Intelligent Search Engine to a Semantic Knowledge Retrieval in the Digital Repositories

Antonio Martín
Department of Electronic Technology
Higher Technical School of Computer Engineering
Sevilla, Spain
toni@us.es

Abstract— **Currently, an enormous quantity of heterogeneous and distributed information is stored in the current digital libraries. This data abundance has made the task of locating relevant knowledge more complex. Such complexity drives the need for intelligent systems for searching and for knowledge retrieval. Access to these collections poses a serious challenge. The present search techniques based on manually annotated metadata and linear replay of material selected by the user do not scale effectively or efficiently to large collections. The Artificial Intelligence and Semantic Web provide a common framework that allows knowledge to be shared and reused. In this paper, we propose a comprehensive approach for discovering information objects in large digital collections. The process is based on analysis of recorded semantic metadata in those objects and the application of expert system technologies. We suggest a conceptual architecture for a semantic and intelligent search engine. We concentrate on the critical issue of metadata/ontology-based search. More specifically, the objective is investigated from a search perspective possible intelligent infrastructures form constructing decentralized digital libraries where no global schema exists. We have used Case Based-Reasoning methodology to develop a prototype for supporting efficient retrieval knowledge from digital library of Seville University. The work suggests a conceptual architecture for a semantic and intelligent search engine and we also have developed a prototype and tested it for supporting efficient retrieval knowledge from digital libraries.**

Keywords-Ontology; Semantic Web; Retrieval; Case-based Reasoning; Digital Library; Knowledge Management.

I. INTRODUCTION

A Digital Library (DL) enables users to interact effectively with information distributed across a network. These network information systems support search and display of items from organized collections. In the historical evolution of digital libraries the mechanisms for retrieval of scientific literature have been particularly important. Traditional search engines treated the information 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. Although search engines have developed increasingly effective, information overload obstructs precise searches. Despite large investments and efforts have been made, there are still a lot of unsolved problems. Thus, it is necessary to develop new intelligent and semantic models that offer more possibilities [1].

There are researchers and works in related fields, which include ontology retrieval methods. The study [2] presents a system, which uses an ontology query model to analyze the usefulness of ontologies in effectively performing document searches. This work proposes an algorithm to refine ontologies for information retrieval tasks with preliminary positive results. [3] uses a medical ontology to improve a Multimodal Information Retrieval System by expanding the user's query with medical terms. The study [4] combines swarm intelligence and Web Services to transform a conventional library system into an intelligent library system with high integrity, usability, correctness, and reliability software for readers. The research [5] proposes meta-concepts with which the ontology developers describe the domain concepts of parts libraries. The meta-concepts have explicit ontological semantics, so that they help to identify domain concepts consistently and structure them systematically. The study [6] presents a formulation and case studies of the conditions for patenting content-based retrieval processes in digital libraries, especially in image libraries. The paper [7] focuses on methods for evaluating different symbolic music matching strategies, and describes a series of experiments that compare and contrast results obtained using three dominant paradigms. The research [8] proposes organizational memory architecture and annotation and retrieval information strategies. This technique is based on domain ontologies that take in account complex words to retrieve information through natural language queries.

There are a lot of researches on applying these new technologies into current information retrieval systems, but no research addresses Artificial Intelligence (AI) and semantic issues from the whole life cycle and architecture point of view [9]. Although search engines have developed increasingly effective, information overload obstructs precise searches. Our work differs from related projects in that we build ontology-based contextual profiles and we introduce an approaches used metadata-based in ontology search and expert system technologies [10]. We presented an intelligent approach for optimize a search engine in a specific domain. This study improves the efficiency methods to search a distributed data space like DL. The objective has focused on creating technologically complex environments digital repositories domain. It incorporates Semantic Web and AI technologies to enable not only precise location of public resources but also the automatic or semi-automatic learning [11].

Our approach for realizing content-based search and retrieval information implies the application of the Case-Based Reasoning (CBR) technology [12]. Thus, our objective here is to contribute to a better knowledge retrieval in DL field. This paper describes semantic interoperability problems and presents an intelligent architecture to address them, called OntoSDL. Obviously, our system is a prototype but, nevertheless, it gives a good picture of the on-going activities in this new and important field. We concentrate on the critical issue of metadata/ontology-based search and expert system technologies. More specifically, the objective is investigated from a search perspective possible intelligent infrastructures form constructing decentralized public repositories where no global schema exists.

The contributions are divided into next sections. In the first section, short descriptions of important aspects in DL domain, the research problems and current work in it are reported. Then, we summarize its main components and describe how can interact AI and Semantic Web to improve the search engine. Third section focuses on the ontology design process and provides a general overview about our prototype architecture. Next, we study the CBR framework jColibri and its features for implementing the reasoning process over ontologies [13]. Finally, we present conclusions of our ongoing work on the adaptation of the framework and we outline future works.

