enVision: An integrated approach towards Semantic Authoring

Tudor Groza, Siegfried Handschuh and **Stefan Decker**
DERI, National University of Ireland
IDA Business Park, Lower Dangan, Galway, Ireland
{tudor.groza, siegfried.handschuh, stefan.decker}@deri.org

Abstract

Writing is one of the most common activities of a researcher. And since, the writing process is not only about writing, but also about researching the background of the domain or collecting the necessary references, we propose an integrated writing environment having semantic web technologies as foundation and encapsulating all the necessities involved in this above mentioned process.

1 Introduction

Writing is one of the most common activities of a researcher. And when it comes to choosing the environment supporting it, we could affirm that LaTeXis one of the most widely adopted. But this is not enough. The scientific writing process as a whole contains also researching the background of the domain, creating the necessary references, generating (or creating) annotations for the written document and much more. By associating actions to these issues we would end up with editing, searching, browsing or annotating. Therefore, in order to provide a full support for the afore mentioned process, we need more than just an editor (having as main function, editing). We need also a search engine (for searching), a web (or reference) browser (for browsing) or an annotation mechanism (for annotating).

Existing approaches focus usually only on one of the actions enumerated above, or maybe combine some of them, but none can be considered an environment supporting all actions. Taking a look at the current LaTeXeditors (e.g. TeXnicCenter[1] or WinEdt[2] in combination with reference managers, such as JabRef[3], we observe the presence of two important actions: editing and reference browsing. But for example, no annotation support is provided, and the whole editing process involves no semantic analysis of the document. On the other side, [5] describes a full-fledged annotation mechanism, but also lacks in semantic analysis besides the fact that is built only for Microsoft WinWord. In a similar category enter also [3] and [4], with the remark that both support semantic analysis, but the first one is oriented more towards *a posteriori* annotations of PDF documents, while the second one towards web pages.

One of the only approaches coming closer to our goals is described in [1]. Although it represents an integrated environment supporting semantic authoring based mainly on NLP techniques, we argue that it would be quite hard to be used in practice, due to its logic programming style of editing. The authors need to declare certain *participants* in the document in an F-Logic manner and add statements about them in chaining style.

Our paper describes the first steps that we want to take in order to face the challenge of building an integrated writing environment supporting semantic authoring and providing support for all the actions performed while in the process of scientific writing. We will start by proposing a workflow for this process and then detail how are we going to solve each issue that appears in the different phases of the workflow.

2 enVision

2.1 The scientific writing workflow

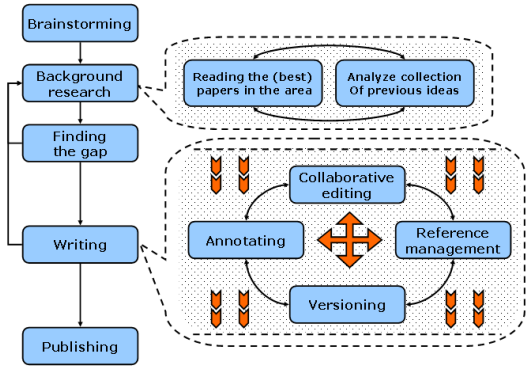

Figure 1: The scientific writing workflow.

Before designing the integrated writing environment we analyzed the necessary actions that could be present in the scientific writing process, having as inspiration the best practices of writing scientific papers. The result of our analysis is depicted in Figure 1 and represents what we call *the scientific writing workflow*. Following we will describe each phase of the proposed workflow.

[1] http://sourceforge.net/projects/texniccenter/
[2] http://www.winedt.com/
[3] http://jabref.sourceforge.net/

Brainstorming represents the phase in which the author(s) establish the topic based on which they will write the paper.

Background research. After having the topic, the next step is to realize an analysis of the background, perhaps by reading the best papers in the area and/or checking the collection of previous ideas on the same topic.

Finding the gap. This phase is probably the most important one, since it represents the starting point when writing a paper and in the end, part of the reasons for the paper's acceptance.

