Revaluation of a Large-scale Thesaurus for Multi-Media Indexing An Experience Report *

Dirk Deridder[1] and Peter Soetens[2]

[1] Vrije Universiteit Brussel (VUB), Programming Technology Lab, Pleinlaan 2, 1050 Brussels, Belgium
Dirk.Deridder@vub.ac.be
http://prog.vub.ac.be/

[2] Vlaamse Radio- en Televisieomroep (VRT), Auguste Reyerslaan 52, 1043 Brussels, Belgium
peter.soetens@vrt.be
http://www.vrt.be/

Abstract. In this paper we provide a preliminary overview of a number of problems we encountered when faced with the revaluation of a large-scale mono-lingual thesaurus. The thesaurus we speak of is used to wade through the vast multimedia archive of the Flemish public radio and television broadcaster (VRT). In order to support advanced and 'knowledgeable' queries on the archive, it became imperative to upgrade the existing infrastructure. In this context we performed an in-depth analysis of the existing legacy situation. This lead to the identification of a number of structural problems as well as problems with respect to content. Solutions to counter some of these have already been established. To support the new search-requirements for the archive, we have migrated the existing system to an ontology-inspired infrastructure.

1 Introduction

The work we present in this paper was performed in the context of the e-VRT MPEG project which consisted of a collaboration between VRT, VUB, and IMEC. The central theme was to investigate and develop the necessary technology to set up an enterprize-wide content management system in the context of a public radio and television broadcaster. In this paper we will focus on a number of experiences obtained in the work-package that concentrated on meta-data management for a multimedia archive. As input for this package we had access to an existing thesaurus that contained 229,860 lead terms. This collection of words continues to grow as we speak, since new entries are added on a regular basis by a team of highly skilled thesaurus administrators. It is a mono-lingual (Dutch) collection that is mainly used to wade through the vast multimedia

* This work was funded by the Flemish government (Belgium).

archive (partly digital) of VRT. To support the archives' search engine, an intermediate database system exists in which the lead term annotations of the archived items reside (over 500,000 digital archive documents with a growth of approximately 30,000 documents each year[3]). At this moment almost 90% of the archive searches are based on these lead term annotations. Besides the trivial use of the archive as a reference work for newscasts, it is also consulted for the purpose of creating game shows and documentaries for instance. It is clear that the multimedia archive as well as the thesaurus are considered as main assets for the broadcaster. Both are used on a daily basis and are under constant pressure to accommodate new demanding search operations. This has lead to usages of the thesaurus infrastructure in a way that was never (or could never have been) anticipated at the time of its conception (around 1986 [4]). It shouldn't be surprising that this resulted in a number of 'inconsistencies' in the data as well as a number of creative abuses of the existing tools. Hence they have a major interest in investigating new ways of organizing and managing the archives' meta-data by revaluating the existing thesaurus. To support advanced and more 'knowledgeable' queries we have opted for an ontology-inspired infrastructure. For this purpose we split up our activities in a content-wise and a structure-wise revaluation process.

In this paper we provide a preliminary overview of a number of problems we encountered when faced with the revaluation of such large-scale thesaurus legacy. In Section 2 we will zoom in on a number of results from our structural and statistical analysis of the thesaurus (content as well as infrastructure). Based on these results we will discuss the conversion and revaluation of the thesaurus into an ontology-based infrastructure in Section 3. To conclude we will also present a number of elements we believe to be important for our future work.

2 Analysis of the Existing Thesaurus

2.1 Numerical Analysis

The thesaurus we analyzed contained 229,860 lead terms (LT). Between these lead terms we counted 147,245 relationships, which boils down to a ratio of 0.64 relationships per LT. At first sight this seems very low, but it is an expected result considering the small set of relationship types available in the thesaurus (broader term - BT, narrower term - NT, use for - UF, use - USE, related term - RT, scope note - SN). In Table 2.1 we present the 30 LTs that have the highest relationships/LT ratio and hence can be considered as main focal points from a numerical perspective. Since cross-language semantical ambiguities are not relevant in the work we present here, we have taken the liberty to translate

[3] This figure excludes the non-digital documents over a 30 year period. Due to resource restrictions, the digitization of these documents is currently put on hold.

[4] Even though the software has evolved since then, the underlying principles have not (i.e. a rigid thesaurus representation with a limited number of relationships to capture lead term semantics).

the lead terms to English. Note that one of the main archive consumers is the newscast department. Consequently, the top level LTs in Table 2.1 clearly reflect their main topics of interest (i.e. countries, sports teams, airline companies, newspapers, ...)

