Semantic Content Annotation and Ontology Creation to Improve Pertinent Access to Digital Documents

Rocío Abascal-Mena[1] and Béatrice Rumpler[2]
[1] Universidad Autónoma Metropolitana - Cuajimalpa, José Vasconcelos 131, Col. San Miguel Chapultepec, Del. Miguel Hidalgo, 11850, México, D.F., México
[2] INSA de Lyon – LIRIS, 7 Avenue J. Capelle Bât 502 – Blaise Pascal, F69621 Villeurbanne cedex, France
mabascal@correo.cua.uam.mx, Beatrice.Rumpler@insa-lyon.fr

Abstract. In order to serve the needs of their current and future users' digital libraries must provide access to the relevant data. Since recent developments are still behind user needs, describing data using metadata has proven to be crucial for building digital libraries and for providing effective access to the information. This paper describes the use of concepts, extracted from the document itself, to annotate documents using them like "metadata tags". In order to suggest new relationships and new terms to seek, we have built also an ontology based on the concepts extracted from the theses. We present the process followed to add new semantic metadata into the digital theses and the methodology followed for the construction of the ontology based on the new metadata.

Keywords: metadata, knowledge markup, ontology, semantic annotation.

1 Introduction

Although there have been substantial advances in the way to structure information, users must still assess the pertinence of documents presented by the web. Generally, users need to get only parts of the pertinent documents rather than the complete documents. It is fastidious to read and evaluate several documents. For this reason, many pertinent documents are always unknown by users. Therefore, we try to propose a solution to enable a better access to pertinent documents or parts of documents in digital libraries.

Our work is situated within the context of a digital library, CITHER of INSA, Lyon. It concerns the online publishing of scientific theses, which is included in this study. As in other digital libraries, we encountered the same difficulties to find pertinent information in the CITHER system. During a search session, it is impossible to extract the pertinent contents of several theses. To evaluate the pertinence of a thesis, users must read several parts of the document. Furthermore, a document may be too long for a quick evaluation. A promising way to solve this problem is to use metadata in order to *"annotate"* and to describe, in a better way, the content of the documents. In our proposal, we have decided to extract the concepts, that best describe the theses, to use them as metadata for *"semantic tags"*. Of course, manual extraction of con-

cepts is a long time-consuming and is an expensive task. Tools for automating the extraction of concepts can overcome these limitations. Another promising way can be to use an ontology based on these concepts used like *"semantic tags"*. In our approach, an ontology is the description of concepts and their relations. We propose the construction of an ontology from digital theses by following a certain methodology.

In our context, which is a digital library that publishes scientific theses, the introduction of new semantic information into documents has clearly for purpose to ameliorate information retrieval. In order to insert new semantic information into digital theses, we have used a tool able to extract concepts from a given document. Section 2, describes how we have chosen this tool. Afterward, we present the system developed to make annotations. Once digital theses are annotated, a search session is based on the new *"semantic tags"*. In order to expand users request and to give to users also the possibility to chose documents that are closer to the pertinent document, we have decide to construct an ontology. The ontology is composed by the terms of a domain, which become, in our proposition, *"semantic tags"* used to annotate theses. In addition, the ontology is composed by the identification of relations between terms. The identification of relations among concepts and the methodology followed to construct our ontology is described and illustrated in Section 3. Section 4 shows the integration, in the CITHER system, of the semantic annotations and the ontology in order to give the user the pertinent information. Afterward, we present a brief summary of related work in Section 5. Conclusions and further research are proposed at the end.

2 Methodology to Annotate Digital Documents

In large document collections, such as digital libraries, it is very important to have mechanisms able to only select the information requested. The use of *keywords* to represent documents is a promising way to manipulate information in order to classify documents like pertinent or not pertinent.

Annotation is the process of adding semantic markup to documents, but determining which concepts are tied to a document is not an easy task. To address to this problem, several methods are proposed to extract concepts from a given document. In the field of extraction of concepts there are two main approaches: *"keyphrase assignment"* and "*keyphrase extraction"*. By the term "*keyphrase"*, we mean a phrase composed by two or more words, which describes in a general way, the content of the document. "*Keyphrases"* can be seen like *"key concepts"* which are able to classify documents into categories. *"Keyphrase assignment"* uses a controlled vocabulary to select concepts or phrases that best describe the document, instead "*keyphrase extraction"* choose concepts from the document itself.

