# Ontology-Driven Semantic Digital Library

Shahrul Azman Noah, Nor Afni Raziah Alias, Nurul Aida Osman,
Zuraidah Abdullah, Nazlia Omar, Yazrina Yahya, and Maryati Mohd Yusof

Knowledge Technology Research Group, Faculty of Information Science & Technology,
Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia
{samn,afni,nurulaida,za,no,yaz,mmy}@ftsm.ukm.my

**Abstract.** This paper discusses an on-going research project in developing a semantic digital library for academic institution. It provides another view of semantic information retrieval for digital library from the perspective of semantic technology and ontology. We proposed an approach for managing, organizing and populating ontology for document collections in digital library. In this sense the document metadata and content are inserted and populated to a knowledge base which allows sophisticated query and searching. The paper also proposed an ontology based information retrieval model which is based on the classic vector space model which includes document annotation, instance-based weighting and concept-based ranking.

**Keywords:** digital library, information retrieval, ontology, semantic technology.

## 1 Introduction

The extensive deployment of digital libraries over the last decades is hardly surprising. They offer remote access to articles, journals and books with many users able to access the same document at the same time. Through the use of search engines, they make it possible to locate specific information more rapidly than ever is possible in physical libraries. Warren et al. [1] describes few challenges of current and future digital libraries, such as interoperability between different libraries or different collection of documents which pose a lot of problems, search and semantic retrieval which need to be enhanced and user interface which need to be improved. Semantic technology seems to offer solutions for the aforementioned challenges in digital library.

Digital libraries contain varieties of documents from newspaper articles to academic journals and even audios and videos collections. These collections of documents are usually described using metadata for easier access, storage and retrieval. Using metadata alone, however, is not enough to describe the semantic of documents and enhanced search is usually not possible. As such ontology is seen potential to support current limitations of digital library. One example of documents is academic thesis which contains numerous knowledge contents. An adaptation of the Dublin core metadata can be used to represent the semantic resources of the theses such as author, title, language etc, but what about the content of the theses, location of which the research presented in the theses were carried out, and how each theses link with the other theses or other resources. Such questions can be potentially answered by an

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 141–150, 2010.

ontology which is considered as a backbone to many semantic applications. A well designed ontology is essential for a successful semantic application. However, the construction of ontology is a complex and tedious process, and further more the management, storing and managing such knowledge resources are often difficult for normal users.

Digital libraries manage various kinds of digital contents and provide services for users to navigate, query, use, produce and disseminate the digital resources [2]. However, in order to provide effective services for users of digital libraries such as conceptual search and semantic navigation, the ontology forms the basis for building such semantic integration functionality. Thus, this paper discusses our on-going research in developing a semantic digital library for academic institution. This paper focuses on the semantic management level of extracting knowledge for the academic theses and populate such knowledge into an ontology in such a way to support the semantic search of digital libraries. We also proposed an ontology-based retrieval model meant for the utilization of complete domain ontologies and knowledge-bases. The search system takes advantage of both detailed instance-level knowledge available in the knowledge base, and topic taxonomies for classification. To manage the large-scale information sources, an adaptation of the classic vector-space model for an ontology-based representation is proposed, upon which a ranking algorithm is defined.

## 2 Related Research

Enhancing the knowledge access to the Digital Library of the British Telecom is the goal of one of the case studies in the EU IST integrated project Semantically Enabled Knowledge Technologies (SEKT). In current interfaces to Digital Libraries, users pose keyword-based queries to perform document retrieval. However, these keywords do not directly represent the semantics of the information need of the user. Therefore, the implementation uses an approach that allows the user to perform structured natural language queries against the information contained in the Digital Library [1]. The semantics of the information and the user queries is defined by an underlying ontology. The implementation of BT digital library shows how semantic technology is being used to enhance the digital library features through richer metadata, enhanced user-profiling, unlocking the documents, enhanced searching and browsing, and displaying the result.

The work of [3] focuses on the verification and the tracing of information using an information dissemination platform and other Semantic Web-based services. Services on the platform include information dissemination services to support reliable information exchange among researchers and knowledge service to provide unrevealed information. It is difficult to support collaboration among users because additional metadata cannot be inserted into documents or containers.

