SALT: Enriching LaTeXwith Semantic Annotations

Tudor Groza, Hak Lae Kim and **Siegfried Handschuh**
DERI, National University of Ireland
IDA Business Park, Lower Dangan, Galway, Ireland
{tudor.groza, haklae.kim, siegfried.handschuh}@deri.org

Abstract

With the emergence of the Semantic Web, encapsulating annotations and metadata into scientific documents became a widely used practice. But in the same time, we observed the lack of well established standard tools supporting this process. There exist a large number of tools used to create *a posteriori* annotations for web pages and scientific documents, with the remark that none of them can be considered a *de facto* standard. When it comes to *a priori* annotations the number decreases significantly. Our paper describes the first steps taken towards a comprehensive and elaborated approach for supporting *a priori* annotation of LaTeXdocuments with the goal of transforming the final result into a semantic document which could be used for different purposes.

1 Introduction

With the emergence of the Semantic Web, encapsulating annotations and metadata into scientific documents became a widely used practice. But, in the same time, we also observed that the current solutions have the tendency to represent compromises rather than well established standards. If we analyze the current status, based on the type of supported annotations, the number of tools creating and maintaining *a posteriori* annotations is significantly larger than the ones supporting *a priori* annotations.

Approaches like [HS02], [Tal03] or [Eri05] represent elaborated solutions supporting *a posteriori* annotations, but none of them can be considered a *de facto* standard. When it comes to *a priori* annotations, the situation is even worse. [Tal03] fits also in this category due to its integrated nature, being embedded in Microsoft WinWord and allowing the user to annotate his documents while in the process of writing. But for LaTeXand then PDF documents there is no tool providing a comprehensive solution for this issue.

Small steps have been made in this direction, one of them being represented by embedding metadata in PDF documents using Adobe XMP [1]. XMP's main issue is that is limited to only some of the DublinCore[2] elements. Although it is extensible, the usual PDF readers are not able to analyze and display the extended part.

Some other means of annotating PDF documents is making use of notes, bookmarks or markups. If the bookmarks are usually created directly from LaTeXwithout much support from the writer, the other two possibilities are not widely used, although they are quite powerful instruments, especially because of their greater visual impact. Our goal is to make the first steps towards providing an elaborated methodology together with the associated tool for supporting *a priori* annotations in LaTeXand then PDF documents.

2 Semantically Annotated LaTex

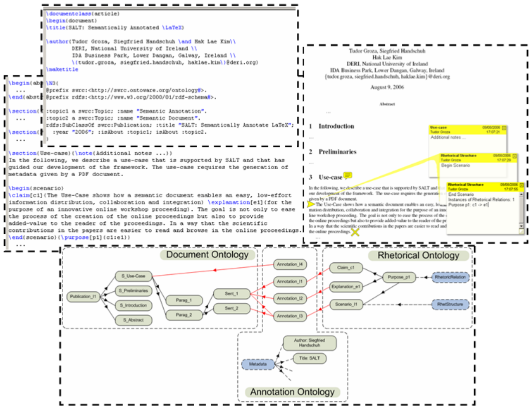
SALT divides the workflow in two processes. The first process, proposes the actual methodology and support for annotating PDF documents, when using LaTeX. The second one was introduced for demonstration purposes, i.e. how the metadata created with SALT can be used in a workshop usecase. For each of the two processes, SALT contains a independent module providing the necessary functionality. The following sections will describe briefly both processes with the associated modules.

2.1 The annotation process

The annotation process takes place during the scientific writing process(*a priori* annotation). This is not mandatory, but we believe that mixing the two processes (i.e. annotation and writing) should be natural. The annotation support provided by SALT is based on a layered organization of three ontologies:
The document ontology - captures the structural information of the document. The population of this ontology is done automatically at compilation time and it is based on the standard LaTeXcommands for organizing the text structure.
The rhetorical ontology - captures the rhetorical structure of the text and the rhetoric relations existing between the annotated parts of the text. The rhetorical structure represents an extension of the ABCDE format [dWT06], while the rhetoric relations have as foundation the Rhetorical Structure of the Text Theory [TM06]. From the annotation process point of view, the writer has to make use of a series of new

