Towards Ontology-based Information Extraction and Annotation of Paper Documents for Personalized Knowledge Acquisition

Benjamin Adrian[1], Heiko Maus[1], Malte Kiesel[1], and Andreas Dengel[1,2]
[1]KM Department, German Research Center for Artificial Intelligence (DFKI)
[2]CS Department, University of Kaiserslautern
{FirstName.LastName}@dfki.de

Abstract: Despite the advent of electronic personal information management (PIM) tools, knowledge workers are still heavily using paper-based information sources. But up to now, even in sophisticated tools for PIM such as the Semantic Desktop, the knowledge workers' paper world is still neglected. Thus, electronic archiving of a web page for later reference is much easier than taking care of an interesting article in a magazine—whose copy might set dust on the user's shelf and will long be forgotten when it would be helpful for a specific task. This paper presents how to use document analysis, ontology-based information extraction, and annotation techniques for personal knowledge acquisition in order to bridge the gap between the user's paper world and his personal knowledge space in the Semantic Desktop. A recent prototype shows the feasibility of the approach.

1 Introduction

Despite the advent of electronic, personal information management (PIM) tools, knowledge workers are still heavily using paper-based information sources such as journals, print-outs, reports, and flyers. On the one hand, some information is available only on paper in the form of, e.g., news articles or presentation hand-outs—on the other hand, usability features of paper as long approved medium keep knowledge workers in the paper world [SH03].

Much to our surprise we found that even sophisticated PIM tools such as the Semantic Desktop [SBD05] neglect the knowledge worker's paper world. These tools promise support for collecting, archiving, and finding digital information sources. Thus connecting both worlds means to heavily reduce the user's current PIM effort in collecting, archiving, and finding paper material.

In previous work [MD07, MHBR05], we already connected paper documents with the user's personal knowledge space (PKS) by scanning, OCRing, and finally processing it with a statistical desktop indexing and recommendation-based classification tool [Den06]. Thus, paper documents are transformed to connected, active objects in the knowledge worker's PKS. Here, documents can either be searched, or recommended by proactive information delivery services [HMBR05]. One major drawback was that these approaches

assigned whole documents as a ”bag of words” to user defined categories—finding passages or facts of interest had to be done manually. Text was analyzed only by means of statistical methods. Besides retrieval efficiency, statistical techniques do not support the acquisition, inference, or association of knowledge.

In this paper, we present the choreography of document analysis, ontology-based information extraction, and semantic annotation techniques that bridges the gap between the knowledge worker's paper world and his PKS in the Semantic Desktop. The knowledge worker is enabled to acquire knowledge by approving recommendations about weighted, semantic annotations in units of sentences and tokens by still interacting with the document image. We introduce the overall approach of the envisioned system as use case in Section 3 and describe how a knowledge worker would interact with it. In Section 4 we explain used components that were composed to reach the feasibility of the envisioned system. Finally, we conclude this approach and provide an outlook on future activities.

2 Related work

Digitalization of documents has been a recent research topic for many years. We were influenced during development of our prototype by OCR approaches from OCRopus [Bre08]. Especially the hOCR format [Bre07]—allowing OCR result markup for reproducing the document image as HTML page—could be an interesting extension of the semantic wiki page for layout preservation of document images. We also considered the influence of OCR errors on information extraction described in [TBC06]. In previous works we showed how to improve performance in document analysis and understanding by using semantic context models [WM00]. One of the first ideas for using domain ontologies in information extraction have been described by [ECSL98]. Information extraction as such has been implemented by regarding [AI99]. A general approach of ontology-based information extraction is provided by GATE [BTMC04] which is a subcomponent of our prototype which is described technically in [AD08]. In [ALM05], the authors presented a way to extract information from text for ontology population, that we adopted. We base the knowledge worker's PKS on the Semantic Desktop [SBD05] and his applications in PIM [SGRB08]. In [Den06], the author showed how to perform personal document management in a Semantic Desktop. As a Semantic Desktop application we used a semantic wiki [Kie06] for annotations inside the Semantic Desktop as described in [KSvEB08]. Regarding semantic annotations, Annotea is one of the first approaches of collaborative semantic web document annotations [KK01]. Cream provided annotation recommendations [HS03] alongside an existing ontology. KIM uses information extraction as core technology for recommendations [PKM+03]. Personal annotation recommendations based on a Semantic Desktop was performed in [ASRB07]. We considered requirements of semantic annotations claimed by [UCI+06] and extended our previous work about semantic annotations of paper-based documents [MD07]. The contribution of our work is a prototype that provides text, passage, and word based annotations instead of annotations on document level only. It visualizes and allows interactions on the document image for assuring an invariant user perception of the document content.