II. MOTIVATIION AND REQUIREMENTS

In the historical evolution of DL, the mechanisms for retrieval information and knowledge have been particularly important. These network information systems support search and display of items from organized collections. Reuse this knowledge is an important area in this domain. The Semantic Web provides a common framework that allows knowledge to be shared and reused across community users [14].

Repositories and digital archives are privileged area for the application of innovative, knowledge intensive services that provide a flexible and efficient method for searching information and guarantee the user with a set of results actually related to his/her interest. Seville University institutional repository is dedicated to the production, maintenance, delivery, and preservation of a wide range of high-quality networked resources for citizens, scholars, and students at University and elsewhere. This repository includes services to effectively share their materials and provide greater access to digital content [15].

Thus, the goal is to contribute to a better knowledge retrieval in the institutional repositories dominium. This scheme is based on the next principles: knowledge items are abstracted to a characterization by metadata description, witch is used for further processing. This characterization is based on a vocabulary/ontology that is shared to case the access to the relevant information sources. This begets new challenges to docent community and motivates researchers to look for intelligent information retrieval approach and ontologies that search and/or filter information automatically based on some higher level of understanding are required. We make an effort in this direction by investigating techniques that attempt to utilize ontologies to improve effectiveness in information retrieval. Thus, ontologies are seen as key enablers for the Semantic Web. We have proposed a method to efficiently search for the target information on a digital repository network with multiple independent information sources [16]. The use of AI and ontologies as a knowledge representation formalism offers many advantages in information retrieval [17]. In our work, we analysed the relationship between both factors ontologies 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 within the information representation [18]. We discuss an opportunity and challenge in this domain with a specific view of intelligent information processing that takes into account the semantics of the knowledge items. In this paper, we study architecture of the search layer in this particular dominium, a web-based catalogue for the University of Seville. The hypothesis is that with a case-based reasoning expert system and by incorporating limited semantic knowledge, it is possible to improve the effectiveness of an information retrieval system [19]. More specifically, the objectives are decomposed into:

- Explore and understand the requirements for rendering semantic search in an institutional repository.
- Investigate how semantic technologies can be used to provide additional semantic properties from existing resources.
- Analyse the implementation results and evaluate the viability of our approaches in enabling search in intelligent-based digital repositories.

To reach these goals we need to consider information interoperability. In other words, the capacity of different information systems, applications and services to communicate, share and interchange data, information and knowledge in an effective and precise way. As well, in order to deliver new electronic products and services, ontologies can be used to integrate with other systems, applications and services. DL initiatives, such as interoperability between public services, require establishing collaborative semantic repositories among public and private sector organizations. Particularly, we require Semantic Interoperability, which is one of the key elements of the programme to support the set-up of the European E-Government services.

III. INTEROPERABILITY REQUIREMENTS

In June 2002, European heads of state adopted the Europe Action Plan 2005 at the Seville summit. It calls on the European Commission to issue an agreed interoperability framework to support the delivery of European E-Government services to citizens and

enterprises. This recommends technical policies and specifications for joining up public administration information systems across the EU. This research is based on open standards and the use of open source software. These aspects are the pillars to support the European delivery of E-Government services of the recently adopted European Interoperability Framework (EIF) [20] and its Spanish equivalent [21]. This document is reference for interoperability of the new Interoperable Delivery of Pan-European E-Government Services to Public Administrations, Business and Citizens programme (IDAbc). European Institutions and agencies should use the European interoperability framework for their operations with each other and with citizens, enterprises and administrations in the respective EU Member States [22]. Member States Administrations must use the guidance provided by the EIF to supplement their national E-Government Interoperability Frameworks with a pan-European dimension and thus enable pan-European interoperability

In this context, interoperability is the ability of information and communication technology systems and of the business processes they support to exchange data and to enable sharing of information and knowledge. The ISO/IEC 2382 Information Technology Vocabulary defines interoperability as the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units. An interoperability framework can be described as a set of standards and guidelines, which describe the way in which organisations have agreed, or should agree, to interact with each other.

Interoperability can be considered on very different abstraction levels, and the distinctions to be made in this respect cut across all the other matrix dimensions. Within a continuum ranging from a very concrete to a very abstract perspective it is possible to distinguish three layers as shown in next Fig. 1.

The aspects of interoperability as a general concept or approach cover technical, semantic, and organisational issues, usually referenced as interoperability layers. Interoperability is conceived on different main abstraction levels:

1) Organisational interoperability level: processes, defined as workflow sequences of tasks, integrated in a service-oriented environment.

2) Technical interoperability level: signals, low-level services and data transfer protocols.

3) Semantic interoperability level: information in various shared knowledge representation structures such as taxonomies, ontologies, or topic maps. Semantic interoperability is not just with about the packaging of data (data format), but mostly focuses into simultaneous transmission of their meaning (semantics). The meaning of the data is transmitted with the data itself, in an "information package" independent of any information system. Semantic interoperability shared vocabulary, and its associated links to an ontology, which provides the basis for machine interpretation and understanding of the logic of the message. This is success by adding metadata (information used to describes other data) and linking each data element to a shared vocabulary.

