Language Technology Support for Semantic Annotation of Iconographic Descriptions

Kamenka Staykova
IICT, BAS, Bulgaria
staykova@iinf.bas.bg

Gennady Agre
IICT, BAS, Bulgaria
agre@iinf.bas.bg

Kiril Simov
IICT, BAS, Bulgaria
kivs@bultreebank.org

Petya Osenova
IICT, BAS, Bulgaria
petya@bultreebank.org

Abstract

The paper describes an approach for semantic annotation of multimedia objects implemented for the purposes of SINUS Project[1]. Semantic annotations are supported by semantic annotation models based on ontological presentation of knowledge concerning Bulgarian Iconography. The process of semantic annotation includes automated data-lifting procedure and user-directed approach. The paper pays attention to a specific variant of the semantic annotation process directed by the user - application of Language Technologies for semi-automated creation of semantic text annotations (tags) based on analysis of descriptive texts. The 'ontology-to-text' approach has been adapted to the needs of the iconographic domain. Initial experiments are established to support the user during the process of manual semantic annotations in the context of SINUS environment.

1 Introduction

The main objective of the research project SINUS is to provide a semantic technology-based environment facilitating development of Technology-Enhanced Learning (TEL) applications, which are able to reuse existing heterogeneous software systems. The SINUS environment has a service oriented architecture allowing unified representation and use of heterogeneous systems as Web services. The environment is tested on a use case, which applies the basic TEL principles to the process of Learning-by-Authoring (Dochev and Agre, 2009). The domain of Bulgarian Iconography is chosen for constructing a SINUS Project scenario, since it provides an interesting example of TEL in humanities. The scenario requires an intensive use of multimedia objects stored in existing heterogeneous digital libraries.

In the SINUS environment a TEL-oriented application is created hierarchically, starting by converting an autonomous system for storing and retrieving a multimedia data (digital library) to a Web service, then transforming this service into a semantically-oriented digital library facilitated by Web services and ontologies, and finally, extending the library into a learning system based on service oriented architecture.

The current paper presents the processes of semantic annotation of multimedia objects (MO) implemented in the SINUS environment. Section 2 describes the basic decisions taken for organizing such annotations. Section 3 presents the first attempts to apply language technologies in order to develop a user-directed approach for semi-automatic creation of annotations. Section 4 discusses the future work.

2 Semantic Annotation of Multimedia Objects in SINUS Project

The domain of Bulgarian Iconography is a fruitful field to show how different multimedia documents (the digital photos of iconographic works, texts, video records, etc.) could be used in TEL applications. The multimedia resources for SINUS demo-examples come from the Multimedia Digital Library "Virtual Encyclopedia of East-Christian Art" described in (Pavlova-Draganova et al., 2007) and marked as "the Library" from here on. Its content is accessible via a special Web service developed in the SINUS environment.

[1] "Semantic Technologies for Web Services and Technology Enhanced Learning" (SINUS) sinus.iinf.bas.bg

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 51–56,
Hissar, Bulgaria, 16 September 2011.

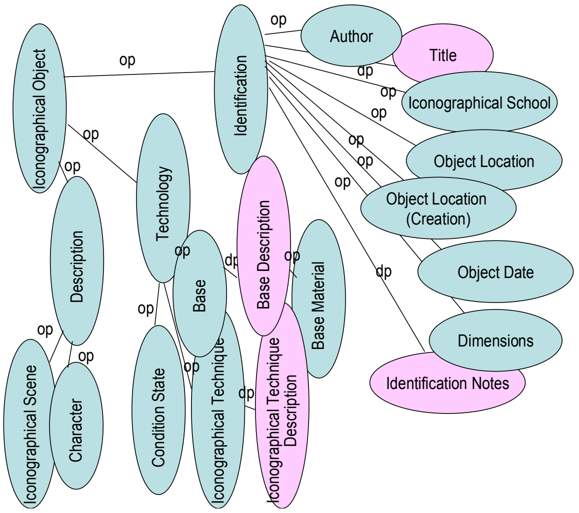
2.1 Resources of Semantic Annotation

The Objects of semantic annotation in SINUS project are multimedia objects presenting information in digital form about icons, wall-paintings, miniatures and other iconographical works; also pictures and different texts concerning the iconographical works; information about authors, places, dating periods, religious characters and so on. The Library uses a fixed annotation schema for organizing all the resources and the available data. In order to allow more flexible and deep reasoning about the iconographical knowledge, the SINUS semantic space extends that schema to present the knowledge in a formalized, ontology-like manner.

