Towards a Better Understanding of the Language Content in the Semantic Web

Pavlin Dobrev[1], Albena Strupchanska[2], Galia Angelova[2]

1 ProSyst Bulgaria Ltd., 48 Vladaiska Str., Sofia, Bulgaria, pavlin@prosyst.com
[2] Bulgarian Academy of Sciences, Institute for Parallel Information Processing, 25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, {albena, galia}@lml.bas.bg

Abstract. Internet content today is about 80% text-based. No matter static or dynamic, the information is encoded and presented as multilingual, unstructured natural language text pages. As the Semantic Web aims at turning Internet into a machine-understandable resource, it becomes important to consider the natural language content and to assess the feasibility and the innovation of the semantic-based approaches related to unstructured texts. This paper reports about work in progress, an experiment in semantic based annotation and explores scenarios for application of Semantic Web techniques to the textual pages in Internet.

1 Introduction and State of the Art

The ultimate aim of the Semantic Web is to make the web resources more meaningful to computers by augmenting the presentation markup with semantic markup, i.e. meta-data annotations that describe the content. It is widely expected that the innovation will be provided by agents and applications dealing with ontology acquisition, merging and alignment, annotation of www-pages towards the underlying ontologies as well as intelligent, semantic-based text search and intuitive visualisation. However, the current progress in all these directions is not very encouraging despite the number of running activities. Isolated results and tools are available for e.g. automatic and semi-automatic annotation of web pages, for knowledge-based information retrieval, for ontology learning and so on but it is still difficult to grasp a coherent picture of how the Semantic Web will drastically change the information age by offering quite new kinds of services. Another discouraging obstacle is that the ontologies in the Semantic Web are not clearly seen at the horizon. Due to these reasons, it makes sense to develop and evaluate experimental settings which might provide on-hand experience with semantic based systems and show the desired benefits as well as the potential gaps in the current research efforts.

A summary of the relevant state of the art should sketch the status in several core Semantic Web directions. One of them is **ontology availability, development and evolution** which is considered as the first challenge for the Semantic Web [1]. Isolated ontologies are available in a number of fields and initiatives like the "standard upper ontology" aim at the design and building of certain acceptable

domain-independent ontologies intended to be reused and extended for particular domain. There is a number of public taxonomies (even very large, in the medical domain for instance) but the most elaborated semantically-based resources like CyC or LinkBase® are not publicly accessible. In fact, the availability of underlying ontologies is an issue at the core of the Semantic Web enterprise. Many tools for building ontologies and reusing existing ones have been developed. Environments like Protégé [2] or Chimaera [3] offer sophisticated support for ontology engineering and merging. Protégé is an integrated software tool used by system developers and domain experts to develop knowledge-based systems. It offers an editing environment with several third party plug-ins. Protégé allows to export the created ontology in a number of formats and has also a flexible plug-in structure that provides modular extension of the functionalities. OntoEdit [4] is another sophisticated ontology editor that supports methodology-based ontology construction and that takes comprehensive advantage of Ontobroker inference capabilities.

Because of the numerous isolated conceptual resources, **ontology mapping** is also a hot area of research. Today matching between ontologies and schema is still largely done manually in a labor-intensive and error-prone process. As a consequence, semantic integration issues have now become a serious bottleneck in a wide variety of applications. The high costs of overcoming this bottleneck have triggered numerous research activities on how to describe mappings, manipulate them, and generate them semi-automatically but unfortunately there has been little cross-fertilization between the results. There are, most generally, two approaches to ontology mappings: based on heuristics that identify structural and naming similarities between models [5] and using machine learning techniques in order to learn mappings between models/ontologies [6]. In both cases, the systems require feedback from the users to refine the proposed mappings. US activities in **ontology alignment and merging** [7] face the fact that different domain-dependent ontologies are organized along different principles for acquisition of concepts and their natural language labels; choosing 42 sample terms from EIA Glossary of Energy terms, it turned out that only 22 of them matched one or more SIMS domain model terms fully or partially, and no single algorithm for automatic label alignment was found [8].