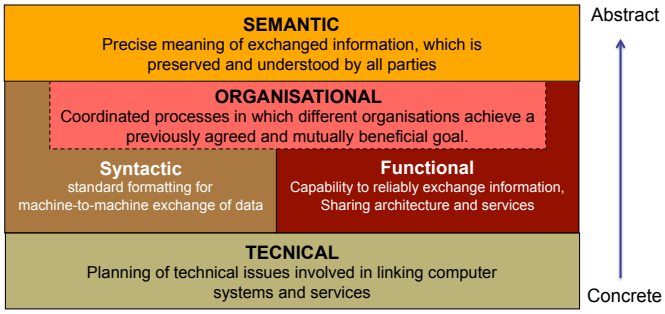

Figure 1. Abstraction layers interoperability

Two or more entities achieve interoperability when they are capable of communicating and exchanging data, which concerns to specified data formats and communication protocols. Exchanging normalized data is a prerequisite for semantic interoperability and refers to the packaging and transmission mechanisms for data. In the semantic interoperability there are concepts and methods available, but which are not yet standardized. However, for organizational interoperability it is by far less obvious what has to be standardized, who could develop and establish appropriate standards, and what is necessary for their operation and maintenance.

In this section, we have focused our work in semantic interoperability analysis. For this purpose, we use ontologies and semantic approach.

This area implies the collaboration of many actors, such as local repositories, information workers and suppliers. For this reason, we can quote the following reasons for the need to develop/define a central ontology:

- Providing a semantic typing for the data distributed all over the repositories in order to facilitate the information request by citizens through efficient search engines. Entities can be assumed to be the institutions offering digital services, digital repositories, public platforms or simply Web services.
- Sharing common understanding of the structure of information among intelligent agents, facilitating the extraction of information and processing of documents. Objects of interaction, the entities that actually need to be processed in semantic interoperability scenarios. Choices range from the full content of digital information objects to mere representations of such objects, which in turn are often conceived as metadata attribute sets.

- Enabling reuse of existing domain knowledge and its further extension, providing a contextual framework enabling unambiguous communication of complex and detailed concepts.

However, semantic interoperability problems emerge as these organizations may differ in the terms and meanings they use to communicate, express their needs and describe resources they make available to each other. Moreover, interoperability can be considered on different abstraction levels, and the distinctions to be made in this respect cut across all the other matrix dimensions. Within a continuum ranging from a very concrete to a very abstract perspective it is possible to distinguish the four layers of technical, syntactic, functional and semantic interoperability. We must bear in mind that interoperability framework is, therefore, not a static document and may have to be adapted over time as technologies, standards and administrative requirements change. In the next sections, we establish the base of all these aspects in our platform OntoSDL.

IV. THE ONTOSDL ARCHITECTURE

In order to support semantic retrieval knowledge in Seville institutional repositories we develop a prototype named OntoSDL based on ontologies and expert system technologies. The proposed architecture is based on our approach to information retrieval in an efficient way by means of metadata characterizations and domain ontology inclusion. It implies to use ontology as vocabulary to define complex, multi-relational case structures to support the CBR processes. Our system works comparing objects that can be retrieved across heterogeneous repositories and capturing a semantic view of the world independent of data representation. The framework presented in the next sections is built on established and widely accepted standards for data transfer and exchange (XML), web services (WSDL, SA-WSDL) and process models (BPMN, BPEL). The main focus of this paper is on semantic interoperability; however, other levels are addressed as well. Use of technological standards enables different kinds of interoperability constitute a major dimension with more traditional approaches geared towards librarian metadata interoperability such as Z39.50 /SRU+SRW or the harvesting methods based on OAI-PMH or again web service based approaches (SOAP/UDDI) and the Java based API defined in JCR (JSR 170/283) as well as GRID based platforms such as iRods.

The architecture of our system is shown in Fig. 2, which mainly includes three parts: ontology knowledge base, the search engine, and the intelligent user interface. Their corresponding characteristics and functions are studied in the following paragraphs.

A. Ontology Knowledge Base

OntoSDL system uses its internal knowledge bases and inference mechanisms to process information about the electronic resources in Seville University repositories. At this stage, we consider to use ontology as vocabulary for defining the case structure like attribute-value pairs. Ontology knowledge base is the kernel part for semantic retrieval information. Ontology is a knowledge structure, which identify the concepts, property of concept, resources, and relationships among them to enable share and reuse of knowledge that are needed to acquire knowledge in a specific search domain. The metadata descriptions of the resources and repository objects (cases) are abstracted from the details of their physical representation and are stored in the Case Base. Ontology provides 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 [23].

