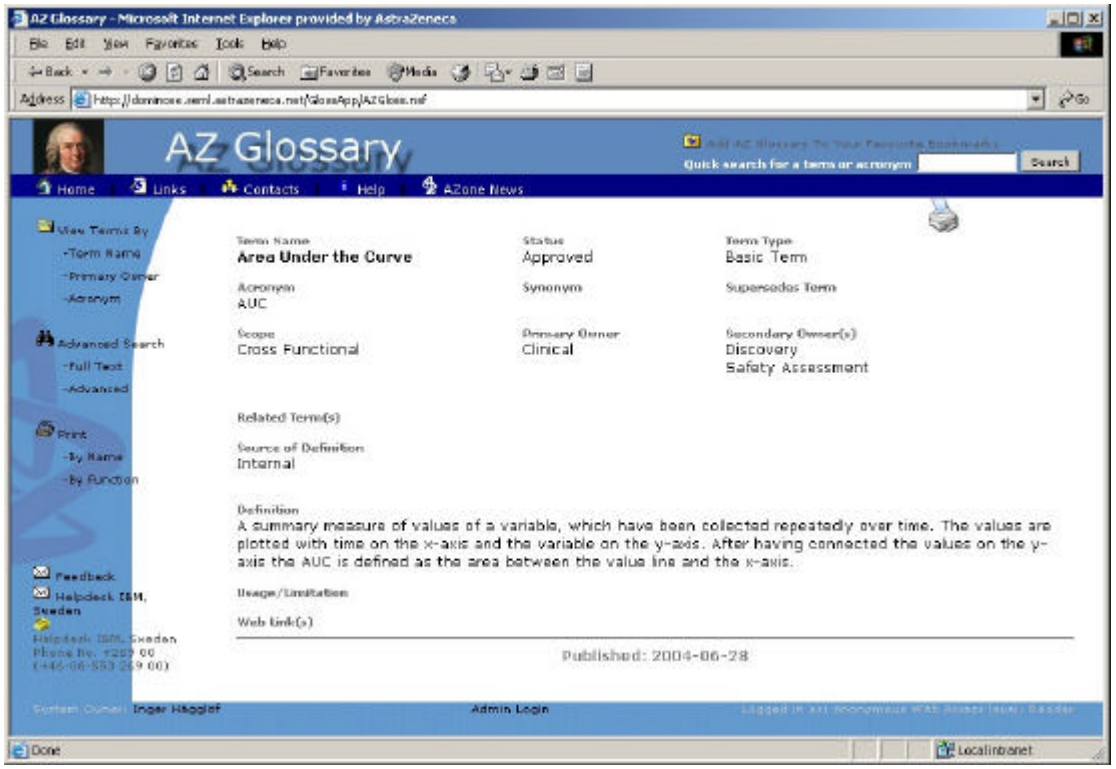


Figure 5: The interface of AZ Glossary on term level.

The first heading “*Term name”* obviously tells the full name of the term and “*Acronym”* gives the initials of the term. The “*Status”* field can be either *approved* or *obsolete*, if the term has the status *obsolete*, another term name under the heading “*Supersedes term”* would be displayed, which informs about the name of the new updated term.

The header “*Scope*” informs if the term is *“Functional”* meaning that it only affects one function or *”Cross functional”* if it affects several functions. *“Primary owner”* and “*Secondary owner”* are two important headings showing which community that are responsible for the term and in what business content it is used. *“Related term(s)”* shows a term or several terms that are used in the same area of work or processes. Together they often show some more content and in that way it makes it easier to see the whole picture.

The *“Source of Definition”* states if the term is either *internal* or *external*. If *internal* it is defined in AstraZeneca or if it is *external* taken from an international standard. The *“Definition”* shows the complete definition of the term that has been decided amongst the communities or functions. Below the heading *”Usage/Limitation”* there can be a recommendation for how to use and how not to use a term. When additional information or value of a term is needed, it is linked under the heading *“Web link”.*

4.1.2 Communities

The maintenance, updates and development of the glossary are performed by a network, consisting of people that represent different functions of AstraZeneca. The organization has eight different communities and every community contains one or more functions, in total there are over thirty different functions together in the eight communities.

For every community or function there is a team representing it, this structure can be different due to some communities might only have one function and thereby only one team. Other communities might have ten functions with a team for each function. Also included in every community or function there is a term manager whose job is to gather and publish new terms.

Every community has a community representative, and these eight representatives together form the core team. In the core team the main owner of AZ Glossary also is represented and that person also handles and manages requests and updates of terms.

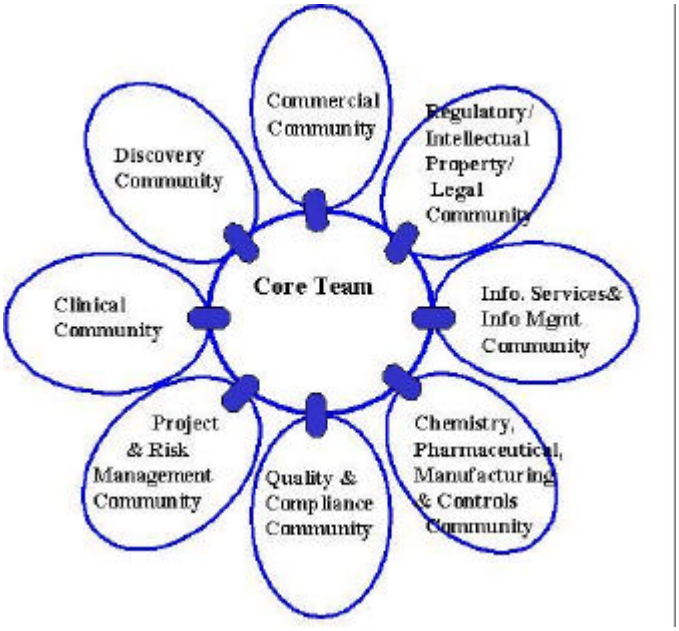


Figure 6: The picture shows how the eight communities are connected to the core team. Under each community a number of different functions are present (not shown in picture).

The purpose with this structure is that the whole organization is represented and every business area reflected.

4.1.3 Term Publishing Process

The main process is the publishing of new terms, meaning from the proposal of the term to the publishing of the term as approved. New terms can be proposed in various ways, for example by sending an email from the application it self, or by phone or through contact with the term managers. When a proposal of a new term is sent, the owner of the glossary determines which community the term belongs to and then passes on the proposal to that community's term manager.

The community team, or if several functions with many teams, then gather around and discuss the proposed term and what the definition should be, for this scheduled meetings are appointed ơr through mail or net meetings. When a decision of the term's definition is reached, the term then get the status ready for approval. When the term is in this stage the other communities have the opportunity to reveal it and check if the specific term involves their function or community as well. However, this phase only occurs if the term is cross functional meaning that it affects other communities or functions, if not cross-functional the term can be approved directly. If the term is cross functional, involved parties should decide who is to be the primary owner. If no other community or function has any objection within the thirty days time the term gets the status functional approved and the term manager from that community or function can publish the term.

Occasionally issues around a new term or terms occur between communities and if no agreement is reached, the case are then handled over to the core team whose decision is final.

The same procedure is carried out if term managers have picked up new terms in his or her function or community, the term is then taken up directly in the community meeting, it does not have to be through the owner. This process can be different from community to community in some small details, but in general this is set to be the standard.

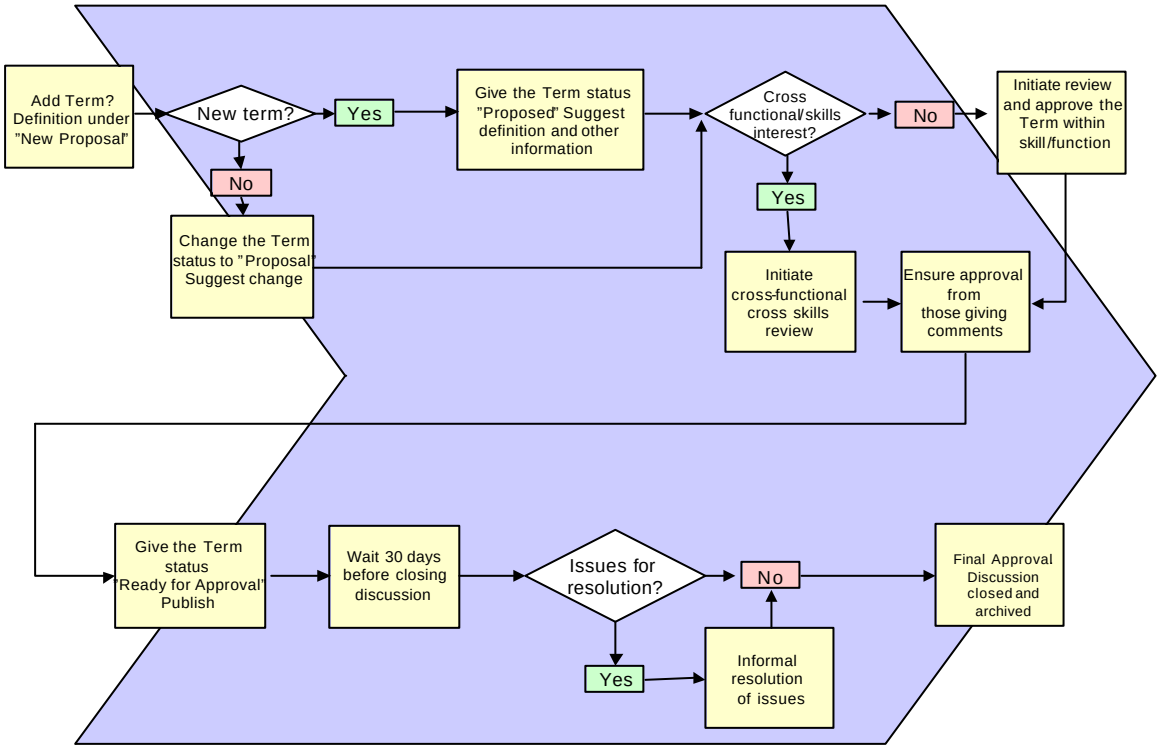


Figure 7: The picture shows the term publishing process.

4.1.4 Ownership

The structure is as follows; there are two sets of ownerships, the first one is the primary owner. The community or function that developed the term and its definition will be the primary owner. This is because to mark which area of function the term belongs to. The term's primary owner together with the term name makes the concept unique. This is to avoid misunderstandings, because there are terms present with the same name but different definitions and owned by different communities and thereby used in different areas. The primary owner's community/function team is responsible for the maintenance of their terms. The other set of ownership is the secondary owner of a term. This type of ownership occurs when a different community or function than the primary owner feels that the term also affects them. Still the primary owner is responsible for the maintenance of the term but when changes are about to be made, they have to confer with the secondary owner first.

Ownership of the terms is needed to sustain the validity of the glossary.

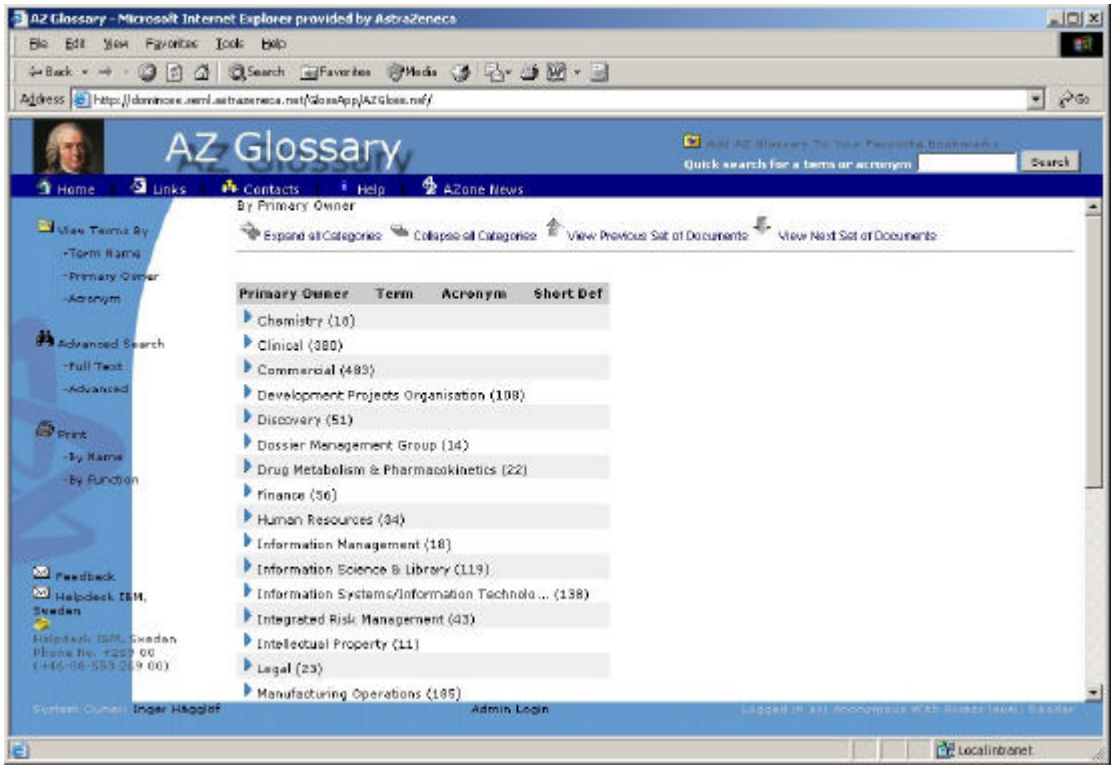


Figure 8: The picture shows a list of all functions and amount of terms they are primary owner of.

