EXPERT KNOWLEDGE MANAGEMENT BASED ON ONTOLOGY IN A DIGITAL LIBRARY

Antonio Martín, Carlos León
Departamento de Tecnología Electrónica, Seville University, Avda. Reina Mercedes S/N, Seville, Spain
toni@us.es, cleon@us.es

Keywords: Ontology, web services, Case-based reasoning, Digital Library, knowledge management, Semantic Web.

Abstract: The architecture of the future Digital Libraries should be able to allow any users to access available knowledge resources from anywhere and at any time and efficient manner. Moreover to the individual user, there is a great deal of useless information in addition to the substantial amount of useful information. The goal is to investigate how to best combine Artificial Intelligent and Semantic Web technologies for semantic searching across largely distributed and heterogeneous digital libraries. The Artificial Intelligent and Semantic Web have provided both new possibilities and challenges to automatic information processing in search engine process. The major research tasks involved are to apply appropriate infrastructure for specific digital library system construction, to enrich metadata records with ontologies and enable semantic searching upon such intelligent system infrastructure. We study improving the efficiency of search methods to search a distributed data space like a Digital Library. This paper outlines the development of a Case-Based Reasoning prototype system based in an ontology for retrieval information of the Digital Library University of Seville. The results demonstrate that the used of expert system and the ontology into the retrieval process, the effectiveness of the information retrieval is enhanced.

1 INTRODUCTION

In the current digital libraries and Internet the access to knowledge depends of the relationship between people, tools and communication devices used. Although search engines have developed increasingly effective, information overload obstructs precise searches. The information is treated as an ordinary database that manages the contents and positions. The result generated by the current search engines is a list of Web addresses that contain or treat the pattern. The useful information buried under the useless information cannot be discovered. It is disconcerting for the end user. Thus, sometimes it takes a long time to search for needed information.

Artificial Intelligent and ontology-based search, from the semantics' perspective, provides added values in searching over documents which are semantically related. Despite large investments and efforts have been made, there are still a lot of unsolved problems. There are a lot of researches on applying these new technologies into current Digital Libraries information retrieval systems, but no research addresses the semantic and intelligent artificial issues from the whole life cycle and architecture point of view (Govedarova & Stoyanov, 2008). Our work differs from related projects in that we build an ontology-based contextual profiles and we introduce an approaches used metadata-based in ontology search and expert systems.

We focus our discussion on case indexing and retrieval strategies and provide a perception of the technical aspects of the application. For this reason we are improving representation by incorporating more metadata from within the information (Ding, 2004). Our approach for realizing content based search and retrieval information implies the application of the Case-Based Reasoning (CBR) technology.

The paper is organized as follows. Next Section describes the setting of Digital Library domain, the research problems and current work in it. Then we present the Ontology design process. Section 3 provides a general overview about our prototype architecture. We summarize its main components and describe how can interact Intelligent Artificial and Semantic Web to enhancement a search engine. Next we study the CBR framework jColibri and its features for implementing the reasoning process over ontologies (GAIA, 2009). Section 4

exemplifies the usage of jCOLIBRI to create the system OntoFAMA, sets out our motivation for choosing this CBR framework, and presents the results of our ongoing work on the adaptation of the framework. Finally we outline the conclusions and future works.

2 DIGITAL LIBRARY DOMINIUM

The Seville Digital Library (SDL) is dedicated to the production, maintenance, delivery, and preservation of a wide range of high-quality networked resources for scholars and students at University and elsewhere. SDL provides tools that support the construction of online information services for research, teaching, and learning. SDL include services to effectively share their materials and provide greater access to digital content (Witten & Bainbridge, 2003).

In this paper we study architecture of the search layer in this particular dominium, a web-based catalogue for the University of Seville. For this purpose we present an ontology-based web architecture for knowledge management in a Digital Library (Stuckenschmidt & Harmelen, 2001). It incorporates ontologies and Artificial Intelligent to enable not only precise location of Web resources but also the automatic or semi-automatic integration of hybrid retrieval knowledge and self-learning.