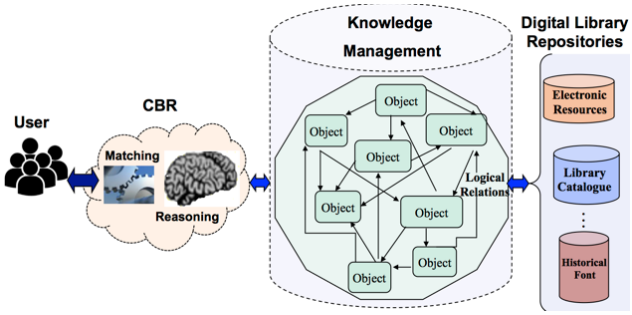

Figure 2. System architecture of OntoSDL

B. The Search Engine

Inference engine contains a CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text [24]. Case Base has a memory organization interface that assumes that whole case-base can be read into memory for the CBR to work with it. Also, we have implemented a new interface, which allows retrieving cases enough to satisfy a SQL query. We used a CBR shell, software that can be used to develop several applications that require cased-based reasoning methodology. We analysed the CBR object-oriented framework development environments JColibri [25]. This framework work as open software development environment and facilitate the reuse of their design as well as implementations. The CBR engine uses an evaluation function to calculate the new case ranking, and the answered question updates the query and the rankings in the displays. The questions are ranked according to their potential for retrieval and matching.

C. The Intelligent User Interface

The acceptability of a system depends to a great extent on the quality of this user interface component [26]. Advanced conversational user interface interacts with users to solve a query, defined as the set of questions selected and answered by the user during conversation. Interface is designed and developed to improve communication between humans and the platform. Interfaces are provided for browsing, searching and facilitating Web contents and services. Interface enhances the flexibility, usability, and

power of human-computer interaction for all users. In realizing the user interface we have exploited knowledge of users, tasks, tools, and content, as well as devices for supporting interaction within different contexts of use. In our system, the user interacts with the system to fill in the gaps to retrieve the right cases. During each search the user selects one item from two displays: ranked questions and ordered cases.

The interfaces provide 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 user interface helps to user to build a particular profile that contains his interest search areas in the DL domain. The objective of profile intelligence has focused on creating of user profiles: Staff, Alumni, Administrator, and Visitor.

We have developed a graphical selection interface as illustrated in Fig. 3.

Figure 3. User profiles interface

In an intelligence profile setting, people are surrounded by intelligent interfaces merged. Rather than building static user profiles, contextual systems try to adapt to the user's current search. OntoSDL monitors user's tasks, anticipates search-based information needs, and proactively provide users with relevant information. 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.

A technical administrator will have a view very different from an end user providing content as an author. Different conceptions, again, will emerge from the perspectives of a digital content aggregator, a 'meta user' or a policy maker. It consists of one user profile, consumer search agent components and bring together a variety of necessary information from different user's resources. Interoperability concepts differ substantially from those of a content consuming end user.

V. CASE-BASED REASONING INTELLIGENT TECHNIQUE

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. CBR is a problem solving paradigm that solves a new problem, in our case a new search, by remembering a previous similar situation and by reusing information and knowledge of that situation. A new problem is solved by retrieving one or more previously experienced cases, reusing the more similar case, revising, and retaining the case. In our CBR application, problems are described by metadata concerning desired characteristics of a library resource, and the result to a specific search is a pointer to a resource described by metadata. These characterizations are called cases and are stored in a case base. CBR case data could be considered as a portion of the knowledge (metadata) about an OntoSDL object. Every case contains both a solution pointers and problem description used for similarity assessment. Description of the framework case, which is formally described in terms of framework domain taxonomy they are used for indexing cases. The possible solutions described by means of framework instantiation actions and additional information to justifies these steps. The following processes may describe a CBR cycle (Fig. 4):

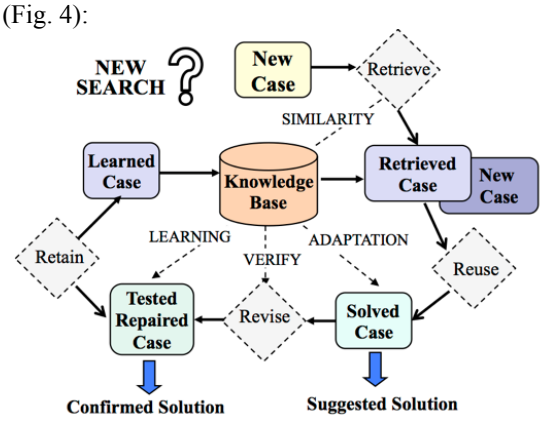

Figure 4. User profiles interface

- 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.

A. CBR Structure

The development of a quite simple CBR application already involves a number of activities. The actions consist on collecting case and background knowledge, modeling 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. This aspect is certainly one of the major reasons for the commercial success of CBR applications. Nevertheless, implementing a CBR application from scratch remains a time-consuming software engineering process and requires a lot of specific experience beyond pure programming skills.

