A versioning-based framework for semantic annotation of Web documents: OVerFA

Francesco Moscato, Beniamino Di Martino, Salvatore Venticinque, Angelo Martone
Second University of Naples, Dip. Ingegneria dell'Informazione
francesco.moscato@unina.it

Abstract: Ontology based annotation is one of the leading trend in creating metadata by using ontologies as their vocabularies. Several languages and tools exist which support annotations of different types of media.

Usually such languages are tightly coupled with specific tools and they are able to describe only those features the tools are designed to manage.

Furthermore very few (or none at all) tools and languages support annotations of multiple media types or documents or ontologies and annotations versioning, so they are not so useful in a collaborative infrastructure which aims to produce annotations of multimedia resources.

In this work the main architecture and the language to create annotations of the OVerFA (Ontology based VERsioned Framework for Annotation) framework are presented. It will be shown how the framework allows for collaborative annotation of documents of different media types and how it is possible to exploit useful features in order to produce clear and complete ontology based annotations.

Keywords: Semantic Annotation, Ontologies

1. Introduction

In the last years, as part of Semantic Web activity, a number of different systems have been designed and developed both for creating ontologies and for annotating documents.

Such systems have been mainly focused on manual and semi-automatic tooling to improve the productivity of human ontologists or annotators. Almost all of them focus on web pages and text documents annotations but some other kind of media have been considered (see for example [Assfalg et al.2003]).

Some examples of systems providing manual or automatic means for annotating media are MnM OntoMat-Annotizer, Onto-H, SHOE, AeroSWARM, Armadillo, Kim or SemTag .

Generally it is possible to distinguish annotations in respect of where they are stored (embedded or in separate files); the language used to define them (OWL, RDF etc.); the type of references to ontologies elements; the type of ontologies used for annotation: lightweight or heavyweight [Studer et al. 1998].

Lightweight ontologies include Dublin Core and thesauri and controlled vocabularies, while heavyweight ontologies include more formal specification of the "shared conceptualization" they describe.

Dublin Core annotations mainly describe properties of the document itself without providing too many details about its content Ontology-based annotations, are instead devoted to describe the content of the document, and not its general properties. Thesauri and controlled vocabularies can be used in both approaches to provide agreed terms in specific domains. Consequently, all the approaches complement each other and a generic annotation should used all of them in order to provide a document description.

Annotations can be used to introduce new elements in the ontologies they use, usually by introducing new relations of the type *isInstanceOf* between document parts and ontologies elements, or can be used to specify that a document part is *one of the* elements in the ontology.

To the best of our knowledge no system exists supporting all these types of annotations.

Availability of annotated content is one of the key challenges to overcome in order to make Semantic Web a reality [Benjamins et al. 2002].

Since annotations themselves can be modified during the time, a versioning system is appealing for managing annotations and ontologies, but at the state of the art very few annotation systems support versioning in their repositories.

A need for repositories (for example SESAME in [Kiryakow et al. 2004]) in annotation systems arises especially when dealing with not embedded annotation.

The aim of this work is to present an annotation framework (OVerFA) able to manage multi-user, collaborative annotation processes, with a repository which supports versioning.

2. OVerFA Architecture

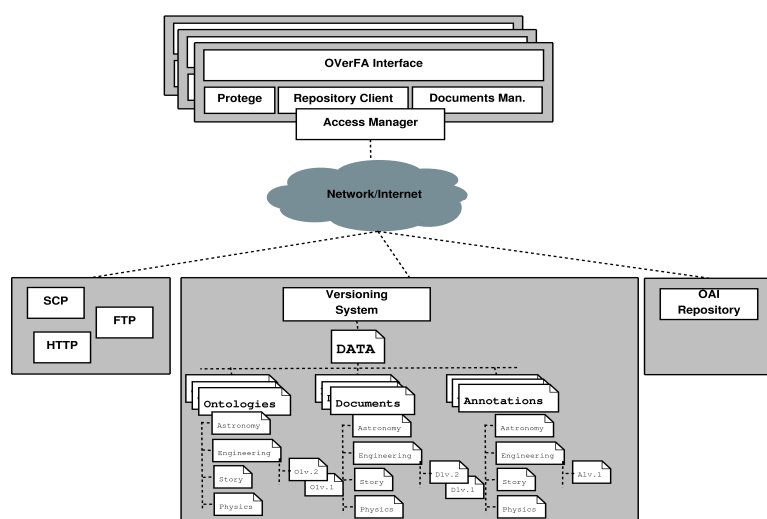The overall architecture of the OVerFA framework is depicted in Fig.1.

