Generation of ontologies for workflows

Per Flensburg

Växjö University

School of Mathematics and Systems engeneering

Per.Flensburg@msi.vxu.se

***Abstract**: The article discusses methods for generation of ontologies in interorganisational workflows. A framework for discussing ontologies is introduced as an ontology to the article. It is argued that ontolgies will play a great role in the future and a growing interest within the scientific community is indicated. Finally a rough sketch to a method is presented.*

***Keywords**: ontology, ontology generation, method for ontology generation*

1 Ontologies within the automotive industry

> "In the end of the last century the automotive industry was focussed on improving processes supported by Information Technology (Kanban, JIT, Lean Production etc.). In the 21st century the automotive industry is faced with challenges requiring novel organisational models and technologies for cross-organisational collaboration, integration and communication [*European Engineering User Conference* (e.e.u.c.) 2002]. Traditional processes, based on division of labour, are replaced by dynamic development and production networks with flat organisational structures, so-called *Networks of Automotive Excellence (NoAE)*. Automotive suppliers will participate in several networks simultaneously, and in each network the role of a specific supplier will be different. Responsibilities at OEMs will shift from purchasing and supplier management to network management, including configuration and termination of networks. True collaboration in NoAE requires the tight integration of coordination and exception handling mechanisms across organisational boundaries. This requires strong interoperability between different local, organisation-specific Workflow Management Systems (WfMSs). The highly dynamic character of NoAE and the necessary required coordination and exception handling mechanisms impose further severe requirements on the flexibility in workflow formation and enactment."

The above is the background for an EU-project within the 6th framework program (Fessl K, 2003). The main idea is using agent technologies for the workflow formation and enactment in the automotive industry. The prerequisite for this to work is of course the same ontology over the whole supply chain, in fact over the whole automotive industry. It has nothing to do with the e-business aspects, here a standard – Odette – have been in use for several years.

Agent-based workflow management requires a machine-readable representation of domain and workflow knowledge. In the project we will create an ontology of the target domain, which includes details relevant for the operation of cross-organisational workflow systems such as formal specification of contracts, business rules and constraints, partner capabilities, the workflow model itself, etc. The ontology ensures the interoperability on a semantic level.

Thus a main research question raise: *How shall an ontology describing workflows be generated?* This is the issue addressed in this paper. The structure is as follows:

- First I provide an ontology of this paper, that is, a definition of central concepts I will use. This is presented in chapter 1.1

- Then I discuss ontologies in the context of workflow management. The conclusion is that we need more ontologies related to processes and hence work flow. This is done in chapter 2.
- In chapter 3 I present a technique for generating ontologies based upon a dialogue with the workers. This is done in chapter 3.
- A severe problem in ontologies and ontology generation tolls is the inability to give a proper description of the context in which the concepts are to be interpreted. This is discussed in chap 4 and some discussions about ways of avoiding these problems are carried out.

1.1 Ontology of this paper

Information system has been a science since the middle of the 60s (Langefors B, 1966). Still there is often confusion about the meaning of central concepts, why I will provide my ontology here. The concepts I will discuss are gathered in the following clusters:

- Data, information, content and knowledge
- Ontology, semantics

1.1.1 Data, information, content and knowledge

Data are defined as symbols without meaning. Example of such symbols are √,Δ❄, ✳↷,➢mh, uht↬bru. These symbols contain a string of signs. However, we can also have symbols, which we associate to familiar phenomenon such as 42, Gothenburg, capitol of Sweden. Still they have no meaning uttered just out in the air. But put into a context, we can give them a meaningful interpretation. If we say "The capitol of Sweden is Stockholm" we achieve *information* as well as if we say "The answer to the question of life, universe and everything is 42" we also have information. The first information provides a common well understood fact, namely the name of the capitol of Sweden. The second is perhaps more tricky if you are not familiar with Douglas Adams' "Hitchhikers guide to the Galaxy" (Adams, 1979). The first could also be tricky if you were not familiar with "capitol" or "Sweden". We see that in order to interpret information we need a context. Information interpreted in a context by a human being is thus considered as *knowledge*.

