Demo Abstract: Semantic Annotation of paper-based Information

Heiko Maus, Andreas Dengel
German Research Center for AI
Knowledge Management Department
Kaiserslautern, Germany
{firstname.lastname}@dfki.de
http://www.dfki.de/km/

Abstract

The demo shows how a user is easily able to semantically annotate and enrich scanned documents with concepts from his personal knowledge space. This bridges the gap between paper documents and the approach of a Semantic Desktop, where each information object on the user's computer desktop is semantically described by means of ontologies.

1. Motivation

The demo of the SCETagTool shows how a user is easily able to semantically annotate and enrich scanned documents with concepts from his personal knowledge space. This bridges the gap between paper documents and the approach of a Semantic Desktop, where each information object on the user's computer desktop is semantically described by means of ontologies.

Thus, a user is able to connect with minimum effort the content of paper documents to concepts or topics he is dealing with because the SCETagTool proposes those concepts for text passages or words from the paper document. The user is then able to interact directly on the document image and accept those proposals.

The resulting document text with its connections to concepts is inserted as a Wiki page in the user's information system representing his personal knowledge space – the gnowsis Semantic Desktop [5]. Thus, the formerly passive paper document is now electronically available and embedded in the user's personal knowledge space, ready for later retrieval when searching, e.g., for documents dealing with specific concepts.

The presented system is realized as a service extending the gnowsis Semantic Desktop for introducing paper-based information objects into the personal knowledge space.

2. Flow of Work

The flow of work is as follows (see Fig. 1): The user scans a document, e.g., a newspaper article, with a document camera, the document image is automatically handed over to the SCETagTool[1], and OCR is applied to the document image.

The document image is shown in the SCETagTool. Now the user is able to select a title for the document. For each text passage selected by the user:

- the proposed concepts are shown directly on the document image by highlighting the words and listing the concepts beneath the mouse cursor, so the user can easily accept them if appropriate[2],
- get and accept tag proposals for the text passage,
- add existing or create new concepts the text passage.

If finished, the system creates a Wiki page[3] with the text of the scanned document and adds it to the gnowsis, including

- embedded hyperlinks from the words to the concepts accepted by the user,
- concepts as tags for text passages (also with hyperlinks to the concepts),
- a hyperlink to the document image (which was saved in the user's file system).

The concepts proposed are taken from the user's PIMO (*Personal Information Model Ontology*) which consist of classes such as Person, Location, Document(-type), or Topics, and their respective instances – e.g., "Andreas Dengel"

[1] done by observing a file folder

[2] Functionality provided by Single Click Entry, see below.

[3] More precisely: it creates an instance of a the class Document and adds the text as content which is then displayed as a Wiki page.

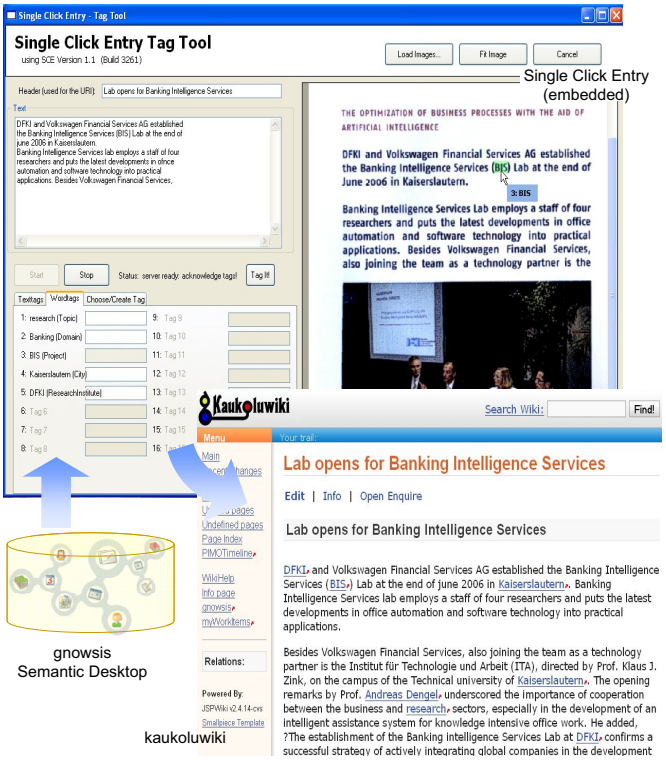

Figure 1. System overview of the SCETagTool

`isa Person` – found on the user's desktop. For instance, all users from the email address book (such as Outlook or Thunderbird) are instantiated as Persons. With this approach a user's personal knowledge space is represented in his PIMO which in turn serves as the base ontology for the gnowsis Semantic Desktop.

