A Knowledge-Intensive CBR Application for Content-Based Retrieval in BULCHINO Catalogue

Stanimir Stoyanov[1], Nadezhda Govedarova[2], Ivan Popchev[2]

[1] University of Plovdiv, eCommerce Lab, Department of Computer Systems,
236 Bulgaria Blvd., 4003 Plovdiv, Bulgaria, www.fmi-plovdiv.org

[2] Bulgarian Academy of Sciences, Institute of Information Technologies,
Acad. G. Bonchev St., Block 2 , 1113 Sofia, Bulgaria, www.iit.bas.bg

stani@uni-plovdiv.bg, nadig@abv.bg, ipopchev@iit.bas.bg

Abstract. Knowledge-based systems (KBS) achieve their reasoning power mainly through the explicit representation and use of different kinds of knowledge about a certain domain. The comparison of different KBS revealed, however, that the design solutions for particular knowledge representation infrastructure differ, depending on the information to be integrated. This observation brings to the conclusion that KBS are often implemented in an ad hoc manner. In this paper we discuss the design of a knowledge-intensive application in the Cultural heritage domain. We present its architecture and underlying development methodology. We pay special attention to the knowledge model that is designed in compliance with the main principles of Semantic Web.

1 Introduction

Our ongoing work [4, 5, 14, 15, 16] on the development of CBR-based approaches for realizing content-based search is mainly targeted at discovering new interesting and powerful technological integrations that permit these technologies to exploit each others` strengths in order to mitigate their weaknesses. Our hypothesis is that case-based reasoning (CBR) [1, 11] supported by a rich knowledge model is a promising approach for achieving semantics aware search and retrieval.

For several years we have been working on the project “CBR Platform for Ontology-Based Knowledge Management”, funded by the Bulgarian Ministry of Education and Science. The project aims at the design and development of a flexible and extensible CBR-based platform for knowledge management and processing. We intend to apply it by the implementation of prototype applications initially in the field Cultural heritage. As a result, we developed a hybrid ontology and CBR-based search engine architecture for the Cultural heritage domain. An essential role in it plays the knowledge model that is designed on the basis of the framework for ontology-based information sharing, presented in [6]. It is structured in three layers: data storage layer (database), metadata repository (i.e. case base); domain conceptualization (ontology).

Our current work is focused on the implementation of a particular application to perform content-based search in BULCHINO catalogue. BULCHINO is a Web-based catalogue for electronic presentation of Bulgarian cultural heritage that has been developed in the E-Commerce Laboratory at the University of Plovdiv. In this paper we present our design solutions and development methodology. We demonstrate, as well, their practical applicability by means of a concrete example from BULCHINO catalogue.

The remainder of the paper is organized as follows: Section 2 describes the setting of Cultural heritage domain, the research problems and current work in it. Section 3 focuses on the application knowledge model. Here we discuss how the main principles, underlying Semantic Web, could be applied in our system where we point out their importance to its design in terms of interoperability and information sharing. The architecture itself is presented in Section 4, where we outline the main components and their functionality. Other technological aspects, concerning the interaction between the components of the architecture, are also concerned. In Section 5 we explain how the representation of a concrete cultural object could be mapped over the knowledge model. Section 6 concludes the paper.

2 Intelligent information processing in Cultural heritage domain

One of most important aspects of intelligent information processing is the ability to understand semantics of the information, i.e. the intended meaning of terms in special context or application [6]. The formalization and representation of the context and taking advantage of it by performing the retrieval process are, therefore, the main research challenges in designing and implementing content-based search engines. Current studies in Knowledge-Based Systems (KBS) [7, 12] focus on the integration of numerous knowledge sources and techniques derived from the fields Artificial Intelligence and Knowledge Management areas to achieve it.

Cultural heritage domain is a privileged area for applying innovative, knowledge intensive applications for two main reasons. On the one hand, the slow digitalization process, the distributed databases and the heterogeneous description of the objects in different schemes significantly hamper the uniform semantic interpretation and hence the information sharing among systems. Interoperability, however, is namely this essential precondition for promotion of cultural diversity that enhances its acceptance and recognition, and enables the dialog and mutual understanding among different cultures. Thus, the digital storage and representation of cultural heritage in compliance with the up-to-date standards has become, according to UNESCO, one of the most pressing contemporary issues. Some of the current projects that address these issues are eCULTURE, HUBUSKA, CATCH, CHIP, STICH, FinnOnto, MICHAEL to mention just a few of them. The long work on the above mentioned problems have led, on the other hand, to the development of numerous domain specific standards, vocabularies, thesauri and common practices for modeling the cultural content. These, along with the knowledge abundance that usually accompanies cultural collections (e.g. archives, images) serve as a sound foundation for KBS in this domain.