Consequently, there is a need for not only a retrieval mechanism, but also for a recommendation system to suggest resources of interest when the resources may be too difficult to locate with traditional retrieval systems. Our system proposes a new form of interaction between people and Digital Library, where the latter is adapted to individuals and their surroundings. For this goal in our work we developed four user profiles based on ontologies: Staff, Alumni, Administration, and visitor, Figure 1.

Figure 1: Teacher Profile and resources associated

These user profiles are representation the user's interests. User profiles are used to specify the search results. This information will satisfy the quality of information for a specific kind of user.

2.1 Motivation and Technical Requirements

We propose a conceptual architecture for a digital library information retrieval system. We discuss an opportunity and challenge in this area of work with a specific view of intelligent information processing that takes into account the semantics of the knowledge objects (Warren, 2005). We concentrate on the critical issue of metadata/ontology-based search and expert systems. More specifically the objectives are decomposed into:

- Explore and understand the requirements for rendering semantic search in a digital library.
- Investigate from a search perspective possible intelligent infrastructures form constructing decentralized digital libraries where no global schema exists.
- Investigate how the semantic technologies can be used to provide additional semantics from existing resources.
- Analyze the implementation results, and evaluate the viability of our approaches in enabling search in intelligent-based digital libraries.

This scheme is based on the next principles: knowledge items are abstracted to a characterization by metadata description witch are used for further processing (Taniar & Wenny, 2006).

3 SYSTEM ARCHITECTURE AND IMPLEMENTATION

We will now discuss the details of providing a CBR recommender system to retrieve the requested metadata satisfying a user query. Following this approach we developed a prototype. For this aim we have used two technologies: JColibri and Protégé. The prototype called OntoFAMA is the main tool to verify that the proposed architecture with ontologies and an expert system is an applicable solution. OntoFAMA is composed of three main functional components: ontology, expert search engine, and user interface. A more detailed description of these components and the interaction between them is presented in next sections. In next figure we can see the architecture of the system, Figure 2.

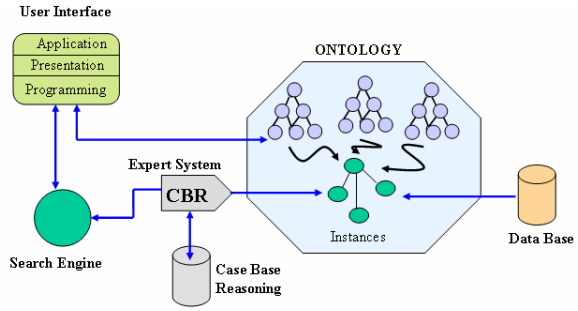

Figure 2: OntoFAMA Search Layer Architecture

The OntoFAMA system uses its internal knowledge bases and inference mechanisms to process information about the electronic resources in a Digital Library. At this stage we consider to use ontology as vocabulary for defining the case structure like attribute-value pairs. First element the Ontology component stores information about resources and services where concepts are types, or classes, individuals are allowed values, or objects and relations are the attributes describing the objects. The metadata descriptions of the resources and library objects (cases) are abstracted from the details of their physical representation and are stored in the case base (Sure and Studer, 2005). CBR case data could be considered as a portion of the knowledge (metadata) about an OntoFama object.

Second element the CBR is widely discussed in the literature as a technology for building information systems to support knowledge management, where metadata descriptions for characterizing knowledge items are used. Current research of distributed CBR shows how CBR systems can benefit from a standardized shared knowledge representation that implies unambiguous interpretation of cases and in this way enable the development of systems that are able to search across multiple case-bases (Toussaint & Cheng, 2006).

In our CBR application, searches are described by metadata concerning desired characteristics of a Library resource, and the solution to the search is a pointer to a resource described by metadata. These characterizations are called cases and are stored in a case base (Luger & George, 2002). Very case contains two slices:

- A description of a framework problem. The possible solutions described by means of framework instantiation actions. These goals will be formally described in terms of framework domain taxonomy and they will be used for indexing cases.
- Solution. Additional information that justifies these steps. Our experience developing has shown that execution graphs are a good technique to represent the list of actions that user should do to reach a solution, so they will be used to represent the solutions in our simple cases.

