Platform for Collection from Heterogeneous Web Sources and its Application to a Semantic Repository Organization at SeDiCI: Preliminaries

Marisa Raquel De Giusti
Comisión de Investigaciones Científicas (CIC)
de la Provincia de Buenos Aires
and Proyecto de Enlace de Bibliotecas UNLP

Ariel Sobrado
Proyecto de Enlace de Bibliotecas UNLP

Agustín Vosou
Proyecto de Enlace de Bibliotecas UNLP
and
Gonzalo Luján Villarreal
Consejo Nacional de Investigaciones Técnicas y Científicas (CONICET)
and Proyecto de Enlace de Bibliotecas UNLP

ABSTRACT

Presentation of a web collection platform designed to relate and unify information available on different standard web sources with a view to creating a user-browseable thematic repository. The platform will be used at the Intellectual Creation Diffusion Service [1] combined with ontologies and thesaurus to provide improved data sorting.

Data is currently spread on web resources and traditional search engines return ranked lists with no semantic relation among documents. Users have to spend a great deal of time relating documents and trying to figure out which ones fully address the issue domain. It is only after locating similarities and differences that information fragments are applied to the user's work, enabling knowledge creation.

The proposed platform sorts out the different theme domain functioning modules to allow their use in various knowledge areas. Development includes two agents that searches data base stored URLs, one is capable of identifying bookmarked pages, interpreting labels and providing rules for extracting information and storing it in a RDF data file; on the other hand, the other agent is in charge of getting related URLs from the given one. After this stage, homogenization is applied and transformed information is sorted out according to domain ontologies.

The platform allows for more efficient automatic extraction processes and information search among heterogeneous sources that represent the same concepts using different standards.

Keywords: SeDiCI, semantic repository, ontology and thesaurus.

1. INTRODUCTION

The Intellectual Creation Diffusion Service [1] was created, initially, to expose the creation of the various academic units of the UNLP as a way of knowledge socialization. SeDiCI offers its contents according to the Open Archives Initiative (OAI) protocol, and simultaneously harvests free external academic information under this protocol. One of the objectives of the service is to provide the users increased and more relevant information. To achieve this goal we have thought how to collect free information from the web about different areas of interest, checking sources and properly structuring this information in the digital library to allow the users to make more accurate searches.

The information on the Internet is normally searched by users through general purpose search engines like Google, Yahoo!, etc. The problem that arises is that these search engines match queries by keywords instead of concepts: thus, important semantic relations are lost and therefore information retrieved may be different even if the keywords used are synonyms [2].

One of the current extensions to the web is the Semantic Web [3][4], a W3C initiative led by Tim Berners Lee [5]. However, in this case, given the nature of the digital library, the proposal is not aimed at generating and sharing ontologies (although the consequences of this work lead to this achievement), but pretends to make a combined use of the ontologies [6] and thesaurus [7] that SeDiCI is using at the moment to increase efficiency of the automatic information extraction processes among heterogeneous sources (web, repositories, etc) and improve user's searches. When speaking of heterogeneous sources, and although we are looking for information in the same domain of interest, different sources use different conventions to represent the same concept. Because of that, we propose to create a platform capable of extracting data from semantically marked websites and unify the format, storing the information in an ontological repository on which users can search semantically instead of using keywords.

2. PLATFORM DESCRIPTION

As shown in Fig. 1, the platform can be divided in two major systems: data collection system and search system. The main tasks of the collection system are visiting a list of web pages, detecting those marked, extracting its content and organizing it taking into account the ontologies defined for the topic and the selected thesaurus. The search system, on which we are not focused in this first presentation, basically is in charge to help users to make intelligent searches (semantic) on the repository (also semantic) that has been populated by the data collection system. The development of the search system has not started yet since we have determined that it is a priority to move forward first with the collection system to obtain a good deal of information that can be searched. In both modules, defined ontologies and thesaurus have a very important role. In this early stage of development, efforts are aimed at the collection system's components whose development has a significant progress but, due to being in a preliminary phase of testing and improvements, is not accessible to the general public yet.