Figure 1: OverFA Architecture

The framework mainly consists of two parts: the annotation front-end and the repositories used to store documents, annotations and ontologies. The *Repository Client* provides the means for the access to the repositories. An *Access Manager* enables single-sign-on for protected resources. The tool presentation layer is finally provided by the *OVerFA Interface*.

The annotation front-end allows for the access to several local or remote repositories. The *native* repository of the OVerFA framework supports data versioning. The versioning system chosen for storing data is *Subversion* (http://subversion.tigris.org).

Finally, the OVerFA architecture is also designed for interact with *OAI repositories* (http://www.openarchives.org/).

3. OVerFA Framework Presentation

Authentication into the framework is the first step of an annotation process (Fig. 2). This is necessary because OVerFA was designed as a collaborative framework. In addition the authentication manager also provides a single-sign-on mechanism for the access to (eventually protected) remote repositories.
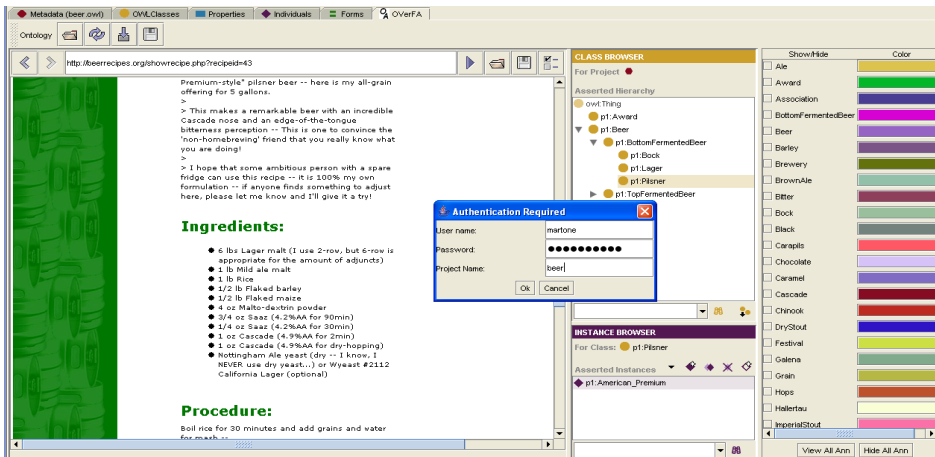
Figure 2: Authentication

The framework at the moment supports access to http, ftp, and protected subversion repositories. Annotation is performed by selecting the part(s) of the document to annotate and the element of the ontology to which annotation is related (Fig. 3). Both annotations with instances and with links to ontologies elements are allowed. Annotations may also refer to the entire document.
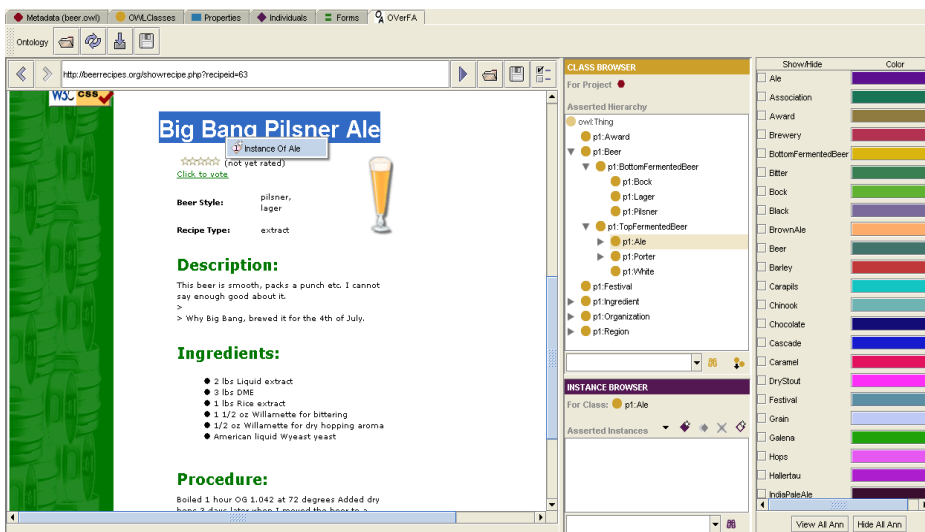

