

# Lightweight Document Semantics Processing in E-learning

Tomáš Gregar, Tomáš Pitner

(Masaryk University, Faculty of Informatics, Brno, Czech Republic

{xgregar, tomp}@fi.muni.cz)

**Abstract:** There are plenty of projects aimed at incorporating semantic information into present day document processing. The main problem is their real-world usability. E-learning is one of the areas which can take advantage of the semantically described documents. In this paper we would like to introduce a framework of cooperating tools which can help extract, store, visualize semantics in this area.

**Key Words:** metadata, e-learning, Web 2.0, tagging, semantics, ontology, RDFS, XML, Wiki, Deep Thought, plug-in, visualization, metadata, Fresnel

**Category:** H.5.1, H.5.2, H.3.2, H.3.3, L.1.5, L.3.0

## 1 Introduction

The concept of the Semantic Web (SW) [Berners-Lee 2001] was introduced nine years ago but the development continued a slightly different way [Horrocks 2007]. There are more causes — unrealistic requirements on the users, excessive requirements on resources, and a theoretical focus of most research projects. A lack of interoperability is also an issue. Real-life applications using semantics are simple, single purpose tools, easy to work with [Pitner et al. 2007].

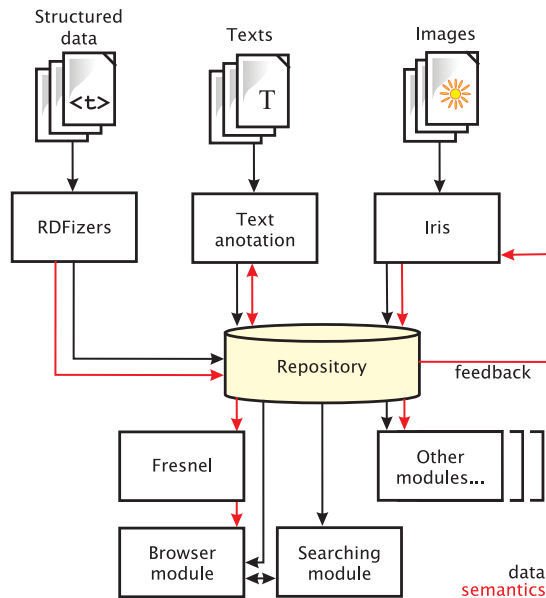
Learning is basically a knowledge transfer process. The student absorbs various explicit knowledge and also gains tacit knowledge via social interaction [Preece 2003]. An e-learning system requires an explicit formalization of the knowledge. However, users (both tutors and students) have limited computer-literacy and limited time available [Aroyo 2004, Zouaq 2007]. LM systems without semantics, like Moodle, are still the most used. We started to develop a lightweight suite of easy, small, and fast cooperating tools providing visualization capabilities that would proxy the semantic processing to a common user.

## 2 E-learning semantic framework

As our team teaches a couple of well-attended courses, we could identify the following key requirements on design and implementation of a prototype e-learning semantic framework:

- The goal is to create an interactive *tool*, not a sophisticated automaton [Berge 2002]. The tool must not be resource demanding thus we reduced requirements on knowledge storage and e-learning ontologies ‘degenerated’ into *hierarchical dictionaries* with just simple relations.

- *Open standards, open source, open interfaces* – it is together with a modular architecture vital for future development. Thus, we store semantics in *RDF Schema* and its visualization is described in *Fresnel*.
- The framework is developed with users in mind and integrated with our project management system. An interface with the university LMS is also prepared. A user tested, internationalized GUI represents another vital part of the development which is itself a part of the learning process — some modules are developed as Master- or Bachelor thesis or interim projects.



**Figure 1:** Framework structure

### 3 Repository

Key part of the prototyped framework is a *repository*, a place to store extracted semantics and documents. It comprises of the RDF(S) database *Sesame*, a *lightweight interface module*, and a *file repository* for documents. For the sake of simplicity, we avoided sophisticated systems like [Kiryakov et al. 2005].

Sesame was selected as a result of our survey [Čížek 2007]. It is an actively developed, standard-aware project. It can work with in-memory ontologies, and is sufficiently fast in management and searching. We had to make some modifications to satisfy our requirements, such as versioning of the repository content, ensuring editing atomicity, and an edit interface. Now, Sesame 2 is used.