Our approach consists in taking a document as input to automatically generate a list of concepts as output. In general, this work could be called "*keyphrase generation"* or "*concept generation"*. However, the tool used in our work performs "*concept extraction"* which means that the concepts extracted always appear in the body of the input document.

2.1 Concept Extraction

In order to choose one tool for the extraction of concepts able to extract the higher number of pertinent concepts, we have evaluated four tools: (1) TerminologyExtractor of Chamblon Systems Inc., (2) Xerox Terminology Suite of Xerox, (3) Nomino of Nomino Technologies and (4) Copernic Summarizer of NRC. To evaluate the output list generated by each tool, we have compared this list with one referring list which contained concepts generated manually. The measure of performance and the method followed for scoring concepts are described in [1]. The results obtained indicate that Nomino is the most interesting tool for our approach because of the high number of pertinent concepts that it can extract.

Nomino is a search engine distributed by Nomino Technologies [15]. Nomino adopts a morphosyntactic approach. The morphological analyzer makes *"steeming"*, which means that the prefix and the suffix are removed to make one single word. Nomino applies empirical criteria to filter the noise associated to the extracted concepts. These criteria include frequency and category, as well as stop lists. Nomino produces two types of interactive index, which contain all the concepts that most accurately summarize the content of a given document. One of the index created is very general, however the other one contains the most interesting concepts for Nomino. This index is based on two principles: the *"gain to express"* and the *"gain to reach"*. The *"gain to express"* classifies concepts according to their location in the given document. For example, if a paragraph is only concerned by one concept then it will be classified as important. The *"gain to reach"* classifies concepts according to the frequency of apparition. So, if a word is very rare, it will be selected as important. For example, if in a given document we find *"computer software"* and *"developing computer software"*, the second phrase is going to be selected as important because it is more complete and describes the document better. Instead, if the frequency of *"computer software"* is higher then both phases will appear in the concept list.

2.2 A Tool to Annotate Documents

Since manually annotation can be time consuming and induce to error, we have developed a tool to add easily knowledge into documents by making selections from one proposed list.

To exploit concepts extracted by the remarkable index of Nomino, we have proposed a tool to *"annotate"* documents [1]. The task we consider here is to take a document as input, in XML format, and to automatically add into it the Nomino's concepts by the way of tags. Usually when the paragraph containing the concept is identified then it is surrounded by a simple tag such as "*<concept-name>*" and "*</concept-name>*" at the end. This annotation scheme is very simple and so it can be easily applied to a text also by using a XML editor. However, by using the annotation tool, users can validate concepts proposed by Nomino or even propose other concepts to be aggregate. This tool allows the management of Nomino's concepts, the indexation and the extraction of pertinent paragraphs of the document according to some

search criteria. During a search session, the system is going to be focus in XML tags in order to retrieve the paragraph(s) containing pertinent information.

New work is taking place in order to improve the annotation tool. In the next paragraph, we describe this work, which concerns the construction of an ontology able to expand requests or categorize documents.

3 Methodology to Construct the Ontology

Gruber has defined ontology like *"an explicit specification of a conceptualization. A conceptualization is defined by concepts and other entities that are presumed to exist in some area of interest and the relationships that hold among them"* [6]. An ontology in the artificial intelligence community means the construction of knowledge models [2], [6], [12], [19] which specify concepts, their attributes and inter-relationships. A knowledge model is a specification of a domain that focuses on concepts, relations and reasoning steps characterizing the phenomenon under investigation.

Our ontology is composed of two elements: the *"domain terms"* and the *"relations"* among them. The "d*omain terms"* are words or groups of words that are used to characterize a specific field. The "r*elations"* among these domain terms are of type associative and hierarchic. Two main approaches can be taken when building an ontology. The first one relies on a *"top-down method"*. Someone may use an existing ontology and specify or generalize it to create another one. The second way to build an ontology is by using a *"bottom-up method"*. This method consists on extracting from the appropriate documents all the elements needed to compose an ontology. We believe that this method is accurate in our case because it does not exist yet an ontology of our domain. This method relies on two main stages: the extraction of domain terms (Section 3.1.1) and the identification of relations among these domain terms (Section 3.1.2).