4.1.5 Technical structure

The technical structure of the organization that surrounds the AZ glossary is of the more advanced kind. Every community has their own database containing all the terms that they are primary owner of. So the clinical community has one database where all the terms that belong to them are stored and commercial has their database and so forth.

However, the only ones that have access to these databases are the specific community team, so Clinical cannot enter Commercial's database and the other way around. When a term gets the Approval status the term manager will add it in the database that belongs to his or her community. All these databases are then linked to a public database, called the Hub. All the community databases connected to it are called Satellites. From the public database the end-users in the AstraZeneca organization are able to reach all the terms from all communities through a common web-interface connected to the Hub. The picture below demonstrates how it works.

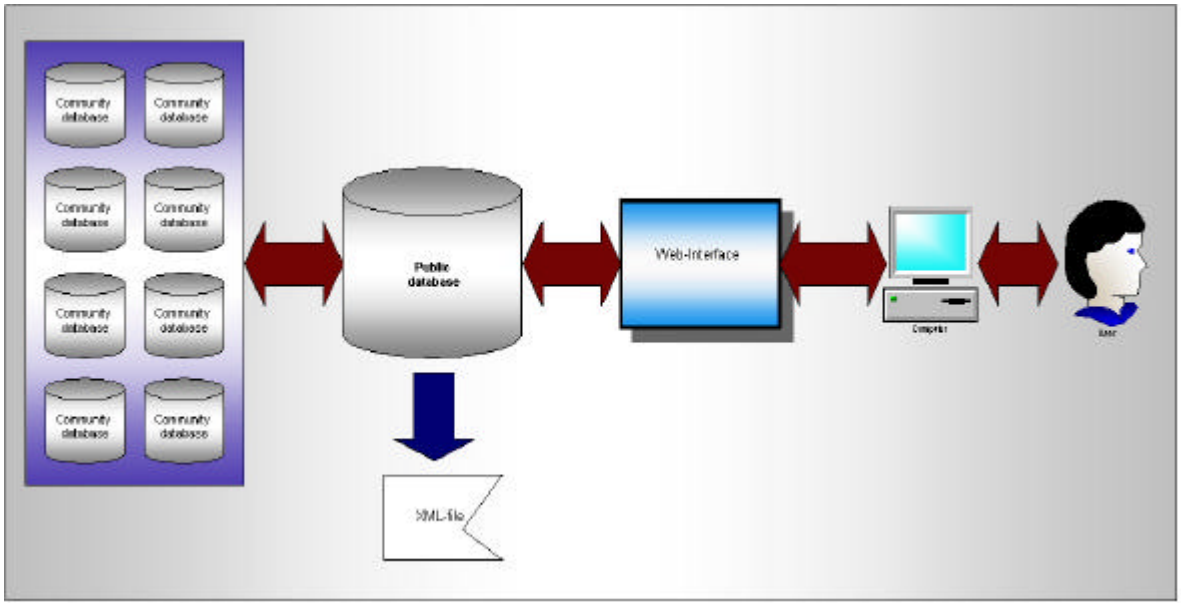


Figure 9: The technical architecture behind the AZ Glossary.

The whole technical structure around the databases is built on a Lotus Notes platform, but due to organizational plans on exchanging everything based on Lotus nodes in the entire company, the platform will be changed within a few years time.

Every night a file with all the terms from all the different communities is created. This file is an extensible mark-up language file, meaning that all the terms in the file are built on an XML syntax. The picture shows the structure how the terms are programmed. Before the implementation of a new web portal on the intranet the file was used to provide other services with the definitions of the terms. But this is not possible anymore and at the moment the XML file is not used for anything, which is a big limitation.

4.2 Summary Interviews

4.2.1 Usage

The usage level of the glossary is spread through the AstraZeneca organization, though some departments use it more than others. Most employees tend to use the AZ glossary as a local dictionary to support their various tasks. One aspect in the line of use is when employees encounter terms or acronyms that they do not recognise, they turn to the glossary for looking up the word.

> *"When I encounter acronyms that I don't know about, I'll enter the glossary and check if they are present, if not, I'll put them on an input list." – Administrator*

One obstacle in the AstraZeneca working environment is that it does not matter in which department one is located, one will still be overwhelmed by the large amount of different, hard-to understand, short lived and mostly by the high numbers of new terms. For the common employee who has been trained in this environment for several years there can still be problems recognizing the terms that are being used. For the newly employed, it is even harder trying to deal with all these terms, the glossary will make itself a truly useful tool.

> *"A more important target group is the newly employed who're facing a labyrinth of acronyms, which makes it hard when coming from the outside." - Administrator*

The glossary does not function only as a dictionary for looking up acronyms and terms that already exist in published information. It is also useful for supporting other every day tasks performed in the company, for instance people that are actually publishing information through various channels.

> *"I write news and so on, and if I encounter some unusual or unfamiliar acronyms, I usually extract them from the glossary and add the explanation in the text."*
> *– Editor of content in the portal*

The glossary should act as the company's single source for its operational terminologies; meaning when an employee is uncertain, he or she should use it, for example in reports, modelling etc.

Working in projects, when focusing on concepts in early stages, the glossary is used to look up the core concepts to eliminate errors from the beginning. More over, in regularly enterprise analysis, or any analyse the glossary is useful for interviewing analysts that are able to go back and look up terms or acronyms that have been said during the interviews.

> "*Often there are packed with different expressions, so in order for me to know I'm using the right version, I check with the glossary. Don't think AstraZeneca would manage without the glossary today." – Editor of content in the portal*

A change for the Clinical department in AstraZeneca has been with the SOP's (Standard Operating Procedures). These documents describe the guidelines how to carry out one's work, and there are a certain numbers of them that are compulsory containing methods, processes and procedures etc. Earlier the terms and acronyms present in these documents were defined in the document itself. The consequence of this was that every single document owner was responsible for each term and its definition. Then different owners could have the same term names in their documents with different definitions, and lacking procedures for updating. The use of the glossary eliminates this problem, by removing the explanations in the SOP's documents to insert them into the glossary, where each term only carries one definition (in every community). This is becoming more standard now and the way of using document specific glossaries are replaced by the AZ glossary.

4.2.2 Awareness

A major challenge for the AZ Glossary is to reach out to the whole organization. Even though the glossary is said to be an official source of knowledge not everybody is aware of its existence.

> *"This glossary… where do I find it?" – Potential end user*

Another user wonders what meaning one of the attribute of a term has, a third user did not know that it was possible to send proposals on new terms. Due to the lack of statistic nobody knows how widely the glossary is spread or how many people that are using it, but it is a fact that the awareness about the application and its functions is too low. The core team says that they have tried to make it a part of the material given out to newly employed, but the HR-department says that there is too much important information already. But this is an important group to reach;

> *"Newly employed get swarmed of new terms and acronyms used instead of the regular words in this company. I think it is important for their confidence to know that they can find it straight away in the glossary" – Administrator*

The core team has also tried to get a link with the logotype on the portal, but without positive result. However, there are a lot of people that do use it, and a lot of them are satisfied. One of the administrators says that they sometimes get mails from happy users saying that it is a very good service and they have not seen it before in other companies.

The AZ Glossary is an official source of knowledge in the company, the use of it is recommended by people high in the organization. But a few of the administrators think that it would be good if directions would come from even higher level in the organization and these directions should be directed to everybody in the organization.

"When reached out it is a very valuable asset, it describes the terminology, what it means and how it should be used". – Administrator

According to one of the administrators the glossary has contributed to homogeneity and it works cross functional, it was of importance especially after the merge. Another administrator says:

"*It has contributed to a more harmonized terminology. In long-term if we put together all our documents with the terms we shall use, then the receiver of the information, both internal and external, knows that the meaning of the terms are always the same, we can trust AstraZeneca"*

4.2.3 Organization and Structure

The members involved in the organization with communities and functions surrounding AZ glossary seem to be quite satisfied with its structure. Though several members in the core team emphasize that it is dependent on the size of the group and the commitment of the people involved. If the core team would be bigger it would be difficult to coordinate meetings. And since nobody of the administrators of the glossary carries it as a main task it never gets top priority but always comes in second hand. It is hard to find enough time and resources to run it in the way wanted.

"The glossary is run by true enthusiasts, if three or four of them quit or must leave the team, things could easily slide away and the glossary might be used for things that it was not intended for" – Administrator

Each community is driven separately; the work and result depend a lot on the individuals running it and there are also cultural differences. The fact that the organization looks the way it does of course affects the processes. The main process is the publishing process, due to the high amount of waiting time it takes a lot of time to get a term published, which means that the glossary is never fully updated. One of the users complained that it took too long time for a new term to show up. According to one of the administrators the time between the meetings in the core team can be really long, and that time is not used in a good way.

"Sometimes you get the feeling that people see the meeting in their calendar the same day and first then starting with the work that should have been done before the meeting" – Administrator

It is not an effect of lack of knowledge but instead lack of time, time to set up meetings, time for consideration, time to take decisions and so on. Another process in the core team is when a term is ready for approval. Then the community leader of the proposed term sends out an email to the whole core team to inform about the new term. Each community leader then has 30 days to make objections or comments. This is not a very efficient process either, it could be done better, says one administrators.

The pioneers that have dedicated them selves for the glossary at their own initiative at their own time also are pressed back by higher management or other working tasks, that makes it hard to keep alive. This increases the risk of value reduction, but on the other hand, the glossary has been such an established tool that the pressure of maintaining it and to uphold the quality is still present.

4.2.4 Development

Most of the people in the organization seem to be quite satisfied with today's version of AZ Glossary, it performs its basic task, but of course there is always room for improvements. The problem with making improvements in the system today is that AZ Glossary is no longer seen as a project. It has reached the administration phase, which means that it has no budget for any system development. However, there are some small changes planned, on the portal there is a site called "Drug Project Operating Model" (DPOM), it works as guideline for how to run drug development projects in AstraZeneca. Before the portal was implemented the DPOM-site worked as normal HTML-page and then it was connected to AZ Glossary and all the terms on the site that occur in the glossary was given an explanation. The explanation was showed in a small pop-up box that occurred when hovering the word with the mouse. That service is wanted back from both the users and the authors of the DPOM-site.

> *"As good as every one of the users wishes this service back".*
> *– Editor of content in the portal*

According to a member of the IS-department that runs the portal this service is coming back, it is on the ToDo-list but because of the complexity and the amount of required time it has been given low priority.

Another change that is coming up is that the company will exchange all Lotus Notes databases for something else. The glossary is built up in such an environment but nobody knows for sure how it will be effected. One of the administrators of the glossary hopes that this opportunity can be used to make some small changes and improvements.

One of the community leaders is planning another update of the glossary, not a technical one but more of the maintenance kind. The terms in that community will be reviewed and updated because some of the old definitions are of poor quality and out of date. From the maintenance perspective there has never been any updating or reviewing of the content at all in that community.

> *"Now we slow down the introduction of new terms and starting to clear out, coordinate and reach for consistency. Review and increase the quality of the content" – Community leader*

A problem with the updating work is that the system does not provide any functions for this at all. All the work must be done manually.

4.2.5 Requests

> *"We want it to be used more efficient, not just the application but also the content must reach both the individual and the whole organization. It must be easy to use the content in different areas." – Administrator*

The quote above is a wish from one of the administrators of the glossary; it is a topic that was coming up in most of the interviews. More distinctively, there is a need for the possibility to make links to the glossary on term level. This means that a term in a text in a document or on a website is marked as a hyperlink and when clicked on you get linked to the glossary and the term. An even better solution would be to get a pop-up box with the explanation on the screen without starting a new session in the browser. Today most of the applications that link to the glossary direct you to the welcoming page, and then you have to perform your own search for the term. There is actually a shortcut today that makes it possible to link directly to a term in the glossary, but is a lengthy procedure and not a lot of people are aware of it.

> *"It must be easy to set links between terms on the Web and AZ Glossary, it shouldn't be necessary for me to work in HTML and put down a lot of effort just for making a small connection. I would like to be able to just click on a term that I've written down and say that I want the definition of this. It should be easy for the editor of the text as well, not just for the reader.*
> *– Editor of content in the portal*

This could be further developed by a connection between the editorial environment in the portal and the AZ Glossary; the authors could then easily insert a definition of a term in their text by just pushing a button. Another suggestion regarding the connection between the portal and AZ Glossary came from one of the workers of the portal; why not integrate the whole portal's search function with the glossary. If the searched term is represented in AZ Glossary its definition would appear on the screen above the search result.

As good as all the administrators said in the interviews that they miss statistic numbers and figures about the usage of the glossary. There used to be a counter on the website that registered all unique visitors, but when an upgrade of a server was done it stopped working. 2003 there were around 25 000 visitors and the year after that 33 000 until it went down, but around 150 a day and after that no numbers are known. One member of the core team specifies the need and says that it would be nice with figures on unique visitors, total number of searches and number of searches on each term.

