SASA- a Semi-Automatic Semantic Annotator for Personal Knowledge Management

Kim Tighe, Sean O'Riain
Semantic Infrastructure Research Group
Hewlett-Packard, Galway, Ireland
Tel.:+353-91-754901
{ktighe,sean.oriain}@hp.com

ABSTRACT
Best practice organisations have realised that people and their knowledge remain their greatest assets and will continue to be the largest contributory factor in obtaining future competitive advantage. Knowledge fundamentally derived from people in the absence of their understanding, personal context and application remains largely as obscure information. Current personal knowledge management (PKM) activities do not adequately support the finding, reminding, reuse and collaboration of information. In this paper we propose a novel PKM tool called SASA, a semi-automatic semantic annotator of PDF documents, which will enable collecting, connecting and collaborating of discovered information to facilitate knowledge sharing and personal content management within a business setting. SASA, a plug-in for Adobe Acrobat Professional, utilises Semantic Web technologies to enable building, augmenting and sharing of ontologies amongst knowledge workers. Within an ontology named entities are connected to additional information such as Web pages, documents, mail messages, personal notes, and wikis. SASA automatically derives the context of the document, highlights named entities and applies the relevant additional information. The business case for such a tool is outlined and user scenario development used to illustrate how SASA will assist Business Client Account Managers in the laborious process of reviewing, annotating and gathering information from customer documentation by enhancing their PKM.

Categories and Subject Descriptors
I.I.7 [**Computing Methodologies**]: Document and Text Processing – *General.* J.0 [**Computer Applications**]: General.

General Terms
Management, Economics, Experimentation

Keywords
Semantic Annotation, Personal Knowledge Management, PDF

1. INTRODUCTION
In a rapidly changing global economy unified by improved communication and transportation, people and their knowledge are an organisation's greatest assets [1]. The constant emergence of new products and competitors combined with an increasing global marketplace are challenges facing an organisation's ability to survive in an increasingly unpredictable and competitive environment. An enterprise's continued existence will increasingly depend upon their ability to becoming a knowledge-rich knowledge managing organisation [2]. Organisations that focus solely on the application of their collective intellectual capital to achieve objectives run the risk of neglecting the fundamental truth that knowledge is derived from people [3]. Lacking the human element of understanding, personal context and application, knowledge within an organisation remains largely as obscure information. Supporting individuals in their PKM is therefore vital and will be the single largest contributory factor in gaining future competitive advantage over the next 25 years [4].

Enabling technologies for the WWW have provided knowledge workers with rich information sources but have also resulted in adding to the existing considerable volume of information that can be searched and queried. The classical Information Retrieval (IR) problem of identification and retrieval of current information for activities such as informed decision making remains problematic. The core focus of PKM is directed at improving individual efficiency. Current activities however remain limited lacking adequate support for the finding, reminding, reuse and collaboration of information [5]. There remains a growing need for intuitive processes and PKM based tools to assist the worker in evaluating not only their own knowledge but also a means to augment it by exploration and learning from additional information sources. Maximising human capital on a personal level leads to enhancing individual effectiveness in a manner that improves productivity for both the individual and enterprise [6]. It is our contention that PKM enhanced with Semantic Web technologies can be used to assist in achieving this productivity gain.

The Semantic Web [7] envisages annotating document content by assigning to entites in the text links to their semantic descriptions from domain ontologies to make it easier for machines to assist humans in finding, sharing, combining, and reusing information. Current semantic annotation tools (e.g. KIM [8], Trailblazer [16][1] and tools based on Annotea[2] or CREAM [9]) cater for document annotation of Web-native formats such as HTML and XML. None however cater for the Portable Document Format (PDF [10][3]), a format prevalent in virtually all market segments and used extensively for document interchange and publishing.

[1] http://www.hp.com/ie/galway/sirg/trailblazer/

[2] http://www.annotae.org/

[3] A de facto standard on the Web alongside HTML.

With the advent of the Semantic Web this paper examines how Semantic Web enabling technologies, namely semantic annotation can be applied to the area of PKM to enhance knowledge worker productivity and efficiency. This paper proposes a novel tool, SASA for semi-automatic semantic annotation of PDF documents, which will enable collecting, connecting and collaborating amongst knowledge workers to facilitate knowledge sharing and personal content management within a business setting.

The remainder of this paper is structured as follows: Section 2 outlines the business case. Section 3 illustrates the scenario development. Section 4 presents our proposed solution. Section 5 compares related work. Section 6 concludes this paper and outlines future work.