The Ontologies. SINUSBasic Ontology is the main conceptual model of the SINUS semantic space. The fixed annotation schema of the Library is taken as a ground for creating this ontology, in order access to be provided to the Library from an upper, semantic level. However, the SINUSBasic Ontology itself (with minor exceptions) is created following the main principles of the standard SIDOC-CRM (Crofts et al., 2010). The SINUSBasic Ontology is implemented in OWL and comprises 58 classes, 38 object properties and 28 data-type properties. Main classes are: *Iconographical Object* with its sub-classes *Icon*, *Wall-Painting*, *Miniature*, *Mosaic*, *Vitrage* and so on, *Author*, *Iconographical Scene*, *Character*, *Iconographical Technique*, *Base Material* and so on.

The SINUS semantic space contains the so called "specialized ontologies", which encode experts' knowledge on particular aspect of the Bulgarian Iconography domain. It is assumed that specialized ontologies represent additional, more specialized domain knowledge that is not contained in the "basic" Library. For example, the specialized ontology on religious characters gives access to such notions as *Canonical Character*, *Apostle*, *Hierarch*, etc., the specialized ontology on iconographical technology gives access to notions as *Soft Material*, *Solid Material*, *Lacquering*, *Resin*, *Primer*, *Plaster*, etc. At the current stage of work the SINUSSpec Technology ontology is implemented in OWL and could be loaded into SINUS semantic space on demand. The ontology contains 16 classes, 14 object properties and 45 ontological individuals. Some concepts of the SINUSSpec Technology ontology represent extensions of concepts introduced in SINUSBasic ontology, and in this way basic domain ontology and specialized ontologies are linked. For example, the root ontological concept of SINUSBasic Ontology is *Iconographical Object*. Such concepts as *Author*, *Iconographical School*, *Collection* are used as root-concepts in SINUSSpec Technology ontology.

Basic Semantic Annotation Model (Basic SAM) is presented in the picture bellow.



Some of the links between concepts represent object properties, others – datatype properties. Some of the object properties are realized as chains of 2 or 3 properties. Many of the datatype properties lead to textual data providing access to the descriptive texts collected in the Library.

Extended Semantic Annotation Model (Extended SAM) adds 14 new features to the Basic SAM of Iconographical Object individuals. All these additional features are supported by SINUSSpec Technology ontology as properties. For example, such features are: *base_has_component*, *gilding_has_type*, *laquering_has_evenness*, *primer_has_filler*, etc. In this way the instances of *Iconographical Object*, class defined in SINUSBasic ontology, is linked to concepts of *Primer*, *Gilding*, *Lacquering*, *Filler*, etc., defined in SINUSSpec Technology ontology.

Semantic Repository. SINUS environment employs SESAME RDF Semantic Repository that provides sufficient reasoning and standard functionalities of semantic repositories for realizing the SINUS scenario. All repository functionalities are accessible through the SINUS User Interface.

2.2 Search Process

The semantic annotation of MO in SINUS is organized as a two step process: at the first step a MO of interest should be found, and at the second step the desired new annotations should be added (manually or semi-automatically) to the object description. Semantic search of multimedia objects starts with preparing a "natural language"-like query, which is constructed on the base of described above SINUS ontologies and presented in user-friendly graphical way. The query is automatically transformed into SPARQL form, which is sent to the Extended Search Engine – a special component of the SINUS environment responsible for searching the information in the SINUS repositories. The component "lowers" the corresponding part of the query to the Library and then "lifts" the answer represented at the semantic level to the semantic repository, where the whole SPARQL query is executed. Practically, during this data lifting process some data from the Library is transformed to several SINUSBasic Ontology individuals that are added to ontologies stored in the semantic repository. The search result, which usually is a set of (identifiers to) multimedia objects, is presented to the user via the SINUS User Interface.

2.3 Semantic Annotation Process

Additional semantic annotations of MO made by the user are also supported. This user-directed semantic annotation process allows the user to add some new (specialized) annotation features to existing MO annotations or to create "basic" annotations for a new MO. The extension to the basic annotation model is supported by the SINUSSpec Technology ontology presented above. The process of user-directed semantic annotation has the following steps enabled by the SINUS User Interface:

1. All properties of a concrete object selected by the user are displayed. The number of properties depends on special ontologies the user is going to use for creating the annotations. Each property could be displayed with particular value (known annotation) or the value could be still unknown. In such case, a list of possible values of the property (stored in the corresponding ontology) is proposed as options to the user.