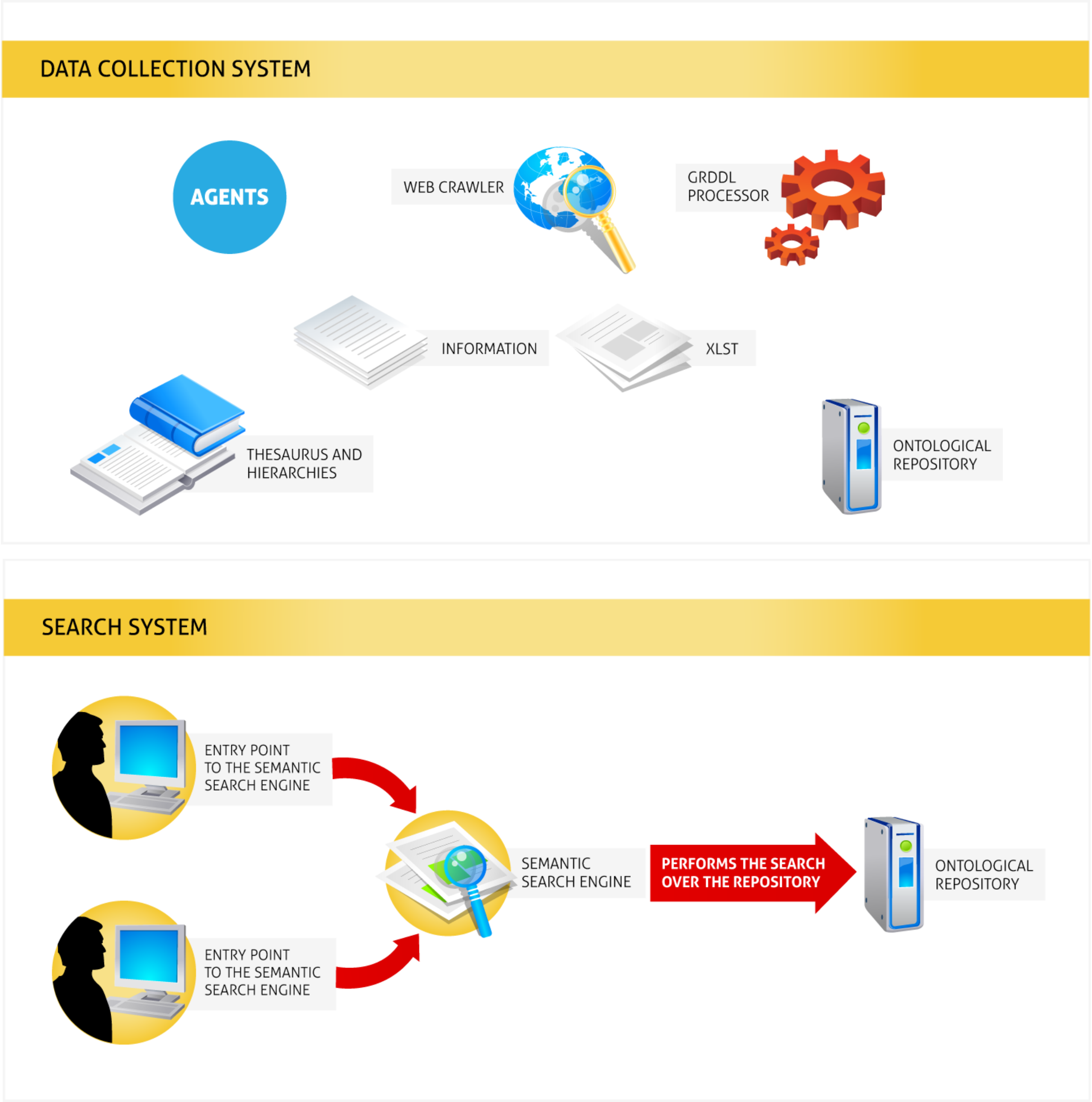

Fig. 1: Platform Architecture

Data Collection System

The collection system is composed by the following elements (they all are described in detail below): 1) a database that contains urls to visit; 2) an agent (Robot 1) that takes non-visited urls from the database and invokes the Web crawler; 3) a Web crawler able to get urls embedded within another url; 4) another agent (Robot 2) that processes urls and populates the ontology; 5) a GRDDL processor that applies transformations to a (X)HTML document; 6) a thesaurus with homogenized terms; and finally 7) an ontology used to represent obtained data.