However, the concept of context has been used in two meanings here. First I said, that data put into context gives information. In fact *context* could here be replaced by "*syntactic structure*". Context in the second case is related to human understanding and is very often impossible to describe. If you havn't read Adam's book, the second example is totally incomprehensible. Yes, even if you read the book, it is not understandable, which in fact is part of the specific humour in this book.

The syntactic structure, in which data are put, does not necessary needs to be a grammatical structure. Suppose we have the following:

Stubbhead	Coneswinger	040307	2
Grimsfeld	Crthw	040306	3
Turbin	Travers	040306	2

Table 1: A syntactic structure

There are several possible interpretations, to this information (because it is data, put into a structure), but none obvious. If we put a header to the table indicating the meaning of every column, then we have something like Table 2.

Customer	**Ordered parts**	**Day of order**	**Number**
Stubbhead	Coneswinger	040307	2
Grimsfeld	Crthw	040306	3
Turbin	Travers	040306	2

Table 2 The table with headings: The content

The name of the columns indicates some kind of meaning of the content in that column. It is *metadata*. With the headings the table is not only information; it is something more. Still it is not knowledge, since we have no background knowledge at all. We could guess that the table is from an order entry system, but it could as well be from a sales statistics system. I also suppose very few have an idea of what "Coneswinger", "Crthw" etc. actually are. Hence I will use the concept *content*, meaning information + metadata.

Concerning the single datum it can have different *formats*. In Table 2 the day of the order has the format YY-MM-DD. However, it cold as well has been DD-MM-YY or even YY-DD-MM. This description must thus be supplied somewhere as *description of metadata*.

1.1.2 Ontology, semantics

The two concepts *semantics* and *ontology* are often used as synonyms. To explain the difference I will use Table 2. The headings for each column, says what kind of information is located in that column.

A row in the table represents *a fact*, namely the fact that a certain customer has ordered a certain number of a specific part. The headings of the columns thus represent metadata according to the facts. The explanation of the meaning of the heading, for instance that the "day of order" should be described as YYMMDD, following standard ISO 321-543-432-645.a, etc, I will call the *semantic description* or simply the *semantics*.

The *ontology*, on the other hand, deals with the value domains of the columns. We know for instance that there are max 12 months a year, we also know the number of days in each month. We also know which parts we are selling and who are our customers. An ontology is thus similar to a dictionary or glossary,

but with greater detail and if machine-readable, it has a structure that enables computers to process its content. The ontology deals with the content and the meaning not with the form, as the semantics does. An ontology consists of a set of concepts, relations, and axioms that formalize a field of interest. We can summarise the ontology of Table 2 in a primitive and homemade description:

Customer: Person or company registered as customer in our customer file.
Ordered parts: The name of the parts the customer has ordered. The names must be the same as in our part database
Day of order: Year (two last digits), month (two digits) and the number of the day in the month.
Number: The number of {parts} the {customer} is ordering.

It is to be noticed that the purpose of an ontology is twofold:

1) To ensure that the data entered in the fields correspond to the reality, the fact that is at hand
2) To provide a comprehensible description of the domain.

The latter can primarily be made for being machine-readable or it can be made primarily for being understood by humans. In most cases the machine-readability is in focus for research, despite the fact that if the users don't comprehend the content, the users of the information system cannot use it effectively and as intended by the designer.

The problem of comprehension can also be put in two ways:

1) To comprehend an existing information system.
2) To make the designer and user share the same comprehension of the domain.

The problem presented in 1) is addressed by many researchers (Gäre, 2003), (Docherty P and A, 1992). The problem presented in 2) is the same as presented in (Flensburg and Milrad, 2003) and in (Flensburg, 2003). In this actual case, generation of workflow ontologies within the European Automotive Industry, case 2) is at hand. Hence I will concentrate my efforts within this area.