2. BUSINESS CASE

HP[4] Services' Managed Services (MS) provides customers[5] with strategic outsourcing services and solutions to manage their IT infrastructures. The MS business unit itself is structured into a number of areas of expertise known as *towers.* Each tower specializes in a particular area such as the End User Workplace Management (EUWM) which focuses on the end users desktop environment. The EUWM Pre-Sales & Implementation Team is assigned Customer Relationship Management and Project Management activities during the pre-sales and transition/transformation stages of any customer engagement.

In each of the above activities, the EUWM consultant's task of understanding, interpreting and producing all relevant support documentation is crucial for the successful proposal, implementation and delivery of any service. Failing to adequately capture all customer requirements, service limitations and any assumptions made will impact customer satisfaction level, the delivery organisations ability to succeed, HP's profitability, and ultimately, HP's ability to win further contracts. Underpinning all activities is the consultant whom has to ensure that services scoped in the solution are delivered efficiently and implemented in adherence to contractual obligations. For that reason, their resulted outputs from reviewing customer documentation such as Project Definition Document or Project Requirements Document are essential for the project to initially commence and to continue on-going successfully.

New EUWM customer undertaking will require the consultant to begin the laborious process of reviewing, annotating and gathering information from on average 50 or more substantial documents which typically are received in either Microsoft Word or Adobe PDF format. At present, each document is manually reviewed and annotated by the consultant. Central document repository systems such as SharePoint[6] are occasionally used for information sharing in addition to documentation notes capture in an associated mail or Word documents. However, it is not a standard practice and can lead to problems of omitting key comments that are difficult to identify and retrieve particularly for new document versions. Increasing the customer base, scope expansion, EUWM organisational expansion and having to comply with standards such as ISO[7] frameworks has led to a considerable increase in documentation volume and the level of manual effort required.

There is a clear opportunity for an intuitive tool that would assist the consultants in performing documentation review and in information gathering process in order to improve both collaboration and traceability of document findings. Currently under active development SASA is such a tool that offers the semi-automatic semantic annotation of PDF documents. Its usage will contribute towards a reduction in the level of effort required in each new project stage, cost reduction and increased team productivity.

3. SCENARIO DEVELOPMENT

Take for example the situation where a EUWM client account manager has to prepare and deliver a Project Definition Document based on the requirements of the U.S. car manufacturing customer "Customer X", and the capabilities of the delivery organisation and support structure. Material used in drawing this information together is contained in a large number of key business documents such as Statement of Work (SOW), HP Overall Scope document, Technical Solution Document (TSD), RCM Model, Overall HP-Customer Contract and Associated Schedules, Requests For Information (RFI), Requests For Proposals (RFP), etc.

The assigned account manager firstly accesses the transition\transformation and delivery documentation, begins the analysis process in an attempt to identify what is of relevance to the EUWM tower and what contractually HP are obliged to deliver. The Overall HP-Customer Contract is opened with Adobe Acrobat Professional and using our plug-in SASA creates the category[8] "Customer X" for the customer and begins reviewing the documentation. With reference to Figure 1, when the account manager identifies an item of interest such as 'Application Packaging' it is added to the category as a named entity. A note of "Due to ITAR[9] U.S. government regulations all Customer X transmissions must be manufactured within North America" is associated with that named entity. As the account managers' analysis progresses, another document, which is part of the Associated Schedules documentation, is found to contain a key stipulation regarding where UNIX application packaging must be performed. Another note is then added to the named entity 'Application Packaging' along with a bookmark to document Schedule B, which was found to have the associated information. In this manner peripheral information obtained from sources such as emails, phones calls and HP-Customer group discussions can be used to filter information and associate it with suitable named entities. Once the document review stage has concluded the Project Definition Document write up commences. Resulting from the review the accounts manager now has in effect a semantically annotated information source.

[4] Hewlett-Packard Ltd.

[5] Telecom/NSP, financial services, manufacturing and government or public sector markets.

[6] SharePoint is Microsoft collaborative management tool for document and information sharing.

[7] International Standards Organisation

[8] 'Category' is used to represent an ontology.

[9] International Traffic in Arms Regulations

Figure 2 shows that to retrieve the information the manager need only click 'Find Entities' to have all named entities such as 'Application Packaging' belonging to the category highlighted.