Various methodologies exist to guide the theoretical approach chosen, and numerous tools for building ontology are available. The problem is that these procedures have not coalesced into popular development styles or protocols, and tools have not yet matured to the degree one expects in other software practices. Examples of methodologies followed for ontology building are described in [4], [8], [10]. In general, the following steps can define the methodology for the ontology building: (1) *"ontology capture"* and (2) *"ontology coding"*. The *"ontology capture"* consists in the identification of concepts and relations. The *"ontology coding"* consists in the definition of concepts and relations in a formal language. These two steps are going to be described in the following paragraphs in order to present the construction of our ontology.

3.1 The Ontology Capture Phase

The ontology capture phase consists in designing the overall conceptual structure of the domain. This will likely involve identifying the domain's principal concrete con-

cepts (Section 3.1.1) and their properties and identifying relationships among concepts (Section 3.1.2).

3.1.1 Concepts Extraction

This section reports on our methodology used towards defining concepts to describe the content of theses. The backbone of our ontology is a hierarchy of concepts, which had been extracted of the theses themselves.

The concepts of the ontology are used to automatically categorize documents and thus to allow a thematic access to documents. The problem of retrieving concepts and their structure come from the using of tools able to retrieve candidate concepts. Like described in Section 2, we have used Nomino for concept extraction. Given a document or a group of documents, Nomino constructs a specific index, which contains phrases composed by two or more words that are supposed to define the field. These concepts are called CNU (Complex Noun Units), series of structured terms composed by nominal groups or prepositional groups [5]. We used the CNU Nomino results as a starting point to construct our ontology. The use of NLP tools (Natural Language Processing), like Nomino, often produces *"errors"* that have to be corrected by a specialist of the field. Some of these *"errors"* include phrases that are not concepts or phrases that do not really describe the document. The *"errors"* found in our work, by using Nomino, were generally about the kind of: (1) verbs frequently used (e.g. *"called"*), (2) abbreviations of names (e.g. *"j."*), (3) names of people, cities, etc., (e.g. *"John"*), and also (4) phrases that were composed like CNU concepts but that they were not interesting (*"next phase of the development"*).

Until now, we have not talk yet about the corpus used to make the ontology. The corpus used was composed of scientific documents. Once, these documents were analyzed by Nomino, we have obtained 78 possible concepts to be included in the ontology. We have gotten concepts like: *"information research"*, *"information system"*, *"research system"*, *"remote training"*, *"abstract ontology"*, *representation of ontology"*, etc.

The next step to construct the ontology is to define the relations between the concepts. In the next paragraph, we describe the process used to find relations by using Nomino's results as input.

3.1.2 Extraction of Semantic Relations

With regard to the acquisition of semantic relationships, there exist several approaches for acquiring semantic information. Once concepts have been retrieved, by using Nomino, they must be structured. One of the best-used techniques to discover relations among terms in documents relies on the number of terms co-occurrences. This technique identifies terms that often occur together in documents.

Different techniques exist to identify relations among terms; they are based on contexts of their co-occurrences. The idea is that two similar terms do not necessarily often occur together, as described above, but occur in similar contexts, they often appear surrounded by the same words. A first method based on this principle is described in [16]. This method represents the contexts in which words occur using a variety of lexical features that are easy to identify in large corpora. These contexts are

then converted into similarity or vector spaces which can then be clustered using a variety of different algorithms. A second method relying on this idea of similarity of contexts of terms occurrences is the one described by [11]. This method combines various text-based aspects, such as lexical, syntactic and contextual similarities between terms. Lexical similarities are based on the level of sharing of word constituents. Syntactic similarities rely on expressions in which a sequence of terms appears as a single syntactic unit. Finally, contextual similarities are based on automatic discovery of relevant contexts shared among terms.

In our approach, we use a NLP tool called LIKES [17], which is able to extract relations among concepts. LIKES (Linguistic and Knowledge Engineering Station) extracts concepts by looking to those concepts that are repeated in the document. LIKES, based on statistic principles, is a computational linguistic station with certain functions able to build terminologies and ontologies. The concepts extracted by Nomino have been paired in order to find relations between them. Thus, we have paired manually all the concepts. These pairs have also been compared in the opposite way, for example for the pair *"knowledge /language"* it has been also evaluated: *"language /knowledge"*. In this way, instead of having for example 200 pairs of concepts, we are going to have 400 pairs of concepts. Identifying relations by using LIKES it is an intense work because it takes a long time to process the corpus and to visualize the possible relationships. Furthermore, sometimes the relationships found are not very pertinent.