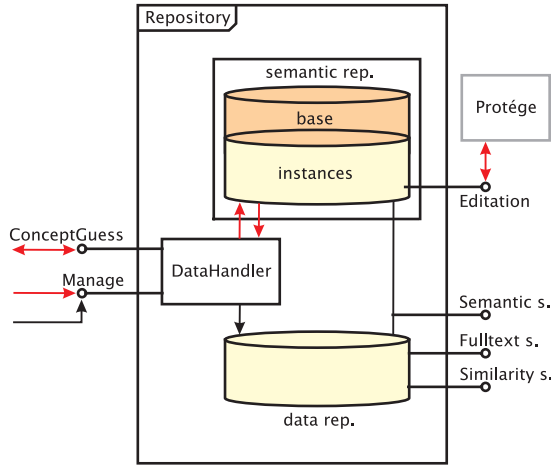


Figure 2: Repository structure

#### 4 Acquiring the semantics

The semantics of a document can be derived automatically or entered manually. For the learning purposes is an *annotation*, i.e. a facultative manual acquisition, probably the best option. It is relatively fast, lucid, and it gives the user full control power. Annotation tools should have no specialized GUI as a domain expert is not necessarily a computer expert.

The area of *text annotation* is relatively well-researched. We try to enrich tools we are using. For example, semantic wiki module of multi-project portal system Deep Thought [Gregar et al. 2009] is under development complying with the requirements in [Klamma et al. 2007].

The following area, *picture semantics acquisition*, has more issues. Project *aceMedia*, <http://www.acedmedia.org/aceMedia> was aimed to create a data framework and a semantic adaptive multimedia content management system. We adopt some of its ideas and modified them for our purposes.

Our framework subsystem for picture annotation is called *Iris*. It suggests concepts for picture segments depending on the selected domain. Users can select the proposal supplied automatically, or assign their own semantics. We proposed the following annotation algorithm in [Gregar and Pospíšilová 2006]:

1. *Segmentation* — the system is prepared to use various segmentation modules. In-house implemented module includes edge-based segmentation based on input settings and generating a multilevel segment structure.
2. *Graphics descriptor acquisition* — segment as a complex structure — must be identified by a vector of descriptors (MPEG-7-based, defined in *Visual*

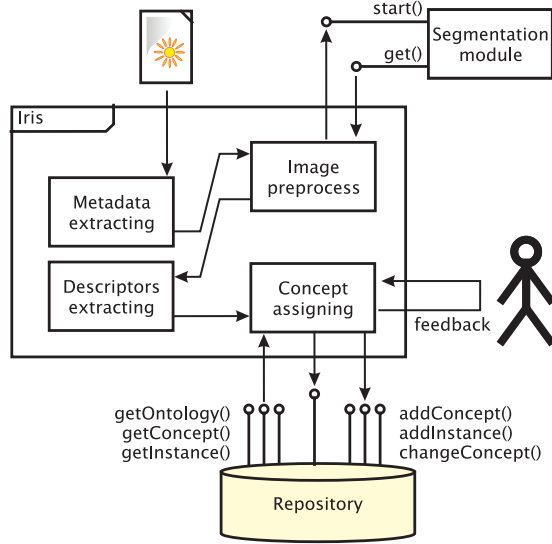


Figure 3: Structure of the IRIS

descriptor ontology, VDO) values. We gather various graphic descriptors: Dominant colors, Texture, Shape (Excentricity, Convexity, Compactness), its Size, Position; and also Structure of the segment.

Descriptors can be described by a single value (excentricity), a vector (shape), a histogram (dominant colors), or a set of histograms (texture).

- 3. Saving the picture structure — the structure is defined as a set of instances of a segment concept in the Media structure ontology and stored. Each segment is described by its VDO-values vectors.
- 4. Automatic concept proposal
- 5. User approval and usage

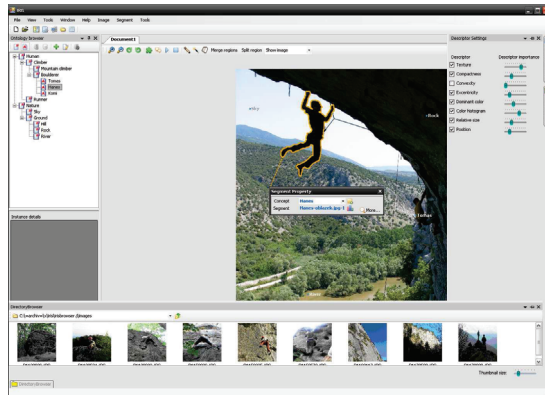
### 4.1 Automatic concept proposal

Furthermore, the domain ontology is modified to incorporate graphical information to make the annotation easier. Every concept in the ontology contains not only the definition or standard properties, but also a link to the archetypal graphical description of the concept. This description is extracted from a set of all depictions of the concept. This descriptions contains boundary values for simple descriptors or a value histogram and its weight. These descriptions allow to reduce resource requirements in the subsequent steps of the algorithm.

- 1. The domain ontology is walked throught from the top of the hierarchy (i.e. the most general terms).