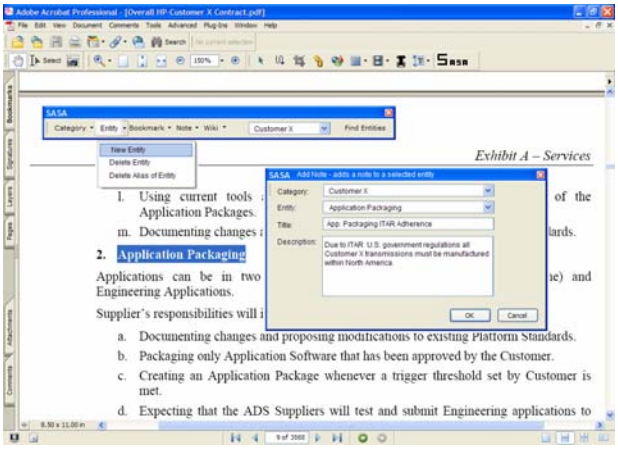

Figure 1. Adding a note to named entity 'Application Packaging'

Selection of the named entity 'Application Packaging' will provide visibility of all additional information and annotations from previous documentation reviews. The client account manager also has the ability to view a summary of all the named entites and their associated information (see Figure 2). This assists the accounts manager in ensuring that issues regarding the like of UNIX application packaging and ITAR regulations are factored in and captured in the Project Definition Document. Otherwise the potential knock on effects of overlooking this information could adversely affect the project timeline, project scope, level of effort required, and delivery model with ultimately negative commercial impact.

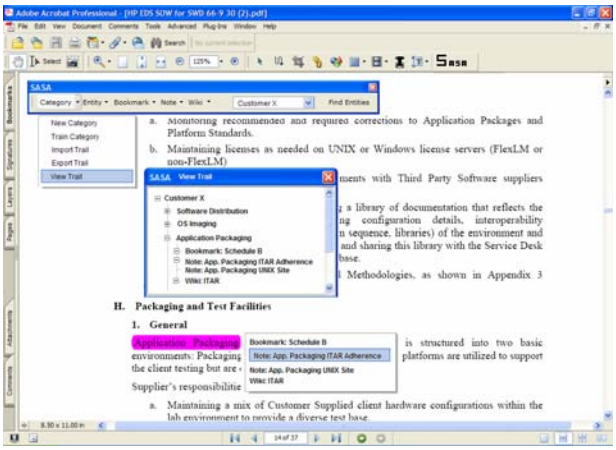

Figure 2. Document annotated showing summary of all named entities and their associated information

4. PROPOSED SOLUTION

SASA is implemented as a plug-in[10] for Adobe Acrobat Professional. SASA adds a toolbar to the standard interface (see Figure 1). The toolbar provides functionality allowing the user to apply or create ontologies and link associated information as semantic annotations to named entities within the PDF document's text content. Dialog windows created using the Adobe Dialog Manager (ADM)[11] are used to facilitate the addition of the named entities and their associated information. SASA stores the ontologies on the users' local machine in Resource Description Format (RDF) [11], which makes them portable amongst groups.

Referring to Figure 3, the SASA application architecture will contain: 1) A **Trainer** component to train SASA using the text and the users' selected ontology about the context of the currently viewed PDF document. SASA will extract the text from the document and use a Vector Space Model (VSM) [12][12] to represent the collected training information by using the words from the document and their frequency of occurrence to augment the existing training data. 2) A **Categorisation** component to derive the context of the currently viewed document using the Cosine Similarity Measure [13] to compare the text of the document with the training data and calculate from the vectors the most likely match to the current ontologies. 3) A **NEIO** component to add and delete named entities and their associated information to and from the ontology. 4) An **Annotator** component to semantically annotate the text of the PDF document by finding and highlighting named entities of interest and applying their relevant additional information. 5) An **Import/Export** component to share ontologies amongst users. 6) A **Viewer** component to view an entire trail of annotations for a selected ontology.

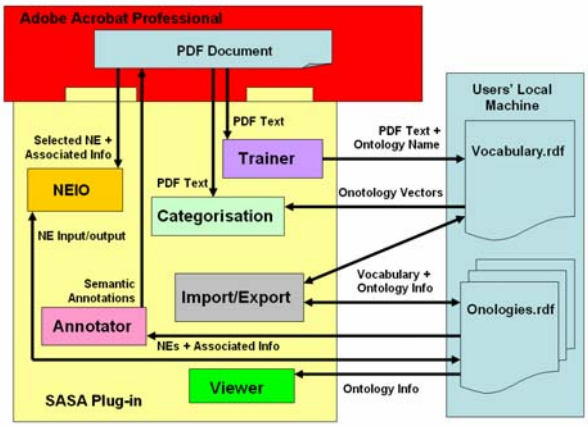

Figure 3. Overview of SASA Architecture

5. RELATED WORK

SemanticWord [14], a Microsoft Word-based environment, adds several toolbars to the interface which support the creation of semantic annotations in documents and templates according to selected ontologies. Magpie [15] is a Web browser extension which uses Named Entity Recognition (NER) based on a supplied

[10] A dynamically-linked extension to Acrobat, which hooks into the user interface and adds functionality to Acrobat Professional, Acrobat Standard, or Adobe Reader.