Another important issue is the **(automatic or semi-automatic) annotation of pages** in the Semantic Web. Annotation is considered as one of the most effort-consuming tasks that has to be performed by especially developed tools. Recent annotation tools support manual and non ambiguous pages annotation according to predefined typology of events, locations, names, and non ambiguous domain terms (see for instance [9]). Selecting manually text fragments as concept labels is not very efficient, moreover different annotators may tend to build the semantic indices in their own way and will need precisely formulated standardization rules how to anchor pieces of terms to the ontological concepts. Please note that many technical terms are made up of multiple words (e.g. *delta doped resonant tunneling diodes*) and the only way to accurately distinguish them is the analysis of qualified noun phrases rather than individual words, which is problematic for automatic recognition and processing by language technologies and even for human beings. So at present it remains unclear how the different annotation practices could be standardized to produce pages annotated in a similar manner.

Another open question is **how many semantic anchors are appropriate per page**: should all possible semantic indices be inserted (even against different ontologies) or only the domain-specific concepts should be anchored to one chosen ontology? . The realistic scenarios are well-defined domain-specific tasks in compact areas - e.g SPIRIT [10] uses a multilingual ontology for mapping geographical concepts and target only annotation of named entities and geographic objects.

At last, the present **language technologies** still did not prove their feasibility for automatic collection of the necessary Semantic Web data; as a result, the majority of running projects in the field still rely on manually-acquired demonstration ontologies in narrow domains. There is still no evidence that **knowledge-intensive information retrieval** provides much better results [11]. To conclude, in our view the development of the Semantic Web is at certain initial phase of data collection, elaboration of design scenarios, development of prototypical applications that may become success stories, and their evaluation and market assessment.

Concerning the **visualisation** in Semantic Web, Kimani et al. [12] classifies existing approaches, where our approach would be sorted as generated rendering with direct manipulation interaction style. Our experimental tool provides visualisation and navigation support. Magpie [13] shows an approach for semantic-based assistance in user's web page navigation. It is a framework supporting the interpretation of web pages that is integrated in the standard web browser. The user chooses particular domain ontology from which ontology-based lexicon is automatically extracted. Then the ontology dependent entities in the web pages are annotated automatically using the *ontology-based lexicon* approach. The user chooses the classes of entities to be highlighted in the web page s/he browses. By right-clicking on a highlighted entity a context dependent menu is shown. The choices in the menu depend on the class of the selected entity within the selected ontology. Magpie services act as an auxiliary knowledge resource, which is at the disposal of users.

In this paper we report about on-going research work addressing the text of the www-pages in the future Semantic Web. We discuss in Section 2 existing semantic structures and annotation according to them, which is called *indexing* in the information science. In fact, while indexing, the present terminological collections - domain nomenclatures, classifications, and ontologies - are applied as semantic backbones in (multilingual) text archives. To know more about automatic annotation is of crucial importance for the appearance of the next generation technologies, since their proper semantic resources are not available yet with the necessary volume, content and format and it still remains an open question when and how they will arise. In Section 3 we present an experiment in manual semantic annotation of real www-pages against a financial ontology in English that is considered as a semantic backbone when building the semantic index. The experiment allowed us to get an idea about the potential difficulties in the annotation process, which include on-the-fly resolution of several kinds of ambiguities like manual choice of the anchored phrases from the natural language pages, choice of ontology node and so on. These ambiguities are embedded elements of the Semantic Web due to the fact that the complexity of the whole enterprise implies multiple choices in many tasks and there will be no (most probably there cannot be) standard solutions for all users. We present as well an experimental visualisation tool, which we develop at present to assist the semantic-based search of annotated web pages. The tool is implemented at concept

demonstration level and allows for browsing of semantically-related pages from the perspective of a single underlying ontology. Section 4 discusses applications that may benefit from the availability of semantic-based search and its visualisation. Section 5 contains the conclusion.

