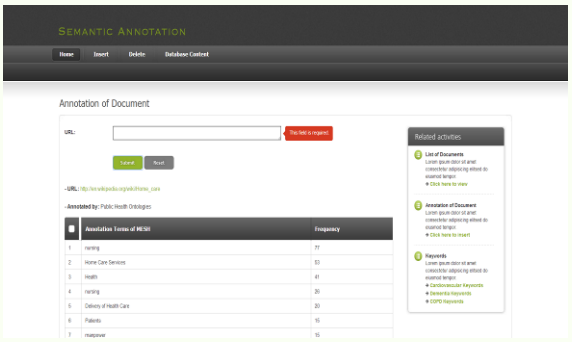
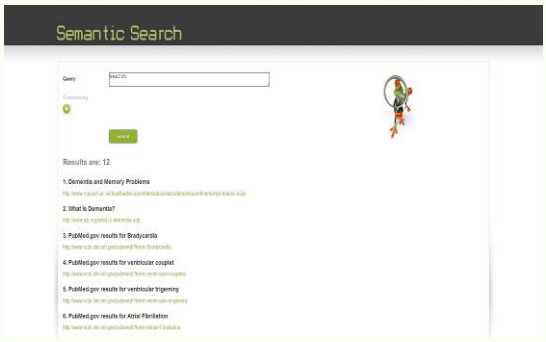
Semantic Integration for Unstructured Sources

Overview

Semantic integration for unstructured sources focuses on the effective exploitation of the numerous textual sources which are available on the web. Textual data sources, qualified as unstructured sources, are accessed through information retrieval techniques. The goal is to construct synthesized, integrated descriptions of the information in order to provide efficient retrieval methods. To meet the requirements of gathering and organizing information in order to understand the contextual meaning of data, it is important to develop techniques to semantic lift data to meaningful information. Unstructured sources are interpreted using their semantic representation, obtained by a semantic characterization process, called **semantic annotation**. Then, the processed data and their metadata can be exploited by information retrieval systems. The utility of such a system is to minimize the time required to locate information and the amount of information which must be consulted. These characteristics have driven us to a system design philosophy that is fundamentally different from that of many conventional systems. **Semantic Search** processes a collection of documents and web pages in order to present accurate results to the users according to their queries. Semantic Integration for unstructured sources focuses on gathering information from on-line sources and characterizing it semantically to find the most reliable and exact answer within disparate sources.



User Interface for Semantic Annotation



User Interface for Semantic Search

Target Domains

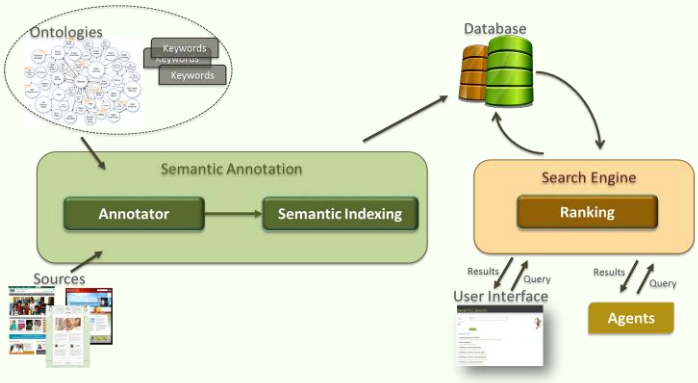
Semantic Integration framework functionalities can be applicable in any domain. For the e-Health Domain there have been the following pilot installations:

- iManageCancer - Empowering patients and strengthening self-management in cancer diseases (H2020).
- MyHealhAvatar - A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information (FP7-ICT-2011-9)
- eHealthMonitor - Intelligent Knowledge Platform for Personal Health Monitoring Services (FP7-ICT-2011-7)

Description

The first part of the Semantic Integration procedure is composed of the **annotation process** that extracts automatically the semantic information of content and the **indexing process** which organizes them efficiently. The annotator can either annotate the documents using a set of ontologies, or it can accept annotations (keywords) provided by the users. The annotations are saved into the database and processed by the indexing component. The purpose of Semantic Indexing is to optimize speed and performance in finding relevant documents for a search query. Since retrieval performance is directly related with the effectiveness of indexing, the Vector Space Model has been considered as an appropriate model to create an optimum semantic indexing vocabulary. The implementation of the Semantic Annotation engine supports the dynamic management of the searchable collection. To accomplish the above objectives and manage the collection, which constitutes the searchable data space, a Semantic Annotation user interface has been designed.

We have developed a Semantic Search system which exploits the collection of documents and their corresponding semantic information maintained in the database. Instead of using textual terms, searching can be done using semantic information. The result is a number of documents which together will provide the desired results/information. The ranking of the resulting documents based on the Vector Space Model guarantees the sorting of the documents based on their importance for a given query. Semantic Search has been designed and implemented as a web service which allows interaction with other tools or software agents.



Semantic Annotation & Semantic Search components and interactions

Additional Information

Scientific Publications

- Semantically-enabled Personal Medical Information Recommender. ISWC, 2015
- Patient Empowerment through Personal Medical Recommendations, MEDINFO, 2015

Semantic Search prototype

Contact details: **Haridimos Kondylakis** **kondylak@ics.forth.gr** *www.ics.forth.gr/isl*

Dimitris Plexousakis **dp@ics.forth.gr**