Although CBR claims to reduce the effort required for developing knowledge-based systems substantially compared with more traditional AI approaches. The implementation of a CBR application from scratch is still a time consuming task. We present a novel, freely available tool for rapid prototyping of CBR applications. CBR object-oriented framework development environments JColibri have been used in this study. By providing easy to use model generation, data import, similarity modeling, 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 [27].

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. 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. 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.

B. Retrieval of similar cases process

The main purpose of establishing intelligent retrieval ontology is to provide consistent and explicit metadata in the process of knowledge retrieval. CBR systems typically apply retrieval and matching algorithms to a case base of past search-result 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. Our system provides multilayer retrieval methods:

1. Intelligent profiles interface: Low-level selection of query profile options, which mainly include the four kinds of user. These users can specify certain initial items, i.e., the characteristics and conditions for a search. For this a statistical analysis has been done to determine the importance values and establishing specified user requirements. User searches are monitored by capturing information from different user profiles. This statistical analysis even can in fact lay the foundation for searches in a particular user profile.

2. Ontology semantic search can query on classes, subclasses or attributes of knowledge base, and matched cases are called back.

3. 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. Similarity measures used in CBR are of critical importance during the retrieval of knowledge items for a new query. Similarity retrieval expands the original query conditions, and generates extended query conditions, which can be directly used in knowledge retrieval. Unlike in early CBR approaches, the recent view is that similarity is usually not just an arbitrary distance measure, but function that approximately measures utility.

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-neighbor technique. 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 neighbor matching is next:

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

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.

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. An important advantage of similarity-cased retrieval is that if there is no case that exactly matches the user's requirements, this can show the cases that are most similar to his query.

VI. Ontology Design and Development

The main objective of our system is to improve the modelling of a semantic coherence for allowing the interoperability of different modules of environments dedicated to E-Government. We have proposed to use ontology together with CBR in the acquisition of an expert knowledge in the specific domain. The primary information managed in the OntoSDL domain is metadata about institutional resources, such as guides, publications, forms,

digital services, etc. We need a vocabulary of concepts, resources and services for our information system described in the scenario requires definitions about the relationships between objects of discourse and their attributes [28]. OntoSDL 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. RDF is used to define the structure of the metadata describing DL resources. Our ontology can be regarded as quaternion OntoSearch:={profile, collection, source, relation), where profiles represent the user kinds. Collection contains all the services and resources of the institutional repository. Source covers the different information suppliers: electronic services, official web pages, publications, guides, etc. Finally, relation element is a set of relationships intended primarily for standardization across ontologies.

We integrated three essential sources to the system: electronic resources, catalogue of documents, and personal Data Base. The W3C defines standards that can be used to design an ontology [29]. We wrote the description of these classes and the properties in RDF semantic markup language. We choose Protégé as our ontology editor, which supports knowledge acquisition and knowledge base development [30]. It is a powerful development and knowledge-modelling tool with an open architecture. Protégé uses OWL and RDF as ontology language to establish semantic relations [31].

Protégé provides an environment for the creation and development of underlying semantic knowledge structures-ontologies and semantically annotated web services. Protégé organizes these elements like a dynamic process workflow. For the construction of the ontology of our system, we followed steps detailed below.

1) Determine the domain and scope of the ontology. This should provide the location of different on-line resources. These are included from different sources: Publications Catalogue, Web Sites, Electronic Resources, etc. Also ontology must be adapted to needs of user kinds.

2) Enumerate important terms in ontology. It is useful to write down a list of all terms we would like either to make statements about or to explain to a user. Initially, it is important to get a comprehensive list of terms without worrying about overlap between concepts they represent, relations among the terms, or any properties that the concepts may have, or whether the concepts are classes or slots.

3) Define the classes and the class hierarchy. When designing the ontology, we first need to group together related resources of the institutional repositories. There are three major groups of resources: users, services, and resources. In order to realize ontology-based intelligent retrieval, we need to build case base of knowledge with inheritance structure. The ontology and its sub-classes are established according to the taxonomies profile. A detailed picture of our effort in designing this ontology is available in Fig. 5. This shows the high level classification of classes to group together OntoSDL resources as well as things that are related with these resources. Profile ontology includes several attributes like Electronic_Resources, Digital_Collections, Publication Catalogue, Public Services, etc.