These components are separated in two modules, searching and processing, which can be executed in parallel without interrupting each other.

Steps followed by the searching module:

1. Robot 1 takes non-visited urls from the database.

2. Robot 1 sends to the Web crawler the obtained url and marks it as visited.

3. The Web crawler checks the url, looks for the embedded urls in HTML code (inside the <a> tag) and adds them to the database.

Steps followed by the processing module:

1. Robot 2 takes from the database urls that has already been visited, but has not been processed. The url is passed to the GRDDL processor.

2. The GRDDL processor applies some XSLT transformations to obtain an XML or RDF document.

3. Robot 2 marks the url as visited.

4. If needed, XML is transformed to RDF.

5. Terms are searched in the defined thesaurus and

instances of the selected ontology are created to populate the repository.

Data collection system is in charge of handling the following tasks: through an agent [8] takes a list of web pages and sends them to a Web crawler [9][10], responsible of collecting other links embedded on the page, and adds them to the previous list. An agent called Robot 2 takes an URL and sends it to a GRDDL processor ("Gleaning Resource Descriptions from Dialects of Languages") [11] that applies transformations, in our case using XSLT, from an XHTML or XML document to an XML [12]. This textual document now contains only tuples of interest, allowing the application to automatically extract information from structured web pages to integrate it into a repository.

The GRDDL processor detects microformats of interest on pages and, by indicating the location of the transformation sheet (XSL) used to capture microformats [13][14] and the URL of the page to transform, returns an XML with the information extracted by the transformation sheet. The XML document is subsequently transformed to RDF [15] to homogenize the data, classify it and populate the ontology. Homogenization is made through conversions (using synonyms) and translations to a single language. The information classification is done through a domain ontology; during this stage individuals are created with attributes and relations from an RDF: each individual belongs to the root class of the ontology. Also the domain relations defined by the ontology are looked for on the RDF.

