A Metadata Generation Framework for Heterogeneous Information Sources

Andreas D. Lattner
(Center for Computing Technologies, University of Bremen, Germany,
adl@tzi.de)

René Apitz
(Institute of Production Engineering and Machine Tools, University of Hannover, Germany
apitz@ifw.uni-hannover.de)

Abstract: The engineering domain features various information sources and is heterogeneous in multiple ways. Managing its specific knowledge requires adapting to these characteristics. In order to ensure the success of software that is to support the exchange and use of information within this domain, the extraction of relevant metadata should be adequately supported. Since the meta information describing the information items is as heterogeneous as they are themselves, various means have to be combined in a modular architecture.
Key Words: Metadata Generation, Knowledge Management
Category: H3.1

1 Introduction

In today's economy, products and processes tend to increase in complexity. On the other hand, increasing competition results in shorter product life cycles and forces companies to cut down development times. In this context, the efficient reuse of existing information as well as the acquisition of new knowledge are crucial for technological and economical success. This is especially important for knowledge intense fields such as the engineering domain.

The mere storage of information is no longer a significant problem. However, finding the right piece of information at the right time is still a challenge. With the amount of available information increasing, the problem of retrieving what is really needed becomes more and more important. Annotating metadata to information objects offers new means of querying and retrieving information items. Metadata is any information describing other information objects. It can be annotated manually, but as this is a time-consuming process and engineers tend to information channels that are easily accessible [Allen and Hall 1993], such a procedure is not practicable – the generation of metadata should be automated.

2 Related Work

Topics related to metadata have been treated by many different researchers from various fields like digital libraries, information retrieval, machine learning and semantic web. [Kashyap and Sheth 1996] describe the function of metadata to "capture the essential information in the underlying data independent of

Lattner, A. D.; Apitz, R.: A Metadata Generation Framework for Heterogeneous Information Sources. Proceedings of the 2nd International Conference on Knowledge Management (I-KNOW '02), Graz, Austria, July 11-12, 2002, pp. 164-169.

representational details". They classify metadata into content independent and content dependent metadata. Referring to [Kashyap and Sheth 1996] content dependent metadata can be directly content-based or content-descriptive. In the latter case a distinction between domain specific and domain independent metadata is possible.

Manually created or learned classifiers can be applied for the automatic creation of metadata. [Sebastiani 2002] gives an overview of machine learning in automated text categorization and discusses the application of various symbolic and sub-symbolic learning algorithms for text categorisation. By using information extraction rules certain values for meta attributes can be extracted from text. The task of creating extraction rules can also be automated by learning [Junker et al. 1999]. Creating metadata for more structured sources than plain text has been subject to prior scientific work as well. For completely structured information like databases, wrappers can be used to extract information. Machine learning can also be applied for the creation of database wrappers [Kushmerick et al. 1997].

Web pages are a popular example for semi-structured documents. The annotation of metadata to web pages has recently become important, as the vast quantity of quite unstructured web information cannot be used reasonably. Two examples for annotating metadata to web pages are the "Simple HTML Ontology Extensions" (SHOE, [Heflin and Hendler 2000]) language and the framework for "Creating Relational, Annotation-based Metadata" (CREAM, [Handschuh et al. 2001]). The classification of web pages has been treated by various authors (e.g. [Pierre 2001]). [Jenkins et al. 1999] have investigated the automatic RDF metadata generation for HTML documents. An approach for ontology-based metadata generation by applying manually created rules is presented by [Stuckenschmidt and van Harmelen 2001]. The application of inductive logic programming for learning structural classification rules for web-page categorization is introduced in [Stuckenschmidt et al. 2002]. The Web→KB project aims at automatically creating computer understandable knowledge bases from the WWW [Craven et al. 2000].

3 Engineering Information Sources

Engineering knowledge is heterogeneous in multiple ways. First of all, it can be tacit or explicit. Tacit knowledge resides in the individual's experience and action, whereas explicit knowledge is codified and communicated in symbolic form or language [Shin et al. 2001]. Tacit knowledge is often regarded to be very important for engineering work, as engineers rely heavily on information such as communication or the examination of physical objects to acquire knowledge [Allen and Hall 1993]. However, the distinction between tacit and explicit knowledge is not as clear as it seems, since tacit knowledge can be created by explicit knowledge and vice versa. Still, it is obviously impossible to add meaningful metadata to tacit knowledge, and in spite of the importance of tacit knowledge, a lot of explicit information remains to focus on.