#	Lead Term	Ratio	#	Lead Term	Ratio
1	FRANCE	1.018	16	RIVER	278
2	USA	748	17	PLANT	277
3	BELGIUM	654	18	RUSSIAN FEDERATION	251
4	GERMAN SITES	592	19	REWARD	246
5	AUTHOR	519	20	NEWSPAPER	245
6	GREAT BRITAIN	488	21	SWITZERLAND	243
7	THE NETHERLANDS	477	22	COMPUTER COMPANY	216
8	ITALY	444	23	AUSTRIA	211
9	TELEVISION PROGRAMME	428	24	AIRLINE COMPANY	208
10	JOURNAL	368	25	DISEASE	189
11	THEATER TROUPE	364	26	BRUSSELS	185
12	SPAIN	325	27	TELEVISION CHANNEL	180
13	MUSEUM	311	28	CHOIR	176
14	BANK (FINANCIAL INST.)	299	29	TURKEY	170
15	SPORTS TEAM	295	30	CYCLE RACING	168

Table 1. Top 30 lead terms based on the number of relationship references

As is to be expected, most of these terms represent high-level concepts and can thus be used to partition the LT-space. Examples of these are *author*, *television programme*, *journal*, *river*, *sports team*, ... Nevertheless the list also contains LTs that easily lend themselves to be grouped in new higher-level concepts. This is especially true for the case of *france*, *usa*, *belgium*, *great-britain*, ... Even though the thesaurus contains a broader term relationship for these LTs to *europe* and to *north-america* respectively, there is no top-level concept that also connects the latter two (e.g. *continent*, *geographic region*).

Delving into the relationships we found that 91,686 (62.27%) of the instances were dedicated to sustaining the hyponymy/hypernymy taxonomy (NT-31.07% / BT-31.20%). This was an expected result since the BT/NT couple is semantically the richest and most useful relationship available. The slight mismatch in percentages was a result of several data-entry errors (e.g. a space added to the countering LT). The synonymy/homonymy pair accounted for 49,018 (33.29%) instances (USE-16.64% / UF-16.65%) which was also an expected result. The most general relationship type available is the RT-relation. We counted 6,234 (4.23%) of the relationship instances that were of this semantically broad type. This low figure seems to indicate that very few relationship occurrences didn't fit into the other categories. As we will explain in the following section this isn't the case since it is mainly a result of several abuses of the semantics of the other relationships. To clarify the intended meaning of LTs one can make use of the

SN-relationship. In this case only 307 (0.21%) of the relationship instances were allocated to this purpose. This is regrettable as scope notes are highly useful for determining the intended sense of a lead term. We have summarized these findings in Figure 1.

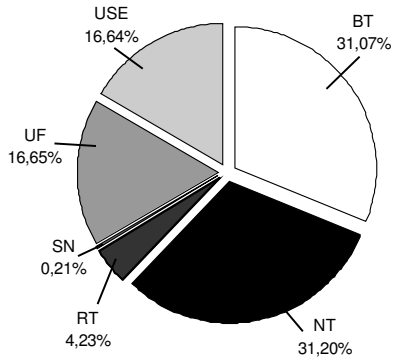

Fig. 1. Distribution of the different relationship types

2.2 Identification of Shortcomings

The shortcomings we present in this section are mainly a result of the evolution of the search-requirements since the conception of the thesaurus infrastructure. As we already mentioned, the application is under constant pressure from the programme makers to accommodate more directed and 'knowledgeable' queries on the archive. This has resulted in a situation where the thesaurus as a medium (not the content!) has become inadequate. Nevertheless to keep supporting the daily operation of the programme makers, it was necessary to creatively bend the rules in anticipation of a newer version. This lead to a number of conceptual lapses and inconsistencies which we will illustrate in the following subsections. For the sake of this discussion we have grouped them into three categories : lead term, taxonomy, and semantical problems.