The work of [4] focus on question answering over heterogeneous knowledge sources that makes use of different ontology management components within the scenario of a digital library application. In particular, ontologies offer a generic solution to the problem of integrating various sources. The documents in the knowledge sources are annotated and classified according to the ontology. The ontology model consists of concepts organized hierarchically in terms of subsumption as well as of

(binary) relations together with appropriate domain/range restrictions. The ontological metadata can then be exploited for advanced knowledge access, including navigation, browsing, and semantic search. Advanced semantics-based mining technology can extract fine-grained metadata from articles contained in the digital library. Finally, current reasoning techniques allow to answer structured queries to access full-text content as well as fine-grained metadata from articles from different sources in a uniform way. The knowledge base of the digital library consists of a number of heterogeneous knowledge sources, partially structured in the form of metadata and topic hierarchies, but largely unstructured in the form of full text documents. All these heterogeneous knowledge sources are integrated using a common ontology, name PROTON. The structured information sources are integrated using a mapping of the underlying structures to the ontology. The mapping for the unstructured sources is not as direct and using the help of the ontology learning tool Text2Onto [5].

The work of [6] describe how ontologies can serve as symbolic tools within a community of practice. It combines the central server which is called knowledge server whose main role is to retrieve appropriate learning initiatives from the database from end-user queries. The knowledge server consists of a customized HTTP server which offers a library of high-level Lisp functions to dynamically generate HTML pages, WebOnto Server, an operational knowledge modelling language which provides the underlying representation for the ontologies and knowledge models., a set of knowledge models which includes the observatory ontology used to index the learning initiatives in the good practice database. Connected to the central server is a database containing several hundred summaries of documented examples of lifelong learning, Named Entity Recognizer, WebOnto Client and a Semantic Search Service.

The research in [3,4] mainly view the semantic digital library at the upper-level ontology with little or none focusing on domain specific ontology. Such specialized ontology of a particular domain deem important to enhance semantic search from the domain perspectives. Research in [6] although not directly related to digital library, it seems appropriate to be considered in this study. Techniques presented in the research which relates to ontology population can be used in digital library.

Furthermore none of the aforementioned approaches look into the possibility of adapting the available information retrieval model to support semantic. As such most of the approach employ exact matching of documents. Apart from that, as the process of inserting and populating ontology is a complicated process, it is deem appropriate to develop a tool which can assist user in managing knowledge of the theses. We proposed a backend process is to annotate the description of the academic theses using GATE (General architecture for text engineering) [7].

## 3 The Approach

The focus of this paper is on academic theses under the computing domain. The aim is to build an ontology driven digital library which can support semantic search and retrieval. The prototype assists users in inserting and extracting knowledge from documents and subsequently populated into a knowledge base.

### 3.1 The Proposed Architecture

Fig. 1 illustrates the general architecture of the proposed digital library. As can be seen, the construction, modelling and querying of documents are all related to the ontology.

