Semantic Browsing of a Domain Specific Resources: The Corese-NeLI Framework

Gayo Diallo[1], Khaled Khelif[2], Olivier Corby[2], Patty Kostkova[1], Gemma Madle[1]
[1]*City eHealth Research Centre, City University, EC1V0HB London, UK*
[2] *INRIA Sophia Antipolis Méditerranée, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis, FRANCE*
{First.Last.1@city.ac.uk, First.Last@sophia.inria.fr, patty@soi.city.ac.uk}

Abstract

In this paper we present the Corese-NeLI semantic web browser dedicated to navigating resources in the infectious disease domain. We describe an overview of the semantic web browser and outline its functionality and the knowledge organization system used as a background knowledge for both the annotation and search processes. The evaluation of the vocabulary-based annotation, essential for the semantic browser, uses the National eLectronic Library of Infection as a test bed and demonstrates over 96% correct annotations
.

1. Introduction

One of the underlying reasons for the success of the Web is that it only uses one type of client, the Web browser, thereby providing a single access point to widely available content and services. However, it is becoming increasingly more dificult to deal with the vast amount of data available in order to make efficient decisions. Moreover, as both search engines and browsers normally follow the "one size fits all" approach, they do not cater for the variety of users and their particular needs.

This is particularly relevant in the medical domain. Despite the internet providing patients and healthcare professionals with access to vast amounts of available information, this tends to result in the inability to find reliable information when required. Medical sites need to be able to support profile-based semantic searches and contextualized browsing in order to integrate knowledge from other medical portals which may be needed by specific groups of users. Furthermore, assistance is needed for the gathering and automated composition of data and services in order to provide an aggregated view of the available information.

Let us imagine a nurse consultant visiting the Internet and searching for information on healthcare associated infections. He/she must use specific search terms such as "MRSA" or "Clostridium difficile", and the search engine will then return a list of documents indexed by these keywords. The user must manually filter through these documents to find the required information. The search engine will not be able to provide any indication of the quality or reliability of the documents retrieved, nor will it discriminate between the types of information, such as fact sheets, clinical guidelines, etc.

It would be more useful to have a single interface to provide users with infectious diseases, their modes of transmission, treatment methods and also dynamic links to organizations such as the Health Protection Agency (HPA, www.hpa.org.uk) in the UK, which contains information on all aspects of infection and also offers authoritative guidelines.

In this context, we have designed and implemented the Corese-NeLI browser developed as s component of the EU funded project SeaLife[1]. Its aim is to build a Semantic Web browser for the Life Sciences. The Corese-NeLI browser is dedicated to the infectious disease domain and aims to improve the browsing experience of the user, with the ultimate goal of providing more relevant and supporting information in a timely manner. It follows the idea behind the Conceptual Hypermedia paradigm which provides navigation between web resources, supported by a knowledge organization system (KOS).

The rest of this paper is subdivided as follows; section 2 gives an overview of the National electronic Library of Infection Portal which is the target application domain of our framework. We also present the Corese engine and the MeatAnnot system used by the semantic web browser. Section 3 details the architecture of the Corese-NeLI semantic web browser and describe its main functionality. In section 4 the

[1] http://www.biotec.tu-dresden.de/sealife/

evaluation of the extraction method we use for the semantic annotation is reported.

2. Preliminaries and background

The objective of the EU-funded Sealife project is the definition and realisation of a semantic Grid browser for the Life Sciences linking the existing Web to the currently emerging eScience infrastructure. This will be accomplished using eScience's Web/Grid Services and its XML-based standards and ontologies. The main target for this application is the National Electronic Library of Infection (NeLI) portal in the UK and is based on two systems: the Corese engine [3] and the MeatAnnot system [2].

NeLI[2] *:* The National electronic Library of Infection is a real-world Internet medical library with over 30000 unique users a month, providing a single access point to high quality evidence on all aspects of infection. NeLI provides the main source of online medical evidence for a wide spectrum of users - clinicians, family doctors (GPs), infection control nurses, and other public health professionals. The main sources of medical evidence included are journals and websites which include: The Health Protection Agency and the Department of Health .