Explicit knowledge can be found in unstructured, semi-structured and structured information items. For example, technical reports are usually rather unstructured files, whereas CAD models are at least structured in terms of representing the whole product as a summary of its components or features. The other extreme of highly

Lattner, A. D.; Apitz, R.: A Metadata Generation Framework for Heterogeneous Information Sources. Proceedings of the 2nd International Conference on Knowledge Management (I-KNOW '02), Graz, Austria, July 11-12, 2002, pp. 164-169.

structured information can be found in data bases. Furthermore, explicit engineering knowledge is also heterogeneous in terms of being derived from various systems, programs and platforms. Finally, these information sources can be internal (i.e. they reside within the company itself) or external. This heterogeneity of information sources calls for an integrated modular approach that combines different technologies.

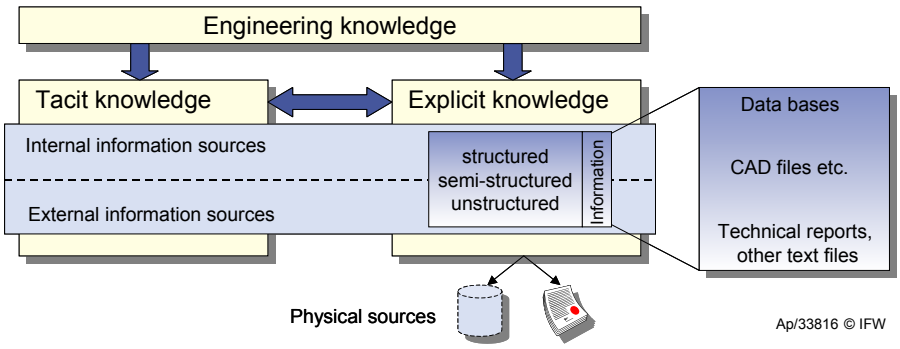

Figure 1: Heterogeneity of engineering information sources

4 Metadata Generation Framework

The aim of the presented approach is to build a unified framework for creating metadata from arbitrary information sources considering the following demands:

- identifying the type of the information source,
- automatically creating properties (attributes and relations to other objects) of the regarded item,
- supporting different kinds of metadata, and
- allowing extensibility by a plug-in mechanism.

As much information as possible like e.g. the content of the information source, existing metadata, context information or a description of the existing information types should be accessible for the modules that generate metadata.

Figure 2 shows the framework for generating metadata. The metadata generation module has access to different sources of background information. Besides the content of the source and existing content independent metadata, context information can be used (e.g. the organisational context of the person who added the information item or the task where the source has been created). The access to ontologies provides information which types of information objects exist and what their properties are. The metadata generation module can also access information sources that have already been annotated and use them for creating new metadata on unknown sources e.g. by rule learning.

Lattner, A. D.; Apitz, R.: A Metadata Generation Framework for Heterogeneous Information Sources. Proceedings of the 2nd International Conference on Knowledge Management (I-KNOW '02), Graz, Austria, July 11-12, 2002, pp. 164-169.

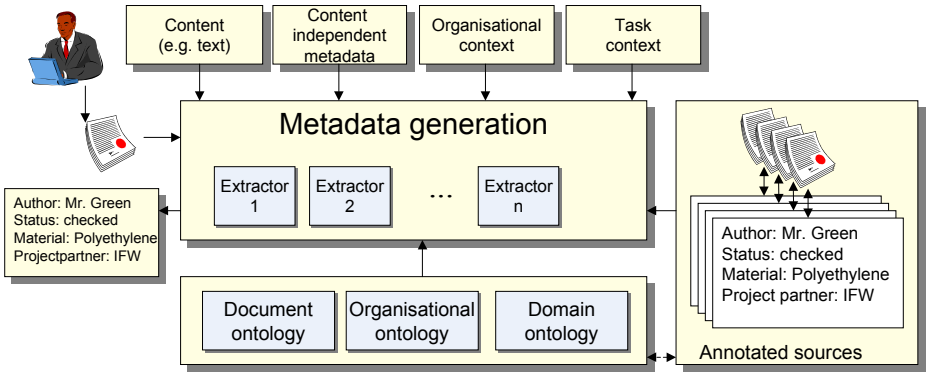

Figure 2: Framework for metadata generation

The presented architecture makes it possible to import arbitrary modules ('extractors') for creating metadata, which allows to create different kinds of metadata. It is also possible to incrementally create metadata by using data generated by previous extractors, which potentially improves effectiveness and allows to validate previously generated values by following extractors. An example is given in Figure 3.