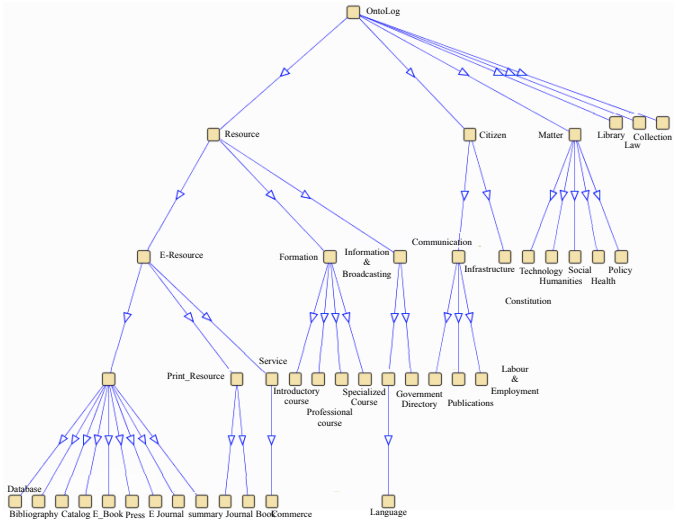

Figure 5. Class hierarchy for the OntoSDL ontology

4) Define the properties of classes and define the facets of the slots. The classes alone will not provide enough information to answer the semantic searches. Once we have defined some of the classes, we must describe the internal structure of concepts. In order to relate ontology classes to each other, we defined our own meaningful properties for the ontology. For this reason, we defined a class hierarchy associated with meaningful properties. Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take. In the following, we give a short RDF description that defined the concept of the user teacher that is a subclass of Member_Community_University.

```
<rdf:Description rdf:about="#Teacher">
  <rdfs:comment rdf:datatype=
  "http://www.w3.org/2001/XMLSchema#string">
    Teacher profile for affiliated colleges
  </rdfs:comment>
<rdfs:subClassOf rdf:resource=
    "#Members of the University community"/>
  <rdf:type rdf:resource=
  "http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
```

5) Generating the ontology instances with SW languages. To provide a conversational CBR system to retrieve the requested metadata satisfying a user query we need to add enough initial instances and item instances to

knowledge base. The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires choosing a class, creating an individual instance of that class, and filling in the slot values.

After designing the ontology, we wrote the description of these classes and the properties in RDF semantic markup language. For this purpose, we have followed next steps. First, we choose a certain item, and create a blank instance for item. Then the domain expert, in this case administrative staff fills blank units of instance according the domain knowledge [32]. 11.000 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. As a plus, domain specific rules defined by domain experts (manually or by tools) can infer more complex high-level semantic descriptions, for example, by combining low-level features in local repositories. On one hand, the rules can be used to facilitate the task of resource annotation by deriving additional metadata from existing ones.

Keeping in mind that our final goal is to reformulate queries in the ontology to queries in another with least loss of semantics, we come to a process for addressing complex relations between two ontologies. As mentioned in previous sections, relations among ontologies can be composed as a form of declarative rules, which can be further handled in inference engines. In our approach, we choose to use the Semantic Web Rule Language (SWRL), which is based on a combination of OWL DL and OWL Lite with the case-based reasoning sublanguages, to compose declarative search rules [33].

VII. EXPERIMENTAL EVALUATION

Experiments have been carried out in order to test the efficiency of AI and ontologies in retrieval information in a DL. 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 DL to be stored in the CBR. The user begins the search devising the starting query. Suppose the user is looking for some resource about "Computer Science electronic resource" in the library digital domain of Seville (Fig. 6).

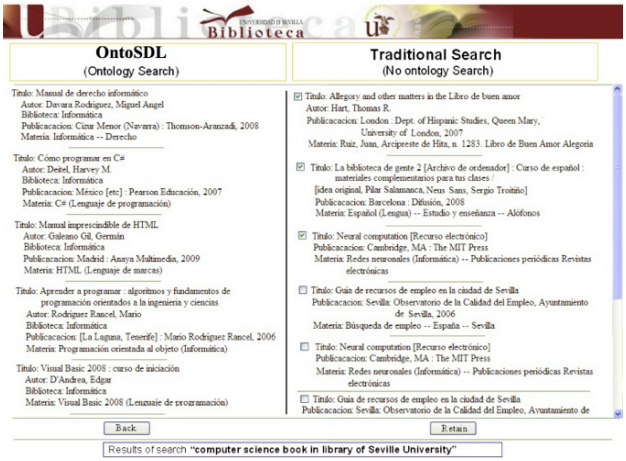

Figure 6. Search engine results page

The user inputs the keywords in the user profile interface. The required resources should contain some knowledge about "Computer Science" and related issues. After searching, some resources are returned as results. 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.

We have compared our prototype with some semantic search engines like Hakia, Lexxe, SenseBot, etc. However, we have focused in Google because is the world's dominant search engine and Google has made significant inroads in semantic indexing in search. It is a fact that deep inside Google is based on breakthrough semantic search techniques that are transforming Google's search results [34].

For our experiments, we considered 50 users with different profiles. Therefore, we could establish a context for the users, they were asked to at least start their essay before issuing any queries to OntoSDL. They were also asked to look through all the results returned by OntoSDL 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 each experiment, we report the average rank of the user-clicked result for our baseline system, Google and for our search engine OntoSDL. In our study, we have agreed different values to measure the quality of retrieved documents, excellent, good, acceptable and poor. Next, 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 I. ANALYSIS OF RETRIEVED DOCUMENTS RELEVANCE FOR SELECT QUERIES

	Excellent	Good	Acceptable	Poor
OntoSDL	7,50%	41,50%	40,60%	10,40%
Google	2,60%	27,90%	43,40%	26,10%

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. These results are presented in Fig. 7.