Finally the acceptability of a system depends to a great extent on the quality of the user interface component (Quan and Karger, 2004). The easiest to implement interfaces communicate with the user through a scrolling dialog as illustrated in figure 3.

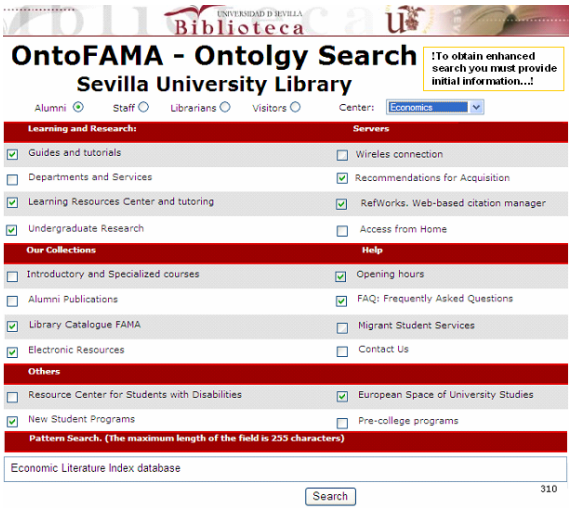

Figure 3: User Profiles, Graphical User interface

The user interacts with the system to fill in the gaps to retrieve the right cases. The interfaces provides for browsing, searching and facilitating Web contents and services. It consists of one user profile, consumer search agent components and bring together a variety of necessary information from different user's resources. The objective of profile intelligence has focused on creating of user profiles: Staff, Alumni, Administrator, and Visitor. The user interface helps to user to build a particular profile that contains his interest search areas in the digital library domain.

In an intelligence profile setting, people are surrounded by intelligent interfaces merged, thus creating a computing-capable environment with intelligent communication and processing available to the user by means of a simple, natural, and effortless human-system interaction. The user enters query commands and the system asks questions during the inference process. Besides, the user will be able to solve new searches for which he has not been instructed, because the user profiles what he has learnt during the previous searchers.

4 DEVELOPING CBR APPLICATIONS

Although Case-Based Reasoning (CBR) claims to reduce the effort required for developing knowledge-based systems substantially compared with more traditional Artificial Intelligence approaches, the implementation of a CBR application from scratch is still a time consuming task. In this section presents a novel, freely available tool for rapid prototyping of CBR applications that focuses on the similarity-based retrieval step. By providing easy to use model generation, data import, similarity modelling, explanation, and testing functionality together with comfortable graphical user interfaces, the tool enables even CBR novices to rapidly create their first CBR applications. Nevertheless, at the same time it ensures enough flexibility to enable expert users to implement advanced CBR applications.

We used a Case-Based Reasoning (CBR) shell, software that can be utilized to develop several applications that require cased-based reasoning methodology. In this study we used the CBR object-oriented framework development environments JColibri a java-based configuration that supports the development of knowledge intensive CBR applications and help in the integration of ontology in them. This framework work as open software development environment and facilitate the reuse of their design as well as implementations. In this section we describe in more detail how JColibri supports rapid prototyping of CBR applications (Bridge & G¨oker, 2006).

Our motivation for choosing this framework is based on a comparative analysis between it and other frameworks, designed to facilitate the development of CBR applications. jColibri enhances the other CBR shells: CATCBR, CBR*Tools, IUCBRF, Orenge. jColibri is and open source framework and their interface layer provides several graphical tools that help users in the configuration of a new CBR system. Another decision criterion for our choice is the easy ontologies integration. jColibri affords the opportunity to incorporate ontology in the CBR application to use it for case representation and content-based reasoning methods to assess the similarity between them.

Our system consists of Query Engine, Inference Engine and Knowledge Base. The mapping between the two layers is realized by connectors. These connectors read the values of the data base columns and ontology and return them to the application, i.e. assign them to the attributes of the case. Query Engine is responsible for the knowledge and queries management. Is a Java library that eases the management of the ontology in an intelligent-based application. It uses Jena library to implement most of the required methods for accessing the ontology, loaded in the reasoner. With this extension the component can acquire domain knowledge from ontology, defined in description logics, and achieve this way uniform case representation, what will enhance the interoperability of the whole system.