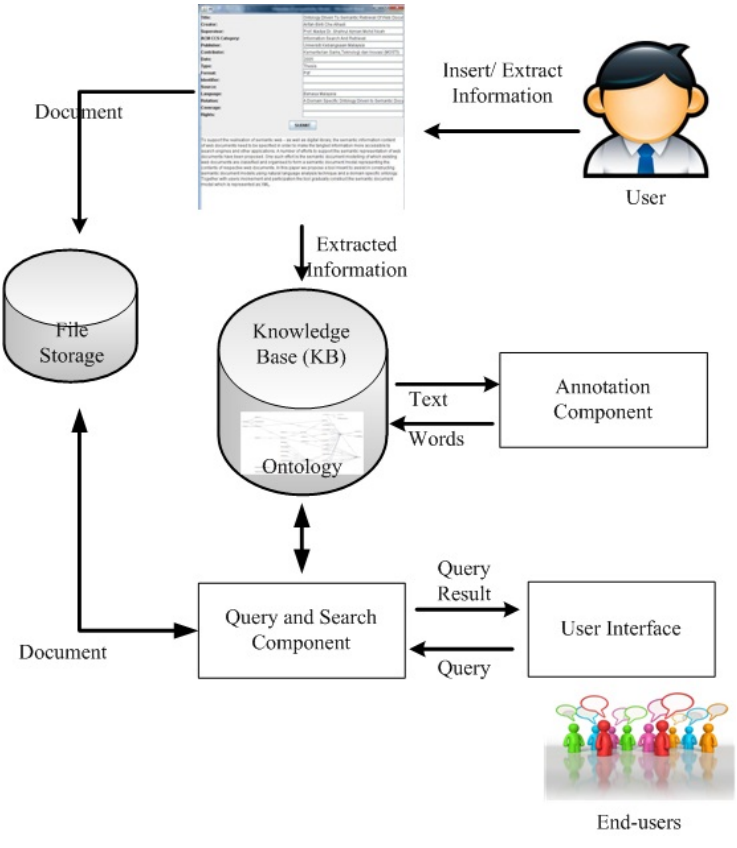


**Fig. 1.** The general architecture of the prototype

There are two (2) types of ontology, i.e. i) standard ontology; and ii) domain ontology. The standard ontology is an ontology which relates to metadata of the document resources such as creator (author), title, date, and language. We assumed that the standard ontology is consistent for all disciplines or subjects of documents. The standard ontology is mainly derived from the Dublin Core metadata and the PROTON ontology. The domain ontology on the other hand is specific to disciplines or subjects such as computing, health sciences, social sciences and medical. It is basically represented as topic hierarchy (or taxonomy). At the moment, only the computing discipline ontology which is the ACM topic hierarchy is available. The domain ontology is meant to semantically describe the content of documents by annotating terms available in the abstract, the whole documents or only keywords. The domain ontology is also used for automatic semantic indexing.

Logically each document is classified under one subject, however terms available in the documents may be annotated by multiple concepts from different domain ontologies. As such the term 'genetic' might be available in the computing domain or biology domain. Similarly for the term 'social network' which may appear in the computing discipline or the social science discipline. When it comes to annotation, the annotation component allows user to manually annotate or accept suggestions provided by the system. We will discuss annotation in more detail in the next section. Apart from that, the standard ontology is also federated with other domain ontology and even extended with more specialized concepts. For instance in the case of academic theses, the standard ontology is federated with the Geo ontology and extended with further specialized concepts such as research, awards and institutions which relate to the theses. Fig. 2 shows a portion of the ontology used to describe the academic theses.

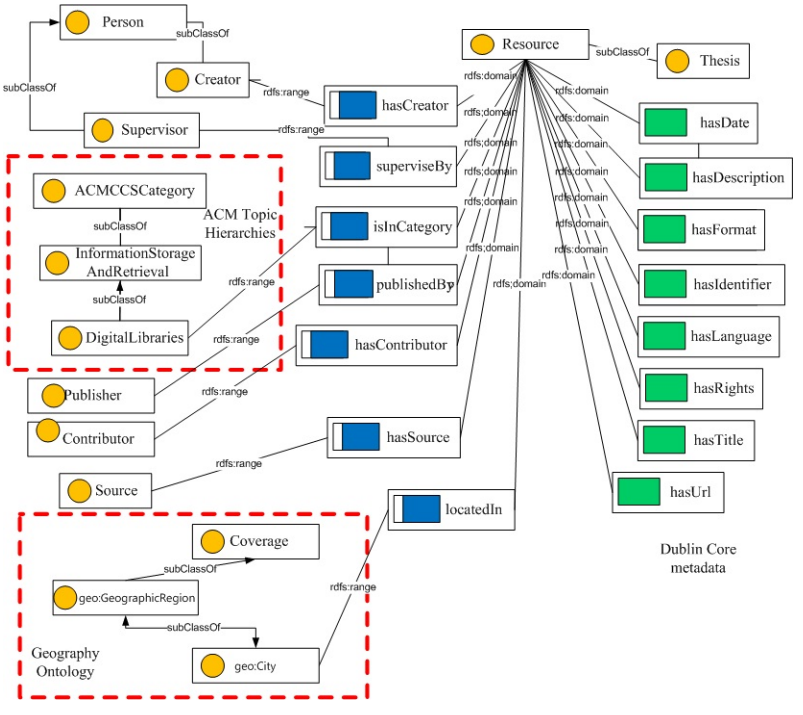


**Fig. 2.** Portion of domain ontology focusing on academic thesis

Referring again to Fig. 1, a user is responsible to insert the document, data and related information (the extracted information) about the digital contents to the knowledge base. Therefore, the prototype acts as a layer or mediator between users and digital libraries. The prototype provides an interface for user to expand the data/instances of the ontology in the digital libraries. The backend of the system consists of an annotation component and a query and search component, while the front end consists of an interface for end-users to access the digital library. As can be seen

in Fig. 1, the process begins with the user insert the extracted information into the prototype. The prototype will populate the information to the ontology based on the information given. Inserted information is the metadata of the thesis which includes the creator, title, year and publisher. These metadata will be populated under the modified Dublin Core ontology represented as OWL. Each inserted thesis will become an instance in the knowledge base.