Figure 3: Annotation Example

OverFA also supports the annotation of images and multimedia documents. In Fig. 4, for example, the annotation of a part of an image is shown. The annotations produced by the OVerFA are stored into XML files. A sketch of the annotation depicted in Fig.3 is reported in the following:
.....

```
        <AnnotationURI URI="http://143.225.250.110/svn/LC3/
                            annotazioni/trunk/Annotation83.xml"
            localID="83" version="1"/>
...
    <Ontologies localID="Ontos0001">
        <Ontology localID="Onto1" version="1">
            <owl:imports rdf:resource="http://143.225.250.110/
                        svn/LC3/ontologie/trunk/beerontology.owl"/>
        </Ontology>
    </Ontologies>
    <Annotation ID="anno0001" version="1">
      <SinglePart id="doc1-spart1" idRef="doc1"> <Part id="1">
        <TextPosition from="658" />
```

```
            <TextValue>American Premium</TextValue>
        </Part>
        <Ontology idref="Ontos0001">
            <isLinkedTo>#Pilsner.American_Preremium</isLinkedTo>
        </Ontology>
    </SinglePart>
</Annotation>
```
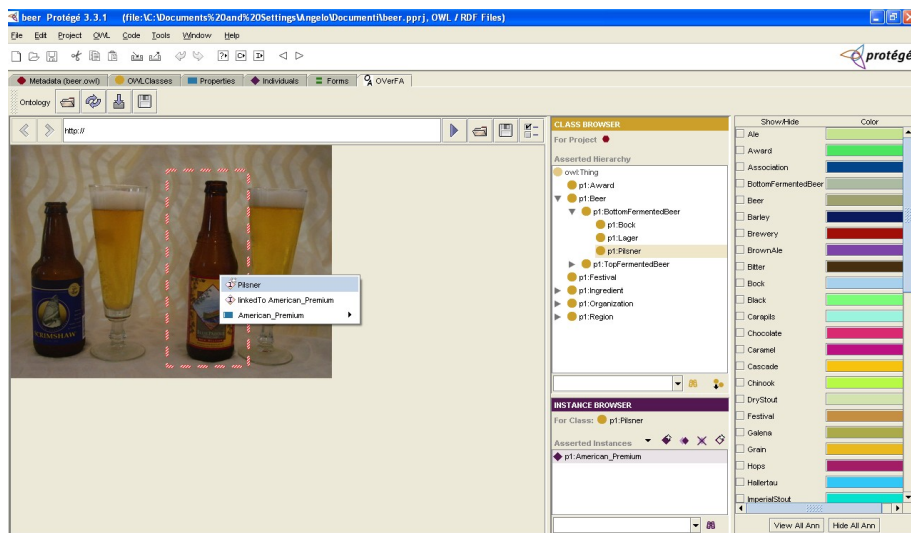

Figure 4: Image Annotation

4. Conclusions

In this work the OverFA architecture has been presented. An overview of its interface and some annotation examples have been shown. The need for repository supporting versioning in a large scale manual or semi-automatic annotation scenario has been introduced. Supporting this kind of repositories also allows for collaborative versioning processes.

As future work we aim to explore automatic merging of versioned ontologies. Furthermore, the OverFA framework will also provide facilities for automatic ontologies extraction and will be integrated with information retrieval systems to provide a unique framework for semantic annotation and data retrieval.

Bibliography

V. R. Benjamins, J. Contreras, O. Corcho, A. Gomez-Perez: "Six challanges for the semantic web" in Cristiani (Ed.): KR2002 Workshop on the Semantic Web, Toulouse. 2002, France.

J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, W.Nunziati: "Semantic annotation of soccer videos: automatic highlights identification". Computer Vision and Image Understanding, Elsevier Vol.92 (2003) pp.285-205.

R. Studer, V. R. Benjamins, D. Fensel: "Knowledge engineering: principles and methods". IEEE Trans. on Data and Knowledge Engineering, Vol. 25, Nos.1-2, pp.161-197, 1998.

A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff: "Semantic annotation, indexing, and retrieval". Web Semantics: Science, Services and Agents on the World Wide Web Volume 2, Issue 1, 1 December 2004, Pages 49-79