Another request that was coming back several times was a closer description on the owner of the terms; sometimes it can be important to know more than the community. That a term is owned by, for example, clinical does not say a lot if you want to know who to ask for more information, a contact person would be good, as said by an end user. There is information about which function that is responsible for each term in the community's groupware but it is not shown on the screen, and the function level is still not accurate enough. This information is also requested by one of the administrators that needs it in an updating purpose. The same administrator thinks that this can also lead to a closer interaction with the users. It could be

complemented with a feedback function on the term page so the users could send queries about and feedback on the term.

Chapter Summary

AZ Glossary is a Web-based glossary placed in the AstraZeneca's intranet. Everybody in the whole company has access to the glossary. The glossary is only containing terms that are of interest for the AZ organization, terms that can be found in other common dictionaries are left out. The organization surrounding the glossary consists of different communities and functions representing different departments in the company. These communities and functions are made up by people who are working with or towards the glossary's maintenance and development.

In the summary of the interviews, different points of views from administrator to end-user level of the employees are represented. The Usage of the glossary points out how different employees are using the glossary when encountering or publishing terms they are unfamiliar with. The Awareness results show the employees insight about the glossary's functions and its existence, and also different attempts to make it more visible such as search functions in the intranet. The Organization and Structure results explain the various opinions and the level of efficiency around the communities and the functions. Further the Development part presents improvements of the glossary, such as hovering functions etc. In the Request part all suggestions and wishes that came up in the interviews are listed.

5 Discussion

This is the analyze part of the thesis. The results are applied to the theories and discussed from the view of the described problems in the introduction chapter. We also discuss AZ Glossary and give recommendations on how to improve the organization and the application and the way it is used.

5.1 Usage

The application AZ Glossary supports the people in the organization in their different tasks. It also serves as a repository of knowledge where one can get help when coming across unfamiliar terms and acronyms. AZ Glossary is an official source of knowledge in the AstraZeneca organization. It is a collection of knowledge from various parts of the organization. The terms with their definitions and attributes come from different communities and they are well reviewed, by experts in respective area, before published. Knowledge appears in different forms and can be seen from different points of views. Does AZ Glossary fit to the description of any of those and can it correctly be referred to as knowledge? Ole Hanseth discusses different aspects of knowledge in his article "Knowledge as Infrastructure". One aspect is that knowledge can be seen as different elements that are stored in our brains or computers, in this perspective AZ Glossary can be seen as knowledge, it consists of different elements like terms and definitions that are stored in a computer database. Another perspective in Hanseth's article is knowledge as a cognitive material described as either explicit or tacit. A more phenomenological perspective is to see knowledge as deeply embodied and embedded into our bodies and our perspectives. Explicit knowledge is knowledge that can be documented, expressed and taught to others, AZ Glossary is a database with documented knowledge, thus explicit knowledge. The AZ Glossary application helps spreading the knowledge throughout the company. The tacit type of knowledge is harder to transform to others, it goes hand in hand with phenomenological view that the knowledge is rooted in our minds and bodies. The last perspective according to Hanseth is that knowledge can also be seen as embedded into institutions or material structures like a house or an Information System. Applied to this theory the AstraZeneca organization can be seen as the institution with the knowledge and it is expressed in the Information System AZ Glossary.

By looking at these different views and theories and comparing them to AZ Glossary we can draw the conclusion that the glossary can be seen as knowledge, or more exactly, the content of the glossary, in form of the terms together with their definitions, is the knowledge and the application is a tool for expressing it.

5.2 Organization and Structure

Further on Hanseth addresses the systemic aspects of knowledge by looking at it as a network; he uses an example about the work and collaboration in modern hospitals. Different doctors have different specialities and they cooperate with each other in different ways, ask each other for advice and send patients from one hospital to another and so on. Hanseth claims that it is crucial with a standardized interface between the practices for the sharing of knowledge. The same situation is current in the AstraZeneca organization. The company is built up by different departments and sub departments that can be seen as a network. The different departments communicate with each other, documents are sent cross the organization and people get together on meetings and phone conferences and so on. To communicate

efficient and painless it is of high importance with a standardized, not only language but also, terminology. For this purpose the company has developed a terminology and created the glossary with explanations for these terms. Now the glossary works as an interface between the different parts of the company to spread the terminology to everybody in the network.

Considering the knowledge in the glossary as a network, it should also, according to Hanseth, imply the theories of network economics such as network externalities and increasing returns. The concept network externalities means that a transaction may have effects for a network of actors not involved in the transaction itself. Hanseth mentions an example about a doctor that adopts new knowledge about a new procedure, the other ones already following this procedure, or compatible ones, will find it easier to collaborate and communicate with that doctor. Consequently the value of this information has increased. The same implies for the terms in AZ Glossary; when somebody looks up a term in the database and adopt that knowledge, others already knowing the definition will find it easier to communicate with him or her. It increases the value of the terminology. This theory also works the other way around, when somebody adds a new term to the database it does not affect the other users directly but it increases the value of the glossary and the terminology gets richer. The users can find one more definition in their glossary. When somebody conducts a search for the new term and finds the definition, the value of that knowledge has increased.

The value of the application AZ Glossary also increases the more users adopting it, which matches the theory of increasing returns. If nobody is using the program it does not matter how many terms and definitions it contains, it is still worthless. But if used by everybody in the entire organization it gives a lot of value to both the users and the company. In consequence the glossary's value is depending on both the number of users that have adopted it as a standard knowledge base and the amount of terms. A high quantity of one is not enough. A glossary with 5000 terms is not giving any value if only two people use it, and a glossary with 5000 users and 5 terms is neither considered a valuable asset.

Hanseth continues his discussion about knowledge and goes from seeing it as a network to consider it as infrastructure. An infrastructure is a standardized network as described before with additional features, for instance infrastructures are shared resources for supporting a wider range of activities for a community. In the interviews it was discovered that the use of AZ Glossary supported a wide range of activities for the company's different departments. Except that it works as the main glossary for looking up unfamiliar terms, it is used as a tool for writers of documents, articles and guidelines to explain terms in their texts and it is also used as a source of knowledge to find definitions to other programs such as BIM (Business Information Modelling) and the DPOM (Drug Project Operating Model). Hanseth's statement, that knowledge because of its systemic character can be seen as infrastructure, is also suitable to AZ Glossary. It has all the characteristics for knowledge and it is a shared resource in the organization for spreading knowledge. The organization and structure is stable and not easy to change. An example of this is the publishing process that is strict and only lets qualified terms and expert definitions through. And once in the terms are not easy to edit or delete, then it has to go through the same process once again. In fact, it is not even possible to delete a term; instead it gets the status obsolete and points to another.

The conclusion of this is that AZ Glossary contains knowledge that is important for the organization and it is spread out and used in a way that makes it comparable to an infrastructure and also a part of the existing knowledge infrastructure in the AstraZeneca organization.

5.3 Recommendations

5.3.1 Administrator View

Today's current term publishing process is supposed to function in the same way in all the different communities, but due to several reasons the level of involvement varies. First off people's personal individual involvement and enthusiasm is a factor that affects the flow in the process of publishing terms. The personal view of the glossary's importance and its relationship towards the person working with it might be of relevance considering the level of participation. Second, this is crucial for the term process because the people working with it do it at their own time, meaning most of them have to squeeze it in under the time they got left during their regular working hours. Most often this does not leave that amount of time left as it requires. This affects the term process in the way that it takes too long between meetings and for setting them up and get people prepared, leading to long periods until new terms are getting proposed or published. This is not the case in every community or function but when affecting one the whole glossary is affected in a less satisfactory way. The fact that time is essential, and most people are in lack of it, is the main reason why the process not always functions to everybody's expectations.

Further on for the people working with the glossary there is a need to make the appointed working task clearer and pass out more precise areas of responsibility in order to make the organization around the glossary more rigid and stable. Of course this is not easy to do and cannot be demanded of the people involved in the work with the glossary, as their time spent with AZ Glossary is unsalaried and not bonus based in anyway.

However, at the present time, in the way that the glossary's organization is living "its own life" might in the long run reduce its level of quality and legitimacy. Therefore the most appropriate solution in the aim to give the glossary more value, influence and impact as being a knowledge resource, is to not treat it as an alternative tool but to make it recognized as a core piece of the organization itself. And to accomplish this, people must be assigned to have the glossary as a part or their main task in the line of work to keep it modern and always up to date, and some guidelines and directions from upper management.

5.3.2 End-User View

The end-users' demands on the glossary is without any exception that it should be easy accessible and easy to use. Some other remarks of improvements that have been made are the possibilities to get in touch with the local skill owner of the terms or even further down in the organization, like the sub-department. At current state the owner of a term is only showed as the community or function, however it is possible to get in touch with the skill owner of the term. One have to go through the owner of the glossary or the community representative and that possibility is relative unknown. The possibility to get in touch with the real owner of the term seems to be of importance to users when the term or terms are to be discussed further in some content for the end-user. Some type of solution or decision might be needed to

consider for making this possibility more visible to the end-user, or some kind of technical solution.

Even though the terms definitions are explained in the dictionary, problems for end-users can occur when some of the definitions are vague or difficult to understand. Points have been made about the possibility of feedback or support in this matter. It does not stop at feedback about the definition of the terms; also help in general and information about the area of usage of the terms. So besides unclear definitions, feedback about in what kind of content the terms can be used or not to be used etc. One solution can be a visible mail link regarding these questions or a real time chat function. Again this type of service demands having a person behind it working full time, which fails on the lack of resources. But meeting it in half way, a rule of some kind of maximum response time when a request is sent in, might in the current state be possible, whatever is one, five or ten hours. The advantage is to make the end-user feel safe through that he or she knows that within some time there will be an answer to the question and that the possibility is there.

Another improvement from the end-user point of view might be some kind of independent representative for the system. This person's task will be to gather the end-users' requests and views or complaints and later discus them with the core team or other persons involved in the glossary. This person will work as a bridge between the core team and end-users.

5.3.3 Organizational view

From the organization's view one of the most important aspects is to increase the knowledge of the glossary's existence. The glossary has to obtain a clearer recognition as the organization's official glossary. As mentioned before this is not easy to achieve, but again some kind of general statement or directions from upper management that tells everybody that AZ glossary is AstraZeneca's official glossary. Also important is to introduce the glossary to the newly employed at an early stage to make them work with it directly.

Together with the statement a reintroduction of the glossary could be set to action. Some new functions could be implemented and perhaps a new interface just to give the impression that it is new and modern. These changes in the interface do not have to be any radical changes but still something fresh. If AstraZeneca performs that change it is of the utmost importance to keep the maintenance of the glossary up to date in order to keep the quality on top.

The quality of the explanations in the glossary varies among the different functions. A lot of terms are old an out of date and some terms only contain the meaning of the acronym. Not all the definitions in the glossary are good enough to use in external reports and presentations. To keep control over the terms and to make it easier for the end-users to know which terms that can be used for external purposes and extra attribute should be added in the glossary. This attribute can be called Classification and tells the user the how the term should be used.

There are other glossaries containing definitions in the organization and one possibility can be to implement a search function in AZ glossary that searches in other databases. The user could just choose in which databases the search will be preformed. This is for making it easier for the employees to orient them selves through the intranet and to deliberate steer them to the AZ glossary in order to make

it the only or at least the obvious and most natural choice. Another option is to introduce a common interface for all glossaries and other knowledge repositories in the organization; the user could then choose to search in all of them or in a specific one. The advantages is that the users only need to remember one URL and one source of knowledge and the terms can still be separated in different databases. Even better, this could be integrated in the portal's search function, that would make it really easy for the users.

5.3.4 Technical view

A suggestion of technical improvements of the glossary at current state is a function that makes the maintenance part easier. A solution of this could be a maintenance notifier attached to terms that are known to be unstable and change over time. The maintenance notifier should be a function that is pre-programmed to the term when it is introduced into the glossary. When the pre-programmed time that had been set in the introduction of the term, has run out, a notification will be sent to the function or community that is primary owner of the term. Then they can check if the term is still up to date and if yes then set a new date for next maintenance check or if no upgrade it. Yet again not all the terms in the glossary would demand this function but only the terms that tend to change over time.

An old function that worked in the past is the counter that kept record of unique visitors. This should be a fairly easy function to install again. The possibility of overlooking the usage of the glossary would always be useful for the administrators. Improvements of this function would be to also count the total amount of performed searches and the number of searches for every term.

The related terms of a term should be linked and clickable. In today's version the user can only read which terms that are related and then new searches on those terms must be performed. This takes a lot of time and effort, which leads to that the users do not read about the related terms.

Regarding the identification schema for the AZ Glossary's URI (see 6.6), we recommend the use of a persistent URL to make it more stable if any updates are done. For example when exchanging the Lotus Notes database it might have effect on the URI but with a PURL the users will not be affected. Further on we think the best solution is to have AZ Glossary as a source that uses the AstraZeneca common name schema. The resources are preferably referred to as concepts, due to the fact that the terms in the glossary are describing the concept. The definition of a concept is "a meaning of a term as agreed upon by a group of responsible persons", and that is what the glossary contains. The last part in the URI that identifies the term level is most appropriate using the terms' numerical id. A lot of the terms in the glossary are acronyms with several characters, if writing the full names the URI they get very complex, long and hard to handle and the acronyms cannot be used because it is not unique. It could be easier for the users to remember a specific term's URI if it was built up by the term name, but we think that the use of this way to enter the glossary is very limited. It should be as quick to enter the glossary as normal and search for the requested term. Considering these thoughts the full address for the term Area Under The Curve should be formulated as follows:
http://purl.astrazeneca.net/azglossary/concept/clinical/MGOR-5BQC32

The technical basis will be further explained in Chapter 6, Techniques.