2 Ontologies and workflow management

The content of the information, the semantic, has been treated within the informatics research area for many years. In database theory, huge amounts of efforts have been invested in the area (see for instance older efforts such as (Griffith, 1982) (Subieta, 1985) (Peckham and Maryanski, 1988), and newer such as (Syu et al., 1996), (Doan et al., 2001), (Ströbel, 2001), (Kim, 2002). In modern time, web-services are claimed as the solution of transferring content from one system to another (Alpher, 2001), (Devendorf, 2001), (Allen and Fjermestad, 2001), (van Hooft F and Stegwee R, 2001). The problem of

reconciliation is also recognised (Embury et al., 2001, Fan et al., 2001), but still, there is a great focus on formatting matters. However, as (Sowa J F, 2000) puts it: "*Formatting is an aspect of signs that makes them look pretty, but it fails to address the more fundamental question of what they mean*." SGML and XML introduce an important separation of semantics and formatting, but the semantics itself must have a semantic.

> A resource can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), and a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources. (Berners-Lee, 1998)

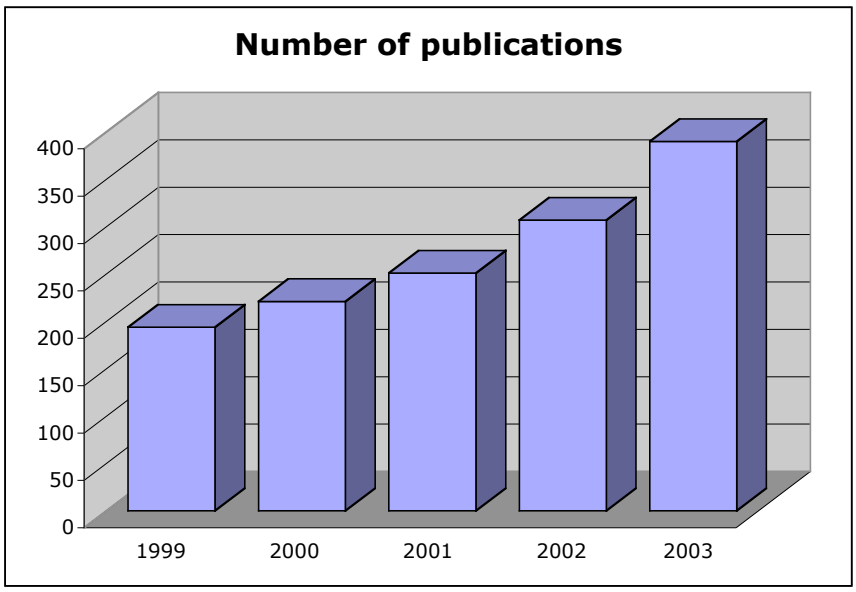
It is to be noted that humans are not considered as "familiar resources". Yet, without people, the document and its content have no meaning.

A specific version of ontology is called *upper ontology* (Schoening J, 2003). An upper ontology is limited to concepts that are meta, generic, abstract or philosophical, and hence are general enough to address (at a high level) a broad range of domain areas. Concepts specific to particular domains are not included in an upper ontology, but such an ontology does provide a structure upon which ontologies for specific domains (e.g. medicine, finance, engineering, etc.) can be constructed.

System integration and information mapping is increasingly more important. The main obstacle for integration is the semantic differences (Embury et al., 2001). These are supposed to be solved by standardisation of the business process, but that possibility is in my mind doubtful, since the customers demand personalised service (Keen, 2001). Standardisation can certainly be possible at the component level, but not for the whole business process.

In order to facilitate integration Web Services are used (Metz, 2001, Willaert, 2001, White and Hall, 2000, Ströbel, 2001). These are web-based services carrying out a small part of a business process in order to making reliable integration. Considering the Web Services we have today, the focus is on transferring the content of the information. It is a transfer on the semantic level, as I have defined it in section 1.1.2. It is of courses important to know that a certain byte-string means "customer", but if "customer" is defined in different ways (as it usually is) the business process mapping will not work as intended. Thus ontology is needed and we see a growing focus on the subject.