2 Annotation of Pages and Conceptual Resources

As we said, the task of linking a free text fragment (a word or a phrase) to a term from certain terminological collection is well known in the computational linguistics and the information science. This is the so-called *indexing*. The most successful indexing applications deal with the medical domain where indexing according to medical classifications is systematically performed in practical settings by software tools and thousands of health professionals who at least have to edit the indexed text. The annotation traditions and standards for specialised texts (patient records) are established since decades, moreover some very large terminological collections are especially designed to spin the indexing process – for instance SNOMED, which among others helps calculating the price of the medical treatments. The more advanced the system is, the more automatic the indexing procedures are, providing high precision of terms recognition. The world leader in indexing software is the company Language and Computing (Belgium) which develops the medical ontology LinkBase® and indexing software against different medical classifications [14]. Annotation with 100% precision is impossible [15]; however, the benefit is that the same document becomes accessible via the synonyms of the annotated terms [14]. So we believe that the annotation task as defined in the Semantic Web could use the indexing experience gathered in medical informatics. There should be domain-dependent conventions how to recognise the ontology terms in the free text to be indexed; please note that the concept and relation labels can be verbalised in a free text in a variety of ways. Another lesson learnt from the medical indexing, which we apply in our current experiment, is to annotate in the text all occurrences of all ontology labels, which we are able to recognise (or the software we develop is able to recognise). However, annotation is not elementary from semantic point of view, even against medical terms that are relatively well-defined collections. Let us illustrate some key semantic problems, which are not frequently discussed.

Consider Figure 1 and the free text uploaded in the leftmost scrolling window. It contains the two subsequent words *investment portfolio*. In the ontology uploaded in the second scrolling window, there are two separate concepts with labels INVESTMENT and PORTFOLIO (they are not visible in Figure 1). Only domain expert who reads and interprets correctly the text will be able to decide how to define the semantic indices from the phrase *investment portfolio* to the concepts investment and portfolio. As natural language is rather vague and the individual reader interprets the meaning, another domain expert could link the phrase and the two concepts differently. In other words, even the manual annotation is a task to be performed with certain (small) percentage of disagreement by highly specialised automatic annotation. Due to this reason, running projects in this field target automatic annotation of some kind of entities only, e.g. named entities.