5.3.5 Administrator View recommendations overview

- The level of involvement varies among the employees.
- Appoint clearer working tasks and pass out more precise areas of responsibility.
- Time invested in AZ glossary is unsalaried and not bonus based in anyway.
- Keep the glossary's up to date to sustain its level of quality and legitimacy.
- Aim to make it recognized as a core source of the organization itself.

5.3.6 End-User recommendations overview

- Easy accessible and easy to use.
- Possibility to get in touch with the local skill owner of the terms.
- Enable general feedback or support of questions around the terms.
- Requests, views or complaints presented from an independent representative of the end users.

5.3.7 Organizational recommendations overview

- Increase the knowledge of the glossary's existence.
- General statement of making the glossary a standard in the organization.
- Classification of terms to tell the user appropriate area of use.
- Multiple search function among the different databases including AZ glossary.

5.3.8 Technical recommendations overview

- A maintenance notifier, to keep track on terms that tends to change over time.
- Implement a counter on the page that registers visitors, searches etc…

6 Techniques

The chapter starts with presenting the different languages that form the base for the new Semantic Web technology. The second part of this chapter explains the concept ontology that is an important part of the Semantic Web thinking. Other form of knowledge repositories and their connection to each other are also explained.

The thesis so far has been focused on the work with and use of AZ Glossary. The use of the glossary and its organization have been mapped and different needs found out. One way of satisfying at least parts of these needs and to improve the glossary's function as a knowledge infrastructure is with the use of semantic technologies. The base in the stack with the semantic languages is XML.

6.1 XML

A common problem in the world of computer systems has been that applications often speak their own language and transfer data that other applications, systems and platforms do not understand. Extensible Mark-up Language, or XML, has evolved to solve this problem. It is a product of a search for a universe standardized file format, which is completely independent from any hardware or software, written language and even independent from itself. XML is used to structure, store and send information and it also acts as a framework for creating a mark-up language or a so-called Meta language, data about data. A mark-up language is a process to identify structures in a document. XML as a mark-up language was designed to describe data and to focus on what data is, unlike HTML, which was designed to display data and focus on how data looks. (W3C, 2006) While for example HTML uses predefined mark-ups, like the letter B defines bold text, in XML the content of the document is undefined and you decide what the mark-ups should represent. The XML-document contains both the content and mark-up of the content, which is the core of the structure. The mark-up words reflect the information it surrounds. Here is an example of how an XML document could look.

```
<Term ID="MGOR-5BQC32">
    <TermName>Area Under the Curve</TermName>
    <Status>
            <Approved />
    </Status>
    <PublicationDate>2004-06-28</PublicationDate>
    <TermType> <BasicType /> </TermType>
    <Acronym>AUC</Acronym>
    <Scope> <CrossFunctional /> </Scope>
    <PrimaryOwner>Clinical</PrimaryOwner>
    <SecondaryOwner>Discovery</SecondaryOwner>
    <SecondaryOwner>Safety Assessment</SecondaryOwner>
    <SourceDefinitionSelection>Internal</SourceDefinitionSelection>
    <Definition>
            A summary measure of values of a variable, which have
            been collected repeatedly over time. The values are
            plotted with time on the x-axis and the variable on the
            y-axis. After having connected the values on the y-axis
            the AUC is defined as the area between the value line
            and the x-axis.
    </Definition>
</Term>
```

The mark-up words are those who are embedded in tags like <TermName>, which defines the start and </TermName>, which defines the end. These mark-up words reflect to the content between them. The mark-up word together with its content form an element, so the line:
<TermName>Area Under the Curve</TermName>
…tells us that it is an element that has the mark-up word "TermName" and contains the content Area Under the Curve. It is also possible to put a characteristic on an element as an attribute, this is used to give the element additional information that is not a part of the data within the element. The element <Term> in this case has an attribute which name is "ID" and has the value of "MGOR-5BQC32". This is often used for the possibility to unique mark an element, and the attribute itself is irrelevant to the data, but might be important for an application that tends to operate the element.

XML makes it easier to for applications to find and process data, it also makes processes more automatized in a larger extent then before. The more processes that are dealt with the greater the need is for a standardized format.

6.1.1 XML Schema

An XML schema is a description of an XML-document. It describes the structure and constrains the contents of the XML-document it is related to. A schema contains a set of rules to which the XML-document must conform to be valid. It can be definitions of the elements and attributes in the XML-document. The process to check if the document is following the standards set in the schema is called validation. That a document is valid is not the same thing as the XML's core concept of syntactic well-formedness. A document does not have to be valid unless it is stated in the XML parser, but all XML-documents must be well-formed.

A schema is built up of two parts, the largest and most complex part structures the relationships while the other part specifies mechanisms for validating the content of simple XML elements by specifying a data type for each element, for example that a certain element must be a two digit number. XML schemas can be expressed by different languages that are developed especially for this purpose; an example of such a language is DTD, Document Type Definition.

6.1.2 DTD

The possibility to decide your own mark-up words in XML is one of the great advantages comparing to HTML. However in a large organization where many employees are using XML-documents to create their own elements to mark-up information, the vast amount of different mark-up words and structures can cause a problem. The strength with XML is that it provides structure to the information. To maintain the structure for specifying the mark-up words in an XML document the structure can be set with a DTD, Document Type Definition. A DTD contains a set of rules that decide how the information in an XML document should be structured. The DTD defines:

- Which elements that are allowed to exist in the document.
- How these elements can occur and how they can be a part of each other.
- Which attributes that are included in an element.

In a small organization or for personal use of XML-documents, a DTD is not really necessary, but in larger organizations it will certainly ease for every person involved to follow the same standard when using different elements and attributes. If the DTD has set a rule that an element should represent information about the brand of the medicine, then no one else can change that in a second document. The DTD helps keeping a consistent structure when creating and using XML-documents throughout an organization.

6.2 Resources

The term resource is widely used in several areas and a lot of definitions exist. In the context of this thesis the extent of the term can be divided into three categories:

- **Information Resources**
- **Resources in the Real World**
- **Conceptual Resources**

Information resources are defined as anything whose essential characteristics can be conveyed in a message. (Halpin, 2006) Information resources can be divided in two classes; network accessible and non network accessible. For example a webpage is a network accessible information resource while a printed book is not. However, information resources that are not accessible on a network can still be represented on one. For instance, a printed book can be represented on an Internet based e-commerce site. But the physical book is not on the Web, it is reflection of it represented by metadata such as title, author, publisher and price.

Resources in the real world are resources which essence is not information, for instance a person, a car or an organization. Resources in the real world are not accessible on the Web. But just like information resources in the class non network accessible they can be represented on it. An organization's webpage can contain all kinds of information about the company such as budget, vision, code of conduct etc., but the organization is not the webpage, the webpage is just data about it.

Conceptual resources; the definition of a concept is *"a meaning of a term as agreed upon by a group of responsible persons"*. (Klein and Smith, 2005) Thus a conceptual resource can be represented as a term and can be explained in a glossary. For example, the concept Area Under the Curve is used in the AstraZeneca organization, the concept is written down and explained, which transforms it from the concept Area Under the Curve to the term Area Under the Curve. The term is then published in AZ Glossary, which is an information resource. But when you look the term up in AZ Glossary you do not see the concept, you see the meaning of it written down as text.

Information Resources	*A resource that has characteristics that can be expressed in text form.*
Resources in the Real World	*Resources from the real world that do not carry information as their main characteristic.*
Conceptual Resources	*A general agreement on something that can be expressed as a term*

6.3 The Semantic Web

The Internet today has been a huge success when it comes to sharing and publishing of information throughout the world. The data on the Internet however do not become information until the data can be combined with some content so it will become useful for humans or computers. Internet today despite its huge success is just in its development phase, to access and understand the data and to make it useful, human involvement is necessary. That is because the language on the Internet is written in a natural way, easy understandable for humans but very hard for machines to understand. (Goble, 2003) To demonstrate this problem searching on the Internet is a good example. One major issue is when searches for information are conducted the response is often inaccurate and irrelevant. This is because a machine cannot understand the content and the meaning of a word; the result is that all the words that match the search are displayed. For instance if a person searches for the car brand Ford it is likely that the results will be mixed with answers about anything from the car brand Ford to the actor Harrison Ford. The machine does not understand the content of the word Ford and it is up to the human to interpret and sort out the relevant information.

The Semantic Web is a development of the current Web that intends to create a more powerful and useful information environment by enabling computers and humans to work better together. (Tim Berners-Lee, 2001) To achieve this, the first step in this progress is to tie machine-processable descriptions to the documents and data that already exist on the Web. (Miller, 2004) This is called metadata, data about data. Machines in form of programs and applications will then be able to know what the content of the word Ford means in the right context. To make this possible the Semantic Web uses descriptive technologies such as Extensible Markup Language (XML) and Resource Description Framework (RDF) and Web Ontology Language (OWL) (see picture). The Semantic Web does not replace the “old Web” it is instead integrated with it and works like a descriptive framework. Expectantly the Semantic Web’s contribution to the information environment will lead to more than effective searches, but also to a new platform of information infrastructure. (Goble, 2003)

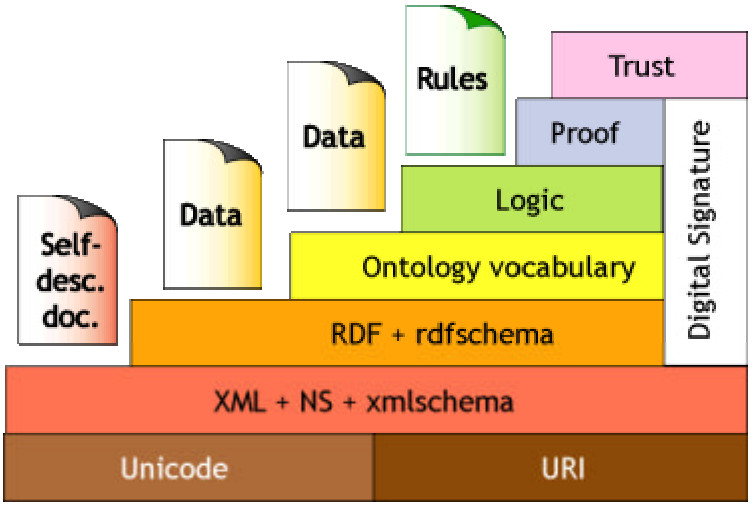
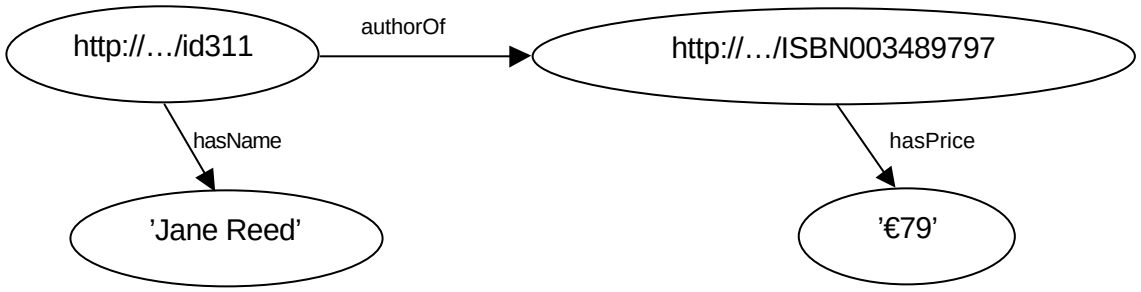


Figure 10: One of the main architectural premises of the Semantic Web is a stack of languages, often drawn in a figure first presented by Tim Berners-Lee. (Davies et al., 2002)

6.4 RDF

RDF (Resource Description Framework) provides a standard way for using XML to represent metadata. It is an extension of the XML and uses the same syntax. The metadata is represented in the form of statements about properties and relationships of items on the Web. These items can be anything as long as they are network accessible and have a web address, a URI. Metadata can be associated with information resources of different kinds, like a webpage or a book. It can also be associated with resources from the real world, for example colour and make of a car or name and age of a person. Of course metadata can also be connected to conceptual resources, for instance definition and term name of a concept.

The basic element in RDF is a triple made of an object, an attribute and a value, usually written as A(O,V). Which means, an object O has an attribute A with the value of V. The relationship can also be seen as a labelled edge between two nodes:
[O] - A ? [V]
Any objects or values can be interchanged. This means that any object can play the role of a value for another object, in a graphic representation they get chained together. (Davies et al., 2002)




```
<Description about="http://…/id311">
     <hasName rdf:resource=”Jane Reed”>
     <authorOf rdf:resource=”http://…/ISBN003489797”
</Description>
<Description about="http://…/ISBN003489797">
     <hasPrice rdf:resource=”€79”>
</Description>
```

The basic concept of RDF is to describe a Resource through a collection of Properties called an RDF Description.
Resource – Anything on the Web that has a URI can be described with RDF, for example all the Web's pages and individual elements in an XML document.
Property – A Resource that has a name and can be used as a Property for another Resource, this could be author or title. Even if only the name of the Property is interesting it has to be a resource because it needs its own properties. Each property has a Property Type and Value.
Statement – States the relationships between a Resource, a Property and a value (object, attribute and value). The value can be just a string, for example "Jane Reed" in the previous example or it can be another resource like "The homepage of http://www.w3.org/employee/id311 is http://www.w3.org".