2. The computed descriptor values of the processed segment and the archetypal ones are compared.
3. If the vector of the values does not correspond with the saved archetypal metadata (i.e. some descriptors are out of range), the whole concept tree is cut off from the search.
4. The pruned tree gives a list of concepts in different levels of the domain hierarchy. To find the most closest concept (i.e. processed segment can be its depiction) we utilize SVM (support vector machines) machine learning. We use all instances of the identified contestant concepts in this learning, i.e. we classify a processed segment among X classes (concepts).
5. Ontology driven revision (from the lowest level to larger segments). Modifications are driven based on the subsegment concept proposal. Other modifications can be driven by the user who can explicitly change the concept.

After the user endorses the assignment, the boundary values, histogram, and weight of the descriptors are modified. This advanced data processing is scheduled to less busy system time.



**Figure 4:** IRIS screenshot

## 5 Visualization module

Applications of the stored semantics in various domains also require some general process to visualize these data. For knowledge captured in RDF there is a proposed standard *Fresnel Display Vocabulary*. We implemented a visualization editor and browser called *Fresnel Editor*. There is currently no other editor with similar capabilities.

## 6 Conclusions and Future Work

Present semantics driven tools usually suffer from their user-complexity and/or a narrow specialization. This paper described design and implementation of a lightweight semantic framework. Its main goal was to facilitate the use of semantics in e-learning processes. Thus, we had to give up some common requirements, like logic inference or automatic semantics acquisition, and we focused on interactivity, processing speed, and user-friendliness.

The documented API of the framework enables further development of semantic tools for visualization of annotated text concepts in full text, tagging assistance, search augmentation using semantic relations, visualization of the output with Fresnel, or further development of the repository.

## References

- [Aroyo 2004] Aroyo, L. and Dicheva, D.: “The New Challenges for E-learning: The Educational Semantic Web”. *Journal of Educational Technology & Society*. October 2004. ISSN: 1436-4522.
- [Berge 2002] Berge, Z. L.: “Active, Interactive, and Reflective eLearning”. *Quarterly Review of Distance Education*, v3 n2 p181-90 Sum 2002. ISSN-1528-3518
- [Berners-Lee 2001] Berners-Lee, T.; Hendler, J. and Lassila, O.: “The Semantic Web”. *Scientific American Magazine*, May 2001. <http://www.sciam.com/article.cfm?id=the-semantic-web&print=true>.
- [Čížek 2007] Čížek, P.: “Knowledge repository”. Diploma thesis, Faculty of Informatics, Masaryk university, 2007 (in Czech).
- [Gregar et al. 2009] Gregar, T.; Pospíšilová R. and Pitner, T.: “DEEP THOUGHT – Web Based System for Managing and Presentation of Research and Student Projects”. In *Proceedings CSEDU 2009*. Lisbon.
- [Gregar and Pospíšilová 2006] Gregar, T. and Pospíšilová R.: “Visual ontology creation and its usage in e-learning”. In *Proceedings DATAKON 2006*. Masarykova univerzita, Brno, 2006. ISBN 80-210-4102-1 (in Czech).
- [Horrocks 2007] Horrocks, I.: “Semantic web: the story so far”. *ACM International Conference Proceeding Series*; Vol. 225. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*. ISBN:1-59593-590-X.
- [Kiryakov et al. 2005] Kiryakov, A.; Ognyanov, D. and Manov, D.: “OWLIM - A Pragmatic Semantic Repository for OWL”. *Web Information Systems Engineering WISE 2005 Workshops*. ISBN 978-3-540-30018-2.
- [Klamma et al. 2007] Klamma, R.; Spaniol, M. and Renzel, D.: “Community-Aware Semantic Multimedia Tagging From Folksonomies to Commonomies”. In *Proceedings of I-MEDIA 07 and I-SEMANTICS 07*.
- [Pitner et al. 2007] Pitner, T.; Derntl, M.; Hampel, T. and Motschnig, R.: “Web 2.0 as Platform for Inclusive Universal Access in Cooperative Learning and Knowledge Sharing”. *Journal of Universal Computer Science*, Graz, Austria : 2007.
- [Preece 2003] Preece, J.: “Tacit knowledge and social capital: Supporting sociability in online Communities of Practice”. *Proceedings of I-KNOW03, 3rd International Conference on Knowledge Management*. Graz, Austria. ISSN 0948-695x.
- [Zouaq 2007] Zouaq, A.; Nkambou, R. and Frasson, C.: “An Integrated Approach for Automatic Aggregation of Learning Knowledge Objects”. *Interdisciplinary Journal of Knowledge and Learning Objects Volume 3*, 2007.