For example in Fig. 3, the name of the theses' creator will be populated under `Student`, and `Student` is a subclass of `Creator` in the ontology and the title of the theses will be related to the `Creator` with the help of the ontology. This ontology also relate the `Supervisor` of the theses to the `Creator`.

```
  <rdf:Description rdf:about="#hasTitleJournal">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:label
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"></rdfs:label>
    <rdfs:domain rdf:resource="#Journal"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  </rdf:Description>
  <rdf:Description rdf:about="#InformationStorage">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Information
Storage</rdfs:label>
    <rdfs:subClassOf rdf:resource="#InformationStorageAndRetrieval"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>
  <rdf:Description rdf:about="#Paper1">
    <rdf:type rdf:resource="#PublishMaterial"/>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">paper 1
semantic digital library paper</rdfs:label>
  </rdf:Description>
  <rdf:Description rdf:about="#Student1">
    <studentOf rdf:resource="#Academic1"/>
    <studyAt rdf:resource="#Faculty1"/>
    <rdf:type rdf:resource="#Student"/>
    <rdfs:label
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Afni</rdfs:label>
  </rdf:Description>
  <rdf:Description rdf:about="#ISR_infoSearchAndRetrieval_Clustering">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">information
search and retrieval,clustering</rdfs:label>
    <rdf:type rdf:resource="#InformationSearchAndRetrieval"/>
  </rdf:Description>
  <rdf:Description rdf:about="#Source">
    <rdfs:label
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Source</rdfs:label>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>
</rdf:RDF>
```

**Fig. 3.** Example of RDF statements

### 3.2 Document Annotation

Document annotation is the most important task for representing the semantic meaning of digital collections. Annotation can be considered as the process of populating ontology with instance or literals. It can be done either in a manual or automatic fashion. We proposed a semi automatic approach by employing the GATE (General Architecture for Text Engineering) engine toolkit [8]. The ACM topic hierarchy ontology developed contain all the concepts related to the ACM classification. We decided to represent

every topic as instances of a generic class topic. The decision is entirely due to implementation issue which makes it possible to annotate related terms with suitable instances. Each of the topic instances is associated with related terms that represent the topic. These terms were extracted from mining the ACM digital library by extracting keywords defined in articles of the specific category. We assumed that the articles in the ACM digital library were assigned with the correct topic and the associate keywords defined by authors are best to represent the topic. At this point we only annotate the abstract of each thesis. Apart from populating the ACM topic hierarchy, terms extracted from the abstract which are related to other ontologies such as the Geo ontology will also be extracted and populated. Fig. 4 illustrates the proposed approach.