In the following we discuss in more details the definitions and standards that we used by designing the particular knowledge model of BULCHINO catalogue, presented in the next Section. We took as a starting point the definitions of UNESKO and CCO (Cataloging Cultural Objects) for cultural field and cultural object (CO) respectively. To ensure the common acceptance of our model we envisage the use of control vocabularies for describing the COs and the integration of ontology for representing the domain context.

2.1 UNESKO Classification

According to the UNESCO classification from 2002 the cultural field is defined as comprising the following categories:

- Cultural monuments;
- Natural sites;
- Traditions;
- Customs;
- Balneotherapy settlements;
- Ethnography and folklore;
- Cultural practices and activities;
- Native crafts;
- Native cuisine and wines;
- Traditional sorts of fruits and vegetables.

We extended this classification with additional categories which are specific for the Bulgarian cultural heritage to achieve more precise subject framework for the concrete system. On its base we developed the conceptual model of the domain ontology, part of which is presented in Section 5 (Fig.3).

Two groups of standards for describing the cultural objects from all of the above listed categories should be differentiated. The first one comprises data content standards (metadata standards) – specify the structure (the set of elements) to be used for representing a given CO. Some examples of such standards are CCO and VRA Core. It should be noted that they are designed for different purposes, CCO for record creation and VRA Core for record sharing. The former focuses on descriptive metadata only, while the latter includes some administrative elements as well. The second group consists of data value standards that are used for guiding the choice of terms and words to fill in the structure. An example of this group is the Getty vocabularies. The CIDOC-CRM could be also considered as data-value standard as far as the use of ontology in terms of a shared vocabulary is concerned.

2.2 Standards

- Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images (CCO) [20] is a manual for describing, documenting, and cataloging cultural works and their visual surrogates. It defines a cultural object as "a distinct intellectual or artistic creation limited primary to objects and structures made by humans, including built works, visual art works and cultural artifacts". According to it every CO could be described by minimal set of elements (Work record), which is a subset of VRA Core Categories. The Work record comprises:
  - *Class*- is used to relate a specific work to others with similar characteristics, often based on the organization scheme of a particular repository or collection;
  - *Work type*-identifies the kind of work being described (e.g. sculpture, altarpiece, cathedral, painting, etc.);
  - *Title*-records the titles, identifying phases or names, given to a work of art or architecture (e.g. Ceramic Bowl);
  - *Creator*-identifies the individual, group of individuals, corporate body, cultural group, or other entity that contributed to the creating, designing, production or altering the work;
  - *Creator role*-records the role or activity performed by the creator in the conception, design or production of the work being catalogued;
  - *Ddate*-records the date or range of dates associated with the creation, design, production, presentation or alternation of the work;
  - *Subject*- contains an identification, description, or interpretation of what is depicted in and by a work;
  - *Style*-identifies the defined style, historical or artistic period, movement, school whose characteristics are represented in the work;
  - *Culture*- contains the name of the culture, people or nationality from which the work originates;
  - *Current location* - includes the geographic location of the work or repository that currently houses the work;
  - *Measurements*-contains information about the dimensions, size or scale of the work;
  - *Materials and techniques*-includes the substances or materials used in creation of a work;
  - *Description*- consist of a declarative note that is generally a relatively brief essay-like text, detailing the content and context of the work;
  - *Description source.*

- VRA Core Categories [18]-data standard for the cultural heritage community. It consists of a metadata element set, as well as an initial blueprint for how those elements can be hierarchically structured. To this end, the VRA Data Standards Committee has developed as XML Schema for the VRA Core 4.0 metadata element set to be used primarily for record sharing and exchange purposes.

We envisage using this standard as a base for defining the case structure. In the current version of the knowledge model, however, only the elements prescribed by CCO are included in it.

- CIDOC Conceptual Reference Model (CRM) [21]-provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage domain. It is intended to promote a shared understanding of cultural information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. Since 9/12/2006 it is official standard ISO 21127:2006.