The development of a quite simple Case-Based Reasoning application already involves a number of steps, such as collecting case and background knowledge, modelling a suitable case representation, defining an accurate similarity measure, implementing retrieval functionality, and implementing user interfaces. Compared with other AI approaches, CBR allows to reduce the effort required for knowledge acquisition and representation significantly, which is certainly one of the major reasons for the commercial success of CBR applications.

4.2 Similar cases process retrieval

CBR systems typically apply retrieval and matching algorithms to a case base of past problem-solution pairs. CBR is based on the intuition that new searches are often similar to previously encountered searches, and therefore, that past results may be reused directly or through adaptation in the current situation.

In our system a new search is solved by retrieving one or more previously experienced cases, reusing the case, revising. The case-based reasoning-cycle in OntoFAMA may be described by the following processes.

- Retrieval. Main focus of methods in this category is to find similarity between cases. Similarity function can be parameterized through system configuration.
- Reuse: a complete design where case-based and slot-based adaptation can be hooked is provided.
- Revise the proposed solution if necessary. Since the proposed result could be inadequate, this process can correct the first proposed solution.
- Retain the new solution as a part of a new case. This process enables CBR to learn and create a new solution that should be added to the knowledge base.

In Jcolibri once the process is modelled, three modules are used for diagnosis: precycle, cycle, and postcycle. The CBR methodology as follows, Figure 4.

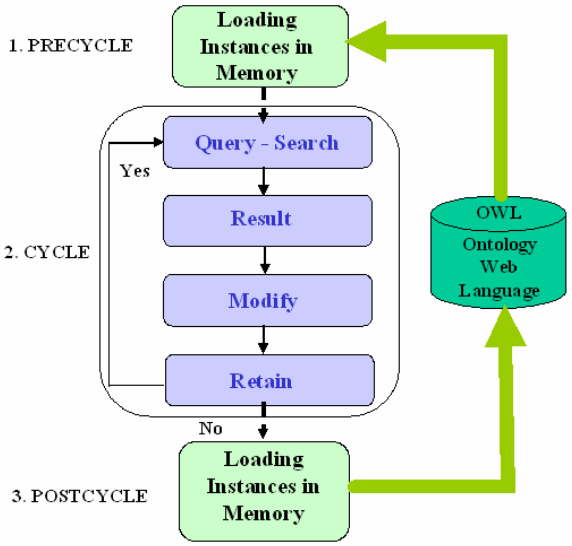

Figure 4: Search solving phases with CBR

Since the problem solving methods are domain independent, the domain specific information should be first loaded from the persistence media so that processing with it is possible (Díaz-Agudo & González-Calero, 2007). The data base connector will read the values in the table and if encounters a concept typed attribute it looks for an instance with the same name in the Ontology. Once found the connector will fill the values of the attribute of each case with the corresponding instances of the ontology, loaded by the Pellet reasoner. It is used as well by the methods to compute the content-based similarity between the concepts typed attributes.

5 ONTOLOGY DESIGN AND DEVELOPMENT

Ontologies are being developed to facilitate knowledge sharing and reuse and are seen as key enablers for Digital Library and Semantic Web (Staab & Studer, 2005). Key benefits of using semantic Web technology in the current digital libraries include:

- An integrated, coordinated and richly-interconnected repository of knowledge of its libraries.
- Transferring knowledge in an economic and scalable way to society.
- Providing a unique point of access for all people interested in information.
- The ontology guarantees interoperation between different applications, allowing easy addition of new ones.
- Possibility to export knowledge and applications to different library areas and dominions.
- Easy interoperation is possible with others services and resources of another digital library.

From a knowledge engineering perspective, ontologies are constructed using specialization generalization relationships to form their taxonomies and using other semantic relationships to extract the meaning of concepts and factual knowledge of a domain. OntoFAMA project contains a collection of codes, visualization tools, computing resources, and data sets distributed across the grids, for which we have developed a well-defined ontology using RDF language (W3C, 2009). RDF is used to define the structure of the metadata describing digital library resources.