Writing. Presuming that the authors had studied the area enough and they discovered the right topic, the next step is to describe their ideas by writing them. The process of writing can be done in a collaborative or individual way and it presumes also keeping track of the different stages through which the document passes or enriching the document with annotations.

Publishing. This last phase includes finding a proper conference, workshop or journal for the paper and submitting it.

2.2 enVision's solutions

Just proposing the workflow it is not enough. That is why, we will now describe the way in which we intend to support its phases. Before doing so, we have to make an observation. The *Brainstorming* and *Finding the gap* phases represent for us, until now, two stages in which no system can provide guidance, and therefore, we will not include them in our analysis.

Background research. The main actions of this phase are represented by searching, ranking and classification. We intend to provide them as means of re-using the knowledge existing on the author's desktop and mainly in his personal (annotated or not) documents. In the same time, making use of existing (semantic) searching APIs, we are considering mixing the author's personal *library* with the Web. The collection of previous ideas could be extracted from annotations present in his documents or notes taken with different occasions. The knowledge re-use or the information extraction can be easily supported by creating a bridge between the environment and the semantic desktop.

Writing. Since it represents the most complex phase, we will split it in 4 possible issues:

- As already mentioned, the actual writing can take place in an collaborative environment or in an individual setting. For the first case, since the second does not represent a problem, we intend to provide synchronization facilities together with a joint mechanism for creating suggestions and taking decisions.
- Semantic annotations will be fully supported based on controlled language, NLP techniques and the document ontology[4]. The document ontology will also be the one helping the creation of the semantic structure of the document, starting from the ABCDE format[2].
- Versioning the document will be supported aswell, in individual or collaborative settings by using the document's logical structure and ontology. One solution could be the use of SemVersion[6].
- Managing and importing references represent another chapter solved by re-using the existing knowledge present on the desktop. We intend to build a small reference manager which will be able to create dynamic views over the author's personal documents and then extract the relevant references. As an addition, the author will be able to include directly quotations or paraphrase a particular paper based on his own notes or existing annotations in the document.

Publishing. This last phase will represent one of the main issues that still need a proper solution. Since there exist no well established service providing descriptions for different conferences or journals based on their domains, we intend to make use, again, of the existing searching capabilities in order to create ourselves such lists and provide proper rankings based on impact factors or other statistics collected from the Web.

3 Conclusions

Although the work described in this paper is only in the conceptual phase, we intend to continue in this direction, taking it step-by-step, in order to make sure that the resulted system supports all the actions present in our proposed workflow. In the same time, we have to take into account that this represents one of the first attempts to create a full-fledged integrated writing environment, and therefore our expectancies could be too high. As an immediate future step, we intend to collect feed-back from the community and integrate it in our conceptual design.

Acknowledgments

This work is funded by the European Comission 6th Framework Programme in context of the EU IST NEPOMUK IP - The Social Semantic Desktop Project, FP6-027705.

References

[1] Ofer Biller. Semantic authoring for multilingual text generation. Master's thesis, Ben-Gurion University of the Negev, 2005.

[2] Anita de Waard and Gerard Tel. The abcde format - enabling semantic conference proceeding. In *Proceedings of 1st Workshop: "SemWiki2006 - From Wiki to Semantics", Budva, Montenegro*, 2006.

[3] Henrik Eriksson. Support for semantic documents in protege. In *Proceedings of 8th Protege International Conference, Madrid, Spain*, 2005.

[4] Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in cream. In *Proceedings of WWW2002, May 7-11, Honolulu, Hawaii, USA*, 2002.

[5] Marcello Tallis. Semantic word processing for content authors. In *Proceedings of Second International Conference on Knowledge Capture, Sanibel, Florida, USA*, 2003.

[6] Max Völkel and Tudor Groza. Semversion: An rdf-based ontology versioning system. In *Proceedings of IADIS International Conference on WWW/Internet*, 2006.

[4] http://research.tudorgroza.org/ontology/paperont20060607.rdfs