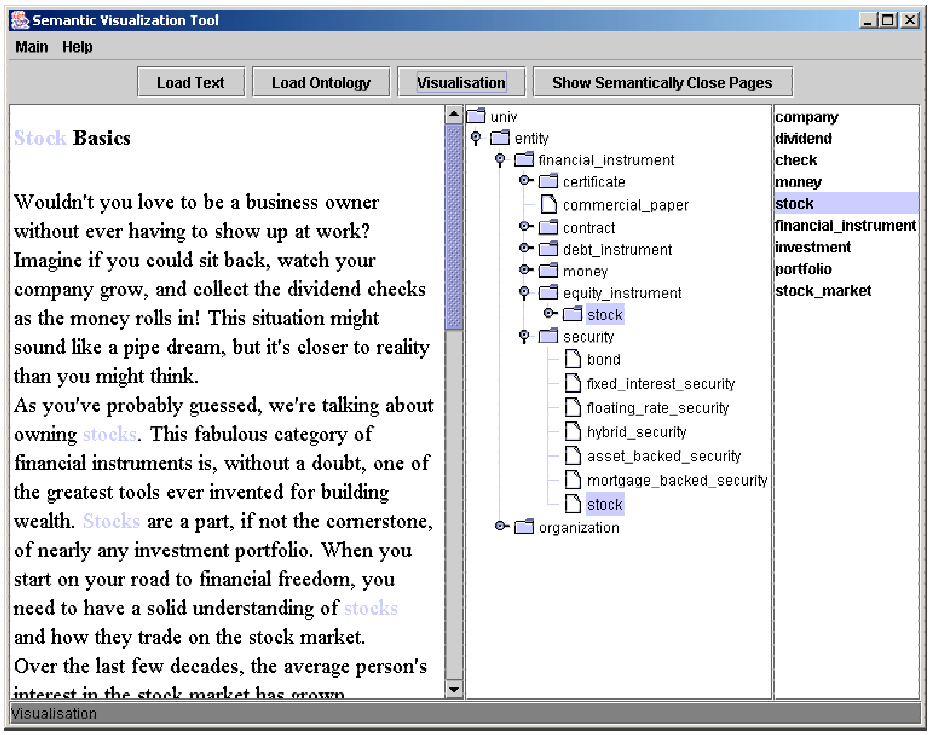

Fig. 1. Visualisation of semantically annotated text and its semantic indices

Regarding automation of annotation in the Semantic Web, a lot of (even commercial) annotation tools have been developed to support manual annotation in the relevant context. Annotea [16] allows annotations in RDF and provides mechanism for publishing annotations on the web. It uses XPointer for locating the annotations in the annotated document and a bookmark schema, which describes the bookmark and topic metadata. Other annotation tools include an ontology browser for the exploration of the ontology and instances and a HTML browser that will display the annotated parts of the text. OntoMat-Annotizer [17] allows the annotator to highlight relevant parts of the web page and create new instances via drag'n'drop interactions. It supports the user with the task of creating and maintaining ontology-based DAML+OIL markups. OntoMat is based on the S-CREAM framework [18], which comprises inference services, crawler, document management system, ontology guidance and document viewers. OntMat supports semi-automatic annotation of web pages based on the information extraction component Amilcare [19]. It provides a plug-in interface for extensions. Another annotation tool that integrates Amilcare is MnM [20]. It can handle multiple ontologies at the same time. MnM makes it possible to access ontology servers through APIs, such as OKBC, and also to access ontologies specified in a markup format, such as RDF and DAML+OIL. The extraction of knowledge structures from web pages is through the use of simple user-defined knowledge extraction patterns.

Regarding the available conceptual resources, the overview of existing public constructions displays primarily the multiple perspectives to classification of reality objects, relations and their attributes and features. The reader may consider as an example three taxonomies in computer science: *(i)* the ACM taxonomy of keywords, designed by the IEEE Computer Society [21], *(ii)* a taxonomy in Computer Graphics Interfaces, developed within a NSF-funded project as a searchable database index for a large part of the ASEC Digital Library [22] and *(iii)* the SeSDL Taxonomy of Educational Technology, which gives links to the British Educational Thesaurus BET [23]. These three hierarchies are built for different purposes and a careful study of their concepts shows that less of 5% of the nodes are common. As a rule, even for these "intersecting" concepts the classification into subconcepts is different. Similar variety exists concerning the emerging ontologies.

We considered in depth the public financial "branch" of the MILO ontology [24] which contains all concepts mentioned at least three times in the Brown corpus and additionally, is relatively easy for people to understand and practical for computational use. MILO is integrated under SUMO [25] and its financial part covers 246 basic entities (terms). In Teknowledge's ontologies, terms belong to one of five basic types: class, relation, function, attribute, or individual. Terms are the vocabulary of the ontology and MILO adopted some naming conventions to differentiate types of terms at sight, which were useful in the annotation process. We mapped MILO to a financial ontology we developed earlier for educational purposes in the LARFLAST project [26]. Three quarters of the terms coincide, however the design decisions are different. For instance, in MILO *dividend* is a relation while for us it is a concept. It looks difficult even for knowledge engineers to formulate precise rules how to align the meaning of *concepts* to *relations*. Mapping of such ontologies will require complex semantic resources, which define better the meanings, and in-depth inference. Most probably, as it happens today with the indexing software, it will be impossible to achieve 100% correctness of mappings and alignment.

3 The Experiment

We annotated manually the text of 132 real www-pages using the LARFLAST ontology that has 280 concepts and relations. While selecting the pages we noticed the following. There are millions of Web pages with financial information in Internet but the majority of them refer to few ontology terms only. This is due to the well-known fact that in general, text meaning *IS NOT* verbalisation of domain knowledge. The pages we found (via Google and searching for keywords) discuss for instance banking and stock exchange information, deals, company descriptions; thousands (newspaper) articles concern financial matters in general and therefore, words like *stock* and *bond* appear here and there in the texts. On the other hand, the ontology is a compact formal encoding of domain knowledge. Only textbooks, manuals, surveys contain very frequently the ontology terms as - by default- these kinds of texts are descriptions of domain knowledge. In other words, the annotation of the static web content at present would face the serious problem that the stories presented in web pages refer to many domain ontologies by addressing only few concepts from several

ontologies. For instance, a document discussing *stocks and bonds for investments in computer and aircraft industry* could concern three domain ontologies. In this way the decision of how to annotate (against which ontology) is rather difficult and we restricted our experiment to strictly specialised texts that belong to the financial domain although these kinds of texts are relatively rare in Internet.