Ontology provides a shared understanding to support communication among human and computer agents, typically being represented in a machine-processable language. To achieve a standard representation we adopt semantic web language such as RDF as the representation syntax of metadata, enabling RDF representation of CBR cases to provide a standard means of representation (Gomez-Perez & Corcho, 2003).

The primary information managed in the OntoFama domain is metadata about library resources, such as books, digital services, etc. We integrated three essential sources to the system: electronic resources, catalogue and personal Data Base. We wrote the description of these classes and the properties in RDF semantic markup language. For the manual generation and modelling of the domain ontology we chose the Protégé editor (Protégé, 2009).

Figure 5 shows the high level classification of classes to group together OntoFAMA resources as well as things that are related with these resources.

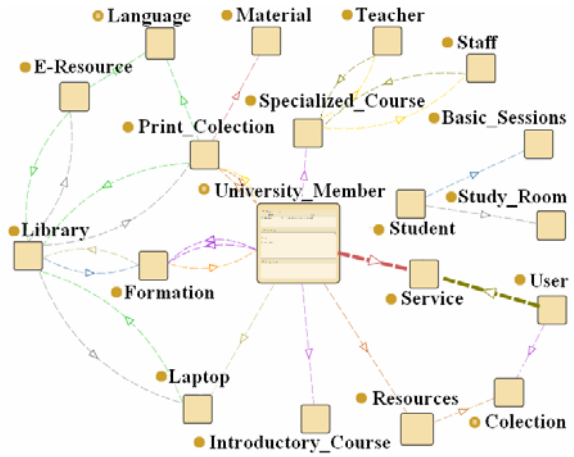

Figure 5: Class hierarchy for the OntoFama ontology

6 SYSTEM FUNCIONALITY

As we have seen in previous sections our system has a graphical user interface for determining initial user requirements early in search. Managing user requirements by placing focus on identifying, gathering, and documenting essential information is a specialized work area or user profiles. This action permits to reduce useless information or completely avoided in the search engine process. There is therefore a need to define, and describe the initial requirements of the user. In the case of not defining user requirements for a search the system presents a default configuration.

Rather than building static user profiles, contextual systems try to adapt to the user's current search. The user's search is monitored by capturing information from the different user profiles. OntoFAMA monitors user's tasks, anticipates search-based information needs, and proactively provide users with relevant information. This configuration contains the user requirements most typically described the relative needs, tasks, and goals of the user for an individual search. For this a statistical analysis has been done to determine the importance values and establishing specified user requirements. This statistical analysis even can in fact lay the foundation for searches in a particular user profile.

The user begins the search devising the starting query Q. In the example shown in the following let us suppose he/she starts with Q = "computer Science books". The outcomes represented in the following table display the number of important documents retrieved in OntoFama and the total number of documents retrieved in a traditional search engine and the values of precision and recall obtained. The results include a list of web pages with titles, a link to the page, and a short description showing where the keywords have matched content within the page. Ordered by relevance with the result that OntoFAMA considers the most important, Figure 6.

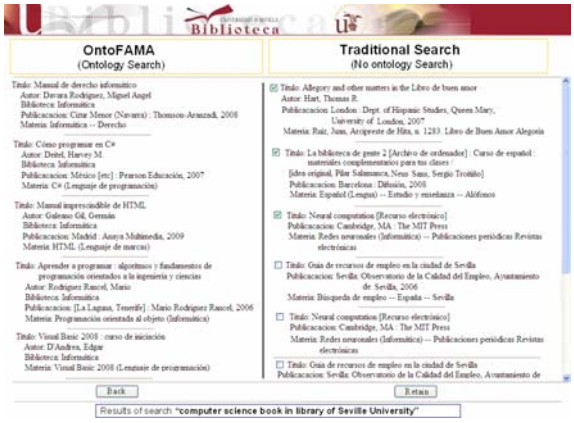

Figure 6: Search engine results page

The retrieval process identifies the features of the case with the most similar query. Our Inference Engine contains the CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text. The system uses similarity metrics to find the best matching case. We used a computational based retrieval where numerical similarity functions are used to assess and order the cases regarding the query. The retrieval strategy used in our system is nearest-neighbour approach. This approach involves the assessment of similarity between stored cases and the new input case, based on matching a weighted sum of features. A typical algorithm for calculating nearest neighbour matching is next:

$$similarity(Case_I, Case_R) = \frac{\sum_{i=1}^{n} w_i \times sim(f_i^I, f_i^R)}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Where w_i is the importance weighting of a feature (or slot), *sim* is the similarity function of features, and f_i^I and f_i^R are the values for feature *i* in the input and retrieved cases respectively.

An important advantage of similarity-cased retrieval is that if there is no case that exactly matches the user's requirements, this can shown the cases that are most similar to her query. The use of structured representations of cases requires

approaches for similarity assessment that allow to compares two differently structured objects, in particular, objects belonging to different object classes.

7 PERFORMANCE TESTING

Experiments have been carried out in order to test the efficiency of Artificial Intelligent and Ontologies in retrieval information in a digital library. These are conducted to evaluate the effectiveness of run-time ontology mapping. The main goal has been to check if the mechanism of query formulation, assisted by an agent, gives a suitable tool for augmenting the number of significant documents, extracted from the Digital Library, to be stored in the CBR.

The library of cases (the "case base") is initially generated from a file store where each case is represented with RDF syntax. 1100 cases were collected for user profiles and their different resources and services. This is sufficient for our proof-of-concept demonstration, but would not be sufficiently efficient to access large resource sets. Each case contains a set of attributes concerning both metadata and knowledge. However, our prototype is currently being extended to enable efficient retrieval directly from a database, which will enable its use for large-scale sets of resources.

Due to the complexity of searches, users may not be able to formulate all the considerations relevant to their resource choices in advance, it is necessary to guide the user at each step of the search. Besides, it has been tested also how many steps are necessary for retrieving the most of the important documents for the user, filtering the queries through the profiles user.

During the experimentation, heuristics and measures that are commonly adopted in Information Retrieval have been used. While the users were performing these searches, an application was continually running in the background on the server, and capturing the content of queries typed and the results of the searches. Statistical analysis has been done to determine the importance values in the results, figure 7.

Figure 7: OntoFAMA search analysis report.

For our experiments we considered 50 users with different profiles. So that we could establish a context for the users, they were asked to at least start their essay before issuing any queries to OntoFAMA. They were also asked to look through all the results returned by OntoFAMA before clicking on any result. We compared the top 10 search results of each keyword phrase per search engine. Our application recorded which results on which they clicked, which we used as a form of implicit user relevance in our analysis. We must consider that retrieved documents relevance is subjective. That is different people can assign distinct values of relevance to a same document. In our study we have agreed different values to measure the quality of retrieved documents, excellent, good, acceptable and poor.

After the data was collected, we had a log of queries averaging 5 queries per user. Of these queries, some of them had to be removed, either because there were multiple results clicked, no results clicked, or there was no information available for that particular query. The remaining queries were analyzed and evaluated. In each experiment we report the average rank of the user-clicked result for our baseline system, Google and for our search engine OntoFAMA. Then we calculated the rank for each retrieval document by combining the various values and comparing the total number of extracted documents and documents consulted by the user (table 1).

Table 1: Analysis of relevance of retrieved documents for select queries

	Excellent	Good	Acceptable	Poor
OntoFAMA	5,5 %	39,3 %	40,6 %	14,4 %
Google	2,7 %	31 %	44,8 %	21,3 %

We can observe the best final ranking was obtained for our prototype OntoFAMA and an interesting improvement over the performance of Google.

