S-CREAM — Semi-automatic CREAtion of Metadata

Siegfried Handschuh[1] and **Steffen Staab**[1] and **Fabio Ciravegna**[2]
[1] Institute AIFB, University of Karlsruhe
sha,sst@aifb.uni-karlsruhe.de
[2] Department of Computer Science , University of Sheffield
F.Ciravegna@dcs.shef.ac.uk

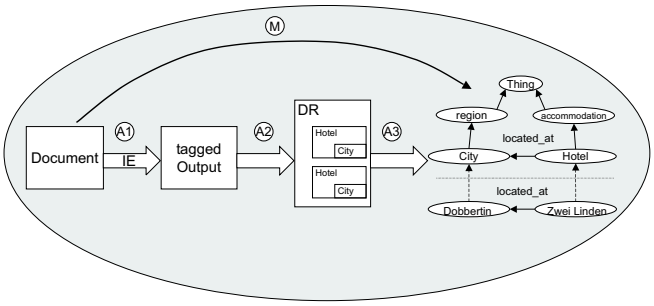

Figure 1: Manual and Automatic Annotation

Introduction. The Semantic Web builds on metadata describing the contents of Web pages. In particular, the Semantic Web requires relational metadata, i.e. metadata that describe how resource descriptions instantiate class definitions and how they are semantically interlinked by properties. To support the construction of relational metadata, we have provided an annotation and authoring [HS02] framework (CREAM — manually CREAting Metadata) and a tool (OntoMat-Annotizer) that implements this framework. Nevertheless, providing plenty of relational metadata by annotation, i.e. conceptual mark-up of text passages, remained a laborious task.

Though there existed the high-level idea that wrappers and information extraction components could be used to facilitate the work , a full-fledged integration that dealt with all the conceptual difficulties was still lacking. Therefore, we have developed S-CREAM (Semi-automatic CREAtion of Metadata), an annotation framework that integrates a learnable information extraction component (viz. Amilcare [Cir01]).

Amilcare is a system that learns information extraction rules from manually marked-up input. S-CREAM aligns conceptual markup, which defines relational metadata, (such as provided through OntoMat-Annotizer) with semantic and indicative tagging (such as produced by Amilcare).

Synthesizing S-CREAM In order to synthesize S-CREAM out of the existing frameworks CREAM and Amilcare, we consider their core processes in terms of input and output, as well as the process of S-CREAM. Figure 1 surveys the three processes. The first process is indicated by a circled M. It is manual annotation of metadata, which turns a document into relational metadata that corresponds to the given ontology. The second process is indicated by a circled A1. It is information extraction, e.g. provided by Amilcare. When comparing the desired relational metadata from manual markup and the semantic tagging provided by information extraction systems, one recognizes that the output of this type of systems is underspecified for the purpose of the Semantic Web. In particular, the nesting of relationships between different types of concept instances is undefined and, hence, more comprehensive graph structures may not be produced. In order to overcome this problem, we introduce a new processing component, viz. a lightweight module for discourse representation. This third process is given in Figure 1 is indicated by the composition of A1, A2 and A3. It bridges from the tagged output of the information extraction system to the target graph structures via an explicit discourse representation. Our discourse representation is based on a very lightweight version of Centering.

Conclusion. The new version of S-CREAM presented here supports metadata creation with the help of information extraction in addition to all the other nice features of CREAM, like comprises inference services, crawler, document management system, ontology guidance/fact browser, document editors/viewers, and a meta ontology. OntoMat-Annotizer is the reference implementation of the S-CREAM framework. It is Java-based and provides a plugin interface for extensions for further advancements, e.g. collaborative metadata creation or integrated ontology editing and evolution.

References

[Cir01] Fabio Ciravegna. Adaptive information extraction from text by rule induction and generalisation. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)e*, Seattle, Usa, August 2001.

[HS02] Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in cream. In *Proceeding of the WWW2002 - Eleventh International World Wide Web Conferenceb (to appear)*, Hawaii, USA, May 2002.