LIKES allows the representation of relationships in order to find similar relations in other pairs of concepts. One example of phrases that contained some relationship among the pair of concept *"knowledge / language"* is the following (we have kept the same sentence structure in English as in French language):

- A core of *knowledge* **is represented by all** *languages*;
- Other *knowledge* **is represented by some** *languages*;
- *Knowledge* **is represented in all** *languages*.

In the next paragraph we present the phase of the ontology coding where we are going to explain how we use the relations, identified by LIKES, to model a formal ontology.

3.2 The Ontology Coding Phase

The ontology coding is defined as the structuring of the domain knowledge in a conceptual model [20]. In our case the concepts are extracted by using Nomino, in some formal language.

To represent concepts and their relationships we have chosen Protégé. Protégé is a knowledge-engineering tool that enables developers to create ontology and knowledge bases [7], [9], [12], [18]. In this way, having the concepts extracted by Nomino and the relationships among concepts identified by LIKES we have used Protégé to model the ontology. Some relationships among concepts were missing and so we have added some relations like *"has"* or *"kind-of"*. Thus, we have constructed a domain ontology able to represent the main concepts included in the corpus. To have a clear idea of

how the ontology is seen, we represent, in the Figure 1, some concepts and their relationships, especially for the concepts *"language"* and *"knowledge"*.

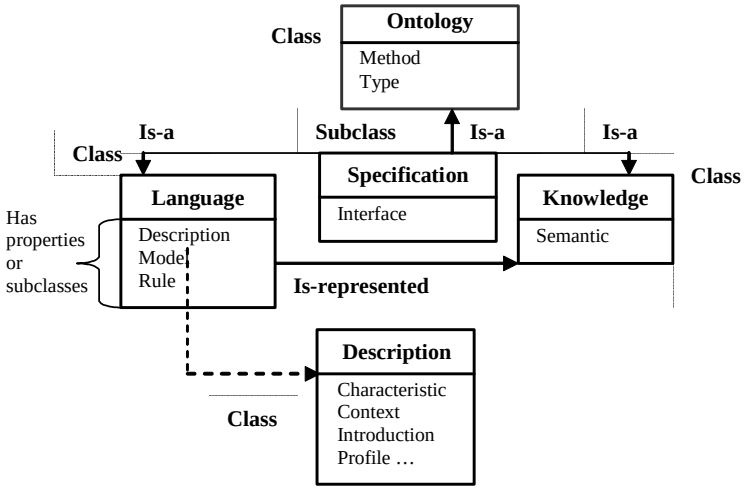

Fig. 1. The classes or concepts *"Knowledge"* and *"Language"* are modeled in order to show their relationships among other classes or subclasses like *"Specification"*

The Figure 1 shows that the class *"Specification"* is a subclass of the classes *"Language"*, *"Ontology"* and *"Knowledge"*. Therefore, *"Specification"* is going to be included in these three classes. The relationship *"Is-represented"* is the one that we have found by using LIKES. Subclasses included in each class also have properties or subclasses. For example in the class *"Knowledge"* we found the subclasses *"Description"* that itself has the subclasses: *"Characteristic"*, *"Context"*, *"Introduction"*, *"Profile"*, *"Theme"*, *"Proprieties"*, etc. In our ontology, we have represented not only concepts with their relations but also "*slots"*. A *"slot"* is an attribute of a class in an ontology. For example, we have the relationship *"Is-represented"*, which can have some values or value types that are typically string type. Some of the values for the *"Is-represented"* relationship are: *"by all"*, *"in some"* and *"in all"*.

4 Semantic Annotation and Ontology Integration

The initial CITHER project proposes the online access to the scientific doctoral theses of the INSA of Lyon, since January 1997. It allows the consultation, the conservation of the theses and the promotion of the research of laboratories. The distribution of theses, in PDF (Portable Document Format) format is done by the way of a server. However, by using the PDF format, it is not easy to automatically exploit the content of the theses. Therefore, we decided to use the XML format to store the theses. The new CITHER system, using XML documents, is under development. The new system's architecture was designed to satisfy the users' requirements. These re-

quirements include selecting pertinent information during a search session. We will briefly summarize the workflow and describe the associated functions.