An indication to this is shown in fig 1 where the numbers of hits on the word "ontology" in the journals in a university library (Växjö University) is plotted for the years 1999-2003. The investigation is not a strict scientific, but the data base covers most journals within informatics, so there should be a pretty good indication of an eventual trend. We also see the contour of an exponential growth.



Figur 1 Number of publications about "ontology" in a university library.

Another indication is the semantic web ((Berners-Lee et al., 2001, Berners-Lee Tim, 1998), which has been around for some years. W3C has put a lot of effort in realising techniques etc. for it, and more is to come (Miller Eric, 2003). Also on the ontologies a lot of efforts have been made and OWL has been recommended as the W3C standard ontology language ((McGuinness and van Harmelen, 2004).

2.1 Scope of the ontology

If we take a look at current ontologies they are usually designed for keeping track of objects (≈things) (Fensel, 2001). However, seen in the perspective of the network society (Flensburg, 2002) there is more need to focus on business processes and on activities. The rest of the paper will focus on this issue.

The reason why we have to develop ontologies also for processes and workflows is the distributed nature we see today in working life and which will probably increase in the future. Supply chain and similar ways of organising works will be more important and thus also external business processes. Within certain areas (for instance car manufacturing, electronic components) ontologies describing the whole supply chain have been developed (Odette, Rosettanet). Due to the flexible nature of industry in the future, I think the need of having several and rather specific ontologies will increase. Thus, in research there is a need for methods of generating a correct ontology for a certain supply chain. As far as I know, little has been done in this area.

Before going deeper into workflow ontologies, we have to discuss how an ontology is created. It is basically the same procedure as when we learn a new language or how we learned to speak when we were children. We start with an everyday language, which covers everything, but with rattle low precision. New concepts are introduced and defined based upon the concepts we already have. Examples are very often given, thus simplifying for humans to understand what it all is about. This procedure works, since the everyday language can transcend itself; that is genuinely new concepts can be introduced within the language.

When we want to develop machine-readable ontologies, we thus have to start with something already existing and then add new concepts by deriving them from the existing ones or by introducing them in other ways. That is why upper ontologies are interesting. They can act as a substitute for the everyday language. However, the basic question is of course to find, identify and describe the new concepts used within the domain.

3 Ontology for processes and workflows

Up to now I have argued that

1. Ontology is one of the most important fields in the coming years
2. There is a need for describing ontologies dealing with processes and workflows and not only e-business.

In order to describe a workflow we have to identify the basic unit of analysis. I argue it is the *activity* or *work task* since all work can be divided into tasks. The level of description of the task is dependent on the work role: A manager gives a less detailed description than a blue-collar worker does on the shop floor. On the other hand, the manager usually considers only one aspect, while on the shop floor; many aspects have to be integrated in the actual work.

If the activity is the basic unit of analysis, what is the upper and lower border of the system (Churchman C W, 1971)? I argue the lower border is *a fact*, as all other descriptions of the reality. The upper border should thus be the supply chain. Supply chain I define as the set of every activity, that increases the value of the product in such a way, that the end customer (consumer) is willing to pay for it.

3.1 Rich description of a fact

If we say “description”, we associate to a verbal description. In fact almost[1] every other type of description could be replaced by a verbal description, however it might in some cases be very lengthy. Let us thus for the moment concentrate on

[1] According to Shannons information theory it can be replaced by ones and zeros.

verbal descriptions. The next question is: “What is a fact?” because if we do not know that, we cannot describe it. On the other hand, as my friend Paul Lindgreen says: If you cannot describe it, you do not know it! Nevertheless, in the context of information systems use, I think we in the workflow context can define a fast as something like:

A fact is an association of a value to a certain attribute of an activity in a workflow.