The model describing the different information types may form an ontology. This ontology has to define object classes and the relations between them for the existing information sources. Referring to [Gruber 1994] an ontology is the "explicit specification of a conceptualization". The ontology specifies what metadata has to be generated. The extractor plug-ins can thus refer to the modelled properties of object types. In order to identify which extractor can create certain values for meta attributes, a mapping between type properties and responsible extractors has to be defined.

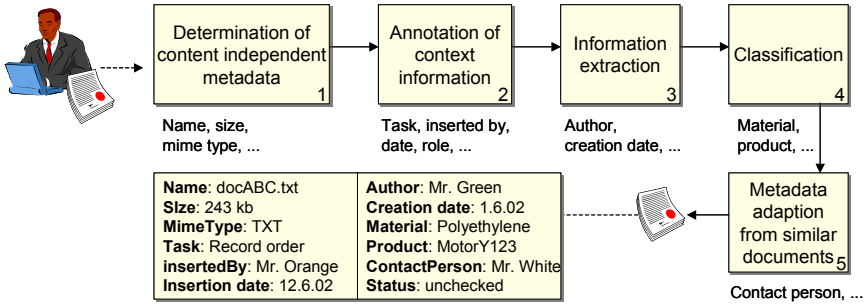

Figure 3: Example for metadata generation

We use an object-oriented representation with a separation between schema and instance layer for describing information types and their instances. At the schema layer the existing types form a class hierarchy with derived types inheriting the properties of their superior types. Each type is described by a set of attributes, that can be atomic (e.g. strings or numbers) or refer to complex information types. The

Lattner, A. D.; Apitz, R.: A Metadata Generation Framework for Heterogeneous Information Sources. Proceedings of the 2nd International Conference on Knowledge Management (I-KNOW '02), Graz, Austria, July 11-12, 2002, pp. 164-169.

modelling of set-valued attributes is possible as well. The instance layer consists of object instances of these information types with concrete values for their attributes.

The possibility to integrate arbitrary extractors allows to treat unstructured, semi-structured, and structured information sources with the same system. The different plug-ins can use different techniques as needed. Figure 4 shows the general structure of the extractor modules. The information source with possible existing metadata is passed to the extractor, that can access previously learned rules and the annotated sources for creating values for attribute of the new information source. Many different implementations for extractors are can be integrated like e.g.:

- Classifiers, that classify objects into one (or more) categories.
- Information extractors, that extract information from the content (e.g. text) of a information source.
- Context annotators, that annotate context information (e.g. task context of the user) from the main application.
- Wrappers, that adapt information from (semi-) structured information sources, possibly with mapping values to other value sets.
- Other extractors, that e.g. copy values with a high probability of appearance.

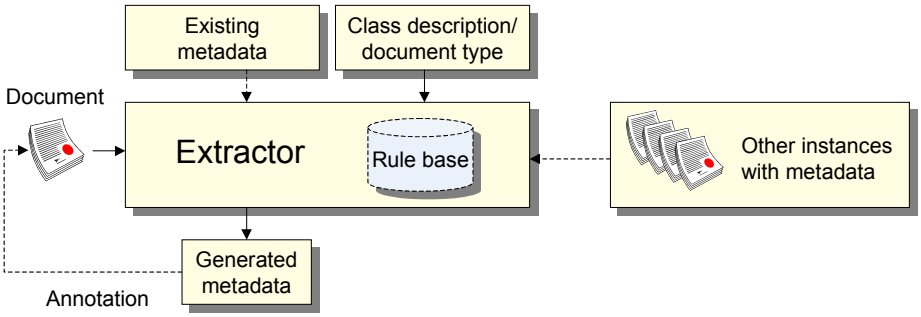

Figure 4: Extractor module

5 Conclusion

Engineering knowledge is heterogeneous in multiple ways. Since the metadata describing the information items is as heterogeneous as they are themselves, the creation of metadata cannot follow a single approach. The presented framework allows to integrate different technologies into a single tool with a modular architecture and to generate and validate metadata from heterogeneous sources in consecutive steps and iterations.