Information Capturing. After the theses are scanned, they are annotated by including "metadata tags". These "metadata tags" come from the concepts extracted from the thesis itself. These "metadata tags" describe the semantic content. During this phase, we use NLP tool to extract concepts from the documents. The meta-information discovered during "*pre-processing"* is then stored with the corresponding documents in the repository. The storage is carried out by the "*XML content manager"*, which adds new information to the theses. Indeed, the domain knowledge contained in the "*ontology manager"* is based on the meta-information contained in the theses.

User Request. Given a search term, the ontology is used to recommend closer terms and to significantly enrich the request of the user. Users will be able to navigate between terms in order to choose pertinent documents. Once, terms are chosen by users, the "XML content manager" search in the "metadata tags" to find and retrieve pertinent fragments of the theses. By this way, if the fragments are pertinent for the user, this one can decide to retrieve the complete thesis.

5 Related Work

The terms are linguistic representation of concepts in a particular subject field [14]. Like this, applications in automatic extraction of concepts, called terms in many cases, include specialized dictionary construction, human and machine translation, indexing in books and digital libraries. Work in this area has been follow in order to produce tools for automatic extraction.

The University Michigan Digital Library (UMDL) ontology [21] delineates the process of publications using six formal concepts: *"conception"*, *"expression"*, *"manifestation"*, *"materialization"*, *"digitization"* and *"instance"*. Each of these concepts is related to other concept by using: *"has"*, *"of"," kind-of"* or *"extends"* relationship. An ontology in the domain of the digital library is presented in the work of [2]. This ontology tries to represent the way in which new work is expressed. As a result, using the ontology researchers will no longer need to make claims about the contributions of documents (e.g. *"this a new theory"*, *"this a new model"*, *"this is a new notation"*, etc), or contest its relationships to other documents and ideas (e.g. *"it applies"*, *" it extends"*, *" it predicts"*, *"it refutes"*, etc).

Some of the methods used to specify ontologies in digital library projects include vocabularies and cataloguing codes such as Machine Readable Cataloguing (MARC). Other projects are based on the use of thesauri and classifications in order to describe different components of a document like the title, the name of the auteur, etc. In this way, some algorithms can make use of already existing thesauri in order to provide the user with useful suggestions in the integration of ontologies [3].

6 Conclusion and Further Research

We have presented an approach to improve the document retrieval by using the semantic content. Our approach has a double advantage, first, it can exploit the entirely content of digital theses by using semantic annotations and it can provide other alternatives to the user requests. We have noticed that by adding related words (concepts words) in a document, it increases the number of relevant documents identified during a search session. In addition, ontologies can be used to support the operation and growth of a new kind of digital library, implemented as a distributed intelligent system [21]. In consequence, an ontology can be used to deduce characteristics of content being searched, and identify operations that are appropriate and available to access content or manipulate it in other ways. We have constructed an ontology by following a methodology. As long as there are not tools able to construct automatically ontologies from documents, the process carried out by using NLP tools will be fastidious and need the help of field experts. The extraction of relations by hand is very complex and by using NLP tools we have noticed that it still remains relations to be instantiated by the expert of the field. It is evident that there are still some needs in the ontology construction domain but at this moment, we are able to build ontology to support an entire domain. The construction of our ontology is only the first step to make a better access to the information in the digital library.

Further research should investigate the use of dictionaries or thesaurus in digital libraries to detect similar and not identical terms. The use of synonyms to complete our ontology could be another attempt.