Lead Term Problems With regard to the lead terms we principally identified two major problems. First of all since a thesaurus follows a term-based philosophy (in contrast to a concept-based philosophy), ambiguities may arise when confronted with homonymous terms. In the thesaurus they have countered this problem by adding a context annotation between braces. An example of such an LT is *casino (film title)*, *casino (superstore)*, *casino (cycling team)*, and *casino (gambling house)*. In our case we counted 2,297 LTs that have such a suffix (approximately 1 % of the LT collection). The major problem with such suffixes is that, as reported in [5], they also can become ambiguous. What would happen if

another film is made with the title *casino* for instance? This is why we have chosen to follow a concept-based approach in the new infrastructure (each concept receives a unique ID to which multiple labels can be attached).

Secondly, a problem poses itself since given names are recorded as 'first-order' lead terms. Unfortunately these are not grouped under a higher-level LT which makes it impossible to distinguish real LTs from these given names. For example there is no broader term connection between *deridder* and *person*[5]. Moreover there is a conceptual misuse of the homonymy/synonymy relationship to capture possible type errors by users of the thesaurus. This manifests itself for example in a 'use for' relation between the given name *d'hooge* and *d'hooghe*. This makes it impossible to distinguish the real LT from the 'fake' LT afterwards.

Taxonomy Problems One of the strengths of a thesaurus is the use of 'broader term' and 'narrower term' relationships [6] to establish a rich taxonomy. This enables the end-user to navigate through the archive by following the different layers of abstraction. A major problem we encountered was the frequent unnatural use of both relationships.

It is clear that an *investor* is a 'broader term' of a *stockholder*. But when we encountered *beenhouwersstraat* and *vrije universiteit brussel* amongst others as 'narrower terms' of *brussels* we couldn't help but frown. This is clearly a deterioration of the knowledge since you lose the information that the first LT is a street and the second LT is a university (both located in Brussels). This is of course a result of the collection of semantically poor relationships in a thesaurus. If it were possible to connect these terms with 'better suited' relationships, this would enable more advanced queries on the archive (e.g. I'm looking for an archive item about a university in Belgium).

We also observed the use of BT/NT relations to indicate an 'instance-of' relationship. Consider for example the NT relation between *author* and *dostojevski fjodor*. This clearly isn't a narrower term but if no other mechanism is available it is an inevitable solution. In relation to this instance-of issue we would like to raise another issue. As we will discuss in Section 4 our work will be used in a general content management system to support the broadcasters daily operation. In this case there will exist different viewpoints on the same set of meta-data. What will be experienced by one user as an instance/class, is not necessarily true for another end-user. In anticipation of the future system we have decided to follow a prototype-based approach to represent the concepts in the new infrastructure. Related problems have already been reported by [11][12].

[5] In some exceptional cases we did find a reference to the fact that the LT was meant to be a given name (in some cases a scope note or context suffix between braces was found). Unfortunately there was no consistent system that would allow us to automatically extract the LTs.

[6] We define a 'narrower term' as a term that has a narrower meaning than the reference term. For example *mosque* is a 'narrower term' of *religious building*. The inverse, a 'broader term', is defined similarly.

Semantical Problems Searching the digital archive is mainly done by composing a set of lead terms that (1) clearly defines the archived items you are interested in, and (2) reduces the size of the search result. Composing a good set of lead terms is done by browsing the thesaurus. Therefore it is crucial to be able to deduce the semantics of the chosen lead terms. 'Scope notes' are particularly useful for this purpose since they could for instance contain a dictionary entry that explains the intended meaning. Unfortunately as we indicated in an earlier subsection they are sparsely used. Another way to discover the meaning of an LT is to look at its relative position to other LTs (i.e. follow the relationships). The most useful for this purpose is the NT/BT couple. But as we already stated, the arbitrary use of this couple could often lead to semantical confusion. The lead term suffixes are in some cases very useful, but they are not consistently used. The entry *antwerp (city)* is a good illustration since most other names of cities are lacking this suffix, and hence obscure this important information. Following the 'use' and 'use for' relations is also an interesting path to follow. It relates the current lead term to synonymous terms, but as we saw earlier it is also abused for other purposes.

Generally speaking, if one wants to deduce the intended meaning of an LT one often falls back on an ad hoc combination of the above. It is only fair to say that this is mainly a result of the shortcomings of the existing thesaurus infrastructure (for this kind of application). Many of the problems we have reported are often the result of wanting to support queries that could only be answered by a detailed knowledge-base. So were is the barrier? Where does the thesaurus end and does the knowledge-base / expert-system / ... begin? It would be nice (to support these so-called 'knowledgeable' queries) for instance to be able to search for a multimedia item about "persons that are in their mid-thirties and that have authored at least two children's books". In this case we would have to be able to attach several attributes to the lead terms. And if we look into the thesaurus we sometimes find a 'hack' to be able to record this kind of information. It is clear that a new kind of infrastructure is needed, which we will discuss in the following section.