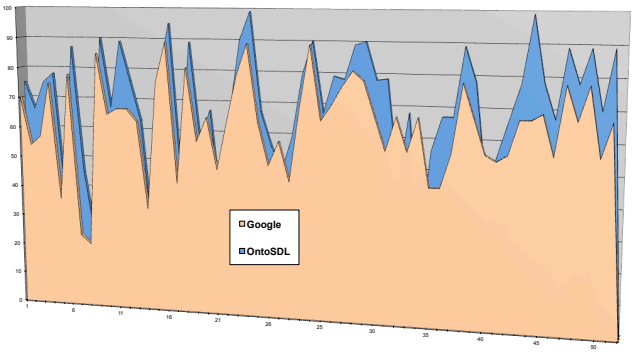

Figure 7. Search engine results page

In the digital library domain we can observe that best final ranking was obtained for our prototype. OntoSDL achieves an interesting improvement over the performance of Google. Other significance test is the analysis of the number of searches that have been resolved satisfactory by OntoSDL. As noted in Table I, our system performs satisfactorily with about a 91.6% rate of success in real cases.

Another important aspect of the design and implementation of an intelligent system is determination of the degree of speed in the answer that the system provides. 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. We can establish that speed in our system improves the proceeding time and the average of the traditional search engine. The results for OntoSDL are 9.15% better than proceeding time and 11.9% better than executing time searches/sec in the traditional search engines.

VIII. CONCLUSION AND FUTURE WORK

We have investigated how semantic technologies and AI can be used to provide additional semantics from existing resources in digital libraries. We described an effort to design and develop a prototype for management the resources in a library such as OntoSDL project, and to exploit them to aid users as they select resources. Our study addresses the main aspects of a Semantic Web knowledge retrieval system architecture trying to answer the requirements of the next-generation Semantic Web user. This scheme is based on the next principle: knowledge items are abstracted to a characterization by metadata description and it is used for further processing.

For this purpose, we presented a system based in ontology and AI architecture for knowledge management in the Seville DL. First of all, to put our aims into practice, we should develop the domain ontology and study how the content-based similarity between the concepts typed attributes could be assessed in CBR system. A dedicated inference mechanism is used to answer queries conforming to the logic formalism and terms defined in our ontology. We have been working on the design of entirely ontology-based structure of the case and the development of our own reasoning methods in jColibri to operate with it. It introduced a prototype web-based CBR retrieval system, which operates on an RDF file store. Furthermore, 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 DL according to his own preferences, which come to represent his interests.

Finally, the study analyses the implementation results, and evaluates the viability of our approaches in enabling search in intelligent-based digital repositories. OntoSDL can be part of a bigger framework of interacting global information networks including e.g., other digital libraries, scientific repositories and commercial providers. The framework relies as much as possible on standards and existing building blocks as well as is based on web standards.

The results demonstrate that by improving representation by incorporating more metadata from within the information and the ontology into the retrieval process, the effectiveness of the information retrieval is enhanced. Future work will concern the exploitation of information coming from others institutional repositories and digital services. Furthermore, we propose 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

[1] A. Martín and C. León, “Intelligent Technique to Accomplish a Effective Knowledge Retrieval from Distributed Repositories,” in Proc. Third International Conference on Intelligent Systems and Applications (INTELLI), pp. 97-102, Seville, Spain, 2014.

[2] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann, “Ontology refinement for improved information retrieval,” Information Processing & Management, Volume 46 (Issue 4), Semantic Annotations in Information Retrieval, 2010.

[3] M.C. Diaz-Galiano, M.T. Martin-Valdivia, and L.A. Urena-Lopez, "Query expansion with a medical ontology to improve a multimodal information retrieval system," Computers in Biology and Medicine, Volume 39 (Issue 4), pp. 96-403, 2009.

[4] L. Chen, "Design and implementation of intelligent library system," Library Collections, Acquisitions, and Technical Services, Volume 32 (Issues 3–4), pp.127-141, 2008

[5] J.S. Cho and K.H. Hyun, "Meta-ontology for automated information integration of parts libraries," Computer-Aided Design, Volume 38 (Issue 7), pp. 713-725, 2006.

[6] H. Sasaki and Y.A. Kiyoki, "A formulation for patenting content-based retrieval processes in digital libraries," Information Processing & Management, Volume 41 (Issue 1), pp. 57-74, 2005.

[7] D. Bainbridge, M. Dewsnip, and I.H. Witten, "Searching digital music libraries." Information Processing & Management, Volume 4, (Issue 1), pp. 41-56, 2005.