Corese : it stands for COnceptual REsource Search Engine. It is an RDF engine based on Conceptual Graphs (CG). It enables processing of a RDF Schema and RDF statements. Corese implements the SPARQL language and allows navigating annotation bases w.r.t the concepts and the relationships hierarchy defined in the used ontology.

MeatAnnot : this system offers a service of free text annotation according to an existing ontology. It relies on analysing scientific articles using linguistic tools in order to automatically generate RDF annotations thanks to the use of both the concepts and the declared semantic relationships between the concepts. In this work, we adapted the term extractor of the MeatAnnot system to find mappings between terms in the free text and the concepts defined in the NeLI vocabulary presented in the section 3.

3. The Corsese-NeLI Semantic Web browser

In this section we describe the approach we propose to implement the Corese-NeLI semantic browser. This approach is sufficiently generic to be adapted to any Web portal or electronic digital library.

A web browser navigates along links between documents while a Semantic web browser navigates along relationships in a web of concepts. We follow this idea which is also implemented in the Tabulator browser [5]. The term *Semantic Web Browser* (SWB) refers to any browser which: i) uses at least one KOS to support the browsing, ii) is able to identify and highlight "useful" terms in the web page being visited,
iii) enables the semantic interpretation of these web pages and adds semantic hyperlinks to their highlighted terms, iv) gathers additional information from the highlighted terms, which may involve access to external services (e.g., EBI or PubMed)

Moreover, to overcome the drawbacks of the "one size fits all" approach for information searching, the SWB need to take into account a bespoke support the variety of the users.

3.1 Architecture

The architecture of the Corese-NeLI SWB is shown in the Figure 1. It takes into account the variety of users by contextualising the information it provides. Context and customisation are some of the key factors for accurate, effective and relevant information access in Internet digital libraries and in the Semantic Web. The main components of the SWB are:

- **A query generator**: this either uses terms entered by the user or extracted automatically from the web pages. It generates several queries in different formats to query search engines (e.g., SPARQL queries for the Corese engine).
- **A profile manager**: this module implements the approach we proposed in [3] to recommend pages according to the detected user's profile. It relies on semantic annotations and navigation logs for learning model profiles and classifying users.
- **A reasoning mechanism**: this consists of a set of rules allowing a (i) building of navigation scenarios, and (ii) use of advanced functionalities of Corese such as the approximate search.
- **An annotation base**: contains ontology-based annotations describing (i) knowledge embedded in web pages; (ii) metadata on web pages including the creation date and the source; (iii) learnt users' profiles, and (iv) information on external sources.
- **Ontologies and vocabularies**: consisting of the NeLI vocabulary and a set of ontologies allowing managing profiles and resources structures.
- **External sources**: this gives additional information on the visited Web page. Terms detected in the page are used to query search engines such as Google, Wikipedia or PubMed.

[2] www.neli.org.uk

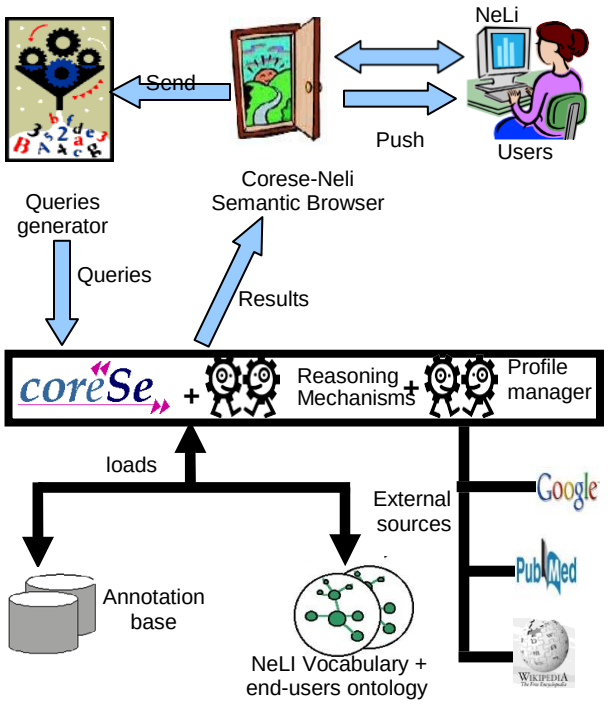

Fig 1. Architecture of the Corese-NeLI browser

The CORESE-based engine supports the navigation of a portal by the use of a structured vocabulary or a domain ontology. It supports two main functionalities:

- **Semantic search of a Web portal:** relying on semantic annotations: the semantic annotations are generated from Web pages using a provided knowledge artefact. It bases its search on the generated annotations. During the search process, Corese uses the taxonomical relationships in the KOS (i.e., narrower, broader, etc.) to retrieve annotated pages which are related to the user's query.
- **Semantic browsing of a Web portal**: the CORESE-based engine offers the possibility to identify and highlight terms retrieved from a structured vocabulary on a visited Web page. From the highlighted terms, it can then create dynamic links to related pages within the portal, thereby enabling the semantic browsing. Moreover, a query can be built from the highlighted terms in order to query external resources such as Google and PubMed.

3.2 The Knowledge Organisation System of the Corese-NeLI browser: The NeLI vocabulary

We have developed a knowledge organisation system (e.g. the NeLI vocabulary) dedicated to the infection disease domain, as the MeSH vocabulary has been judged inappropriate for indexing and navigating the NeLI portal [6]. Whilst the structure of MeSH is appropriate and its linguistic representations are very useful, some areas of NeLI's domain are not covered at all. Moreover we need a vocabulary suitable for the Core-NeLI SWB. To this end, the NeLI vocabulary musts fulfill a set of requirements inluding an *explicit* and *natural language naming of concepts*, *multiple labels for concepts* and a *mapping to other biomedical vocabularies or ontologies* to allow interoperability and gathering additional information and services while browsing with the Corese-NeLI SWB.

The NeLI vocabulary[3] is represented in the SKOS language. We used the SKOS plugin[4] for Protégé 4[5] which offers support for viewing and editing SKOS vocabularies.

The vocabulary has 630 concepts organised around 12 Top entities. They include Population, Prevention of Infection, Symptom, Transmission Mode and Treatment. Each concept of the hierarchy has an ID, a preferred label, one or more alternative labels and a definition in natural language text. It may also have one or more semantic relationships to other concepts of the hierarchy. The semantic relationships include Caused by, Affects, Is Transmission Mode Of, Has Symptom.

3.3. Principle of the Semantic Web Browser

Semantic browsing provides users with dynamically selected concepts or links from an ontology. This is enriched by the profile-based customization which selects and integrates web portals by working as a "semantic recommender" system. The Corese-NeLI SWB enables users to semantically browse the Web by highlighting ontology concepts and providing dynamic access to Web servers or knowledge portals semantically related to the Web content being visited. It relies on semantic annotations generated from the visited web pages in order to implement this functionality..

3.3.1 Generating Semantic Annotations: the NeLIAnnotator

In this step we used an adaptation of the MeatAnnot system (section 2.3) (a.k.a the NeLIAnnotator) in order to generate semantic annotations of the NeLI portal. These annotations are used later for the semantic browsing of the library.

[3] http://topcat2.soi.city.ac.uk:9090/ontologies/neli.rdf
[4] http://code.google.com/p/skoseditor/
[5] http://protege.stanford.edu/

It starts from text and relies on Corese, GATE[6] and the NeLI vocabulary to generate an RDF annotation describing the semantic content embedded in the web page. After tokenizing and Pos-tagging texts, the NeLIAnnotator tries to match candidate terms with the NeLI vocabulary . To query the vocabulary, it sends queries to Corese in order to compare the candidate terms with the labels of the concepts w.r.t their linguistic variations (e.g. *lung development* vs. *development of lung*), plural forms or varying spelling.