We implemented the ViSem tool supporting the visualisation of semantic indices and providing some kinds of semantic-based text search. ViSem works over an archive of html-pages annotated according to an ontology encoded in DAML+OIL/OWL-format. Figure 1 shows the tool interface in "Visualisation" mode. Texts are loaded by the button "Load text" and appear in the leftmost text window. The ontology is open by the "Load ontology" button, in the second scrolling window. After loading a www-page and its ontology, ViSem analyses the semantic index and displays the list in the rightmost side of its main window: in the order of appearance, the text phrases from the leftmost window which are semantically anchored to terms from the chosen ontology (please note the list can be empty). In "Visualisation" mode, after selecting an item from the list in column 3, ViSem shows (colours) in the leftmost text all words and phrases linked to the highlighted term. The same colour is used for colouring the ontology label too. Originally our idea was to link the text and ontology term by a line, as we believe the visualisation by straight line will be the best intuitive illustration. However, supporting a second presentation layer over the text layer would require too much implementation efforts worth to be invested in a bigger project only. Please note that the visualisation of links by colouring is the most often approach at present and almost all annotation tools work in this way.

ViSem is implemented in Java using Swing graphical library. In ViSem we reuse ideas and code from our tool CGWorld which supports the graphical acquisition of conceptual graphs in Internet [27]. Ontology processing is built on top of Jena's API. Jena is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL [28].

4 Application Scenarios

If the Semantic Web makes the web resources more meaningful to computers, then what is the benefit for human users, given the fact that humans develop ontologies and annotate (millions of) web-pages? In the context of our small experiment we try to answer this questions by exploring different application scenarios where the semantically indexed archive and its ontology improve the information services.

The first useful application is the semantic-based text search. It is illustrated in Figure 2. The user selects a concept from the ontology and clicks the button **Show semantically close pages**. Then the most relevant page, which is semantically linked to the highlighted concept, is loaded in the leftmost window. Belonging to the same domain and annotated against the same ontology, this page obligatory contains many terms from the ontology. Please note that in this case, the present information retrieval techniques will not distinguish between the pages shown in Figure 1 and Figure 2. They are 100% similar, as they practically contain the same words, and ranking them in a list would depend on the algorithms of the searching engine. So the only way of

finer ranking is to have semantic-based indices as they are suggested today in the Semantc Web.

The second application scenario we envisage is support of comprehension while reading www-pages which could be useful for domain novices or for foreigners who read a page in unknown language and do not grasp the meaning of all specific terms. ViSem supports at the moment the following services: *(i)* after highlighting a phrase in the left-most window, via the right button menu, the user can view **properties and attributes**, **relations**, and **multiple inheritance**. This may help understanding the meaning of the semantically-anchored phrase (in a earlier project, which investigated knowledge-based machine-aided translation, we integrated for similar purposes a module for natural language generation to provide more readable explanations for non-professionals [29]).