2. The user can either change a displayed value of a selected property (if this annotation has been created earlier by him or semi-automatically) or the user can create a new annotation by selecting a value from the corresponding list, if the current value of this property is empty.

3. After completing the annotation process and the user can save the new annotations in the SINUS semantic repository.

Opportunities for semi-automatic semantic annotation by use of descriptive texts analysis. The semantic annotation model of MO contains several links to descriptive texts concerning the MO. For example, each individual of the class Base of SINUSBasic Ontology is connected through the datatype property *has_Base_Description* to the particular text kept in the Library. An example of a short text describing the base of a particular Iconographical Object is given bellow.

```
BG: Основата е от иглолистна дървесина с
два кошака, добре запазена. Гипсов
грунд, нанесен тънко и равномерно.
```

```
EN: The base is of softwood with two
keys, well kept. Plaster ground coat,
applied thinly and evenly.
```

Most of the descriptive texts contain a lot of terminological notions of a particular domain and many of the terms are defined in the corresponding specialized ontologies. The main idea of semi-automatic semantic annotations is to help substantially the user in his/her attempt to annotate MO with notions presented in Extended Semantic Annotation Model. The support consists of access to preliminary created semantically annotated texts, which makes some (ontological) notions visible and sensitive, and also "technically" prepared to be used further in the process of used-directed semantic annotation.

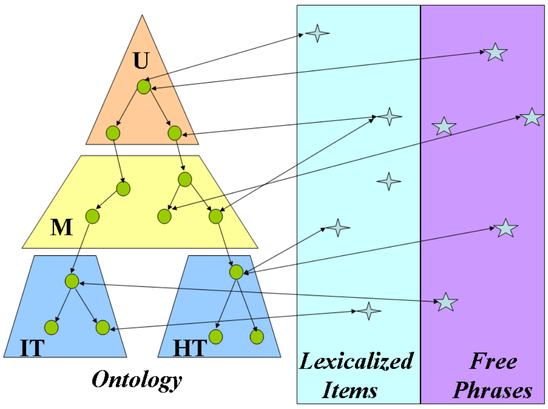
The (preliminary) semantic annotations of texts are created off-line and stored in such a way that they can be seen as indexes to MO and used for on-line searching and retrieving the objects. The text annotation procedure is implemented as a special Web service accessible from the SINUS environment. The output of this process is a set of XML files, so in order to use them in the SINUS environment they have to be accessible during the on-line process of creating new annotations. The annotations (tags) in the texts are treated as parts of preliminary semantic annotations of particular MO. They could be acknowledged, extended or denied by the user during the semantic annotation process. The annotations suggested in texts are shown to the user as "default" values of the corresponding properties of the MO. SINUS platform has to be equipped with a special procedure that "translates" the annotated text into form of Extended Semantic Annotation Model.

The task of text annotation could be formulated in this way: given an ontology and text, return annotated text, which is sensitive to the ontological notions. This general task is known as *Ontology-to-Text* relation and is still a research challenge in the crossroad of Language Technologies and Semantic Technologies. Language Technologies operate with specific methods and tools to annotate text documents semantically.

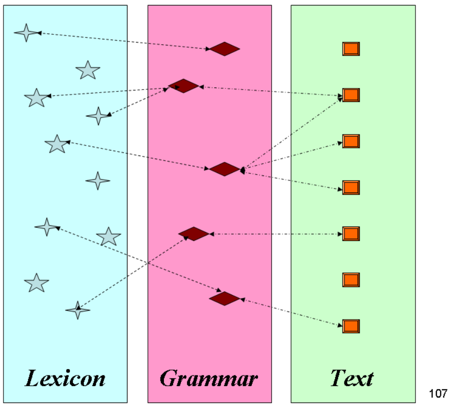
3 Semantic Text Annotation for SINUS Project

Semantic text annotation presented here is based on a model of *Ontology-to-Text* relation developed within (Simov & Osenova, 2007; Simov & Osenova, 2008). *Ontology-to-Text* relation is defined with the help of two intermediate components: (terminological) lexicon and concept annotation grammar.