6.4.1 Characteristics

There are a few characteristics that distinguish RDF from other languages and make it unique and flexible:

Independence – The fact that a property is a resource makes it possible for people to invent their own properties. One person might need a property called Author for books while another person is describing movies and needs one for Director. This is necessary since the Web does not provide an already finished database with all existing properties in the world.

Interchange – RDF statements are easily converted into XML that makes them easy to interchange.

Scalability – RDF statements are simply constructed and thereby easy to handle even when they come in large numbers. Since the Web is so big and is still growing there will probably be billions of these out there, scalability is needed to keep control.

Properties are resources – So properties can have their own properties, that is important because there will be a lot of them on the Web, way too many to be looked at one by one. Imagine that somebody is looking for a Property that describes the genre of a movie, with values like thriller, comedy and horror, then they can find an appropriate genre by searching on its properties.

Values can be resources – For example, a webpage has a property Home-Page which points to the home page of the site. The value of that Property is a resource itself thus it includes its own values, like title, Webmaster, last update and so on.

Statements can be resources – Sometimes Statements need their own Properties, for example “creatorOfStatement” and “dateOfStatement”. By adding this metadata to a Statement it enables people to perform searches of all the Web’s Statements, this can tell more about the Statement’s credibility and so on. (Bray, 2001)

Example

RDF extends the XML model and syntax to be specific for describing resources. The Namespace facility of XML, which is pointing to a URI, is used by RDF to scope and uniquely identify a set of properties, also known as a schema. This schema can be accessed at the URI identified by the namespace.

The namespace for RDF is shown as:

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

This declaration sets RDF as the default namespace. All the other namespaces will be declared as attributes within the RDF tag.

A full example of an RDF description is shown below:

```
1   <RDF
       xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
2      xmlns:dc="http://purl.org/dc/elements/1.1/">

3      <Description about="http://www.w3.org/Press/99Folio.pdf">
4        <dc:title>The W3C Folio 1999</dc:title>
5        <dc:creator>W3C Communications Team</dc:creator>
6        <dc:date>1999-03-10</dc:date>
7        <dc:subject>Web development, World Wide Web
8         Consortium, Interoperability of the Web</dc:subject>
9      </Description>
10  </RDF>
```

In this example, RDF is used to express data about the W3C Folio, the Consortium's Prospectus. The basic concept is that metadata about this item on the Web is described through a collection of properties.

Line 1: This line declares that the code is an RDF expression and that it uses the format defined by the RDF Model and Syntax specification on the given URL.
Line 2: This line indicates where on the Web the vocabulary can be found and how it should be used. The location http://purl.org/dc/elements/1.1/ is the Dublin Core, a vocabulary associated with bibliographic information.
Line 3: Shows the URI for the described resource. In other words the metadata descriptions will be about the Web resource http://www.w3.org/Press/99Folio.pdf, which is the W3C Prospectus in on-line form on the Web.
Lines 4,5,6, and 7: These lines show the metadata. The properties used are; title, creator, date, and subject. These refer directly to properties defined as part of the Dublin Core RDF vocabulary. When the metadata is processed, software will recognize these property names and deal with the metadata accordingly.
Line 8 and 9: Description end and RDF end. (W3C, 2000)

6.4.2 RDF Schemas

The properties defined in RDF can be seen as attributes of resources that are given a value, or they can represent relationships between resources. RDF can in neither of these cases provide any mechanism for describing the properties nor the relationships, that is the role of RDF Schemas (RDFS). RDFS is a semantic extension of RDF; it is used to describe vocabularies in RDF. This may be definitions of the characteristics and relationships of a set of properties and it can include constraints on potential values and inheritance of properties from other schemas.

An RDF-Schema does not have the same function as the name related XML Schema, while an XMLS prescribes the order and combination of tags in an XML document, RDFS gives information about the statements in an RDF data model and how it should be interpreted. It does not restrict the syntactical content in an RDF description.

```
<rdf:Property rdf:about="http://purl.org/dc/elements/1.1/title">
    <rdfs:label xml:lang="en-US">Title</rdfs:label>
    <rdfs:comment xml:lang="en-US">A name given to the
     resource.</rdfs:comment>
    <dc:description xml:lang="en-US">Typically, a Title will be a name
     by which the resource is formally known.</dc:description>
    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/elements/1.1/" />
    <dcterms:issued>1999-07-02</dcterms:issued>
    <dcterms:modified>2002-10-04</dcterms:modified>
    <dc:type rdf:resource=
          "http://dublincore.org/usage/documents/principles/#element" />
    <dcterms:hasVersion rdf:resource=
              "http://dublincore.org/usage/terms/history/#title-004" />
</rdf:Property>
```

The code above is an extract from the Dublin Core RDF Schema referred to in the prior example.

6.5 Ontology Languages

The Resource Description Framework (RDF) is an ideal and very powerful language for making and describing statements about web resources and their metadata. But it only provides the low level semantics needed to form metadata statements, which means that an RDF vocabulary must be built on top of existing RDF to support the expression of more specific forms of information within metadata (see picture). This

can be done with an ontology language like OWL (Web Ontology Language). OWL adds a layer of expressive power to RDF where it is possible to define complex conceptual structures that can be used to generate rich metadata. This technology is, however, very advanced and great expertise and knowledge is needed to make useful ontologies and this is superfluous for small applications like thesaurus and glossaries. (Alistair et al., 2005)

6.5.1 SKOS

SKOS Core is an application of RDF; it is used as a tool for publishing descriptions of concepts and concept schemes. A concept is a simple knowledge structure, it can be defined as "An abstract idea or notion; a unit of thought". A concept scheme is "A set of concepts, optionally including statements about semantic relationships between those concepts. (Alistair Miles et al., 2005) Examples of concept schemes are glossaries, taxonomies, terminologies and other types of controlled vocabularies. SKOS is a lighter version of the other ontology languages, compared to its big brother OWL. Compared to each other SKOS can be seen as a nutcracker while OWL is a big sledgehammer. (Alistair et al., 2005) The SKOS Core Vocabulary is a set of predefined RDF properties and classes. With these classes it is possible to express the basic concept and structure of a concept scheme as an RDF graph. The SKOS Core Vocabulary contains the most common and used relationships and attributes that are used in concept schemes context. To illustrate how it works an example is shown below, the graph represents an extract from the UK archival Thesaurus. (Alistair et al., 2005)

Term: *Economic Cooperation*
Broader terms: *Economic Policy*
Narrower terms: *Economic integration, European economic integration, European industrial cooperation, Industrial cooperation*
Related Terms: *Interdependence*
Scope note: *Includes cooperative measures in banking, trade, industry etc., between and among countries.*

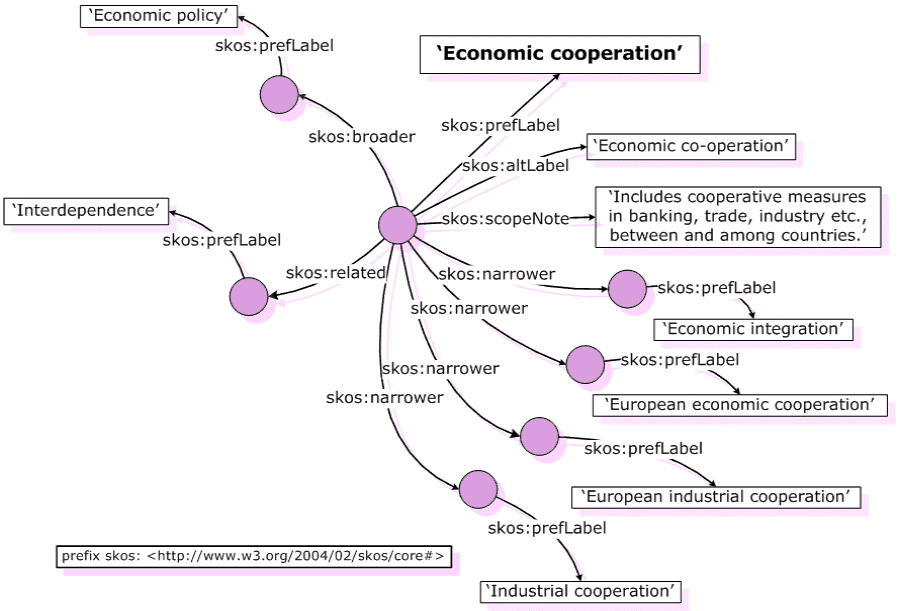


Figure 11: An RDF Graph from the UK Archival Thesaurus (Alistair et al., 2005)

The graph focus on the term Economic cooperation, all the circles on the graph represent another concept in the thesaurus and they are all related to the term Economic cooperation. The *prefLabel* indicates that the preferred name of the term is Economic cooperation and the *altLabel* shows that it is also known as Economic co-operation, which means that a search on any of these terms will lead to the same resource.

A SKOS serialisation of the RDF description above would look like this:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">

    <skos:Concept rdf:about="http://www.ukat.org.uk/thesaurus/concept/1750">
        <skos:prefLabel>Economic cooperation</skos:prefLabel>
        <skos:altLabel>Economic co-operation</skos:altLabel>
        <skos:scopeNote>Includes cooperative measures in banking, trade, industry
                        etc., between and among countries.</skos:scopeNote>
        <skos:broader rdf:resource="http://www.ukat.org.uk/thesaurus/concept/4382"/>
        <skos:narrower rdf:resource="http://www.ukat.org.uk/thesaurus/concept/2108"/>
        <skos:narrower rdf:resource="http://www.ukat.org.uk/thesaurus/concept/9505"/>
        <skos:narrower rdf:resource="http://www.ukat.org.uk/thesaurus/concept/15053"/>
        <skos:narrower rdf:resource="http://www.ukat.org.uk/thesaurus/concept/18987"/>
        <skos:related rdf:resource="http://www.ukat.org.uk/thesaurus/concept/3250"/>
    </skos:Concept>

