Watson: A Gateway for Next Generation Semantic Web Applications

Mathieu d'Aquin, Claudio Baldassarre, Laurian Gridinoc,
Sofia Angeletou, Marta Sabou, and Enrico Motta*

Knowledge Media Institute (KMi), The Open University, United Kingdom
{m.daquin,c.baldassarre,l.gridinoc,s.angeletou,r.m.sabou,e.motta}@open.ac.uk

The amount of knowledge published on the Semantic Web – i.e, the number of ontologies and semantic documents available online – is rapidly increasing. Following this evolution, changes appear in the way semantic applications are designed: instead of relying on one, self engineered base of semantic data, these *next generation Semantic Web applications* [2] assume the existence and availability of a large scale, distributed Web of data. In order for these applications to be able to exploit and combine the increasing amount of semantic data and ontologies available online, a *gateway to the Semantic Web* is required.

1 Watson: a Gateway to the Semantic Web

In an earlier paper [1], we presented the design and architecture of WATSON (`http://watson.kmi.open.ac.uk`), a gateway to the Semantic Web that collects, analyzes and gives access to ontologies and semantic data available online with the objective of supporting their dynamic exploitation by semantic applications. Here we summarize the technical elements of the current implementation, before briefly discussing the analysis of the content it collected.

Crawling. Collecting semantic content online is based on a set of dedicated crawlers, exploring known repositories and search engines to discover locations of semantic documents and to collect them. These crawlers are implemented using Heritrix, the Internet Archive's Crawler.[1]

Metadata extraction. Many different elements of information are extracted from the collected semantic documents: information about the entities and literals they contain, about the employed languages, about the relations with other documents, etc. The extracted information provides a valuable base of metadata, exploited for indexing, searching and characterizing the content of the Semantic Web.

User interface. Even if the first goal of WATSON is to support semantic applications, it is important to provide Web interfaces that facilitate the access to ontologies for human users. WATSON provides different "perspectives", from the most simple keyword search, to sophisticated queries using SPARQL. These interfaces are implemented in Javascript, using the principles of AJAX.

* This work was funded by the Open Knowledge and NeOn projects sponsored under EC grant numbers IST-FF6-027253 and IST-FF6-027595

[1] `http://crawler.archive.org/`

Web services. Finally, a set of Web Services and an associated API have been developed to facilitate the programatic access to the semantic content collected by WATSON for applications. This API implements access functionalities, like searching for an ontology by keywords, retrieving the entities corresponding to these keywords, the relations between these entities, or the metadata associated to ontologies and entities.

2 Characterizing the Semantic Web

Besides enabling the exploitation of the Semantic Web, WATSON can be seen as a research platform supporting the study of its content, i.e., to better understand how semantic technologies are used to publish knowledge online, what is the level of quality of online ontologies, and how these ontologies form a network of duplicated, extended, reused knowledge that can be exploited. Such a large scale analysis have been conducted on a sub-set of the content of WATSON, containing about 25 500 semantic documents. Due to space limitations, it would not be possible to detail all the results of this analysis. We briefly summarize the main conclusions.

Usage of semantic technologies. The analysis of the languages used to describe ontologies has shown that, while OWL is clearly taking over as the standard ontology representation language, the expressiveness it provides is only rarely exploited. Moreover, it appears that the division of the language into the three sublanguages Full, DL and Lite, being based on complexity considerations, is not followed in practice (most of the ontologies are OWL Full, even if they do not exploit even the expressiveness of OWL Lite).

Knowledge quality. Looking at the size and the richness of the collected documents, it appears that the Semantic Web is characterized by a very large number of lightweight ontologies, and a small number of huge semantic resources. By analyzing, the coverage of such ontologies, a *knowledge sparseness* phenomena appears: while some topic are particularly well covered (computers, society) some are almost ignored by the Semantic Web (adult, home).

The knowledge network. The main conclusion on this aspect is about the inability of tools to handle networked ontologies. Indeed, ontologies are very often duplicated, extended, revised, with these semantic relations between different (version of) ontologies not being made explicit. Another issue here concerns ontologies using the same URI without being intended to: the URI that is the most often duplicated is the default URI of ontologies edited using Protégé.

References

1. M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. WATSON: A Gateway for the Semantic Web. In *Proc. of European Semantic Web Conference, ESWC, Poster Session*, 2007.
2. E. Motta and M. Sabou. Next Generation Semantic Web Applications. In *Proc. of the 1st Asian Semantic Web Conference (ASWC)*, 2006.

3 Demo of Watson Related Technologies

WATSON exposes its automatically gathered and validated data through a number of query interfaces, among which we distinguish between the Web user interface and Web services.

The Watson Web Interface. Users may have different requirements and different levels of expertise concerning semantic technologies. For this reason, WATSON provides different "perspectives", from the most simple keyword search, to sophisticated queries using SPARQL.

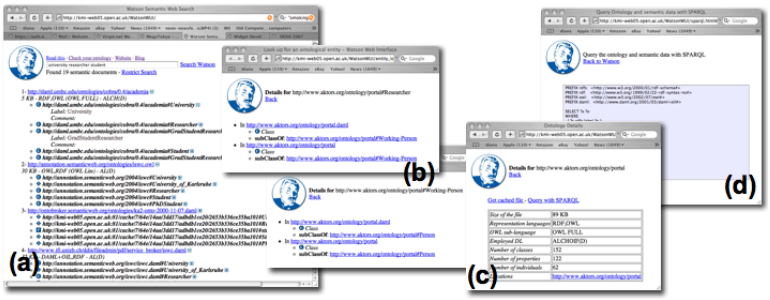

Fig. 1. The WATSON Web interface (`http://watson.kmi.open.ac.uk/WatsonWUI`).

The **keyword search** feature of WATSON is similar in its use with usual Web or desktop search systems. The result is a list of semantic documents with, for each of them, the list of entities matching each keyword (Figure 1(a)). The search can also be restricted to consider only certain types of entities (classes, properties, individuals) or certain descriptors (labels, comments, local names, literals).

A URI displayed in the result of the search is a link to a page giving the details of either the corresponding ontology or a particular entity. Since these descriptions also show relations to other elements, this allows the user to **navigate** among entities and ontologies ((Figure 1(b,c)). A page describing a semantic document provides a link to the interface for the **SPARQL endpoint** (Figure 1(d)) of this semantic document.

Applications of Watson. The Watson Web services and API can be demonstrated by some of the next generation semantic applications that employ them. For example, a **relation discovery** service explore the ontologies collected by WATSON to extract relations existing between two terms. This service is used for example in ontology matching, or for introducing ontological knowledge in folksonomies. A **Protégé plugin** has also been implemented using the WATSON API. It allows the user to retrieve existing descriptions of the entity considered in the edited ontology, and to integrate them, implementing the idea of knowledge reused in a large scale, but yet simplified way.