2.3 Getty vocabularies

The Getty vocabulary databases are compliant with ISO and NISO standards for thesaurus construction. They contain terms, names and information about people, places, things and concepts relating to art, architecture and material culture.

- The Art &Architecture Thesaurus (AAT);
- The Union List of Artist Names (ULAN);
- The Getty Thesaurus of Geographic Names (TGN).

3 Knowledge Framework

As discussed in the previous Section, Knowledge-Based Systems (KBS) achieve their reasoning power mainly through the explicit representation and use of different kinds of knowledge about a certain domain. The main principle underlying their organization and management consists in abstracting the knowledge items to characterizations (metadata descriptions), which are used for further processing. The representation of these characterizations is based on some vocabulary that is a shared conceptualization of the domain (ontology). This approach is presented in more details in [6].

The comparison of different KBS revealed, however, that the design solutions for particular knowledge representation infrastructure differs, depending on the information to be integrated (domain knowledge) as well as on the intended use of the application. This observation brings to the conclusion that KBS are implemented in an ad hoc manner what excludes the possibility for collaboration with other systems and information sharing among them – the main ideas, underlying Semantic Web [2, 6]. As a preventive measure to this tendency, a set of standards for metadata descriptions (Dublin Core and domain specific standards, based on it) and ontology representation (W3C standards [2,6]) have been defined to facilitate the designers by providing them with common accepted practices. Yet, they solve only partially the problem. Still remains the question concerning their relevance to the problem at hand and the possibility to apply them for modelling it. Another problematic issue is the choice of software tools to support the use of these standards with regard to their integration and compatibility with existing technologies. The development of new reasoning techniques that take advantage of the explicitly represented context is also an open topic of research. Taking into consideration these issues, as a general conclusion imposes the assumption that the definition of design methodologies is an essential phase in the development of KBS.

In this Section we present a framework for information representation in Cultural heritage domain (Fig.1). The design of the framework is motivated by the need of a knowledge model for our application that takes into consideration the basic principles, underlying Semantic Web. Our approach implies the application of the CBR technology for metadata characterization of the cultural objects (presented as cases and stored in a case base) and the integration of ontology for semantics formalization of these characterizations.

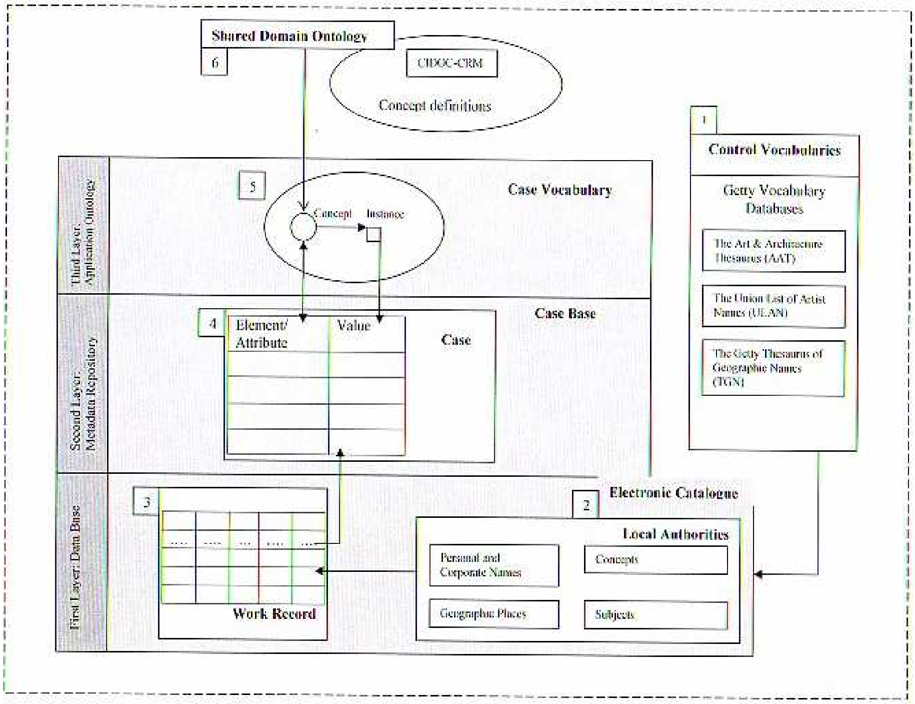

Fig.1. Knowledge representation framework for Cultural heritage domain

3.1 First layer-Database

The Database is designed on the basis of CCO. Apart from defining the basic element set for each CO (3: Work record), it recommends the development of local authorities (2: Subject, Concept, Geographic, Personal and corporate authorities) that are populated with terminology from standard published controlled vocabularies (1: Getty vocabularies) as well as with local terms and names. The terminology related to the works and images and used for describing them is stored namely in these authority files.