Put in a more formal way it can be expressed as <value>; <attribute name>; <object name>, <object type>, <system>. In a verbal description, this is manifested as a simple sentence, saying for instance, “42 gearshafts were put into the store“. In a formal way it could be described as:

<42>, <put into>, <gear shafts>, <the store>

(Langefors B, 1966) pointed out that this is not enough. You must also know the time when this was correct. Then (Ivanov, 1972) pointed out that even this is not enough, since you must give an estimation of the uncertainty of the value. This estimation cannot be done by any other than the actual users, and in fact, Invanov proved a necessity for a human judgement just in the middle core of the system. Principally, it is the same as when Markku Nurminen proposed HIS (Nurminen, 1988).

Nevertheless, this does not provide a description of the context. It is just a more accurate description (or estimation) of the value. In the context, we have other types of attributes. In fact, what we can say about an activity is determined by our grammar. Expressed in grammatical terms a fact can be descried as

<verb phrase>, <noun phrase>, <accusative attribute>, <amount attribute>, <time attribute>

It is to be noted that the activity is denoted by the verb phrase. The main idea now emerges: If we assign every possible attribute to a verb phrase, we then obtain as rich description as possible of the reality described by that specific verb phrase under the specifically given circumstances. Let us now skip a 30 pages excursion in dependency grammar and just present the result.

4 The Socrates-technique

This description technique was named after a project, which ended 1984.

The first problem is: How do you identify an activity? This in fact rather easy because you have a certain word class, the verbs, denoting activities. Obviously it must be the events concerning the system users that are interesting. So, by

asking the simple question: "What do you do?" or "what happens?" you can get a good description of the actual events as the actual person perceives them.

The next problem is the dependency between the events. Here the grammar of the natural language comes to our help. It exists in fact just a few possible types of qualifiers to a verb and with their help you can describe everything(!) about it. You catch these qualifiers by asking certain questions. These questions are of course dependent on the grammar and the language you use. For the Swedish language we have so far used the following set of questions. Items in brackets < > should be replaced with the actual item. They should not be considered as XML-tags even if this is possible with some modifications.

1. Who performs the activity? (The subject)
2. What <activity> <who>? (Direct object)
3. To whom <activity> <who> <what>? (Indirect object)
4. When <activity> <who>?
5. Where <activity> <who>?
6. How <activity> <who>?
7. What's the intention with the activity?
8. Which are the conditions for performing the activity?

The questions give an exhaustive enough description of the actual event but one problem remains and that's the problem of how to know that you haven't missed any events. The last two questions in the scheme above indicate which events that are immediately before and immediately after and in the answers to the other questions may other events pop up.

The next step in the development process is to detect certain "information activities", that is activities in which you deal with information in some way. You could produce new information, you can retrieve stored information, you can modify information and so on. The identification of these activities may be a little tricky to identify, but it has been possible to do by using common sense hitherto. After having identified the information activities you describe them in a sort of "almost natural programming language" using common program language concepts such as IF ... THEN ... ELSE, AND, OR, DO etc. In the language you use name on processes, other activities and data items. This processes and data items should be defined when so is possible.

With this semi-formal description as background you construct the data model. This is very simple and has two basic concepts, information object and attributes of this information object. The attribute may be another information object, which in turn can have other information objects as attributes and so on.

This technique will be a good description of the "linguistic reality" of the person answering the questions. If we make such descriptions for all people coming into contact with the computerized information system it will be possible to construct "the objectified reality" for these people.

The model (data system) is to be used by people in the work situation. As activities are of central importance in the work situation the description will concentrates on words denoting activities, i e on verbs.