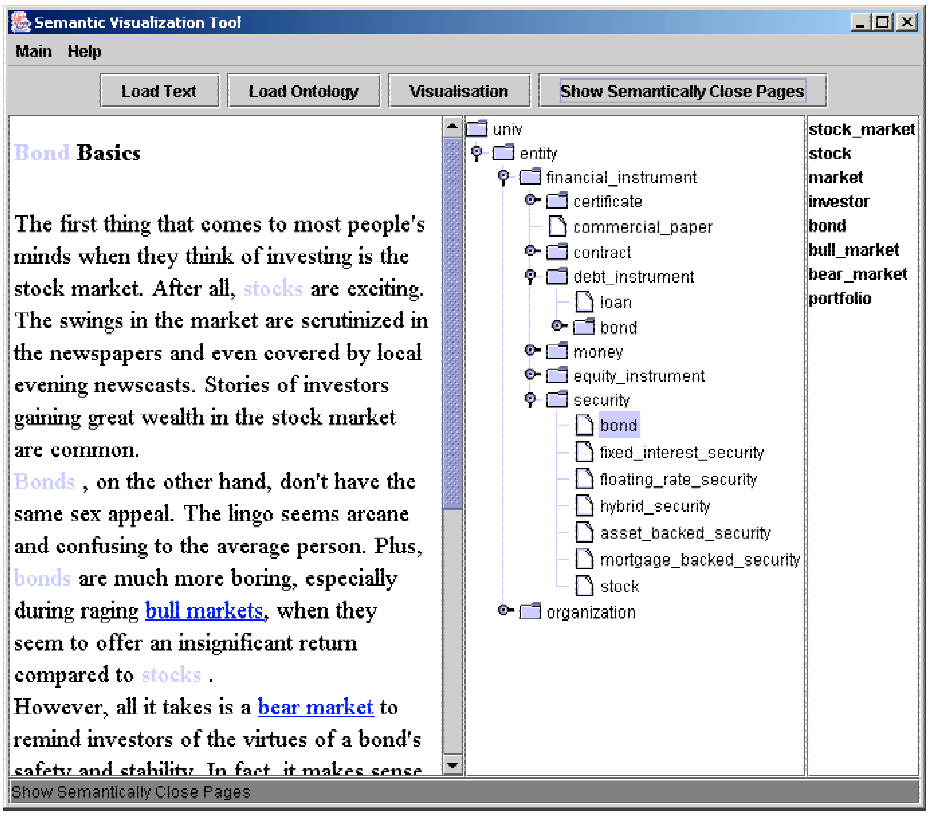

Fig. 2. The most relevant corpus page, which is semantically linked to *bond*

Another useful application of semantically-indexed archives of www-pages is the emerging "educational Semantic Web". Current eLearning systems need increased abilities in quality, quantity, and interaction and the Semantic Web agents (i.e. content, that is aware of itself) will play crucial role in the next generation learning management systems which take content from various sources and aggregate it. In our previous project [26] we implemented personalised information retrieval as the

system was offering to the learner dynamically retrieved readings, which are "most relevant" to the concept he/she does not know well or know wrongly. Unfortunately, information retrieval by keywords does not work well for educational purposes as it does not recognise the text genre and the majority of the "relevant" pages contain much technical data (e.g. banking information) which is not a good teaching source. Semantic indexing, however, works via the ontology thus proposing content to students who need it. In this way the Semantic Web can be used as a technology for realising sophisticated eLearning scenarios for all actors involved in the learning process, including teachers in collaborative courseware generating system.

5 Conclusion

In this paper we report about on-going experiment designed after a careful study of the recent developments in the Semantic Web. As our focus is on the text of the Internet pages, we explore scenarios for annotation and semantic-based search with compact ontologies. It seems the annotation as such will be problematic, as the indexing is problematic even in the well-defined domains. It also looks likely that in the foreseeable future, we will have numerous non-coherent ontologies, as they exist today in the domain of medicine, developed by different organisations and in different natural languages, with manual alignment defined by explicit correspondences. On the other hand we believe that the semantic-based text processing is feasible at least in narrower, well-defined domains. So we plan to continue our experiment by annotating financial pages against MILO as well and exploring the semantic search with several ontologies.

References

[1] Benjamins R., and Jesús Contreras, The six challenges for the Semantic Web. White paper 2002, http://www.isoco.com/isococom/whitepapers/files/SemanticWeb-whitepaper-137.pdf

[2] PROTEGE, http://protege.stanford.edu/

[3] McGuinness D., Fikes R., Rice J., and S. Wilder. An environment for merging and testing large ontologies. In *Proceedings of KR 200*0, pp. 483–493. Morgan Kaufmann, 2000.

[4] Sure Y., Angele J.and S. Staab OntoEdit: Guiding Ontology Development by Methodology and Inferencing. In Proc. of the International Conference on Ontologies, Databases and Applications of SEmantics ODBASE 2002

[5] Noy N, and Musen M. PROMPT:Algorithm and Tool for Automated Ont. Merging and Alignment. In Proc. Of the 17th National Conference on Artificial Intelligence (AAAI-2000),Austin, TX, 450-455.