3 Approach

This section details foundations on ontology-based information extraction (OBIE) for personalized knowledge acquisition within a Semantic Desktop. The aspired goal is to bridge the media gap between paper-based information and the digital PKS. That means, text on paper is transformed into semantically annotated and hyperlinked hypertext inside the Semantic Desktop.

This *semantification* rests on OBIE techniques which retrieve symbolic occurrences in text of concepts (e.g., persons, companies, topics), relations (e.g., *attended-conference*), and facts (e.g., `Dengel` *attended-conference* `WM 09`). These are ontologically grounded by the *Personal Information Model Ontology* (PIMO) [SvED07] that is a major component of the Semantic Desktop. PIMO consists of an upper ontology describing common PIM concepts (Person, Location, Project, Organisation, etc.). Users are free to extend and subsume these concepts with personal hierarchies and concept maps and instantiate these with personal individuals.

The goal is to instantiate knowledge from documents in terms of PIMO to be reused in everyday work. Therefore, documents are semiautomatically annotated with PIMO knowledge. Finally, these documents are open for further processing such as browsing, hyperlinking, semantic querying, and manually annotating. We claim to reduce effort of transferring paper documents to PIMO because it is as fast and simple as an abstract drag & drop operation. Thus, semantification of paper content has to preserve the knowledge worker's initial cognition of identified information on a document image. Therefore, we preserve the document image as such inside the PKS and link it with its semantic representation. The intuition of manual information extraction from paper documents (e.g., annotating or underlining) should also be represented inside the user interface.

The consequence of using the Semantic Desktop with its applications is an immediate response as soon as the knowledge worker adds annotations about text. Inferred, related, and existing knowledge in form of retrieved and extracted concepts and relations can be recommended automatically. The semantification process results in an integration of original paper information inside the knowledge worker's PKS. It enables knowledge workers to perform intuitive knowledge acquisition of paper-based information by working directly on the accustomed document image. Layout preserving document analysis bridges the media gap as close as possible. As side effect, the ongoing knowledge workers' perception during the semantification step facilitates management and reduction of OCR errors. As a final result, our approach provides quick navigation to PIMO knowledge in terms of concepts and facts occurring inside the document content. As being part of PIMO, extracted and annotated knowledge of all documents can be queried and visualized dynamically.

In order to increase transparency of benefits given to knowledge workers, we describe user interactions with the system:

Introducing a paper document
Assuming a knowledge worker finds an interesting newspaper article and he likes to keep this article for further reference. Instead of copying and—finally—forgetting it in a pile on his desk, now he scans the article. Currently, various input devices are in use ranging from