3.3.2 Semantic Browsing of Resources

The Corese-NeLI SWB is available as a firefox extension. Semantic browsing is enabled by navigating the graph of the NeLI vocabulary according to the user's actions. Figure 2 presents a screenshot of the interface for semantic browsing. The different components are detailled in the following section.

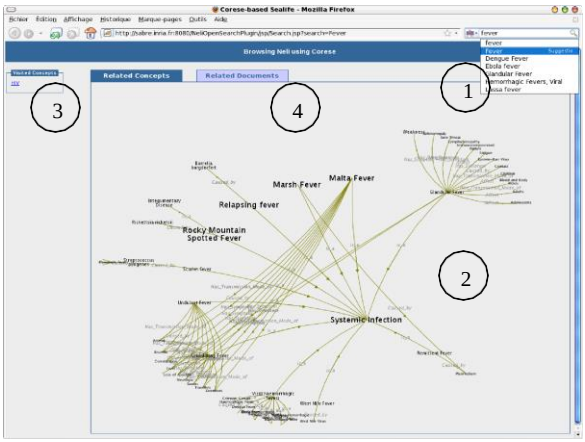

Fig 2: Semantic browsing of the NeLI vocabulary

Panel 1: allows the user to query the NeLI portal using their own keywords. These keywords are mapped into the NeLI vocabulary concepts (w.r.t synonyms and linguistic variations). Suggestions of terms related to the user's entered keywords are also available (in this example the user is entering the term *fever* and the system proposes *Dengue fever*, *Ebola Fever*, etc.).

Panel 2: Gives a view on concepts related to the user query (in this example there are different kind of fevers and fever symptoms, etc.). We adapted the hyperbolic navigation [4] used in the navigation of websites for the SPARQL results navigation. Hyperbolic navigation has the advantage of giving an overall view which is well suited to a user that does not know the hierarchy of concepts. The hypergraph portlet allows navigation through the result and the choice of one or several concepts from the NeLI vocabulary.

Panel 3: Stores the navigation history of the user in order to go back and find previous results.

Panel 4: Gives the list of NeLI pages annotated by the chosen terms. This list can be refined by the source of the document and by the date of publication. User can also have an idea about other terms annotating the selected page. By clicking a link the user is forwarded to the actual NeLI portal in order to have more information about the selected document.

4. Preliminary Evaluation: extraction method for the semantic annotation

As an efficient Semantic Web browser relies on accurate semantic annotations, we performed an initial evaluation of the extraction method which provides these semantic annotations. To do the validation, we adopted a user-centered approach and then chose a test corpus randomly from the NeLI portal. We ran the NeLIAnnotator in order to extract the annotations. The resulting annotations were presented to the NeLI information scientist via a dedicated interface for quality assessment (results in Table 1).

Table 1. Results of the extraction method evaluation

Suggestions	Corrects	Missing	Useful
451	453	9	344
Measures	**Precision**	**Recall**	**Usefulness**
Results	0.96	0.97	0.79

Precision relates to the absence of noise (terms correctly extracted) in the extraction, *recall* relates to the absence of silence (ratio of correctly extracted terms to the terms that should be extracted), *usefulness* measures the number of useful suggestions.

The second column describes the number of terms correctly extracted from the texts. The difference between this number and the number of suggestions proposed by the system is mainly due to the errors generated by the NLP tools (e.g. wrong grammatical category) and to the terms missing in the vocabulary.

Nevertheless, these methods showed a good level of precision, since 96% of the suggestions were correct. The third column describes the number of terms not extracted by the NeLI Annotator. These missing suggestions are also due to the errors generated by the NLP tools and result from terms deduced by the domain expert when s/he reads the sentence and makes a semantic indexing which cannot be extracted automatically.

[6] http://gate.ac.uk/

Finally the extraction method has a good level of usefulness, since 79% of the correct suggestions were considered to be useful by user.
In order to complete the evaluation of the Corese-NeLI browser, we have designed a global end user centred evaluation framework which is also part of the evaluation of the SeaLife Semantic browser. Its aim is to demonstrate the advantage of using the semantic browser in terms of i) mobility and travel within the system and ii) user attitude and satisfaction. The study participants will consist of a set of NeLI users separated in two groups; those taking part in a workshop-based evaluation and those taking part remotely online (for 3 months).