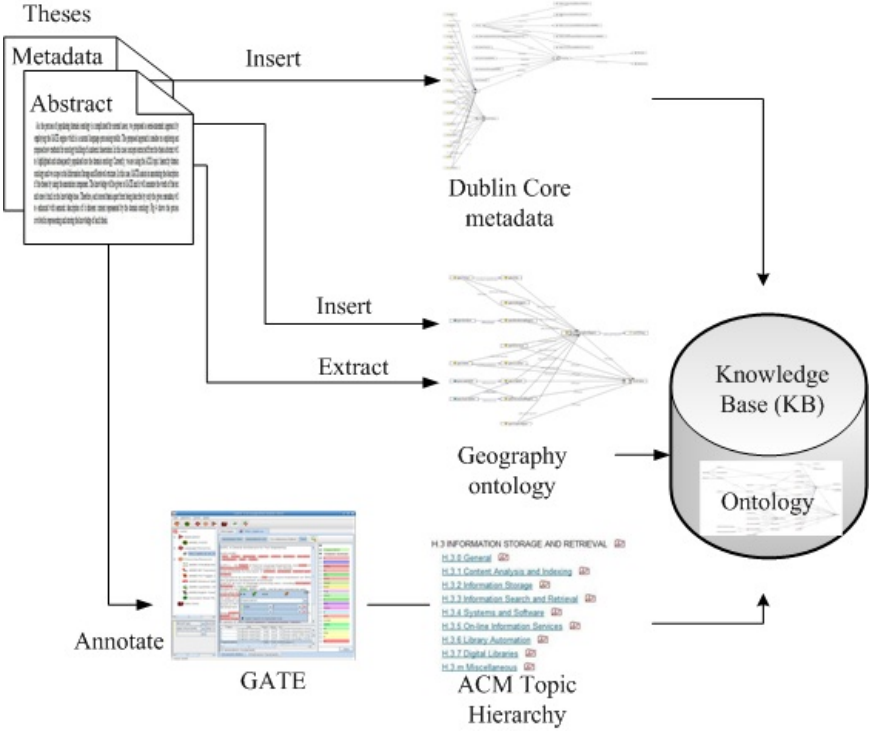


**Fig. 4.** Document Annotation

The annotation of terms or concepts which relate to the standard ontology is rather a straight forward process whereby if such concepts are found matched with the concepts in the ontology then a new triplet will be added to the knowledge base. However, as users are still favoured keyword search which results a ranked list of documents, few consideration is required. We decided to implement an adaptation of a vector space model to support a so-called ontology-based information retrieval. The model is quite similar to the one that was proposed in [9]. In this case annotation of the terms in the abstract (and may be the whole documents) is not stored as triplets as previously described but instead are represented as a vector space model. This would

allow weighting and ranking of retrieved documents. Text in the abstract will undergo the normal indexing process such as tokenisation, frequency calculation and weight assignment.

Terms in the documents are then annotated with concept instances from the existing knowledge base by creating instances of the Annotation Class. Annotation Class is purposely created to facilitate the semantic search. It is a part of the ontology which stored the annotated documents separately in a different database. Annotation Class will link between the knowledge base and the index upon the executed query. Annotation class has two properties which are instance and document, where the concepts and documents are related together. Whenever the label of an instance in the ontology is found, an annotation is created between the instance and the document. It then will be stored in the annotation class under the property of term (instance), concept and document by which are related to each other. Thus, whenever a user sent a search query, the searching will be run upon the ontology first. Whenever the satisfied query found in the domain ontology, it then will be referred to the annotation class and then the documents will be retrieved and presented to the user.

### 3.3 Semantic Search and Processing

The overall semantic search and processing is as illustrated in Fig. 5.

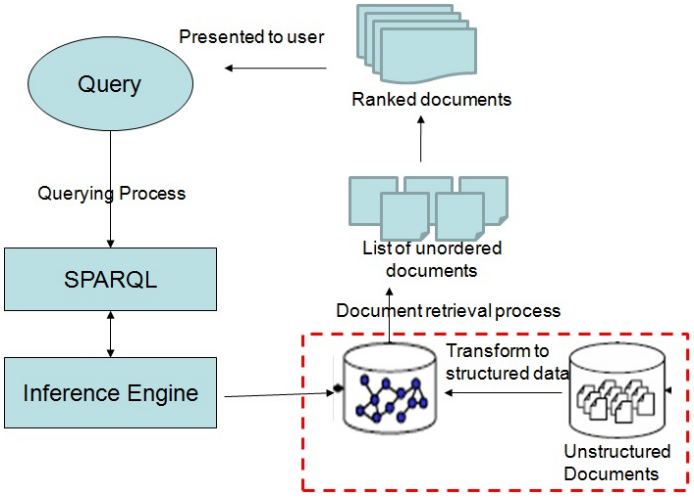


**Fig. 5.** An Ontology Based Information Retrieval Process

The query process takes an input as a user search request. The search request can be either a list of keywords or a complex natural language query. The search request will be first analysed by a query parser and will be parsed into SPARQL. These queries are then sent to the inference engine which will return a set of RDF (Resource Description Framework) triples containing the related concepts or instances in the knowledge base. For example a query on “digital library technology” will first be transformed into a SPARQL query which will be then submitted to the inference

engine. The inference engine will find instances and other "related instances" concerning the submitted terms. The retrieval process will then retrieved all documents which have been annotated with all the instances related to the query terms and subsequently ranked them.