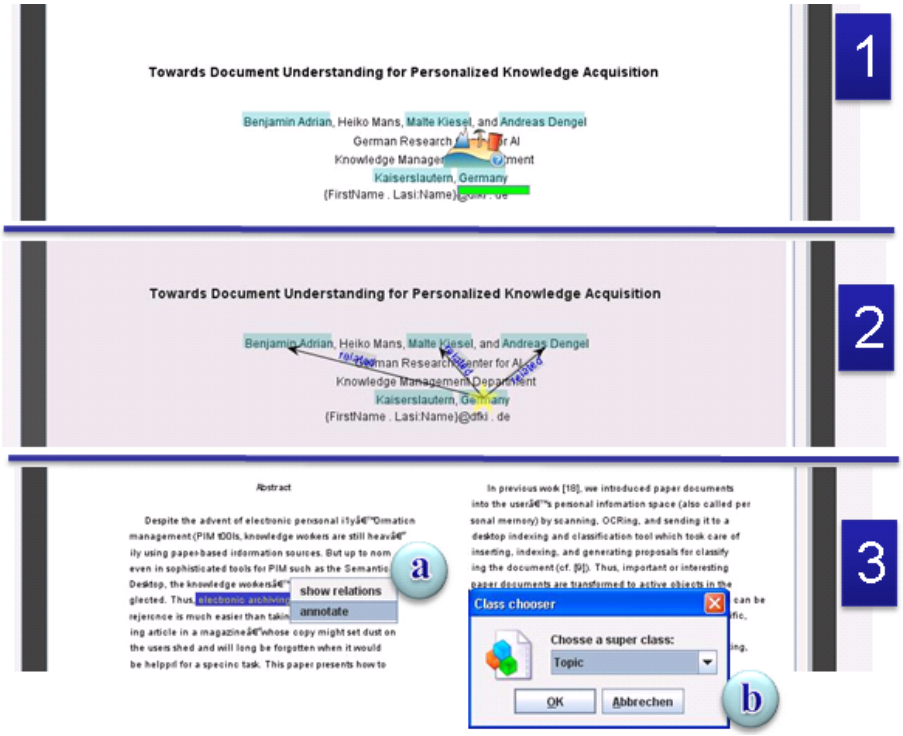

Figure 1: Interacting with recommendation from iDocument

multifunctional devices (cf. [MHBR05]) over to mobile digital document cameras (cf. [MD07]) up to consumer digital cameras as used in DFKI's iDesk [KKK+07] for desktop observation. From the input device, the document is send to the Semantic Desktop and works as input for the semiautomatic information extraction process. First, the image gets pre-processed by applying skew detection and OCR. The result of this pre-processing is an optimized document image and the extracted text. Then ontology-based information extraction is applied to the text for understanding the content in context of the knowledge worker's actual PIMO ranging from instance- to fact resolutions.

As a result, the application comes up with its initial screen, showing the document image with the extracted text as overlay on the image in the center. Extracted information is highlighted as image either *(as shown in Fig.1)*:

1. A resolved instance is annotated as PIMO class `Location` (here, illustrated with a landscape icon, the question mark denotes that is is not yet accepted),
2. Known relations between resolved instances are rendered as directed graph (starting from Germany with edges to the authors),
3. The knowledge worker annotates a token sequence manually as instance and classifies it (by choosing a PIMO class for the selected text).

These results comprise instances found in the document and relations the specific instance is associated with. The instance is highlighted in the document image—users see the origin for the proposal directly in the document image (see Fig.1). At this point, knowledge workers are enabled to perform the following actions for knowledge acquisition purpose:

Accept resolved instances Existing PIMO instances that are successfully resolved from text are highlighted in terms of color with an increasing intensity regarding the amount of certainty. The knowledge worker is enabled accepting these by simply clicking on them.

Accept resolved classifications Yet unknown instances that have been resolved from text are recommended to be classified with existing PIMO classes. Again, the instance is highlighted and additionally tagged with the recommended class. One click accepts the proposal (the question mark in the icon changes to a green check mark).

Annotate and create missing instances In cases of missing recommendations, the knowledge worker is able to select a certain text area on the document image (e.g., two words resembling a name of a person) and to choose an instance or class (e.g., `Person`) out of existing PIMO instances or classes.

Accept resolved relations between instances Relations not existing inside the PIMO, but extracted from text are listed depending on the currently focused instance. The knowledge worker can accept the recommendations by one click.

Annotate missing or new relations between instances The knowledge worker is able to create new relations between extracted instances by selecting two instances as subject and object choosing one valid and existing relation as predicate.

Finally, the knowledge worker completes annotating the document image by pressing a commit button. As a result, the application creates an instance of the PIMO class `Document`, renders a semantic wiki article for this instance that is filled with the document content, and finally stores it in the Semantic Desktop's PIMO. The resulting semantic wiki article is now comprised of:

Extracted text with paragraph preservation The text and the original paragraphs of the document are preserved in the wiki article. A future goal is to preserve the full layout of the original document image and to include pictures and figures.

Semantic text annotations The wiki article displays semantic annotations in text as embedded hyperlinks around symbolic words (e.g, person names) that link to concepts that were accepted by the knowledge worker. As well as annotations about words, annotations about whole text passages (also with hyperlinks to the concepts) may exist.

Link to the document image The document image is stored on the file system as a file and attached to the wiki article. The storage directory is selected by the user, however, potential locations are proposed from the system based on the file directories that have been connected to PIMO instances as containers.

Concluding this visualization and interaction setting, the knowledge worker is enabled to view the document inside his Semantic Desktop as annotated semantic wiki article with a link to the original document image. Moreover, now it is possible to perform semantifications:

Browsing/navigational search All identified concepts and facts in the document can now be browsed by following the links of the annotations leading to the concepts mentioned in the document. Thus, it provides also a means for navigational search because the document is also reachable from those concepts.