Acknowledgements

The content of this paper is an outcome of the KnowWork project [Tönshoff et al. 2001], which is funded by the German Office for Education and Research (BMBF) and is supervised by the German Aerospace Center (DLR). The authors wish to acknowledge these institutions for their support. We also wish to acknowledge our gratitude and appreciation to our projects partners for their contribution during the development of ideas and concepts presented in this paper.

References

[Allen and Hall 1993] Allen, K.; Hall C. M.: "The information-seeking behaviour of engineers", Encyclopedia of Library and Information Science, Vol. 52. New York, Basel, Hong Kong: Dekker, 1993, p. 167-201.

[Craven et al. 2000] Craven, M.; Dipasquo, D.; Freitag, D .; McCallum, A.; Mitchell, T.; Nugam, K.; Slattery, S.: "Learning to Construct Knowledge Bases from the World Wide Web", Artificial Intelligence, 118(1-2), 2000, p. 69-113.

[Gruber 1994] Gruber, T. R.: "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", In: International Journal of Human-Computer Studies, Vol. 43, No. 3, 1994, p. 907-928.

[Handschuh et al. 2001] Handschuh, S.; Staab, S.; Maedche, A.: "CREAM – Creating Relational Metadata with a Component-based, Ontology-driven Annotation Framework", In: Proceedings of the First ACM Conference on Knowledge Capture (K-CAP'01), 2001.

[Heflin and Hendler 2000] Heflin, J.; Hendler, J.: "Searching the Web with SHOE", In: Artificial Intelligence for Web Search. Papers from the AAAI Workshop. WS-00-01. AAAI Press, 2000, p. 35-40.

[Jenkins et al. 1999] Jenkins, C; Jackson, M.; Burden, P.; Wallis, J.: "Automatic RDF Metadata Generation for Resource Discovery", Computer Networks, 31, 1999, p. 1305-1320.

[Junker et al. 1999] Junker, M.; Sintek, M.; Rinck, M.: "Learning for Text Categorization and Information Extraction with ILP", Proceedings of the Workshop on Learning Language in Logic, 1999.

[Kashyap and Sheth 1996] Kashyap, V; Sheth, A.: "Semantic Heterogenity in Global Information Systems: The Role of Metadata, Context and Ontologies", In: Papazoglou, M.; Schlageter, G. (eds.): Cooperative Information Systems, Academic Press, 1996, p. 139-178.

[Kushmerick et al. 1997] Kushmerik, N.; Doorenbos, R.; Weld, D.: "Wrapper induction for information extraction", Proceedings of the 15th International Joint Conference on Artificial Intelligence, 1997.

[Pierre 2001] Pierre, J. M. : "On the Automated Classification of Web Sites", Linkoping Electronic Articles in Computer and Information Science, Vol. 6, 2001.

[Sebastiani 2002] Sebastiani, F.: "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 34(1), 2002, p. 1-47.

[Shin et al. 2001] Shin, M.; Holden, T.; Schmidt, R. A.: "From knowledge theory to management practice: towards an integrated approach", Information Processing and Management 37 (2001), p. 335-355.

[Stuckenschmidt and van Harmelen 2001] Stuckenschmidt, H.; van Harmelen, F.: "Ontology-based Metadata Generation from Semi-Structured Information", Proceedings of the 1st International Conference on Knowledge Capture (K-CAP 2001), Morgan Kaufmann, 2001.

[Stuckenschmidt et al. 2002] Stuckenschmidt, H.; Hartmann, J.; van Harmelen, F.: "Learning Structural Classification Rules for Web-page Categorization". Accepted for Special Track on the Semantic Web at FLAIRS'02, 2002.

[Tönshoff et al. 2001] Tönshoff, H. K.; Apitz, R.; Lattner, A. D.; Schlieder C.: "KnowWork – An Approach to Co-ordinate Knowledge within Technical Sales, Design and Process Planning Departments", Proceedings of the 7th International Conference on Concurrent Enterprising, Bremen, Germany, 27 – 29th June 2001, p. 231-239.

Lattner, A. D.; Apitz, R.: A Metadata Generation Framework for Heterogeneous Information Sources. Proceedings of the 2nd International Conference on Knowledge Management (I-KNOW '02), Graz, Austria, July 11-12, 2002, pp. 164-169.