- Subject authority–contains subjects from the iconographic terminology, names from literature, mythological and religious nature, historical event;
- Concept authority–contains terminology, needed for the description of work of art, the material that the work is made of, the activities involved with the work, style, role of the creator, etc;
- Geographic place authority–contains information about geographical locations of the cultural works and their creators;
- Personal and corporate name authority–contains information about architectural work, the individuals and corporations, connected with the cultural-historical work;
- Image record–contains an image of the cultural work. It should be noted we do not take into consideration the visual representations of the works.

We have been, currently, working on the development environment BECC (Bulgarian Electronic Cataloguing Cultural) for building the authority editors.

3.2 Second layer-Metadata Repository

The metadata repository, i.e. the Case Base contains cases, whose structure is defined on the basis of the CCO Work record. A case, therefore, consist of attribute-value pairs where the attributes are defined by the elements listed in Section 2.2 and their corresponding values are filled from the database. We studied as well the possibility for using ontology by designing the case structure.

The usage of ontology is useful for the CBR community regarding different purposes [9,12]: persistence of cases and/or indexes using individuals or concepts that are embedded in the ontology itself; as the vocabulary to define the case structure, either if the cases are embedded as individuals in the ontology itself, or if the cases are stored in a different persistence media as a data base; as the terminology to define the query vocabulary; retrieval and similarity, adaptation and learning. At this stage we consider using ontology only as vocabulary for filling in the case structure.

It should be noted that different case structures could be defined over this model:

- Standard cases-are composed only by attributes with simple data types (e.g. String, Integer). An example is presented in [13];
- Hybrid cases-are composed by attributes some of which represent concepts from the domain ontology;
- Pure ontology-based cases-consists only of attributes, representing concepts from the ontology.

Subject of discussion in this paper are only hybrid cases. We use Concept data type (supported in the application architecture- Section 4) to indicate that an attribute is going to represent a concept of the ontology. The values of this attribute are going to be the corresponding instances of the linked concept. More about the mapping between the different layers and the instantiation of the case is explained in Section 4.

3.3 Third layer-Ontology

Most of the ontology-based integration approaches use ontologies for the explicit description of the information-source semantics. There are, however, different ways of how to employ them. In our model we applied the hybrid approach (Fig.2), proposed by Frank van Harmelen in [6]. According to it the semantics of each source (cultural collection in our case) is described by its own ontology (local ontology-LO). In order to make the source ontologies comparable to each other they are built upon global shared vocabulary, which contains basic terms of a domain.

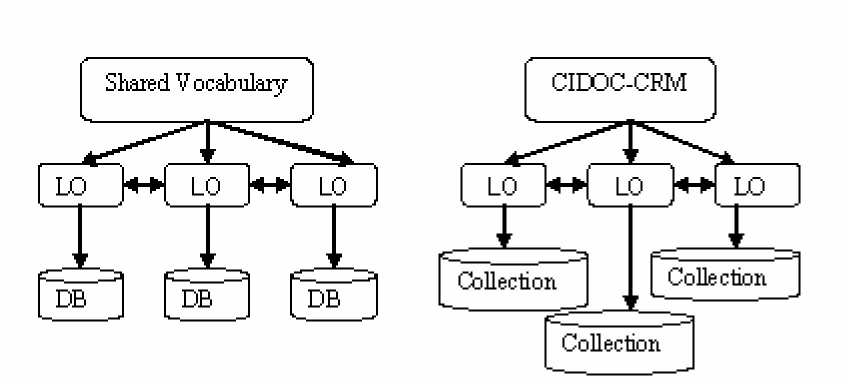

Fig. 2. Ontology design approach

For shared vocabulary we chose the CIDOC referent model, and the different collections are defined in accordance to the application ontology model, part of which is presented in Section 5. It is based on the UNESCO classification, where a single collection corresponds to a particular category.