Semantic representation As the document is now embedded in the Semantic Desktop and key facts are expressed by semantic statements, the document is more 'understandable' by the Semantic Desktop using its reasoning capabilities. Thus, the document is also searchable by semantic queries—i.e., documents can be found not only by words found in the text but also by inferences based on facts inside documents. For instance a query expressing, *'get all articles where research institutes are mentioned'* would also deliver documents mentioning, e.g., 'DFKI' and not the terms *'research institute'*. That articles mentioning DFKI should be part of the results is inferred based on the fact that DFKI is an instance of the class `ResearchInstitute`.

Add annotations and connect information objects As the document is an instance in the PIMO and the wiki article is freely editable, the knowledge worker is able to add more annotations to the text, or link to other documents or web pages with the same topic.

Add comments to text passages A feature of the used semantic wiki is to allow to annotate text passages without changing the actual content or wiki markup—i.e., the text stays the same and comments can be precisely embrace the passages they address.

With all these possibilities, the formerly inactive and passive paper document is now an active part of the knowledge worker's PKS represented in the Semantic Desktop.

4 Implementation

For reaching our goal of embedding paper-based office information in structured domains, we built a system upon five pillars, namely: (i) a scanning device, (ii) a document image processing and OCR service, (iii) a structured knowledge repository for personal and business knowledge, (iv) a document and semantic annotation system, and finally, (v) an ontology-based information extraction system. The implementation is based on the following components:

Document Scanner In business, we identified two possibilities for delivering document images, namely (i) a shared scanning device outside the knowledge worker's office and (ii) a personal desktop scanner inside the office. We provided the first possibility by using Multifunctional Devices or Peripherals (MFPs) (cf. [MHBR05]). The latter solution is delivered by the portable, digital document camera *sceye*[1] 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.

OCRopus and hOCR The OCR is provided by the OCRopus Open Source OCR System [Bre08]. Interaction on a layered OCRed document image was implemented by using hOCR [Bre07] that is a format for representing both intermediate and final OCR results.

[1] `http://www.sceye.biz`

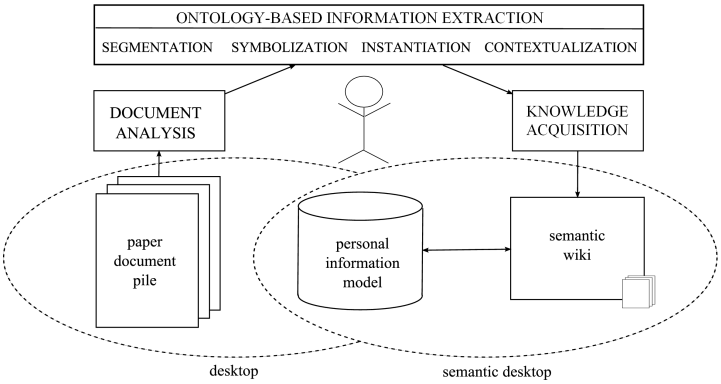

Figure 2: iDocument's extraction process and architecture

The format is defined as a micro format on top of the HTML and CSS standards.

iDocument In order to extract information, we developed the OBIE system called iDocument[2]. It uses existing knowledge inside a PIMO as context for extracting relevant and new information from text. Along an extraction process (see Fig.2), iDocument enriches information bits from documents until reaching the semantic level of PIMO concepts and relations between. Besides the segmentation of sentences and paragraphs, it begins with:

(i) Known phrases (e.g., names, identifiers) or patterns (e.g., email addresses, URLs, phone numbers) are recognized in a succeeding *symbolization step*.

(ii) Resulting symbols are matched against PIMO. Thus, positive matches lead to retrieved or classified concepts in an *instantiation step*.

(iii) Relations between concepts (i.e., facts) in context of the document's content and the current information state of PIMO are resolved in a contextualization step.

(iv) Finally, the knowledge acquisition step provides the insertion of new facts and concepts into the knowledge worker's PIMO.

Document Annotator The Document Annotator is a further development of the SCETagTool [MD07]. The main enhancement is the visualization of document images in different layers that contain the plain text and a set of annotations and recommendations concerning iDocument's results. Similar to SCETagTool, the Document Annotator produces a semantic wiki article from the annotated document.

Nepomuk, Semantic Desktop The Semantic Desktop[3] is a system and framework for expressing semantics in personal information management. The foundational idea is to let knowledge workers weave their personal knowledge space on their desktops by using the PIMO vocabulary and means of the Semantic Web. It enables users to connect bits of information from native information sources (e.g., file system, emails, business repositories, semantic web ontologies, etc.) by semantic relations in style of a semantic network.