The lexicon plays twofold role. First, it interrelates the concepts in the ontology to the lexical knowledge used by the grammar in order to recognize the role of the concepts in the text. Second, the lexicon represents the main interface between the user and the ontology. This interface allows the ontology to be navigated or represented in a natural for the user way. For example, the concepts and relations might be named with terms used by the stakeholders in their everyday activities and in their own natural language. This could be considered as a first step to a contextualized usage of the ontology in a sense that the ontology could be viewed through different terms depending on the context. For example, the material names will vary from very specific terms within the domain of iconography to more common names used when a set of icons are exhibited to a wider audience. As the image depicts it, the lexical items contain the following information: a term, contextual information determining the context of the term usage, grammatical features determining the syntactic realization within the text. In the current implementation of the lexicons the contextual information is simplified to two values, *common term* and *others*.



The second component of the *Ontology-to-Text* relation, the concept annotation grammar, is ideally considered as an extension of a general language deep grammar which is adapted to the concept annotation task. Minimally, the concept annotation grammar consists of a chunk grammar for concept annotation and (sense) disambiguation rules. The following picture demonstrates this part of the *Ontology-to-Text* relation.



The chunk grammar for each term in the lexicon contains at least one grammar rule for recognition of the term. As a preprocessing step we consider annotation with grammatical features and lemmatization of the text. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context, such as topic of the text, discourse segmentation, etc. Currently we have implemented chunk grammars for Bulgarian and English.

For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System (Simov et al., 2001). The structure of each grammar rule in CLaRK is defined by the following DTD fragment:

```
<!ELEMENT line (LC?, RE, RC?, RM, Com-
ment?) >

<!ELEMENT LC (#PCDATA)>

<!ELEMENT RC (#PCDATA)>

<!ELEMENT RE (#PCDATA)>

<!ELEMENT RM (#PCDATA)>

<!ELEMENT Comment (#PCDATA)>
```

Each rule is represented as a line element. The rule consists of regular expression (*RE*) and category (*RM* = return markup). The regular expression is evaluated over the content of a given XML element and could recognize tokens and/or annotated data. The return markup is represented as an XML fragment which is substituted for the recognized part of the content of the element. Additionally, the user could use regular expressions to restrict the context in which the regular expression is evaluated successfully. The *LC* element contains a regular expression for the left context and the *RC* for the right one. The element Comment is for human use. The application of the grammar is governed by *Xpath* expressions which provide additional mechanism for accurate annotation of a given XML document. Thus, the CLaRK grammar is a good choice for implementation of the initial annotation grammar.

The creation of the actual annotation grammars started with the terms in the lexicons for Bulgarian and English. Each term was lemmatized and the lemmatized form of the term was converted into regular expression of grammar rules. Each concept related to the term is stored in the return markup of the corresponding rule. Thus, if a term is ambiguous, then the corresponding rule in the grammar contains reference to all concepts related to the term.

The relation *Ontology-to-Text* implemented in this way provides facilities for solving different tasks, such as ontological search (including cross-lingual search), ontology browsing, ontology learning. In order to support multilingual access to semantic annotations we could implement the relation for several languages using the same ontology as starting point. In this way we implement a mapping between the lexicons in different languages and also comparable annotation of texts in them.

Within SINUS Project we have started the implementation of the *Ontology-to-Text* relation on the basis of the terms included in the ontology. In contrast to past applications where the concept grammars included only the concepts themselves, here also properties have been added. The relation from these terms to conceptual information is represented in two ways – direct terms for a given concept and terms for some property of a given concept. In order to keep this information within the annotation we keep it in the model. Thus, we annotated not only concrete concepts, but also fragments of conceptual information comprising a property and a concept (in the domain or the range of the property). In this way we provide annotation appropriate for future recognition of relations in the text.

The terms extracted from the ontology are lemmatized by the Bulgarian Morphological Lexicon. The lemmatized versions of the terms are converted automatically into CLaRK regular grammars which are used for the actual document annotation. In the following we present the example text from above annotated by the system. The actual annotation is done by the following format:

```
<OntoAnnotation>
… Term …
    <OntoFragment>
      … Ontology Fragment
    </OntoFragment>
</OntoAnnotation>
```

The Term is presented as a sequence of <tok> elements for each token of the term. Each token is annotated with the appropriate grammatical features. These features are used in the concept annotation grammars. The Ontology Fragment is represented by a set of <class> and <property> elements. Both kinds of elements have attribute @uri which represents the corresponding class or property identifier. This attribute is obligatory. Additionally the <property> element has @domain, @range and @value attribute. They determine the domain, range and the value of the attribute when recognized uniquely from the ontology and the annotation within the text. Bellow is given the resulting annotation for a part of our text example.