References

1. Abascal-Mena, R., Rumpler, B. Adaptive Hypertext Annotations for a Digital Library. International Transactions on Computer Science and Engineering. Vol. 32, No. 1. pp. 7-14. ISSN: 1738-6438, ISBN: 89-953729-5-8. July 2006. (2006)
2. Benn N., Buchingham S., Domingue J. Integrating Scholarly Argumentation, Texts and Community: Towards an Ontology and Services.5th Workshop on Computational Models and Natural Argument. IJCAI'05: International Joint Conference on Artificial Intelligence, Edinburgh, July 2005. (2005)
3. De Bo J., Spyns P., Meersman R. Assisting Ontology Integration with Existing Thesauri. On the Move to Meaningful Internet Systems 2004: CoopIS, DOA and ODBASE. Lecture Notes in Computer Science. pp. 801-818. ISSN: 0302-9743. (2004)
4. Fortuna B., Mladenic D., Grobelnik M. Semi-Automatic Construction of Topic Ontology. Proceedings of ECML/PKDD Workshop KDO 2005. October 2005. (2005)
5. Golebiowska, J.: SAMOVAR – Knowledge Capitalization in the Automobile Industry Aided by Ontologies. In Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2000). Juan_les-Pins, France, October 2, (2000)
6. Gruber, T. R.: A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5, 2, pp. 199-220 (1993)

7. Hogeboom M., Lin F., Esmahi L., Yang C. Constructing Knowledge Bases for E-Learning Using Protégé 2000 and Web Services. 19th Conference on Advance Information Networking and Applications (AINA 2005). Vol. 1. pp. 215-220. ISSN: 1550-445X. (2005)
8. Kim J-M., Choi B-I, et al. A Methodology for Constructing of Philosophy Ontology Based on Philosophical Texts. Computer Standards & Interfaces. Vol. 29. Issue 3. March 2007. pp. 302-315. (2007)
9. Musen, M. A., Fergerson R. W, Grosso W. e., Noy N. F., Crubézy M., Gennari J. H.: Component-Based Support for Building Knowledge-Acquisition Systems. In Proceeding of the Conference on Intelligent Information Processing (IPP 200) of the International Federation for Information Processing World Computer Congress (WCC 2000), Beijing, (2000)
10. Nanda J., Simpson T. W., Kumara S. R. T. A Methodology for Product Family Ontology Development Using Formal Concept Analysis and Web Ontology Language. Journal of Computing and Information Science in Engineering. June 2006. Vol. 6. No. 2. pp. 103-113. (2006)
11. Nenadic G., Ananiadou S. Mining Semantically Related Terms From Biomedical Literature. ACM Transactions on Asian Language Information Processing (TALIP). Vol. 5. pp. 22-43. ISSN: 1530-0226. (2006)
12. Noy N. F., Musen M. A. Ontology Versioning in a Ontology Management Framework. IEEE Intelligent Systems. July/August 2004. Vol. 19. No. 4. pp. 6-13. ISSN: 1094-7167. (2004)
13. Noy, N. F., Fergerson, R. W., Musen, M. A.: The Knowledge Model of Protégé-2000: Combining Inter-operability and Flexibility. In Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France, October 2, (2000)
14. Pazienza M. T., Pennacchiotti M., Vindigni M., Zanzotto F. M. AI/NLP Technologies Applied to Spacecraft Mission Design. Proceedings of the 18th International conference on Innovations in Applied Artificial Intelligence. Lecture Notes in Computer Science. pp. 239-248. (2005)
15. Plante, P., Dumas, L., Plante, A.: Nomino version 4.2.22 updated the 25 July 2001. http://www.nominotechnologies.com. (2001)
16. Purandare A., Pedersen T. Discriminating Among Word Meanings By Identifying Similar Contexts. Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04). (2004)
17. Rousselot, F., Frath, P.: Terminologie et Intelligence Artificielle. In Traits d'Union, G.Kleiber and N. Le Queler, dir., Presses Universitaires de Caen. pp. 181-192, (2002)
18. Taboada M., Martínez D., Mira J. Experiences in Reusing Knowledge Sources Using Protégé and Prompt. International Journal of Human-Computer Studies. Vol. 62. pp. 597-618. (2005)
19. Vallet D., Fernández M., Castells P., Mylonas P., Avrithis Y. A Contextual Personalization Approach Based On Ontological Knowledge. International Workshop on Context and Ontologies (C&O 2006) at the 17th European Conference on Artificial Intelligence (ECAI 2006), (2006)
20. Wache H., Vogele T., Visser U., et al. Ontology-Based Integration of Information – A Survey of Existing Approaches. Proceedings of IJCAI 2001. Workshop "Ontologies and Information Sharing". (2001)
21. Weinstein, P.: Seed Ontologies: Growing Digital Libraries as Distributed, Intelligent Systems. In Proceedings of the Second International ACM Digital Library Conference, Philadelphia, PA, USA, July, (1997)