4 Application architecture

For the implementation of the application prototype we chose the framework jCOLIBRI [8,9,10]-a Java-based framework that supports the development of knowledge intensive CBR systems and help in the integration of ontology in them. 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: CAT-CBR, CBR*Tools, IUCBRF, Orenge in several aspects: availability (open source framework), implementation (the Java implementation is one of our main requirements with respect to the easy integration in the BULCHINO system which is implemented in J2EE environment), GUI (the provided graphical tools facilitate the system design). Another decision criterion for our choice is connected with the fact that jCOLIBRI affords the opportunity to incorporate ontology in the CBR system to use it for case representation and content-based reasoning methods to assess the similarity between them. The ontology support of jCOLIBRI is built arround the OntoBridge library to easily manage ontologies and DLs (Description Logics) reasoners. It is based on the Jena framework for development of Semantic Web applications [17] and makes possible the connection with several DLs reasoners. Some of the actively supported once are presented in [19]: CEL, FaCT++, fuzzyDL, KAON2, MSPASS, Pellet, QuOnto. We chose Pellet as inference engine in our architecture for several reasons. Apart from being open source and Java based, it supports Jena interface what enables the direct connection with the OntoBridge library of jCOLIBRI. For the manual generation and modeling of the domain ontology we chose the Protégé editor.

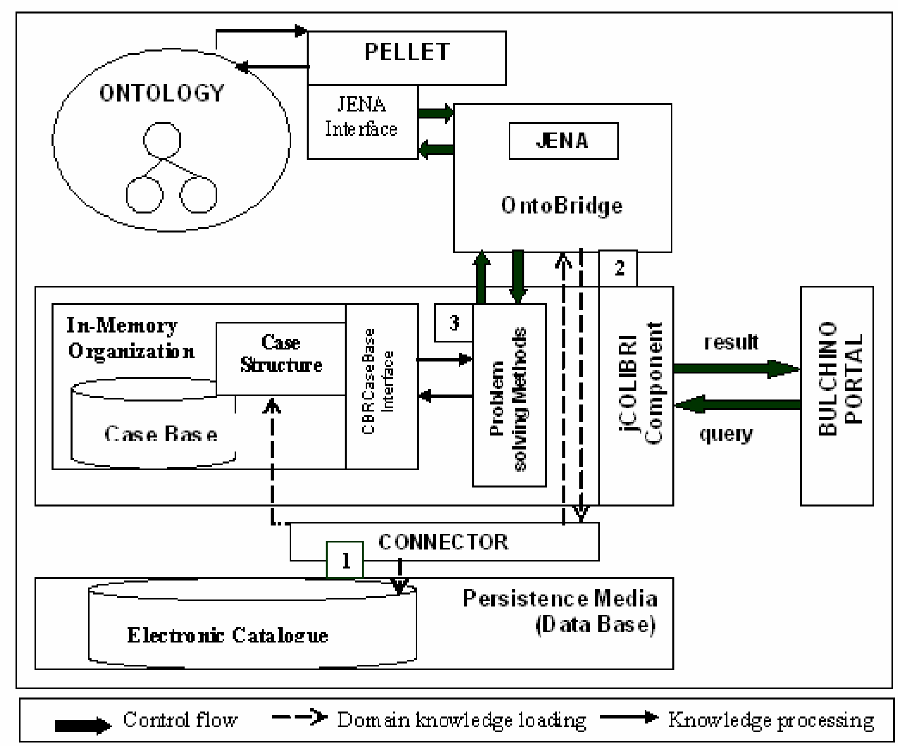

Fig.3. Application Architecture

- Persistence Media - in this case we are not storing the whole case base into the ontology. Data Base continues storing the values of the attributes. But now the column that stores the attribute linked with a concept is going to have the names (stored as strings of characters) of the instances of this concept. Ontology stores information about cultural objects where concepts are types, or classes, individuals are allowed values, or objects and relations are the attributes describing the objects. Knowledge is formally represented and stored in OWL (Web Ontology Language) ontology that is managed with the Pellet reasoner;
- Connectors- objects that 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. They realize the mapping between the layers;
- jCOLIBRI Component - is designed on the basis of jCOLIBRI framework.It is the core of the architecture and is responsible for the knowledge and queries

management. 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. The data base connector will read the values in the table (1) and if encounters a concept typed attribute it looks for an instance with the same name in the Ontology (through OntoBridge) (2). 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 (3) to compute the content-based similarity between the concept typed attributes;

- OntoBridge - is a Java library that eases the management of the ontology in a jCOLIBRI-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 jCOLIBRI component can acquire domain knowledge from ontology and achieve this way uniform case representation, what will enhance the interoperability of the whole system;
- Pellet – is an open source Java-based OWL-DL reasoner.