Two terms are recognized in the text extract: *Основа* (Base) and *Дървесина* (Wood). The first is annotated with one class and two properties, the second – with two classes and one property. The property in the second case received also a concrete value *дърво* (wood). At later stage the user can add a statement that the base, mentioned in the text, is made of *wood*. The user intervention is important in cases when the text contains ambiguity. The sublanguage of descriptive texts from the Library gives us the possibility to write rules for automatic addition of such statements in the future.

```
<OntoAnnotation>
  <tok ana="Ncfsd">Основата</tok>
  <OntoFragment>
   <class uri="sinus:OWLClass_Base"/>
   <property
      domain="sinus:OWLClass_Base"
      range="sinus:OWLClass_Primer"
      uri="sinus:OWLObjectProperty_base_
has_Component"/>
  <property domain="sinus:OWLClass_Base"
      range="owl:DataRange"
      uri="sinus:OWLDataProperty_base_ha
s_Cloth"/>
    </OntoFragment>
</OntoAnnotation>
    <tok ana="Vxitf-r3s">e</tok>
  <tok ana="R">от</tok>
  <tok ana="Afsi">иглолистна</tok>
<OntoAnnotation>
  <tok ana="Ncfsi">дървесина</tok>
  <OntoFragment>
    <class uri="sinus:OWLClass_BaseMa-
terial"/>
    <class uri="sinus:OWLClass_SolidMa-
terial"/>
    <property range="owl:DataRange"
      uri="sinus:OWLDataProperty_baseMa-
terial_has_Name"
      value="дърво"/>
  </OntoFragment>
</OntoAnnotation>
```

A Web service is implemented for the text annotation purposes. The input to it is a plain text. The output is an XML document according to the above format. The communication of Web service is made possible with the adoption of a RESTfull approach to the service communication with a simple but effective use of output XML files. In future the Web service will be integrated in the overall architecture of SINUS platform interacting directly with the Library and semantic repository.

4 Future Work

The experiment to support the user during the semantic annotation process with information extracted from texts is established to estimate the efforts against the benefits, and price of preliminary work on texts. The process of texts tagging (semantic text annotation) is applied for purposes of particular use-case suggested by SINUS platform for Bulgarian texts. The future work on SINUS project includes the usage of the pre-prepared annotations in texts and extensive tests on the semantic annotation process. The results will be analyzed in detail and compared to some related works as those reported in (Hare et al., 2006), (Ossenbruggen et al., 2007) and others. Another interesting topic arising here is the multilinguality and possible cross-references if the experiment is provided with texts in different languages (English, for example).

Acknowledgments

The research is supported by the National Science Fund of Bulgaria under the project No. D-002-189.

References

Crofts N., Doerr M., Gill T., Stead S., Stiff M. (editors). 2010. *Definition of the CIDOC Conceptual Reference Model*.

Dochev D., Agre G. 2009. *Towards Semantic Web Enhanced Learning*, In Proceedings. of Int. Conference on Knowledge Management and Information Sharing, Madeira, pp. 212-217.

Hare, J. S., Sinclair, P. A. S., Lewis, P. H., Martinez, K., Enser, P. G. B. and Sandom, C. J. 2006. *Bridging the Semantic Gap in Multimedia Information Retrieval: Top-down and Bottom-up approach*,. In: Mastering the Gap: From Information Extraction to Semantic Representation, 3rd European Semantic Web Conference, Budva, Montenegro.

Ossenbruggen J., Amin A., Hardman L., Hildebrand M., Assem M., Omelayenko B., Schreiber G., Tordai A., de Boer V., Wielinga B., Wielemaker J., de Niet M., Taekema J., van Orsouw M.-F., and Teesing A. 2007. *Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques*, In: Museums and the Web, pp.11-14.

Pavlova-Draganova L., Georgiev V., Draganov L. 2007. *Virtual Encyclopaedia of Bulgarian Iconography*, Information Technologies and Knowledge, vol.1, №3, pp. 267-271.

Simov K., Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, A. Kiryakov. 2001. *CLaRK - an XML-based System for Corpora Development*. In: Proc. of the Corpus Linguistics 2001 Conference, pp. 558-560.

Simov K. and P. Osenova. 2007. *Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects*. In: Proceedings of the Workshop on NLP and Knowledge Represenattion for eLearning Environments, RANLP-2007, pp. 49-55.

Simov K. and P. Osenova. 2008. *Language Resources and Tools for Ontology-Based Semantic Annotation*, In: Proceedings of OntoLex 2008 Workshop at LREC 2008, eds. Al. Oltramari, L. Prévot, Chu-Ren Huang, P. Buitelaar, P. Vossen, pp. 9-13.