3 Towards an Ontology-based Infrastructure

In this section we will briefly sketch the basic infrastructure we created as a proof of concept as well as a number of conceptual insights. The experimental prototype we created consists of an ontology-environment in Java that uses an RDBMS (PostgreSQL) to store the concepts. It contains an importer which makes it possible to convert the data dumps of the existing thesaurus. We have also included an OWL-exporter [10] for compatibility with other tools. A primitive graphical representation of the concept networks was also implemented. For the ontology we have chosen to follow a prototype based view. In practice this means that all concepts are instances. To create a new concept you can either clone an existing one or build one from scratch. It should be mentioned that we never had the intention to build a full-fledged ontology tool. We merely tried to

explore the possibilities of an ontology-based infrastructure.

To get a better idea of what was needed we initially performed a number of small test-runs with existing ontology tools such as Protégé2000 [6], SoFaCB [2], ... A major issue we were confronted with was scalability. Loading the 229,860 lead terms ('naively' promoted to concepts) immediately made clear that some of these tools were not up to the task (at that time [7]). This was not only an issue of the (in memory) data storage but also of the user interface. Take for example the standard tree-view component for showing concepts : it simply didn't scale up. A solution for this was to introduce an intelligent caching mechanism that loaded a limited range of concepts on a need-for basis. Also when zooming in on a highly connected concept (e.g. *belgium*) it became clear that visualizing (e.g. in DHTML) all connections on-the-fly is infeasible. Summarizing the number of connections (e.g. "this concept has n NTs and m RTs") and caching the visual representation on disk appeared to be a workable approach.

In our context an ontology-based application could be used in two different ways : as a *controlled vocabulary*, and as a *provider of concept templates*. The first corresponds to the use of the existing thesaurus for annotating the multimedia items in the archive. To enable this it is of vital importance to be able to upgrade the existing thesaurus content to an ontology. After all it is unfeasible to perform a manual re-indexation of the archived items. A meticulous restructuring and cleansing of the lead term collection imposes itself. A very promising approach to support the thesaurus administrators in this process is [7] [3]. However we still have to evaluate the actual use on such a huge collection. The existence of a controlled vocabulary that goes beyond the enterprize-boundaries of the broadcaster is already envisioned. This should enable the external content providers (e.g. independent production houses) to accompany their products with a set of annotations in this "unified vocabulary. This would greatly reduce the chore of manual in-house annotation, and would certainly improve the quality of the meta-descriptions. In order to succeed in setting up such an ontological commitment between the different media partners, it is crucial to create a good set of upper ontologies (cf [4]). As we already mentioned in an earlier section, we believe that an in-depth analysis of the existing thesaurus could identify possible candidate candidate concepts for this level. For this purpose we are currently including a number of analysis functionalities in the experimental environment (e.g. on-the-fly generation of Table 2.1, extraction of LT suffixes, ...). This should help the thesaurus administrators to make sound choices to compose the set of core concepts.

The second use of an ontology as a concept-template provider is a separate issue. To accommodate the richer set of archive queries one also has to provide a richer set of knowledge in the annotations. For this purpose it is necessary to

[7] Since the start of this project, a number of ontology tools have been greatly improved. Based on our own recent findings and the evaluation results reported in [1] we currently consider using the KAON tool suite [8] for future experiments.

be able to 'instantiate' concepts in the ontology. In our approach this is done by cloning existing prototypical concepts (i.e. these form the ontology for the 'instantiated' concepts) . An example of such a concept could be *person* in which we foresee slots for *name*, *date-of-birth*, ... This is very similar to the traditional class/instance creation in knowledge-bases. It is clear that this will certainly improve the support for advanced queries. However it remains unclear how far the broadcaster should go in setting up such a broad encyclopedic knowledge-base. Also in this case the need for a shared ontology between the different media partners arises. The knowledge-base would consequently contain pointers to the relevant 'out-house' knowledge bases.