5 Example

In this Section we explain briefly how the representation of a concrete cultural object could be mapped over the knowledge model, where we take into consideration only the upper layers – metadata repository and ontology.

5.1 Scope

We chose the Boyana church object. It is one of the Bulgarian cultural sight and part of our cultural heritage. It could be classified as a cultural monument (category in UNESCO classification) and as a temple, more specifically (kind of cultural monument). In compliance with CCO it could be described by means of class, type, location, creator, date, etc. It is obvious that the CCO “class” element corresponds to the UNESCO classification category “cultural monuments” and “type” to “temples” respectively. We made use namely of this observation by designing the application ontology. We defined ID, Class, Type as concept typed attributes of the case (Fig.3).

5.2 Application ontology

By designing the application specific ontology we had to take into consideration the domain ontology, on the one hand, to achieve interoperability of the application, and the case structure on the other hand to ensure the exact matching between its concepts and concept typed case attributes. As a result we defined:

- Thing – the root of the ontology;
- CBR_Index concept – is described by means of sub-concepts that correspond to case elements (this way we ensure the matching between the case attributes and application ontology concepts).

- CBRCase concept – has only one subconcept – CulturalCase, which instances are the concrete cultural objects (e.g. case1-Boyana church);
- Relation – defines the relation between the two concepts (e.g. cultural case has-type type, i.e Boyana church has-type temple)

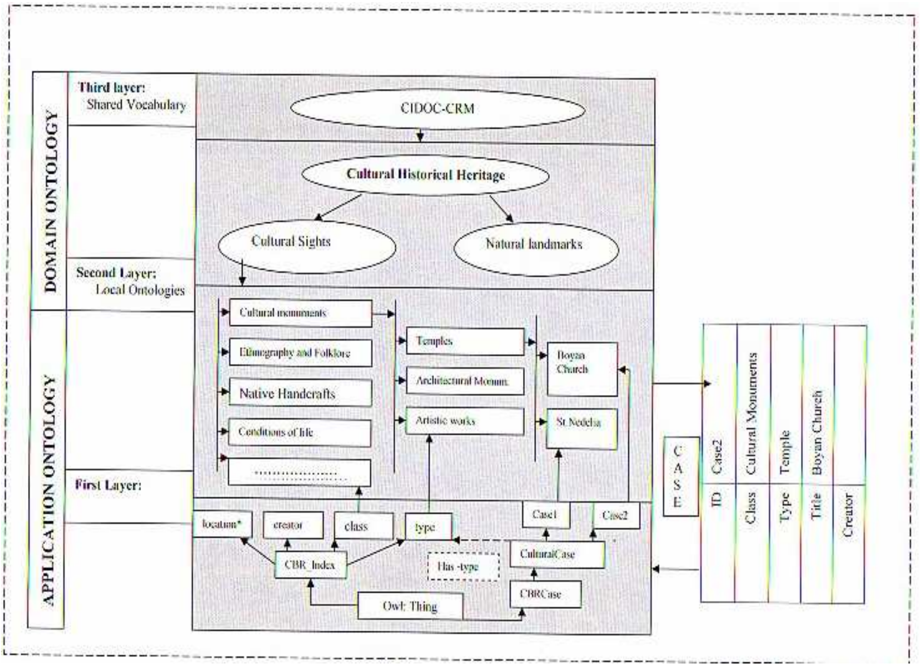

Fig.3. Example of matching between domain and application ontologies and case structure

6 Conclusion

We believe that by integrating case based reasoning with extensive knowledge bases, structured and implemented in compliance with the up to date Semantic Web technologies, we can develop tools for intelligent information retrieval. The main contribution of this paper is the presented search engine architecture that adopts an approach for knowledge representation, based on case-based metadata characterisations of the information and ontology conceptualisation of the domain at hand - Cultural heritage domain. Future work includes the specification of the architecture and the integration of actual cultural content in the developed knowledge models. This will help us to gather the experience and the experimental data about the architecture that will be necessary to evolve it, as well as discovering what new problems lie ahead.