</rdf:RDF>
```

An example of an organization that is already using SKOS is the European Environment Information and Observation Network (EIONET), they have built up their online thesaurus GEMET with RDF and SKOS (see picture). The thesaurus is using RDF as backbone to categorize the concepts into different thems and groups, and SKOS to relate the concepts to each oter. The web service used in GEMET to reach the data in the SKOS-file is available on the EIONET homepage. The web service provides the basic functions such as search for concept and get concept.

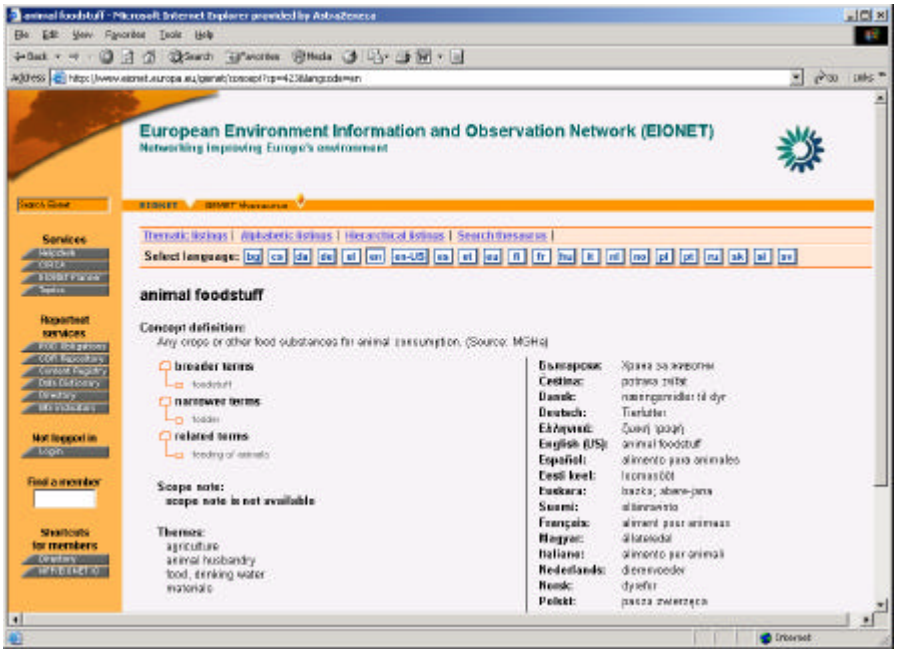


Figure 112: Screenshot from the GEMET online thesaurus.

6.6 Uniform Resource Identifiers

A uniform Resource Identifier (URI) is a string of characters for, just like the name indicates, identifying resources on the Web. Any abstract or physical resources on the Web like documents, web pages, images, downloadable files, electronic mailboxes can be identified with a URI. Resources are available under a variety of naming schemes and access methods like FTP, Internet mail and HTTP, but they are all addressable in the same way. In contrast to web protocols and web data formats were you can find several techniques, not just HTTP and HTML, there is only one technology for naming and addressing on the Web, and that is URI.

A URI consists of a sequence of characters that matches the syntax rules in a certain scheme, how the identification is done depends on the scheme specification that is used. The most well known form of URIs is probably the Uniform Resource Locator that locates resources on networks, for example on the World Wide Web.

6.6.1 PURL

One disadvantage with the addresses on the Web is that they are unstable and changes easily. A hardware reconfiguration or a simple modifying of the file system can make the old address out of date and all the links that point to that particular URL do not work any more. One solution to this can be to use a Persistent URL (PURL). A PURL stays the same even if the real address to the page changes. Instead of pointing to the location of an Internet resource the PURL points to an intermediate resolution server. When the PURL is entered in a browser, the browser sends the page request to a PURL server that returns the real URL of the page, and directs the user straight to that page. (Weibel et al.)

6.6.2 Identification Schema

There are several possibilities to build up the identification schema for publishing resources on the Web, and the result can be seen in the URI. For example, the URI for AZ Glossary can be structured in the following ways.

First part of the URI
AZ Glossary could use the common AstraZeneca namespace or be a stand-alone source.

http://www.astrazeneca.net/azglossary/ or http://purl.astrazeneca.net/azglossary/

The addresses above indicate that AZ Glossary is a source using an AstraZeneca common name schema for persistent identifiers.

AZ Glossary can also lie as a stand-alone source:

http://www.azglossary.astrazeneca.net/ or http://purl.azglossary.astrazeneca.net/

Second part of the URI
The type of resource could be integrated in the ID. For AZ Glossary the resources can be seen as both terms and concepts.

http://purl.astrazeneca.net/terms/azglossary/
or
http://purl.astrazeneca.net/terms/azglossary/

Third part of the URI
The specific resource could be identified with either its full name or its numeric id.

http://purl.astrazeneca.net/concept/azglossary/clinical/area_under_the_curve
or
http://purl.astrazeneca.net/concept/azglossary/clinical/MGOR-5BQC32

6.7 Structuring knowledge

A prerequisite for the improved functions that come with the Semantic Web is that the knowledge is structured in a way that makes it processable for machines. The meanings of the terms and possible relationships in between them are stated in an ontology.

6.7.1 Ontology

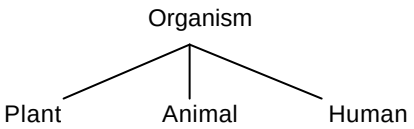
The term ontology has different meanings in different contexts. In general ontology is the science or study about existence and beings, about what different kinds of things or entities that exist in the universe. The word derives from the Greek *onto* (being) and *logia* (written or spoken discourse). It is a branch of metaphysics, the study of first principles or the essence of things. In Information Technology ontologies are used as a form of knowledge representation about the world or about just a part of the world. It is seen as a data model representing a domain, it is used to reason about the entities in that domain and the relationships between them. Ontologies consist of the following parts:

- **Concepts –** The objects and the sets of objects (classes or categories)
- **Characteristics –** The properties and attributes of the object.
- **Relations –** Models that show the relationships between the concept and characteristics.

Ontologies play a very important role in the Semantic Web framework. They have the ability to both classify data and store reasoning rules about the data that helps the computer to conclude new knowledge from the knowledge already represented in the ontology.

6.7.2 Taxonomy

Just like ontology the term taxonomy derives from the Greek, it is built up by the words *tassein* (to classify) and *nomos* (law). That is also the meaning of the word, to classify items into wider categories. One of the most famous taxonomies is biology classification that divides organisms in different categories, by the Swedish scientist Carolus Linnaeus. The science of classifying living organisms used to be the only meaning of the word taxonomy, but later on the word has applied in a wider sense and now it also refers to either a classification of things, or the principles underlying the classifications.



Taxonomies are often structured in a hierarchical order, like a tree, but they can also refer to other relationship schemes, like network structures. The categories in the taxonomy must be mutually exclusive, which means that a concept cannot belong to more than one category. Taxonomies are also exhaustive, which means that they must include all possibilities. The classes inherit all the properties from other classes above them in the tree.

When making an ontology the concepts that make up the domain needs to be divided into different classes and relationships must be set up in between them, therefore the work with the ontology includes making a taxonomy as well. One cannot study the concepts without knowing how they are related to each other.

6.7.3 Glossary

A glossary is a list of terms, often difficult or specialised, with the definitions of those terms. A common use of glossary is in the end of a book where the uncommon and newly introduced words are explained. Generally a glossary contains explanations of concepts relevant to a specific area of study or science, which makes the term contemporaneously related to ontology.

6.7.4 Thesaurus

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 are made explicit".* (Prüller, 2003) AF thesaurus is used for finding relationships and associations between terms within a particular domain. Given a certain term the thesaurus indicates the terms that have the same meaning and the ones that denote the broader and narrower categories. Compared to a dictionary that gives information to the users about unfamiliar concepts, a thesaurus provides the right words for the users when they just have concept in mind.

6.7.5 Controlled vocabularies

A vocabulary is a set of words known to a person or an other entity, or that are a part of a specific language. The vocabulary of an organization can be defined as the set of all the words that the organization is familiar to or is likely to use in speech, writing or other communication.

A controlled vocabulary is a set of word or phrases that are used to tag information so that a more precise retrieval of the content can be given in a search. The content in controlled vocabularies is referred to as metadata; it describes the data in databases, documents or other knowledge repositories. Thesaurus and taxonomies can for example be used as controlled vocabularies.

6.7.6 The connection

All the different ways of representing knowledge and its structure described above are closely related to each other, even though they work in different levels of the knowledge representations. On top is the ontology; it describes what concepts that exist in a world or domain and define the relevant ones. On the level below comes the taxonomy, here all the relationships between the concepts are clarified and they get categorized. The definitions of the concepts and their relationships to each other are put together in the glossaries and thesaurus. In each of them different aspects

are implemented. The thesaurus shows the knowledge structure from the ontology and the taxonomy, while the glossary gives the definitions of the concepts.

Chapter Summary

The Semantic Web is an idea to make the documents on the Web machine processable, which means that programs and applications can understand the meaning of the content and not only humans, benefits of this are that relationships between resources can be set up and the search accuracy will be enhanced. The Semantic Web is built up by a stack of languages with XML as a base further developed by RDF that makes statements about resources, and SKOS and OWL to define the meanings and relationships between resources.

There are three different kinds of resources; information resources, resources from the real world and conceptual resources. An information resource can be based on the Web, but even if it is not it can still be represented on the Web or another network with the help of metadata. The same implies for conceptual resources and resources from the real world

An ontology can classify data and store reasoning rules about the data, it describes the concepts that exist in the represented world and how they can be related to each other. In the taxonomy they get categorized and related to each other and then they can be explained in a glossary or a thesaurus.

7 Design

This chapter explains how the Semantic Technologies can be implemented in AZ Glossary and how it can improve the sharing of knowledge. In the second part of this chapter we present a few use case scenarios based on requests found out in the interviews and the possibilities enabled by the making the terms machine processable.

By making relatively small adjustments in the technical architecture of AZ Glossary, it can be improved a lot and it can be used in a more efficient way and providing several new important functions. Actually, the technical structure does not need to change, the solution with the Lotus Notes database will work fine, and it should not become a problem when it will be switched out in 2008.

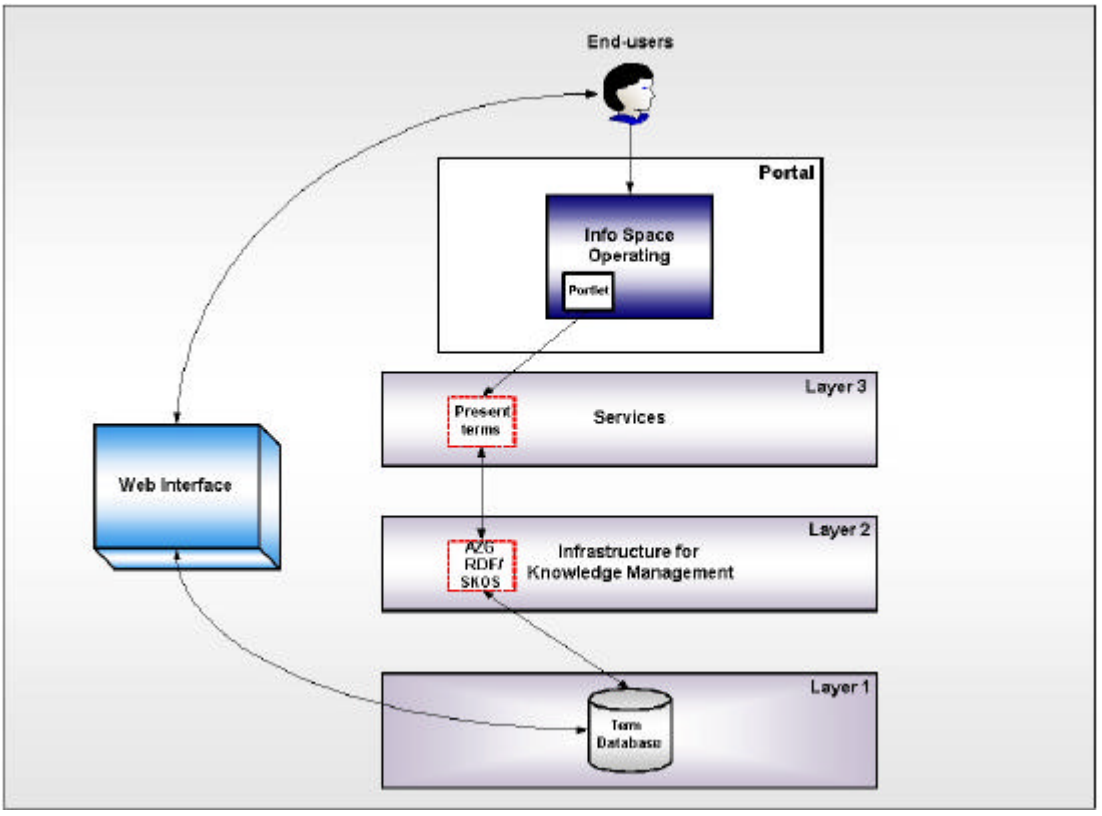


Figure 123: The picture shows how an extra layer can be added to the existing technical infrastructure of AZ Glossary.

The base will still be the same, what needs to be done is to add another layer on top of it (see picture). A service layer with the terms represented in a machine processable way. This layer will work against the other services and provide them with information from the database, like a definition on a requested term.

This new service layer will consist of all the terms formatted in SKOS. The setup will be the same as the XML-file that is produced automatically from the system today. The SKOS-file should be produced in the same way (see picture).

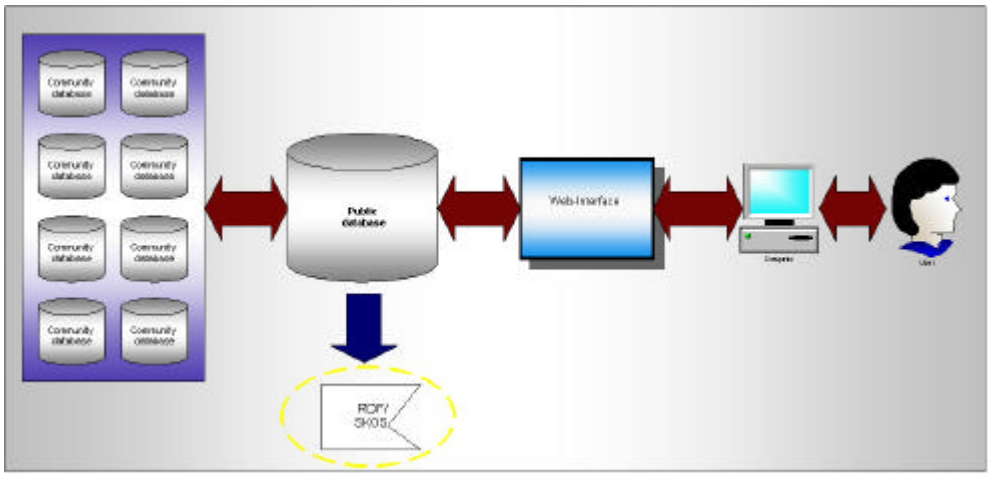


Figure 134: The picture illustrates how the new semantic file will be created.

In SKOS the information is tagged and built up in a way that makes it possible for machines to process the information. Every attribute for a resource is predefined in the SKOS Core Vocabulary and there are also possibilities to create your own attributes. When the information is structured like this other applications can easily extract data from the file. For instance, when clicking on a term in DPOM on the portal, the portal calls a web service that picks up the term's definition from the SKOS-file and shows it on the screen.

An example of how a term will be connected to its attributes is shown in the picture below. The attributes not available in the SKOS Core Vocabulary are specified in an RDF-schema named AZG or in the metadata initiative Dublin Core.

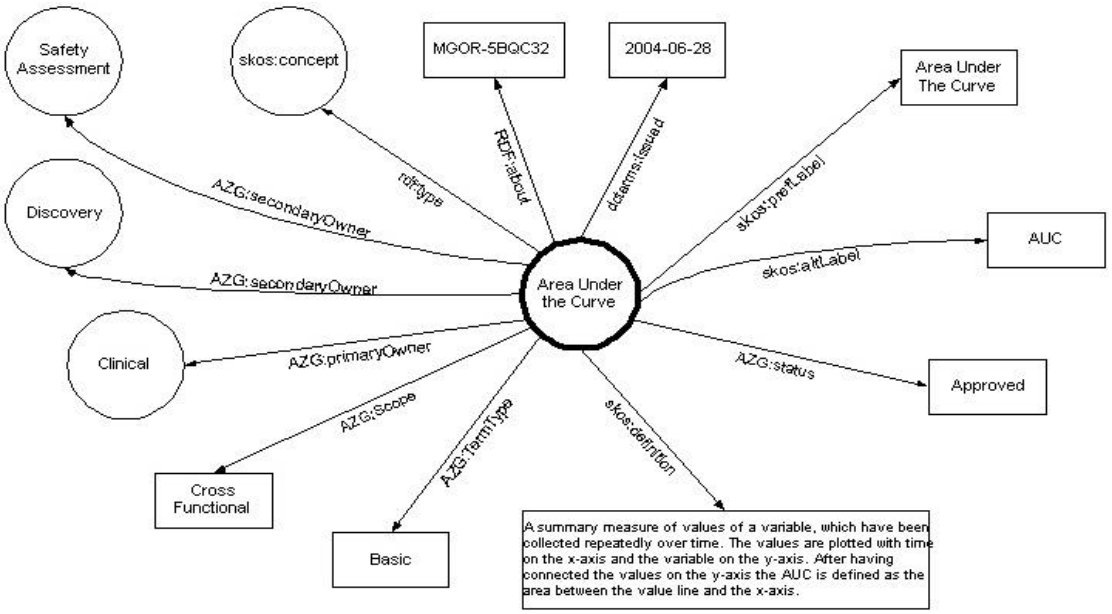


Figure 15: A graph showing the term Area Under the Curve and its attributes

The same term represented in SKOS code format would look like:

```
<!DOCTYPE skos [ <!ENTITY skos "http://www.w3.org/2004/02/skos/core#"> ]>

<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:AZG=" http://purl.astrazeneca.net/azglossary/AZG_schema">
    xmlns:dcterms="http://purl.org/dc/terms/">

    <skos:Concept rdf:about=
                "http://purl.astrazeneca.net/azglossary/concept/clinical/MGOR-5BQC32">
         <skos:externalID>MGOR-5BQC32</skos:externalID>
         <skos:prefLabel> Area Under The Curve </skos:prefLabel>
         <skos:altLabel> AUC </skos:altLabel>
         <dcterms:issued>2004-06-28</dcterms:issued>
         <AZG:status> "Approved" </AZG:status>
         <AZG:TermType> "Basic" </AZG:TermType>
         <AZG:Scope> "Cross Functional" </AZG:Scope>
         <skos:subject> "Clinical" </skos:subject>
         <AZG:SecondaryOwner> "Discovery" </AZG:SecondaryOwner>
         <AZG:SourceDefinitionSelection> "Internal" </AZG:SourceDefinitionSelection>
         <skos:definition>A summary measure of values of a variable, which have been
          collcted repeatedly over time. The values are plotted with time on the x-
          axis and the variable on the y-axis. After having connected the values on
          the y-axis the AUC is defined as the area between the value line and the x-
          axis.</skos:definition>
    </skos:Concept>

</rdf:RDF>
```

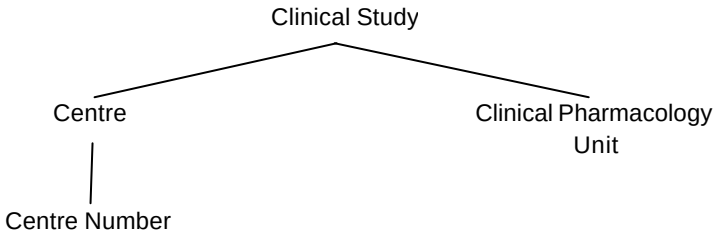
As seen in the code, SKOS is used as the core vocabulary, but also other vocabularies are used. All attributes used in AZ Glossary do not fit the existing ones in SKOS. These attributes are defined and explained in the RDF-schema AZG (see example below) and also the Dublin Core, that is another schema for providing metadata for bibliographic information.

```
     <!DOCTYPE skos [ <!ENTITY skos "http://www.w3.org/2004/02/skos/core#"> ]>

     <rdf:RDF
       xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"


       <!-- This is the extension of SKOS-Core for the AZ glossary -->
       <rdf:Property rdf:ID="primaryOwner">
          <rdfs:label>Primary Owner</rdfs:label>
          <rdfs:comment>Function/Discipline within AZ who have the main responsible on
                    the term,
                    also gives a hint in what business context the definition are
                    valid.</rdfs:comment>
       </rdf:Property>

       <rdf:Property rdf:ID="Scope">
         <rdfs:label>Scope</rdfs:label>
         <rdfs:comment>Functional (affects only one Function/ Discipline) or
                    Cross-Functional (affects several Functions)</rdfs:comment>
       </rdf:Property>