[2] `http://idocument.opendfki.de`
[3] `http://nepomuk.semanticdesktop.org`

Kaukolu Semantic Wiki Kaukolu [KSvEB08] is a semantic wiki. While regular wikis have semi-structured text and untyped hyperlinks, semantic wikis provide a semantic model behind text for capturing further information about the article's information entities and their relations. As part of Nepomuk, Kaukolu provides the ability to annotate text with information from PIMO. We use Kaukolu as a back-end storage for document text content and image. It serves as a platform for visualizing and interacting with semantic annotations that have been extracted from text manually (by users) or semi-automatically (e.g., by the information extraction system *iDocument* explained above). Kaukolu allows navigation based on text annotations ("Show me text passages that also mention *this* entity") as well as embedding autogenerated content ("List of all pages mentioning *concept*") and semantic search based on annotations.

The final prototypical implementation of this work is a composition between the upper described components. The current process between these components is:

1. The knowledge worker scans a document with a scanner device. This device creates a document image and sends it automatically to *OCRopus*.
2. *OCRopus* takes a document image and creates image preserving OCR results in *hOCR*.
3. The hOCR-annotated text is automatically send to *iDocument* that performs information extraction with respect to the knowledge worker's current *PIMO*. *iDocument* creates weighted recommendations about occurring *PIMO* knowledge in style of annotations of terms in text passages.
4. The knowledge worker interacts by using the *Document Annotator* GUI in order to acquire and connect *PIMO* knowledge by creating, correcting, or accepting recommended annotations.
5. After accepting the recommendations, the system produces annotated wiki article in *Kaukolu* within the Semantic Desktop. The article is attached with the original document image.

5 Conclusion and Outlook

We presented the use of document analysis, ontology-based information extraction, and annotation for bridging the gap between the knowledge worker's paper world and his PIMO in a Semantic Desktop in order to provide personalized knowledge acquisition. We extended previous work [MHBR05] by adding semantic facilities by means of the ontology-based information extraction system iDocument.

This enables knowledge workers to introduce paper documents into their Semantic Desktop, without loosing layout information or the document image. Paper documents are transformed to PIMO objects and thus can be searched, navigated, hyperlinked, or annotated. It heavily reduced the knowledge worker's effort in collecting, archiving, and remembering paper material. Additionally, existing knowledge inside the Semantic Desktop

is used for personalizing extracted bits of information and embedding documents inside the personal knowledge space. Extracted knowledge from text is formalized and transfered to PIMO.

First evaluation results about iDocument are described in [AD08]. In future work, we plan to increase durability and usability of the system for using it in further evaluation scenarios and use cases. We plan to integrate the functionalities of this prototype in the Nepomuk Semantic Desktop.

Thanks to DFKI Image Understanding and Pattern Recognition department for their support and tools used. Thanks to Jan Frederic Meyer for his implementation effords. This work was supported by "Stiftung Rheinland-Pfalz für Innovation".

References

[AD08] Benjamin Adrian and Andreas Dengel. Believing Finite-State cascades in Knowledge-based Information Extraction. In *KI 2008: Advances in Artificial Intelligence*, volume 5243 of *LNCS*, pages 152–159. Springer, 2008.

[AI99] Douglas E. Appelt and David J. Israel. Introduction to Information Extraction Technology. A tutorial prepared for IJCAI-99, Stockholm, Sweden, 1999.

[ALM05] Florence Amardeilh, Philippe Laublet, and Jean-Luc Minel. Document annotation and ontology population from linguistic extractions. In *K-CAP '05: Proc. of the 3rd Int. conf. on Knowledge capture*, pages 161–168, New York, USA, 2005. ACM.

[ASRB07] Benjamin Adrian, Leo Sauermann, and Thomas Roth-Berghofer. ConTag: A semantic tag recommendation system. In Tassilo Pellegrini and Sebastian Schaffert, editors, *Proc. of I-Semantics' 07*, pages 297–304. JUCS, 2007.

[Bre07] Thomas Breuel. The hOCR Microformat for OCR Workflow and Results. In *ICDAR '07: Proc. of the Ninth Int. Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 1063–1067, Washington, DC, USA, 2007. IEEE Computer Society.

[Bre08] Thomas M. Breuel. The OCRopus Open Source OCR System. In *Proc. IS&T/SPIE 20th Annual Symposium 2008*, 2008.