Acknowledgements. The authors wish to acknowledge the financial support of the Bulgarian Ministry of Education and Science. Research Project Ref.No.MI-1502/2005.

References

1 Aamodt, A., Plaza, E., Case-base reasoning: Foundational issues, methodological variations, and system approaches, AI Communications, 7(1), 1994, 39-59

2 Antonio, G., F. Harmelen, A Semantic Web Primer, ISBN 0-262-01210-3

3 C. Doulaverakis, E. Nidelkou, A. Gounaris, Y. Kompatsiaris, An Ontology and Content Based Search Engine For Multimedia Retrieval, 10th East-European Conference on Advances in Databases and Information Systems, ADBIS 2006, Thessaloniki, Hellas, 3-7 September, 2006

4 Govedarova N., Stoyanov S., Popchev I., An Ontology-based CBR Architecture for Knowledge Management in BULCHINO Catalogue, In Proc. of the Conference "Computer Systems and Technologies 2008", 12-13 June, Gabrovo (Best paper award)

5 Govedarova N., Stoyanov S., Popchev I., Hybrid Ontology and CBR-based Search in BULCHINO Catalogue, In Proc. of the Conference "Informatics in the Scientific Knowledge", 26-28 June, Varna

6 Heiner Stuckenschmidt, Frank van Harmelen, Information Sharing on the Semantic Web, Springer Verlag, 2005, ISBN 3-540-20594-2

7 Díaz-Agudo, B., González-Calero, P.A., An Architecture for Knowledge Intensive CBR Systems, In Blanzieri, E., Portinale, L., (Eds.): Advances in Case-Based Reasoning (Procs. of the 5th European Workshop on Case-Based Reasoning, EWCBR 2000), Lecture Notes in Artificial Intelligence, 1898, Springer, 2000

8 jCOLIBRI Theoretical Foundations, http://gaia.fdi.ucm.es/projects/jcolibri/docs.html/

9 Juan A. Recio-García, Belén Díaz-Agudo, Pedro A. González-Calero, and Antonio Sánchez, Ontology based CBR with jCOLIBRI, Applications and Innovations in Intelligent Systems XIV.Proceedings of AI-2006, the Twenty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, pages 149–162, Cambridge, United Kingdom, December 2006. Springer

10 Juan A. Recio-García, Belén Díaz-AgudoPedro, González-Calero, jCOLIBRI2 Tutorial, Document version 1.1, Januery 22, 2008

11 Lenz, M., Bartsch-Sporl, B., Burkhard, H., and Wess, S. Case-Based Reasoning Technology – From Foundation to Applications,Lecture Notes in Artificial Intelligence 1400, Springer Verlag, 1998

12 Ralph Bergman, Martin Schaaf, On the Relation between Structural Case-Based Reasoning and Ontology-Based Knowledge management, In Proc. of German Workshop On Experience Management, April, 2003

13 Stoyanov,S., N.Govedarova, I.Popchev, CBR-based Search in BULCHINO Catalogue, In Proc CS&P`07, vol. 2, Pp 521-533

14 Stojanov,S., M. Trendafilova, CBR-Search in Electronic Catalogs, In Proc. of the International Conference "Automatics and Informatics '03", vol.1, Pp.65-68, 6-8 October,2003, Sofia, Bulgaria. ISBN 954-9641-34-1

15 Stojanov, S., D. Chaushkova, M. Trendafilova, Applying Case Based Reasoning for Generation of Tourist Routes, In Workshop "Concurrency, Specification & Programming", vol.3: Multiagent Systems and Applications, Pp. 563- 575, Caputh, September 24-26 2004,

16 Stojanov S., I. Popchev, D. Chaushkova, M. Trendafilova, A Case based reasoning Approach for Development of Intelligent Services, Journal "Information Technologies and Control", No. 3/2004, Year II, Pp. 31-34, ISSN 1312-2622

17 Jena Framework, http://jena.sourceforge.net/index.html

18 VRA Core elements http://www.vraweb.org/projects/vracore4/

19 Description Logics Reasoners, http://www.cs.man.ac.uk/~sattler/reasoners.html

20 CCO (Cataloging Cultural Objects), http://vraweb.org/ccoweb/cco/index.html

21 CIDOC-CRM, http://cidoc.ics.forth.gr