[8] C.M. Toledo, M.A. Ale, O. Chiotti, and M.R. Galli, "An Ontology-driven Document Retrieval Strategy for Organizational Knowledge Management Systems," Electronic Notes in Theoretical Computer Science (Vol. 281), pp. 21-34, 2011.

[9] D. Govedarova, S. Stoyanov, and I. Popchev, "An Ontology Based CBR Architecture for Knowledge Management in BULCHINO Catalogue," in Proc. International Conference on Computer Systems and Technologies (CompSysTech), 2008.

[10] P. Warren, "Applying semantic technologies to a digital library: a case study", Library Management Journal, Emerald, 2005.

[11] H. Stuckenschmidt and F.V. Harmelen, "Ontology-based metadata generation from semi-structured information," K-CAP, pp. 163–170, ACM, 2011.

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

[13] GAIA - Group for Artificial Intelligence Applications. *jCOLIBRI project - Distribution of the development environment*, [Online]. Available from: http://gaia.fdi.ucm.es/research/colibri/jcolibri/ 2015.04.25

[14] Y. Sure and R. Studer, "Semantic Web technologies for digital libraries," Library Management Journal, Emerald, Vol. 26, pp. 190-195, 2005.

[15] I.H. Witten and D. Bainbridge, "How to Build a Digital Library" Morgan Kaufmann, 2003.

[16] H. Ding, "Towards the metadata integration issues in peer-to-peer based digital libraries," GCC. H. Jin, Y. Pan, N. Xiao, and J. Sun, (eds.) (LNCS), Vol. 3251, Berlin, Germany, Springer, 2004.

[17] R. Guha, R. McCool, and E. Miller, "Semantic search," In Proceedings of WWW2003, 2003.

[18] G.F. Luger, "Artificial Intelligence, Structures and Strategies for Complex Problem Solving," 4ª edition. Ed. Pearson Education Limited, 2002.

[19] Z. Sun and G. Finnie, "Intelligent Techniques in E-Commerce: A Case-based Reasoning Perspective," Heidelberg: Springer-Verlag, 2004.

[20] SEC. Commission Staff Workking Paper: linking up Europe, *the importance of interoperability for egovernment services*, [Online]. Available from: http://europa.eu.int/ISPO/ida/export/files/en/1523.pdf, 2015.05.3

[21] MAP. Aplicaciones utilizadas para el ejercicio de potestades. Criterios de Seguridad, Normalización y Conservación. *Ministerio de Administraciones Públicas*. [Online]. Available from: http://www.csi.map.es/csi/criterios/index.html, 2014.03.05

[22] EIF. *European Interoperability Framework Version 2*. [Online]. Available from: http://ec.europa.eu/isa/strategy/doc/annex_ii_eif_en.pdf, 2015.04.19.

[23] S. Staab and R. Studer, "Handbook on Ontologies," International Handbooks on Information Systems, Springer, Berlin, 2005.

[24] M. Bridge, H. G¨oker, L. McGinty, and B. Smyth, "Case-based recommender systems," Knowledge Engineering Review, 2006.

[25] B. Díaz-Agudo, P.A. González-Calero, J.Recio-García, and A Sánchez-Ruiz, "Building CBR systems with jColibri," Journal of Science of Computer Programming, Volume 69, Issues 1–3, 1 December 2007, pp. 68–75, doi: dx.doi.org/10.1016/j.scico.2007.02.004.

[26] D. Quan and D.R. Karger, "How to make a semantic web browser," I Proc. of Thirteenth International World Wide Web Confecence (WWW), pp. 17–22, New York, New York, USA, Vol. 12, Issue 1, pp. 1- 40, 2004.

[27] L.A. Breslowm and D.W. Aha, "Simplifying decision trees: A survey," The Knowledge Engineering Review archive, Cambridge University Press New York, NY, USA, 1997.

[28] D. Taniar And J.W. Rahayu, "Web semantics and ontology," Hershey, PA: Idea Group, 2006.

[29] W3C. *RDF Vocabulary Description Language 1.0: RDF Schema*. [Online]. Available from: http://www.w3.org/TR/rdf-schema/, 2015.02.10.

[30] PROTÉGÉ. *The Protégé Ontology Editor and Knowledge Acquisition System*. [Online]. Available from: http://protege.stanford.edu/, 2015.04.05.

[31] J. Heflin, "OWL Web Ontology Language Use Cases and Requirements," W3C Recommendation, 2004.

[32] M. Horridge and H. Knublauch, "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools," The University Of Manchester, United Kingdom, 2004.

[33] S. Bechhofer, F.V Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F Patel-Schneider, and L.A. Stein, "OWL web ontology language reference", W3C recommendation. Volume 10 (Issue February), Publisher W3C, 2004.

[34] D. Amerland, "Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact and Amplify Your Online Presence", Que Publishing Kindle Edition, July, 2013.