5. Conclusion and future work

We have described in this paper the Corese-NeLI Semantic Web browser which is one of ways in which the SeaLife browser dedicated to the Life Sciences has been implemented. We have also presented the preliminary evaluation of the extraction method used to generate the semantic annotation which the SWB is based on, and the designed framework for the evaluation of the entire browser. We believe that our approach offers several key benefits to the end users and can help them easily find the information they are looking for. One of the novelties of the approach is the detection of the user's profile on the fly.

As mentioned in [5] a challenge for Semantic Web browsers is to bring the power of domain-specific applications to a generic program when new unexpected domains can be encountered in real time. Our approach can be generalized to other scientific domains and electronic libraries. Indeed, the components of the framework described here are sufficiently generic and are also reusable. Moreover, they rely on standard technologies; (i) the NeLIAnnotator requires a KOS covering the domain studied; (ii) the Corese engine can load any annotations base and allows navigation of the annotated resources;
(iii) the queries generator and the information extraction modules are application-independent; and
(iv) the interface can be adapted to other scenarios.

From a dynamic linking perspective, the Corese-NeLI framework follows the same principle as COHSE[7] and PowerMagpie [7], while for navigating over RDF linked data, it is close to Tabulator [5] and the Disco Hyperdata Browser[8]. However, our approach can be distinguished by its ability to cater to a variety of users by capturing their behavior during the course of their browsing. Moreover, due to the use of the adapted MeatAnnot system, it is able to store and reuse any annotation generated from visited Web pages Regarding future work, there are several ways to improve the relevance of the ranking of documents according to the user profile. It would also be beneficial to be able to use several domain ontologies at the same time in order to broaden the browsing perspective.

Funding by the Sealife project (IST-2006-027269) is kindly acknowledged. We thank Faiza Hansraj for her help in validating the extraction method and Gawesh Jawaheer for providing us the NeLI web logs.

6. References

[1] Corby O., Dieng-Kuntz R., Faron-Zucker C., Gandon F., Searching the Semantic Web: Approximate Query Processing Based on Ontologies, IEEE Intelligent Systems, Vol. 21, No. 1, pp. 20-27, 2006
[2] Khelif, K., Dieng-Kuntz, R., and Barbry, P., An ontology-based approach to support text mining and information retrieval in the biological domain. In Journal of Universal Computer Science (JUCS), Vol. 13, No. 12, pp. 1881-1907, 2007
[3] Mrabet Y., Khelif K., and Dieng-Kuntz R., Recognising professional-activity groups and web usage mining for web browsing personalisation. In International Conference on Web Intelligence, 2007
[4] Munzner T., Burchard, P.: Visualizing the structure of the World Wide Web in 3D hyperbolic space., Computer Graphics, pages 33–38. ACM Press, 1995
[5] Berners-Lee, T., Hollenbach, J., Lu, K., Presbrey, J., Pru d'ommeaux, E. and schraefel, m. c. , *Tabulator Redux: Writing Into the Semantic Web*, 2007
[6] Diallo G., Kostkova P., Jawaheer G., Jupp S., Stevens R. Process of Building a Vocabulary for the Infection Domain,, 21st IEEE Int. Symposium on Computer-Based Medical Systems, pp. 308-313, 2008
[7] Gridinoc L., Sabou M., D'Aquin M., Dzbor M. and Motta E., Semantic Browsing with PowerMagpie, , ISBN:978-3-540-68233-2, pp. 802-806, 2008
[8] Schroeder M., Burger A., Kostkova P., Stevens R., Habermann B. and Dieng-Kuntz R.. Sealife: A Semantic Grid Browser for the Life Sciences Applied to the Study of Infectious Diseases. HealthGrid'06, 120:167--78, 2006

[7] http://cohse.cs.manchester.ac.uk/
[8] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/