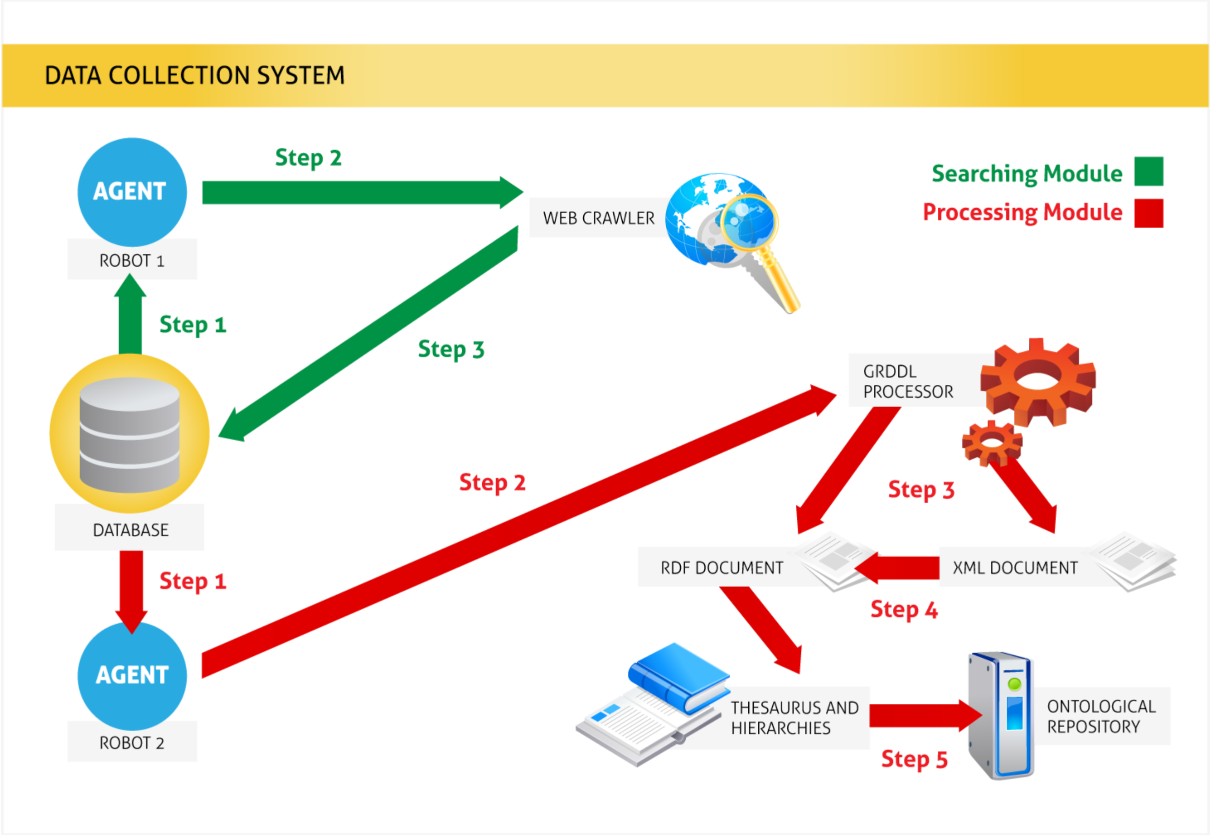

Fig. 2: Data Collection System

The next step is checking that the individuals satisfy all its class restrictions through a reasoning module in charge of validating individuals that has been added to the ontology. If the individual in question belongs to the class, the heritable attributes are added. Once this stage is done, storage on a semantic repository accessible via web is performed.

3. CASE STUDY

The choice of the case was subjected to restrictions on the existence of marked pages on the web today and required to make changes on the fly to reliably validate the data collection system's platform. As a first exercise we worked directly with the digital library SeDiCI (which would operate as a mixed site). The inherent features of SeDiCI required to modify the source code to start using Dublin Core microformat [16][17] to represent records [18] on the repository. The chosen search example consisted in finding material that in the field "descriptors" contains: "Física del estado sólido" [Solid state physics].

4. ONTOLOGY DEFINITION

Our ontology SediciON modeled in Protégé [19] has a class called MATERIAL which is a superclass of the type of document we are analyzing. Our ontology reuse the Dublin-Core Ontology [20].

There are two levels in the definition of the ontology:

1. Syntactic level: For the representation of SediciON ontology we used the OWL-DL language, as W3C recommends [21]. The features of OWL-DL ensure interoperability with other systems and formats. Other features of OWL-DL, such as its ability to infer in conceptual organization systems based in hierarchies, will be used to provide more functionality to the system and describe in a richer way the resources involved.

2. Semantic level: We used a high-level ontology for the overall organization of academic repositories to ensure

interoperability with other similar systems. Not only it is necessary to use a syntactic format for standard representation, but also to ensure future semantic compatibility with other extensions, in case it is necessary to add information about new resources related to the digital library SeDiCI; hence the use of an high-level ontology.

Reused entities from the DC ontology appear very briefly referenced above, but readers are encouraged to access the original description for further clarification.

5. EXTRACTION EXAMPLE

In our example we will extract information from marked documents in digital library SeDiCI. During the process, the data collection system processes only the pages that contain the Dublin Core microformat. The created RDF file contains information about the documents related to the subject of interest, i.e. the title, author and a number of descriptors. The population module of the ontology creates individuals from the data contained in the RDF file. In our case, before storing the individuals (in the near future) we could look through the Description attribute values, those which correspond to alternatives terms in a thesaurus other than operating in SeDiCI. The population module could also make some language transformations.