The weighting scheme is based on the modified TFIDF as follows:

$$TFIDF = \frac{freq_{ij}}{max_e freq_{ej}} \times \log \frac{N}{n_i}$$

where $freq_{ij}$ is the total number of instance $i$ in document $j$; $max_e freq_{ej}$ is the maximum number of any instance $e$ in document $j$; $N$ is the total number of annotated documents collection and $n_i$ is the total number of documents that contained instance $i$. The similarity measures is based on the standard cosine measures.

The semantic search facility also allow sophisticated query such as "*Find the supervisor of Arifah Alhadi and the title of her thesis*". Such a query can be represented as a SPARQL statement as follows:

```
SELECT ?student ?thesis ?academic ?studentname ?svname ?title
WHERE {?student rdf:type :Student.
       ?student rdfs:label ?studentname.
       FILTER(REGEX(?studentname,"Arifah"))
             ?academic rdf:type :Academic.
             ?academic :hasFirstNameAc ?svname.
             ?thesis rdf:type :Thesis.
             ?thesis :supervisedBy ?academic.
             ?thesis :hasTitle ?title}
```

The result for the query will return the following tuples which shows that `student001`, who is Arifah Alhadi is supervised by `supervisor067` who is Shahrul Azman and return the title of the thesis.

| student | thesis | academic | svname | title |
|---|---|---|---|---|
| student001 | thesis008 | academic067 | "Shahrul Azman" | "Semantic Document Modelling" |

## 4 Conclusion

This paper has presented our on-going work in implementing an ontology driven to digital library. Ontology proof to be a powerful tool for supporting complex querying and semantic search of digital library collections. However, to manage a knowledge rich digital collections is difficult for most users. The inherent idea of the proposed approach in this paper is to engaged user in implementing a digital library by inserting, updating and populating the knowledge bases.

A semantic retrieval framework is also proposed in this paper which aims to improve the precision of search results by concentrating on the context of concepts. Document annotation is represented as an extension ontology and stored in a relational database. The triple searching and semantic matching is performed by the inference engine and results are passed to the ranker to sort them according to their relevancy to

user's queries. In the current framework we focused on academic theses. Our near future is currently focusing on the aspect of document annotation. Current annotation is purely based on exact match by referring to the labels of each instances stored in the knowledge base. We look into the possibility of doing document annotation by means inexact match or contextual term matching. Other current research work is on evaluating the effectiveness of the proposed instance-based weighting scheme. We are in the process of compiling decent number of test documents in order for the evaluation to be carried out.

**Acknowledgments.** We would like to thank Universiti Kebangsaan Malaysia for supporting this research project.

## References

1. Warren, P., Thurlow, I., Alsmeyer, D.: Applying Semantic Technology in Digital Library: a Case Study. Library Management 26(4/5), 196–205 (2005)
2. Yeh, C.L.: Development of an Ontology-Based Portal for Digital Archive Services. In: Presented in International Conference on Digital Archive Technologies (2002)
3. Jung, H.M., Lee, M., Sung, W.K., Park, D.I.: Semantic Web-Based Services for Supporting Voluntary Collaboration Among Researchers using an Information Dissemination Platform. Data Science Journal 6, S241–S249 (2007)
4. Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I., Völker, J.: Ontology-Based Question Answering for Digital Libraries. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 14–25. Springer, Heidelberg (2007)
5. Cimiano, P., Volker, J.: Text2onto-a framework for ontology learning anda data driven change discovery. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 227–238. Springer, Heidelberg (2005)
6. Domingue, J., Motta, E., Shum, S.B., Vargas-Vera, M., Kalfoglou, Y., Farnes, N.: Supporting Ontology Driven Document Enrichment within Communities of Practice. In: Proceedings of the 1st International Conference on Knowledge Capture, pp. 30–37. ACM Press, New York (2001)
7. Cunningham, H.: GATE, a general architecture for text engineering. Computers and Humanities 36(2), 223–254 (2002)
8. Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P., Humphreys, K.: GATE: An Environment to Support Research and Development in Natural Language Engineering. In: Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence, pp. 58–66. IEEE Press, New York (1996)
9. Castells, P., Fernandez, M., Vallet, D.: An Adaptation of the Vector Space Model for Ontology-Based Information Retrieval. IEEE Trans. on Knowledge and Data Engineering 19(2), 261–272 (2007)