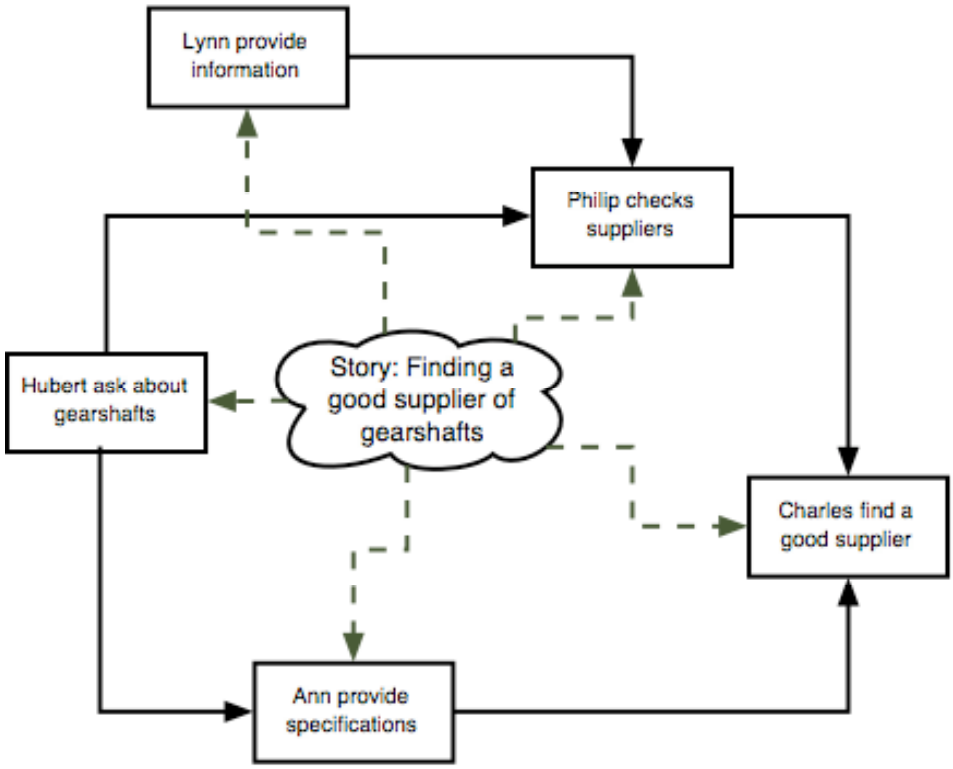
By asking these eight questions, I claim it is possible to obtain a rich description of a certain activity <verb>. No 1 identifies the starting conditions and no 8 identify what is happening when <verb> is completed. In fact, the method has been tested in practice and here is a translation of how a sales representative with the name Carl describes part of the selling process:

If the customer is satisfied with the offer of the Company, Carl makes an agreement with the customer, which means that the customer shall buy a certain quantity of a certain product to a certain price.

This is a description on the meta-level of a single fact, what is needed is a description of the total collection of facts. But if we put the descriptions of all facts involved in a certain process and order them according to 1) and 8), well we then have a description of the total process in form of a story! The style is somewhat mechanical, but still: It is a story. Reading it, makes the reader create an interpretation of what it means. Seen into a context of other such descriptions, we will create a context for interpreting the story in a correct way.

However, the story is not very readable! In order to make it readable, we have to rephrase the sentences, maybe put them in another order and maybe add some stylistic flavour. Ways of doing this can be fetched from literature science and specifically literature analysis. How this can be done, I do not know – yet, but I am pretty convinced that telling a story is the best description of a business process.

Telling stories is by the anthropologists recognised as one of the main processes for keeping the society together. In our days, stories are used for the same purpose in enterprises. Telling a story, being a part of a story and acting in a story is a deep human need. I have now indicated that this need can be used in order to create better business processes and understanding the supporting systems we have around those processes. In fact, we then have true human scale information system!



Figur 2. This is part of a process "buying gearshafts". The part is "Finding a good supplier". In doing so some people are involved in information gathering and information processing. The whole process is guided by a story called "Finding good suppliers of gearshafts" which everybody is acquainted with.

In Figur 2 I have sketched such a system, which is hold together by a common story, well know by all involved. It demands high professional skills, since the people involved must know not only stories about finding suppliers of gearshafts but also of steering wheels, wheel centres, differential houses, cup holders etc. It also put high demands on those who write the stories, since they must be understandable in a wide cultural context and besides being easy to remember. We see finally a market for those humanists, cultural workers and artists going around and having no work. In the new economy, they have their central positions!