[6] Doan A, Madhavan J. Domingos P., and Halevy A. Learing to Map between Ontologies on the Semantic Web. In Proc. 11th Int. World Wide Web Conf. (WWW2002)

[7] The Energy Data Collection (EDC) project: deep focus on hydra-headed metadata. www.digitalgovernment.org/news/stories/2002/images/metadatafinal.pdf

[8] Hovy E., and J. Clavans. Comparison of Manual and Automatic Inter-Ontology Alignment, 2002, http://altamira.isi.edu/alignment

[9] Vargas-Vera M., Motta E., Domingue J., Shum S. B. and M. Lanzoni. Knowledge Extraction by using an Ontology-based Annotation Tool. Proc. 1st Int. Conf. on Knowledge

Capture, (K-CAP'01), Workshop on Knowledge Markup & Semantic Annotation, Victoria, B.C., Canada, 2001
[10] SPIRIT, Spatially-Aware Information Retrieval on the Internet, IST FP5 project in Semantic Web, http://www.geo-spirit.org/
[11] Sparck-Jones, Karen, "What is the Role of NLP in Text Retrieval?," in Tomek Strzalkowski (ed.), Natural Language Information Retrieval, Kluwer, 1999, pp. 1-24.
[12] Kimani St., Catarci T., and I. Cruz. Web Rendering Systems: Techniques, Classification Criteria and Challenges. In Vizualizing the Semantic Web Geroimenko, V.Chen Ch. (Eds.), Berlin: Springer 2002, pp. 63 – 89.
[13] John Domingue, Martin Dzbor, Enrico Motta: Semantic Layering with Magpie. Handbook on Ontologies 2004, pp.533-554
[14] TeSSI®: Get more out of your unstructured medical documents. Language & Computing, White paper April 2004, see www.landc.be
[15] Natural Language Processing in Medical Coding. Language & Computing, White Paper April 2004, www.landc.be
[16] http://www.w3.org/2001/Annotea/
[17] http://sourceforge.net/projects/ontomat
[18] Handschuh, S., Staab, S. and F. Ciravegna, "S-CREAM - Semi-Automatic Creation of Metadata," Semantic Authoring, Annotation and Markup Workshop, 15th European Conference on Artificial Intelligence, (ECAI'02), Lyon, France, 2002, pp. 27-33.
[19] Ciravegna F., Dingli A., Petrelli D. and Yorick Wilks: "Document Annotation via Adaptive Information Extraction" Poster at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval August 11-15, 2002, in Tampere, Finland
[20] Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A. and Fabio Ciravegna. MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup", In Proc. of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002), ed Gomez-Perez, A., Springer Verlag, 2002.
[21] ACM Computing Classification System, www.computer.org/mc/keywords/keywords.htm
[22] CGI Taxonomy, http://www.siggraph.org/education/curriculum/projects/Taxonomy2001.htm
[23] SeSDL Taxonomy, www.sesdl.scotcit.ac.uk/taxonomy/ed_tech.html
[24] Nichols D. and A. Terry. User's Guide to Teknowledge Ontologies. Teknowledge Corp., December 2003, ontology.teknowledge.com/Ontology_User_Guide.doc
[25] Pease A., Niles I., and J. Li, 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada, July 28-August 1, 2002. http://projects.teknowledge.com/AAAI-2002/Pease.ps
[26] Angelova G., Boytcheva S., Kalaydjiev O., Trausan-Matu S., Nakov P. and Albena Strupchanska. Adaptivity in Web-Based CALL, In Proc. ECAI-2002, July 2002, Lyon, France, pp. 445-449. (see the LARFLAST ontology at http://www.larflast.bas.bg)
[27] Dobrev P. and K. Toutanova, CGWorld - Architecture and Features, In Proc. of the 10th International Conference on Conceptual Structures: Integration and Interfaces, pp. 261 – 270; http://www.larflast.bas.bg:8080/CGWorld
[28] JENA, http://jena.sourceforge.net
[29] Projects DB-MAT and DBR-MAT, 1992-1998 : knowledge-based machine aided translation, see http://nats-www.informatik.uni-hamburg.de/~dbrmat and http://www.lml.bas.bg/projects/dbr-mat.