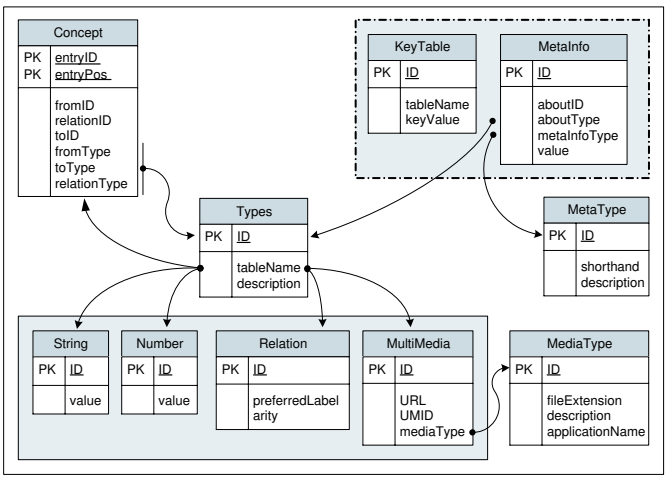

Fig. 2. Conceptual schema of database

The database we use to store the concepts in the ontology is based on the schema presented in Figure 2. Central in this schema is the table *Concept* which is used to establish the concept network. A concept could be related to another concept (e.g. a BT relationship) or to a terminal type such as *String*, *Number*, ... To distinguish between these we have included a foreign key (*fromType*, *toType*, *relationType*) that points to the corresponding type-table. This makes it possible to support new types without changing the existing conceptnetwork. An example conceptnetwork is shown in Figure 3. The *MetaInfo* table is used to record all kinds of meta-information such as the date an entry was created/updated/.... We have found that this scheme is quite robust to change since it is very generic. The downside to this is that a lot of the semantic behavior has to be captured outside the database (such as the interpretation of a BT/NT relationship).

As we already stated we initially converted the thesaurus in a naive way. This resulted in three *Concept* table entries per LT, which boils down to ap-

proximately 700,000 tuples. Adding new relationships between LTs will surely augment this number.

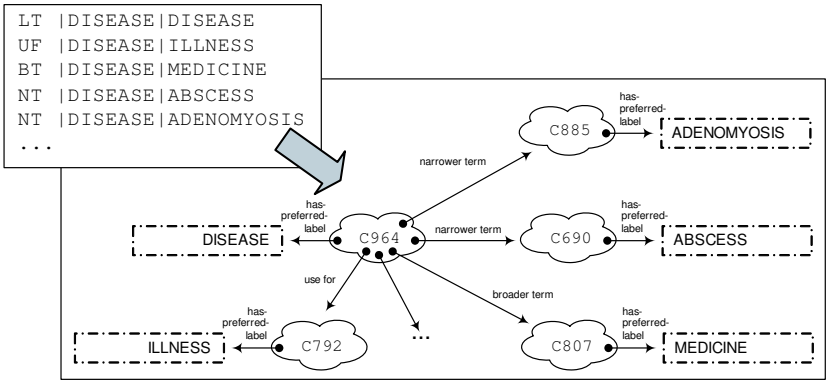

Fig. 3. An example conceptnetwork for *disease*. A cloud represents a concept, the arrows indicate the relationships

4 Future Work

Search Heuristics We have already experimented with the use of search heuristics on the ontology. In the particular experiment we kept track of the frequency that certain lead terms were used. This was consequently used to compose 'virtual' groupings of terms in the ontology. These virtual groupings were initially implemented in an extensional way by enumerating the different LTs that belong to it. This kind of grouping is mainly interesting to record static information (e.g. a thematic grouping). In the future we want to explore intentional groupings which provide a way to dynamically group LTs.

Temporal Information Currently the thesaurus doesn't contain references to temporal information. The main reason for this is that there was no 'clean' way to include it in the existing infrastructure. It is however seen as highly relevant for the archive queries. An example of this is information about the period when a certain person was president of a country. The availability of a mechanism to store temporal information would also make it possible to get rid of situations with unclean LTs. We find for instance the LTs *germany1* and *germany2* in the existing thesaurus to indicate germany before and after the unification of east and west. As a consequence all links that are relevant in both situations are recorded twice (in some cases this isn't the case and hence entering the search path from one or the other could lead to different results!).