8 CONCLUSIONS AND FUTURE WORK

In this study, we addressed the main aspects of a semantic Web information retrieval system architecture trying to answer the requirements of the next-generation semantic Web user. An ontology and integrated intelligent system architecture for search operation support system and its implementation platform have been developed in this paper. We presented a system based in an Ontology and Artificial Intelligent architecture for knowledge management in the Seville Digital Library. It introduced a web-based CBR retrieval system which operates on an RDF file store. This system combines RDF representation and CBR recommendation methodology to do code selection for the resources codes; thus it applies a CBR approach with RDF data model.

A prototype implementation that uses caching and fat operations is implemented and an intelligent agent was illustrated for assisting the user by suggesting improved ways to query the system on the ground of the resources in a Digital Library according to his own preferences, which come to represent his interests.

Evaluation results have illustrated the feasibility of our approach. The test results show that the proposed service is a feasible solution that fields predictable performance in terms of response time and scalability.

A decisive role in it plays the jColibri-based and Protégé components that are the principal elements in the proposed architecture. Because jColibri is domain independent, and the domain-specific information for the system is captured entirely in the RDF ontology and ontology instances, the developed system could be easily transferred to other domains as well.

Future work will concern the exploitation of information coming from others libraries and services and further refine the suggested queries, to extend the system to provide another type of support, as well as to refine and evaluate the system through user testing. It is also necessary the development of an authoring tool for user authentication, efficient ontology parsing and real-life applications.

REFERENCES

Govedarova, D., Stoyanov S., Popchev, I., 2008. *An Ontology Based CBR Architecture for Knowledge Management in BULCHINO Catalogue*. International Conference on Computer Systems and Technologies.

Ding, H., 2004. *Towards the metadata integration issues in peer-to-peer based digital libraries*. GCC (H. Jin, Y. Pan, N. Xiao, and J. Sun, eds.), vol. 3251 of Lecture Notes in Computer Science, Springer.

GAIA - Group for Artificial Intelligence Applications, 2009. *jCOLIBRI project - Distribution of the development environment with LGPL*, http://gaia.fdi.ucm.es/grupo/projects/. Complutense University of Madrid.

Witten, I. H., and Bainbridge, D., 2003. *How to Build a Digital Libary.* Morgan Kaufmann.

Stuckenschmidt, H., and Harmelen, F. van., 2001. *Ontology-based metadata generation from semi-structured information.* K-CAP, pp. 163–170, ACM.

Warren, P. 2005. *Applying semantic technologies to a digital library: a case study*" Library Management Journal, Emerald, vol. 26, no. 4/5, pp. 196–205

Taniar, D., Wenny Rahayu, J., 2006. *Web semantics and ontology*. Hershey, PA: Idea Group Pub, 2006.

Toussaint, J., Cheng, K., 2006. *Web-based CBR (case-based reasoning) as a tool with the application to tooling selection.* International Journal of Advanced Manufacturing Technology.

Luger, George F., 2002.*Artificial Intelligence, Structures and Strategies for Complex Problem Solving*. 4ª edition. Ed. Pearson Education Limited.

Sure, Y., and Studer, R., 2005.*Semantic web technologies for digital libraries*. Library Management Journal, Emerald, vol. 26, no. 4/5, pp. 190–195.

Quan, D., and Karger, D. R., 2004. *How to make a semantic web browser.* Proceedings of WWW2004.

Bridge, M., G¨oker, H., McGinty,L., Smyth, B. 2006.*Case-based recommender systems.* Knowledge Engineering Review.

Díaz-Agudo, B., González-Calero, P.A., Recio-García, J., Sánchez-Ruiz, A., 2007.*Building CBR systems with jColibri.* Journal of Science of Computer Programming.

Staab, S., Studer, R., 2005. *Handbook on Ontologies.* International Handbooks on Information Systems, Springer, Berlin.

W3C, 2009. *RDF Vocabulary Description Language 1.0: RDF Schema*. http://www.w3.org/TR/rdf-schema/.

Gomez-Perez, A., Corcho, A., O., Fernandez-Lopez, M., 2003. *Ontological Engineering. Advanced information and knowledge processing*, Berlin: Springer.

PROTÉGÉ, 2009. *The Protégé Ontology Editor and Knowledge Acquisition System.* <http://protege.stanford.edu/>.