[1] http://www.adobe.com/products/xmp

[2] http://dublincore.org/



```
\documentclass(article)
\begin(document)
\title(SALT: Semantically Annotated \LaTeX)

\author(Tudor Groza, Siegfried Handschuh \and Hak Lae Kim\\
        DERI, National University of Ireland \\
        IDA Business Park, Lower Dangan, Galway, Ireland \\
        \(tudor.groza, siegfried.handschuh, haklae.kim\)@deri.org)
\maketitle

\N3(
@prefix swrc:<http://swrc.ontoware.org/ontology#>.
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>.

:topic1 a swrc:Topic; :name "Semantic Annotation".
:topic2 a swrc:Topic; :name "Semantic Document".
rdfs:SubClassOf swrc:Publication; :title "SALT: Semantically Annotate LaTeX";
  :year "2006"; :isAbout :topic1; isAbout :topic2.
)
```

```
\begin(ab
  ...
\end(abst
\section(
  ...
\section(
  ...
\section(Use-case)(\note(Additional notes ...))
In the following, we describe a use-case that is supported by SALT and that has
guided our development of the framework. The use-case requires the generation of
metadata given by a PDF document.

\begin(scenario)
\claim[c1](The Use-Case shows how a semantic document enables an easy, low-effort
information distribution, collaboration and integration) \explanation[e1](for the
purpose of an innovative online workshop proceeding). The goal is not only to ease
the process of the creation of the online proceedings but also to provide
added-value to the reader of the proceedings. In a way that the scientific
contributions in the papers are easier to read and browse in the online proceedings.
\end(scenario)(\purpose[p1](c1:e1))
  ...
```

Figure 1: The result of the annotation process using SALT.

LATEXcommands and of a N3-like syntax in order to be able to create rhetoric elements and the appropriate relations between them. Therefore, the impact on the writers already familiar with LATEXis minimal, since the newly introduced commands respect the same pattern.

The annotation ontology - represents the semantic bridge between the two afore-mentioned ontologies. The instantiation of this ontology is done automatically.

Figure 1 shows the final result of the annotation process. When compiling the enriched LATEXdocument, the annotations are extracted, the following step being the creation of the appropriate ontology instances in RDF format and the creation of the visual annotations that will be represented in the PDF document as notes. The resulted chunk of RDF data is embedded in the final PDF file in the XMP field. A more detailed description of the ontologies and of the entire annotation process can be found in [GHK06].

2.2 The workshop usecase

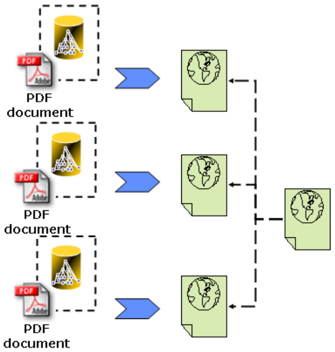


Figure 2: Information workflow in the workshop proceedings publication scenario.

We started modeling the scenario from the presumption that a workshop needs a collection of web pages containing information about the papers accepted at the workshop, together with an associated index page. In order to automate the process as much as possible, the information should be extracted directly from the PDF files. The second process provided by SALT deals with the functionality needed by this scenario and even more, it takes advantage of the rich annotation support created previously. The final result is achieved in two steps. Figure 2 depicts the information workflow. The first step takes each PDF file annotated with SALT, extracts the metadata and thus the embedded instances of the document and rhetorical ontology, processes the information extracted and creates the appropriate HTML file. The second step is represented by an iteration over all HTML files in order to create the index.

3 Conclusion and future work

The current status of SALT represents the first step towards the development of an integrated writing environment supporting semantic authoring. As an intermediate phase, we intend to establish a standard for describing *a priori* annotations using the document ontology, by improving its definition and representation with the goal of capturing more semantics in it. In terms of tool development, we will continue the integration of useful functionalities and in parallel putting an accent on efficiency and user-friendliness.

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. Special thanks to Big Faceless Organization (`big.faceless.org`) for providing the PDF library used in the metadata analysis process.

References

[dWT06] 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.

[Eri05] Henrik Eriksson. Support for semantic documents in protege. In *Proceedings of 8th Protege International Conference, Madrid, Spain*, 2005.

[GHK06] Tudor Groza, Siegfried Handschuh, and Hak Lae Kim. Salt: Semantically annotated latex. In *Proceedings of 1st Semantic Authoring and Annotation Workshop, co-located with ISWC 2006, Athens, Georgia, USA*, 2006.

[HS02] Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in cream. In *Proceedings of WWW2002, May 7-11, Honolulu, Hawaii, USA*, 2002.

[Tal03] Marcello Tallis. Semantic word processing for content authors. In *Proceedings of Second International Conference on Knowledge Capture, Sanibel, Florida, USA*, 2003.

[TM06] Maite Taboada and William C. Mann. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8, No. 3:423–459, 2006.