Enterprize-wide Content Management System As we mentioned in the introduction, the broader context of this work is the installation of a general content management system to support the activities of a radio and television broadcaster. This is strongly related with Enterprize Application Integration. The major difference is that a lot of work involved in the production of radio and television has a creative (artistic) nature. Consequently each team has its own approach and set of tools. It is clear that it is out of the question to enforce a unified view / tool suite on these production activities (this would restrain their creativity!). Moreover such a de facto central alignment is unmanageable in this context as a result of the scale and scope of this business's nature. Nevertheless from a management perspective it is crucial to have an integrated and controllable system of data and processes. We believe we can find a compromise in the semantic web vision to meet these contradicting requirements. Instead of a tightly coupled system (e.g. ERP systems) this would enable setting up a loosely coupled 'intranet-alike' system (cf the internet). For our work package this would mean that a much richer set of meta-data could already be captured at production-time instead of at archival-time (e.g. GPS data indicating the recording location, lenses used, participants, scenario, ...)

5 Conclusion

In this paper we have reported on a number of experiences when confronted with the revaluation of a large-scale thesaurus. The revaluation of this thesaurus was mainly driven by an urgent need to support more advanced and 'knowledgeable' queries on a vast multimedia archive in the context of a radio and television broadcaster.

During our analysis of the existing infrastructure we found that there were several conceptual lapses and inconsistencies in this thesaurus. These were mainly the result of its inadequacy as a medium to support the daily operation of the programme makers. The problems we identified were related to lead terms (suffixes to resolve ambiguity, given names as first-order lead terms), to the taxonomy (misuse of broader term / narrower term), and to the semantics of the lead terms (intended meaning of a lead term). This identification has resulted in a number of guidelines which will be used to support the restructuring work of the thesaurus administrators.

To experiment with the possibilities of a next generation infrastructure we have approached the thesaurus from an ontology perspective. For this purpose we have built an experimental environment into which we converted the thesaurus to a prototype based concept network. Even though preliminary, this has resulted in insights with respect to scalability issues (in memory storage, user interface) and more conceptual themes. With respect to the latter, it became clear that in our context, an ontology-based application could be used as a controlled vocabulary as well as a provider of concept templates. Moreover exploring the use of the conceptnetwork schema has lead to a deeper understanding of issues related to

genericity and meta-circularity (e.g. defining the semantics of the relationships inside the database itself)

Specifically in the context of an enterprize-wide content management system these insights will certainly influence our future directions.

References

1. Angele, J. and Sure, Y. (eds): EON 2002 - Evaluation of Ontology-based Tools. In: EKAW 2002 Workshop Proceedings (2002)
2. Deridder, D. : A Concept-Oriented Approach to Support Software Maintenance and Reuse Activities. In: Proceedings of the 5th Joint Conference on Knowledge-based Software Engineering. IOS Press (2002)
3. Gangemi, A. and Guarino, N. and Oltramari, A. and Borgo, S.: Cleaning-up WordNet's top-level. In: Proc. of the 1st International WordNet Conference (2002)
4. Gangemi, A. and Guarino, N. and Masolo, C. and Oltramari, A. and Schneider, L.: Sweetening Ontologies with DOLCE. In:Proceedings of the International Conference on Knowledge Engineering and Knowledge Management. AAAI (2002)
5. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers (1994)
6. Grosso, W.E. and Eriksson, H. and Fergerson, R.W. and Gennari J.H. and Tu, S.W. and Musen, M.A. :Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000). SMI Report SMI-1999-0801 (1999)
7. Guarino, N. and Welty, C. A. : Evaluating ontological decisions with OntoClean. In: Communications of the ACM 45-2(61-65) (2002)
8. KAON - The Karlsruhe Ontology and Semantic Web Framework. FZI (Research Center for Information Technologies) and AIFB (Institute of Applied Informatics and Formal Description Methods) - University of Karlsruhe. http://kaon.semanticweb.org/
9. Maedche, A. and Motik, B. and Stojanovic, L. and Studer, R. and Volz R.:Ontologies for Enterprise Knowledge Management. In:IEEE Intelligent Systems. 1094-7167/03. IEEE Computer Society (2003)
10. Patel-Schneider, P. F. and Hayes, P. and Horrocks, I. : OWL Web Ontology Language Semantics and Abstract Syntax. W3C Working Draft 31 March 2003. http://www.w3.org/TR/owl-semantics/
11. Welty, C.A.: Towards an Epistemology for Software Representations. In: Proceedings of the 10th Knowledge-Based Software Engineering Conference. IEEE Computer Society Press (1995)
12. Welty, C.A. and Ferruci, D.A.: Classes in Software Engineering. In: Intelligence. Summer (1999) 24–28