Both data and information from the different resources are stored in the SeDiCI ontological repository: FisSol. Entities of this database must be mapped to RDF instances, which are organized according to the conceptual model of the SediciON ontology.

6. CONCLUSIONS AND FUTURE WORK

We have presented the first outlines of a semantic search platform in heterogeneous sources. This platform combines the use of ontologies and thesaurus which will provide greater relevance to the searches performed by visitors in SeDiCI. For the PrEBi work group this is a first experience that has led to understand the work cycle and its applicability in the digital library according to a primary goal: provide better and more structured information. The data collection system must be analyzed in terms of efficiency: the usage of two robots, selected tools and libraries, specially considering that after an initial stage of semantization of the library, it must be complemented with new content from the web. In this sense, the search of marked pages, in the way it is being currently performed, can be a limitation and we will have to consider other extraction techniques. Finally, it seems necessary to consider alternatives to handle other relationships defined in more complex thesaurus and ontologies.

The search system should be developed completely given that the idea of this platform is to let users access the repositories and make more relevant searches, for which the search system should provide a web-based entry point for users to look for concepts, select attributes and choose restrictions to finally obtain as a result a list of elements that meet the given restrictions.

7. REFERENCES

[1]: SeDiCI, Servicio de Difusión de la Creación Intelectual (SeDiCI) http://www.sedici.unlp.edu.ar/

[2]: Abian, M.A, El futuro de la web. Xml,rdf/rdfs, ontologías y la web semántica, 2003 http://www.javahispano.org/contenidos/es/el_futuro_de_la_web/

[3]: W3C Semantic Web Activity [Online], 2009 http://www.w3.org/2001/sw/grddl-wg/td/grddl-tests#spaces-in-rel/

[4]: Wikipedia, Web Semántica [Online], 2009 http://es.wikipedia.org/wiki/Web_semántica

[5]: Berners-Lee, T. y Fischetti, M, Weaving the Web: The original Design and Ultimate Destiny of the World Wide Web by its Inventor. San Francisco:Harper, 1999

[6]: Gruber, T. R., “What is an Ontology?”.[Online], 1992 http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[7]: Wikipedia, Tesauro. [Online], 2009 http://es.wikipedia.org/wiki/Tesauro

[8]: Wikipedia, Agent [Online], 2009 http://es.wikipedia.org/wiki/Agente_inteligente_(Inteligencia_Artificial)

[9]: Wikipedia, Web crawler [Online], 2009, July http://en.wikipedia.org/wiki/Web_crawler

[10]: Sun, Writing a Web Crawler in the Java Programming Language [Online], http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/

[11]: W3C, Gleaning Resource Descriptions from Dialects of Languages (GRDDL) [Online], 2007 http://www.w3.org/TR/grddl/

[12]: W3C, XSL Transformations (XSLT) Version 1.0 [Online], 1999 http://www.w3.org/TR/xslt

[13]: Wikipedia, Microformatos [Online], 2009, July http://es.wikipedia.org/wiki/Microformato

[14]: Microformats, Microformats.[Online], 2005 http://microformats.org/

[15]: Wikipedia, Resource Description Framework. [Online], 2009 http://es.wikipedia.org/wiki/Resource_Description_Framework

[16]: Dublin Core Metadata Initiative, Expressing Dublin Core in HTML/XHTML meta and link elements [Online], 2003 http://dublincore.org/documents/dcq-html/

[17]: Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Initiative, 2009 http://dublincore.org.

[18]: Mendez, E, DCMF:DC and microformatos, a good marriage”. International Conference on Dublin Core and Metadata Applicationes. [Online], 2008 http://dc2008.de/wp-content/uploads/2008/09/dc2008_mendezetal.pdf

[19]: Stanford Center for Biomedical Informatics Research, Welcome to protégé [Online], 2009, July http://protege.stanford.edu/

[20]:Dublin Core Ontology, http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl

[21]: W3C, Web Ontology Language [Online], 2004 http://www.w3.org/2004/OWL/

Received: Aug. 2009. Accepted: Sep. 2009