[BTMC04] Kalina Bontcheva, Valentin Tablan, Diana Maynard, and Hamish Cunningham. Evolving GATE to meet new challenges in language engineering. *Nat. Lang. Eng.*, 10(3-4):349–373, 2004.

[Den06] Andreas 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.

[ECSL98] David W. Embley, Douglas M. Campbell, Randy D. Smith, and Stephen W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *CIKM '98: Proc. of the 7th Int. Conf. on Information and knowledge management*, pages 52–59. ACM, 1998.

[HMBR05] Harald Holz, Heiko Maus, Ansgar Bernardi, and Oleg Rostanin. From Lightweight, Proactive Information Delivery to Business Process-Oriented Knowledge Management. *Journal of Universal Knowledge Management*, 0(2):101–127, 2005.

[HS03] Siegfried Handschuh and Steffen Staab. CREAM - Creating Metadata for the Semantic Web. *Computer Networks*, 42:579–598, 2003. Elsevier.

[Kie06] Malte Kiesel. Kaukolu: Hub of the Semantic Corporate Intranet. In *SemWiki Workshop, ESWC 2006*, pages 31–42, 2006.

[KK01] José Kahan and Marja-Ritta Koivunen. Annotea: an open RDF infrastructure for shared Web annotations. In *WWW '01: Proc. of the 10th Int. Conf. on World Wide Web*, pages 623–632. ACM, 2001.

[KKK+07] Christian Kofler, Daniel Keysers, Andres Koetsier, Jasper Laagland, and Thomas M. Breuel. Gestural Interaction for an Automatic Document Capture System. In Koichi Kise and David S. Doermann, editors, *Proc. of 2nd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR07)*, 2007.

[KSvEB08] Malte Kiesel, Sven Schwarz, Ludger van Elst, and Georg Buscher. Using Attention and Context Information for Annotations in a Semantic Wiki. In *3rd Semantic Wiki Workshop co-located with ESWC 2008*, 6 2008.

[MD07] Heiko Maus and Andreas Dengel. Semantic Annotation of Paper-Based Information. In Koichi Kise and David S. Doermann, editors, *Proc. of 2nd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR07)*, 2007.

[MHBR05] Heiko Maus, Harald Holz, Ansgar Bernardi, and Oleg Rostanin. Leveraging Passive Paper Piles to Active Objects in Personal Knowledge Spaces. In *1st Workshop on Intelligent Office Appliances (IOA 05) at the WM 2005, Kaiserslautern. Germany*, volume 3782 of *LNAI*, pages 50–59. Springer, 2005.

[PKM+03] Borislav Popov, Atanas Kiryakov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. KIM - Semantic Annotation Platform. In *Second Int. Semantic Web Conference (ISWC2003), Proc.*, volume 124 of *LNCS*, pages 834–849. Springer, 2003.

[SBD05] Leo Sauermann, Ansgar Bernardi, and Andreas Dengel. Overview and Outlook on the Semantic Desktop. In *Proc. of the 1st Semantic Desktop Workshop at ISWC*, 2005.

[SGRB08] Leo Sauermann, Gunnar Grimnes, and Thomas Roth-Berghofer. The Semantic Desktop as a foundation for PIM research. In Jaime Teevan and William Jones, editors, *Proc. of the PIM Workshop at the CHI 08*, 2008.

[SH03] Abigail J. Sellen and Richard H. R. Harper. *The myth of the paperless office*. MIT Press, 2003.

[SvED07] Leo Sauermann, Ludger van Elst, and Andreas Dengel. PIMO - A Framework for Representing Personal Information Models. In *Proc. of I-MEDIA '07 and I-SEMANTICS '07*, J.UCS, pages 270–277. Know-Center, Austria, Online-Proceedings, 9 2007.

[TBC06] Kazem Taghva, Russell Beckley, and Jeffrey S. Coombs. The Effects of OCR Error on the Extraction of Private Information. In Horst Bunke and A. Lawrence Spitz, editors, *Document Analysis Systems*, volume 3872 of *LNCS*, pages 348–357. Springer, 2006.

[UCI+06] Victoria Uren, Philipp Cimiano, Jose Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, 2006.

[WM00] Claudia Wenzel and Heiko Maus. An approach to context-driven document analysis and understanding. In *4th IAPR Int. Workshop on Document Analysis Systems, Rio de Janeiro, Brazil*, 2000.