     </rdf:RDF>
```

The code above shows an extract from the AZG RDF-schema.

7.1 Term Relationships

AZ Glossary today can be referred to as a controlled vocabulary. The terminology in the glossary is a set of terms that are used in communication in the company. But to capture all the advantages that comes with the Semantic Web structure and get the maximum out of an implementation of SKOS a little bit more work should be done. A more accurate categorization of the terms than primary owner is desirable. The terms should get connected to each other, relations set up and networks be formed. This would take the glossary from a controlled vocabulary to taxonomy. All strict rules that a taxonomy implies must not be obeyed, for example the terms in the glossary must not be mutually exclusive, exhaustive or inherit all properties from their parent. A structure in the glossary with synonyms, narrower and broader terms (see picture), is all that is needed.



It is hard to say how much work this would imply for AstraZeneca, everything does not have to be done at once though. If starting now with defining relationships for all new terms added and adding relationships gradually when they show up, and when maintaining the terms also take a look at the relationships, the taxonomy starts to take shape and after a while it gets useful.

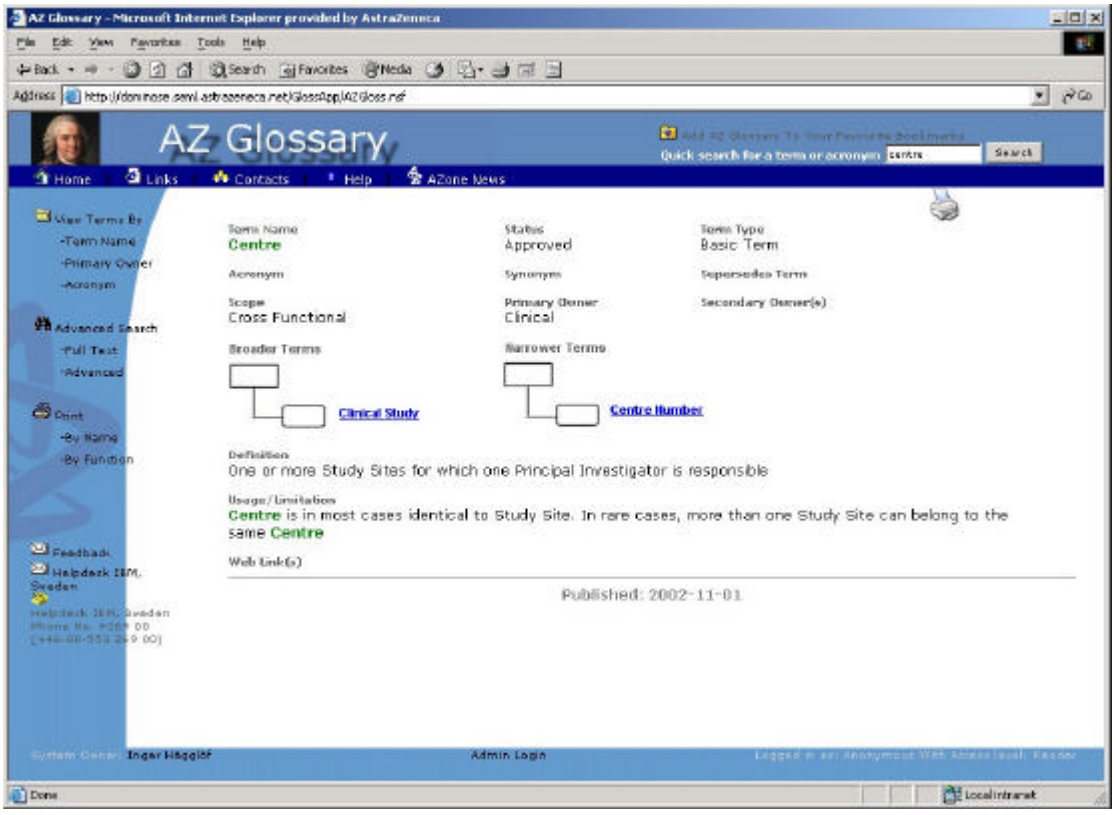


Figure 16: With SKOS it is possible to set relationships between the terms and link them to each other.

7.1.1 Use case scenarios

By using semantic technology in the AZ Glossary it will be possible for other applications to get information from the glossary and show directly on the screen. An example where this can be useful is on the Drug Project Operating Model webpage. The DPOM is a part of the portal and gives directions and guidelines for how to run drug related projects. The site contains a lot of terms that are found in AZ Glossary. Today the DPOM provides a link to the glossary and recommends the users to visit the glossary when they are unfamiliar with a term. But when using the link they get directed to the glossary in a new session in their browsers, and in the glossary they have to conduct the search themselves.

If implementing RDF/SKOS it will be possible to create functions for presenting the definition of a term in a pop up box directly on the screen, in the same browser session. Just by clicking or hovering with the mouse on the term.

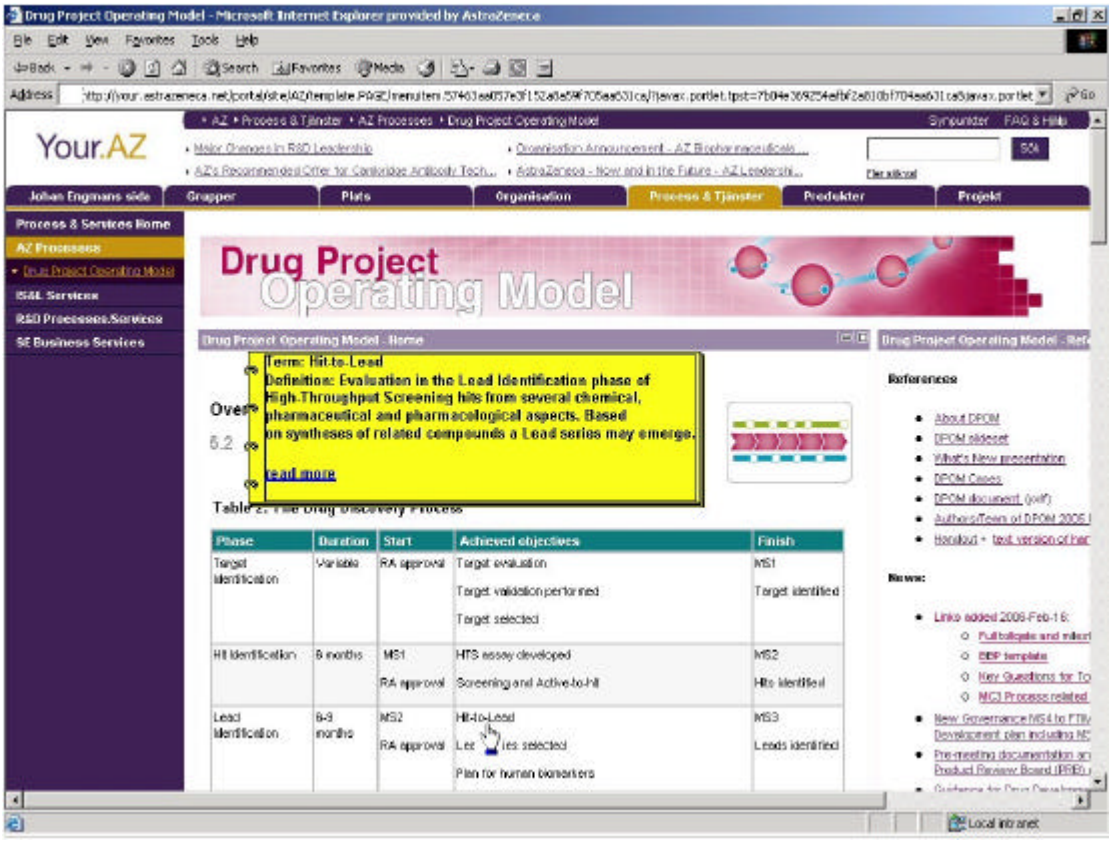


Figure 17: A proposal on how the content of the glossary can be used in DPOM.

The picture above is an example of how it can look with an on-screen-definition in a pop up box. In this example the box will appear when hovering with the pointer over the term in the text, in the box the full term name and its definitions appear. If clicking read more the box enlarges and shows all the attributes for the term.

This will save the users a lot of time and make their work more efficient. It will lead to a higher use of the content in the glossary. By positioning the content of the glossary closer to the users it will result in more people reading the correct definitions of the terms. That helps to prevent misunderstandings and by using the same terminology the staff will come closer to each other and the communication will improve.

The same type of function used in another way could be in the editorial environment of the portal. A method could provide an easy way for inserting definitions of terms in the text by just a few clicks. One possible solution, as seen on the picture below, is to put the insert definition-button in the menu that appears when clicking the right button on the mouse. The application then sends away the term via a web service call that picks up the definition from the SKOS-file and sends it back and the application inserts it in the text.

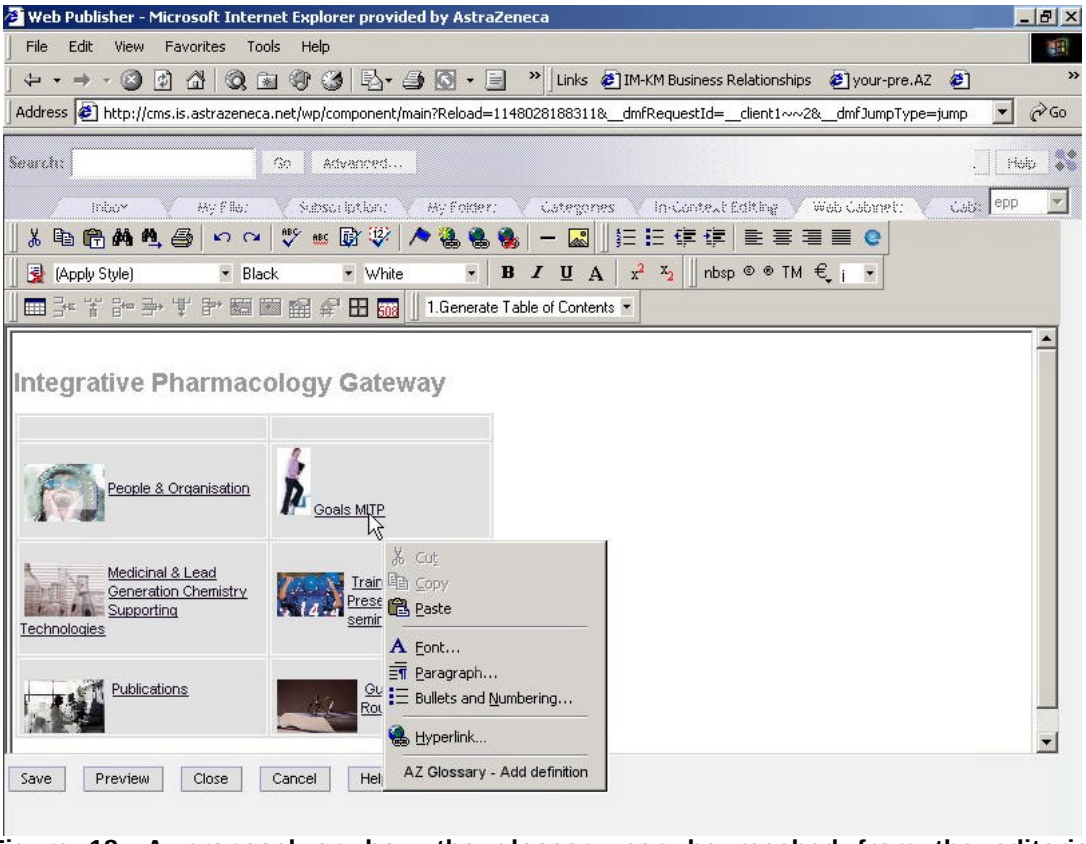


Figure 18: A proposal on how the glossary can be reached from the editorial environment of the portal.

The picture shows a possible way of inserting definitions from AZ Glossary in the texts on the portal. This function will make the work easier for the editors working with the portal. It will lead to more definitions in the texts, which spreads the knowledge to more readers.

By using RDF/SKOS to link the glossary with other tools in the organization it will improve the knowledge infrastructure in the company. The content will reach more people and improve the spreading and sharing of knowledge. As found out in the theory chapter, *"a powerful knowledge infrastructure strengthens the capabilities of the organization".*

For an easy, quick and smooth implementation of SKOS it is recommended to look at former SKOS related projects in other organizations. An example of one is EIONET's online thesaurus GEMET. They have put all their RDF and SKOS files available for the public on their webpage. They also provide a web service that can be used by programs and applications to access the data in the RDF/SKOS files. Using these files, or parts of them, can save a lot of time and effort. A bit modification is needed but it serves as a good ground to base the further work on.

8 Conclusion

This chapter brings all the parts together. We answer the questions and problems from the first chapter and summarize the outcome from the discussion and discuss which implications they might have.

In our examination of AZ Glossary we performed interviews with different users and administrators of the glossary. We found out that the glossary is an important tool for sharing of knowledge in the AstraZeneca organization. It is seen as an official source of knowledge and helps spreading knowledge in the company. From the view from our question at issue in the first chapter:

"How can AZ Glossary be seen as a part of AstraZeneca's knowledge infrastructure and how can the implementation of semantic technologies improve its function as one?"

The content in AZ Glossary can be seen as knowledge and the application as a tool for the sharing of it. Both to be considered infrastructures but in different ways, according to Ole Hanseth knowledge should be seen as infrastructure due to its systemic and stable characteristics and that it consists of various numbers of shared resources to support the activities in the company. The application itself is one of the resources that supports various activities in different communities by transporting knowledge through the users. Therefore AZ Glossary is a part of the knowledge infrastructure. With the help of the semantic techniques this infrastructure would be improved. Today's version is not perfect, it lacks a function to link users straight to a term from other programs and applications, a function like this is demanded all over the company. An implementation of RDF/SKOS would make the terms machine processable, then the glossary could be linked to other tools in the organization. More exactly, other programs and applications could extract information from the glossary with the help of web services. This information can for example be used to provide users with definitions of terms when working in certain programs and applications, by hovering with the mouse over a term the user could see the definition directly on the screen without having to leave the program or starting a new session. This would be even better and more efficient than linking to terms. To get the maximum out of the SKOS implementation, relationships between the terms should be set up; this will take the glossary from a controlled vocabulary to a taxonomy. By making the content easy accessible and bringing it closer to the user, the knowledge will reach more people in the company. This will improve the communication in the company and prevent misunderstandings. A powerful knowledge infrastructure increases the capacity of an organization.

References

ALISTAIR, M., BRIAN, M., DAVE, B., DAN, B., MICHAEL, W. & NIKKI, R. (2005) SKOS: A language to describe simple knowledge structures for the web. **URL**: http://www.idealliance.org/proceedings/xtech05/papers/03-04-01/ Last accessed: 2006-05-29

ALISTAIR MILES, C., BRIAN MATTHEWS, C., MICHAEL WILSON, C. & DAN BRICKLEY, W. C. (2005) Simple Knowledge Organisation for the Web. Madrid, W3C.

BACKMAN, J. (1998) *Rapporter och uppsatser*, Studentlitteratur.

BLACKLER, F. (1995) Knowledge, knowledge work and organizations: An overview and interpretation. *Organization studies***,** 1021-1046.

BRAY, T. (2001) What Is RDF. XML.com. **URL**: http://www.xml.com/pub/a/2001/01/24/rdf.html Last accessed: 2006-05-29

CHEN, I.-X., YANG, C.-Z. & HSU, T.-L. (2005) Design and evaluation of a panoramic visualization environment on Semantic Web. *Information and Software Technology,* In Press, Corrected Proof.

CONNOLY, D. (2006) Naming and Addressing: URIs, URLs, ... W3C.

DAN BRICKLEY, W. C. & GUHA, R. V. (2004) RDF Vocabulary Description Language 1.0: RDF Schema. IN MCBRIDE, B. (Ed.), W3C.

DAVIES, J., FENSEL, D. & HARMELEN, F. V. (2002) *Towards the Semantic Web*, Wiley.

DESOUZA, K. & EVARISTO, R. (2003) Global Knowledge Management Strategies. *European Management Journal,* 21**,** 67.

EASTERBY-SMITH, M., THORPE, R. & LOWE, A. (1991) *Management research : an introduction*.

GOBLE, C. (2003) The Semantic Web: an evolution for a revolution. *Computer Networks,* 42.

GUPTA, B., IYER, L. S., & ARONSON, J. E. (2000) Knowledge management: Practices and challenges. Industrial Management and Data Systems,. 100(1), 17–21.

HALPIN, H. (2006) Identity, Reference, and Meaning on the Web. Edinburgh, Scotland, United Kingdom, University of Edinburgh. **URL**: http://www.ibiblio.org/hhalpin/irw2006/hhalpin.html Last accessed: 2006-05-29

HAMMERSLEY, M. & ATKINSON, P. (1983) *Ethnography: Principles in Practice*.

HANSETH, O. (2004) KNOWLEDGE AS INFRASTRUCTURE. IN AVGEROU, C., CIBORRA, C. & LAND, F. (Eds.) *The Social Study of Information and Communication Technology: Innovation, Actors and Contexts.* Oxford: Oxford University Press.

HARPER, R. (2000) The Organisation in Ethnography - A Discussion of Ethnographic Fieldworks Program in CSCW.
URL: http://lair.indiana.edu/courses/l701/papers/harper.pdf
Last accessed: 2006-05-29

IANELLA, R. (1999) An Idiot's Guide to the Resource Description Framework. *The New Review of Information Networking,* 4.

KING, W. R. & MARKS, J. P. V. (2005) Motivating knowledge sharing through a knowledge management system. *Omega,* In Press, Corrected Proof.

KLEIN, G. O. & SMITH, B. (2005) Concept Systems and Ontologies.
URL: http://ontology.buffalo.edu/concepts/ConceptsandOntologies.pdf
Last accessed: 2006-05-29

LANGEMAR, P. (2005) ATT LÅTA EN VÄRLD ÖPPNA SIG - Introduktion till kvalitativ metod i psykologi. Stockholm, Psykologiska Institutionen, Stockholms Universitet.

LASSILA, O. & SWICK, R. R. (1999) Resource Description Framework (RDF) Model and Syntax Specifikation.

LEE, J. & GOODWIN, R. (2006) Ontology management for large-scale enterprise systems. *Electronic Commerce Research and Applications,* 5**,** 15.

LIEBOWITZ, J. (2003) A knowledge management implementation plan at a leading US technical government organization: A case study. Knowledge and Process Management. 10(4), 254–259.

MILLER, E. (2004) Den semantiska webben.
URL: http://www.w3c.se/nl-news/nl-2004-08-25.html
Last accessed: 2006-05-29

PRÜLLER, R. (2003) KEYWORD LISTS after WALLINGFORD.
URL:http://www.efi.fi/attachment/f5d80ba3c1b89242106f2f97ae8e3894/16041eb2606438a55f9e2e40846e16f0/table1_keywords_part.xls
Last accessed: 2006-05-29

SHAPIRO, C., AND VARIAN, H. R. (1999) Information Rules: A Strategic Guide to the Network Economy. *Boston, MA: Harvard Business School Press.*

SIVAN, Y. Y. (2001) Nine Keys to a Knowledge Infrastructure:A Proposed Analytic Framework for Organizational Knowledge Management. Harvard University.

BERNERS-LEE, T., Hendler, J. & Lassila Ora The Semantic Web (2001) Scientific American
URL:http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21
Last accessed: 2006-05-29

W3C (2000) Metadata activity. World Wide Web Consortium.
URL: http://www.w3.org/Metadata/Activity.html
Last accessed: 2006-05-29

W3C (2006) Introduction to XML.
URL: http://www.w3schools.com/xml/xml_whatis.asp
Last accessed: 2006-05-29

WEIBEL, S., JUL, E. & SHAFER, K. PURLs: Persistent Uniform Resource Locators. OCLC Online Computer Library Center.
URL: http://purl.oclc.org/docs/new_purl_summary.html
Last accessed: 2006-05-29

WONG, K. Y. & ASPINWALL, E. (2006) Development of a knowledge management initiative and system: A case study. Expert Systems with Applications

Appendix 1 – Interview Guide, Core team

1 Define your part in the AZG organization?

2 What different processes exist in the work with AZ Glossary?

3 What type of terms do you regard should exist/not exist in the glossary?

4 How do you use AZG in your work? Give example .

5 How do you think the system with the communities functions?

6 Do you have any plans on further development with AZG?

7 Is there anything you would like to change immediately in AZG?

8 What effects has AZG contributed with to the AstraZeneca organization.

9 How do you think AZG contributes to the sharing of knowledge in the company?

10 Who is the primary user of the AZG?

11 How do you think AZG contributes to the work of the common employee?

12 How do you measure the use of AZG?

13 What do you miss in AZG?

14 How do you think the awareness of AZG can be increased in the organization?

15 What possibilities for the development of AZG can you see?

16 What obstacles for the development of AZG can you see?

17 How has your perspective of AZG changed since the implementation?

18 What do you think AZG would look like in ten years?

Appendix 2 – Interview Guide, End-Users

1 What is your part in the AstraZeneca organization?

2 How do you use AZG today?

3 How often do you use AZG to solve a problem?

4 What sort of problems do you solve with AZG?

5 In what other service do you think the content in AZG can be utilized?

6 How do you think AZG contributes to the sharing of knowledge in the company?

7 How did you come across AZG for the first time?

8 Which alternative sources of knowledge do you use for solving problems?

9 Vilka alternativa kunskapskällor använder du för att ta reda på något du inte vet?

10 What do you miss in AZG?

Appendix 3 – Interview Guide, People with interests to use the glossary's content in other services

1 What is your part in the AstraZeneca organization?

2 What experiences do you have with AZG?

3 How do you use AZG today?

4 What kind of services are offered at the portal today?

5 How do you wanna use AZG in your services?

6 What other type of services do you want to provide?

7 Who are using your services?

8 What is needed to enable these services?

9 Can you see any obstacles for the development of AZG?

10 How do you think an end-user can utilize AZG through your services? Do you know anybody who does or wants to today already?

11 Have you done any attempts of using AZG in your services before? How? What services?