The resulting Wiki page is thus connected to the user's concepts such as persons, locations, companies, and topics. If the user browses, e.g., a person in the gnowsis, all connected Wiki pages are listed, i.e., also the page from the previously scanned paper document.

As the gnowsis is a Semantic Desktop, it is then possible to exploit those documents also by means of semantic search, e.g., provide all documents where persons from the company DFKI are mentioned.

3. System overview

The SCETagTool is a Java application consisting of several components which are commercial as well as open source:

The document image is delivered by the portable, digital document camera *sceye*[4] from the company *silvercreations*. With its pivot arm it can be quickly placed on a desktop and the document camera is ready to scan within seconds, thus it supports also mobile workers.

The OCR and the document image interaction is provided by the tool *Single Click Entry*[5] from *ODT* (*Océ Document Technologies*) which is a system for guided data capture from document images. Its function is as follows: Assume an insurance form which requires the input from an insurance claim letter. The required input of the specific fields are described by data format and regular expression (e.g., "US phone number"). Now, Single Click Entry guides the user through the displayed image and highlights the data found (e.g., a telephone number) for the current field (e.g., "phone number of policy holder") as well as lists the data in an information box at the mouse cursor. In case the data is correct the user accepts the proposal by simply clicking on it. In that case the data is transferred to the input data field and the data for the next field is highlighted. Single Click Entry provides several SDK's (e.g., VBS, C#) for including this into own applications what has been done for the SCETagTool.

The gnowsis Semantic Desktop is an open source[6] system which provides a semantic layer to information objects on the user's computer desktop. Embedded in the gnowsis is the kaukoluwiki – a Semantic Wiki which enables semantically enriched Wiki pages [3] and is based on the PIMO.

4. Outlook

This first prototype serves as a basis for further investigation of the easy document capturing interaction method provided by Single Click Entry and how to embed this more tightly into the gnowsis as a means for evolving the user's personal knowledge space with paper documents in order to evolve the approach presented in [4].

Furthermore, more work has to be spent on proposing adequate concepts for document text. Here, we will use the personalized, multi-perspective document classification approach as explained in [1] where the documents belonging to a concept are used to learn a document similarity vector for that specific concept. This is used in order to propose the concept for suitable text passages, i.e., if the passage is similar to the learned vector. Furthermore, more sophisticated proposals will be applied with a collection of specialized services which analyse the text and propose concepts from the PIMO [2].

Acknowledgements The work has been funded by the Rhineland-Palatinate cluster of excellence "Depend-

[4] `http://www.sceye.biz`
[5] `http://www.odt-oce.com/SCE/`
[6] Download at `http://www.gnowsis.org`

able adaptive systems and mathematical modeling" (DASMOD), project ADIB (Adaptive Information Delivery). We thank Andreas Pfau for implementing the system.

References

[1] A. Dengel. Six thousand words about multi-perspective personal document management. In *Proc. EDM, IEEE Int. Workshop on the Electronic Document Management in an Enterprise Computing Environment, Hong Kong, China*. IEEE Computer Society, 2006.

[2] B. Horak. ConTag - A Tagging System linking the Semantic Desktop with Web 2.0. Diploma thesis, University Kaiserslautern, August 2006.

[3] M. Kiesel. Kaukolu: Hub of the semantic corporate intranet. In *SemWiki Workshop, ESWC 2006*, pages 31–42, 2006.

[4] H. Maus, H. Holz, A. Bernardi, and O. Rostanin. Leveraging Passive Paper Piles to Active Objects in Personal Knowledge Spaces. In *Professional Knowledge Management. Third Biennial Conference, WM 2005, Kaiserslautern, Germany, April 2005. Revised Selected Papers*, volume 3782 of *LNAI*, pages 50–59. Springer, 2005.

[5] L. Sauermann, A. Bernardi, and A. Dengel. Overview and Outlook on the Semantic Desktop. In S. Decker, J. Park, D. Quan, and L. Sauermann, editors, *Proc. of the First Semantic Desktop Workshop at the ISWC Conference 2005*, 2005.