5 References

Adams, D. (1979) *The hitchhiker's guide to the galaxy,* Harmony Books, New York.
Allen, E. and Fjermestad, J. (2001) *Logistics Information Management,* **14,** 14–23.
Alpher, D. (2001).
Berners-Lee, T., Hendler, J. and Lassila, O. (2001) http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809.
Berners-Lee, T., R. Fielding, & L. Masinter, (1998), Vol. 2004 The Internet Engineering Task Force.
Berners-Lee Tim (1998), Vol. 2003 W3C.
Churchman C W (1971) *Design of inquiring systems,* Basic Books, New York.
Devendorf, G. (2001) *Group Computing Magazine*.
Doan, A., Domingos, P. and Halevy, A. (2001) In *ACM SIGMOD 2001*.
Docherty P and A, D. (1992) *Lärande med förhinder,* Arbetsmiljöfonden, Nutek, Stockholm.
Embury, S. M., Brandt, S. M., Robinson, J. S., Sutherland, I., Bisby, F. A., Gray, W. A., Jones, A. C. and White, R. J. (2001) *Information Systems*, 657–689.
Fan, W., Lu, H., Madnick, S. E. and Cheung, D. (2001) *Information Systems*, 635–656.
Fensel, D. (2001) *Ontologies: a silver bullet for knowledge management and electronic commerce,* Springer, Berlin.
Fessl K (2003) 6th EU Framework.
Flensburg, P. (2002) In *IRIS 25*(Ed, M, K. P.) Bautahøj, Denmark.
Flensburg, P. (2003) In *People and Computers: Twenty-one Ways of Looking at Information Systems*, Vol. 26 (Eds, Järvi T and P, R.) Turku Centre for Computer Science, Turku, pp. 332.
Flensburg, P. and Milrad, M. (2003) In *26th Information Systems Research Seminar in Scandinavia*(Eds, Laukkanen, S. and Sarpola, S.) Haikko Manor, Finalnd, August 9-12 2003.
Griffith, R. L. (1982) *ACM Transactions on Database Systems,* **Vol. 7,** 417-442.
Gäre, K. (2003) *Tre perspektiv på förväntningar och förändringar i samband med införande av informationssystem,* Univ., Linköping.
Ivanov, K. (1972) *Quality-control of information : on the concept of accuracy of information in data-banks and in management information systems*, Stockholm,.
Keen, P. G. W. (2001) In *Information Technology and the Future Enterprise,*(Ed, Dickson G W, d. G.) Prentice-Hall Inc.
Kim, H. (2002) In *Communications Of The Acm*, Vol. 45.
Langefors B (1966) *Theoretical Analysis of Information systems, I & II,* Studentlitteratur, Lund.
McGuinness, D. L. and van Harmelen, F. (2004), Vol. 2004 W3C.
Metz, D. (2001) *DeveloperWorks, IBM*.
Miller Eric (2003) W3C.

Nurminen, M. I. (1988) *People or computers : three ways of looking at information systems,* Studentlitteratur ;
Chartwell-Bratt, Lund
Bromley.
Peckham, J. and Maryanski, F. (1988) *ACM Computing Surveys,* **20**.
Schoening J (2003) IEEE.
Sowa J F (2000) In *Conceptual Structures: Logical, Linguistic, and Computational Issuse*(Ed, Ganter B, M. G. W.) Springer-Verlag, Berlin, pp. 55-81.
Ströbel, M. (2001) In *WWW10*Hong Kong.
Subieta, K. (1985) *ACM Transactions on Database Systems,* **10,** 347-394.
Syu, I., Lang, S. D. and Deo, N. (1996) In *CIKM 96*Rockville MD USA.
van Hooft F and Stegwee R (2001) *Logistics Information Management,* **14,** 44-53.
White, T. and Hall, K. (2000) Giga Information Group.
Willaert, F. (2001) *Xml-Based Frameworks And Standards For B2b Ecommerce, PhD Thesis,* Departement Toegepaste Economische Wetenschappen, Katholieke Universiteit Leuven, Leuven.