[11] A cross-platform API for implementing dialog interfaces for Adobe applications such as Acrobat, etc.

[12] An algebraic model used for information filtering and information retrieval.

ontology of the user's choice to highlight and add links to named entities on a Web page. Table 1 shows an extract from a recent survey of semantic annotation tools. It was found that they cater primarily for Web native formats such as HTML and XML. SASA caters for PDF and can be integrated with HPs Mozilla Firefox extension Trailblazer [16] to allow for HTML also.

Table 1. Extract from comparison of annotation tools for requirements 1-7 [17].

Annotation Tool	Semantic Word	Magpie	Amaya
Standard formats	DAML+OIL	HTML OCML	RDF(S) XLink, XPointer
User-centered design	Microsoft Word GUIs	Web browser plug-in	Web browser & editor
Ontology support	-	-	Annotation server
Document formats	Word	HTML	HTML, XHTML and XML
Document evolution	Mark-up tied to text regions	-	XPointer
Annotation storage	-	None, real time	Local or annotation server
Automation	Yes	Yes	No

6. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a plug-in for Adobe Acrobat Professional called SASA, a novel PKM tool for semi-automatic semantic annotation of PDF documents utilising Semantic Web enabling technologies. SASA allows the user to build, augment and share ontologies amongst knowledge workers. Within an ontology named entities are connected to additional information such as Web pages, documents, mail messages, personal notes, and wikis. SASA automatically derives the context of the document, highlights named entities and applies the relevant additional information. The business case for such a tool is outlined and user scenario development used to illustrate how SASA will enhance PKM. Our future work plans, aside from continued implementation of our SASA plug-in, include detailed definition of the case study. We also plan to carry out a systematic user evaluation – with the help of Client Account Managers at HP Galway. Lastly, we are also working on semantically annotating a number of PDF documents at the one time and researching sub section document training.

7. ACKNOWLEDGMENTS

We would like to thank Robert Connolly and Richard Joyce from the EUWM Pre-Sales & Implementation Team, Dara Keogh, and Colman O'Dywer from the Solutions Management Services Team at HP Galway for their time and expertise in framing the business case.

8. REFERENCES

[1] Nonaka, I., (1991). *The Knowledge Creating Company*. Harvard Business Review.

[2] Davenport, T. H., Prusak L. "*Working Knowledge, How Organisations Manage What They Know*" Harvard Business School Press 1997.

[3] Nonaka, I., Takeuchi, M. (1995). *The Knowledge Creating Company – How Japanese Companies Create the Dynamics of Innovation.* Oxford, The Oxford University Press.

[4] Drucker, P. F. "*Managing Knowledge Means Managing Oneself*" Leader to Leader. 16(Spring 2000):8-10.

[5] Volkel, M., Oren, E. *Personal Knowledge Management with Semantic Wikis* Technical Report, AIFB Karlsruhe. December 2005.

[6] Ernst & Young Center for Business Innovation. (1995). *The Financial and Non-Financial Returns to Innovative Work Practices.* New York: Ernst & Young. March.

[7] Berners-Lee, T, Hendler, J & Lassils, O. *The Semantic Web*, Scientific American, May 2001.

[8] Popov, B., Kirayakov, A., Ognyanoff, D., Manov, D., Kirilov, A. *KIM-a semantic platform for information extraction and retrieval*, Nat. Lang. Eng. 10 (3/4) (2004) 375-392

[9] Handschuh, S., Staab, S., Studer, R. *Leveraging metadata creation for the Semantic web with CREAM*, KI '2003-advances in artificial intelligence, in: Proceedings of the Annual German Conference on AI, September 2003, 2003.

[10] *Portable Document Reference Manual, Fifth Edition*, Adobe Systems Incorporated. http://partners.adobe.com/public/developer/pdf/index_reference.html

[11] Brickley, D., Guha. R.V. 2004. *RDF Vocabulary Description Language 1.0: RDF Schema.* W3C Recommendation 10 February 2004.

[12] Salton G., McGill, M. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[13] Van Rijsbergen. C.J. *Information Retrieval*, 1979.

[14] Tallis, M. *SemanticWord processing for content authors*, in: Proceedings of the Knowledge Markup and Semantic Annotation Workshop (SEMANNOT 2003) at 2nd International Conference on Knowledge Capture (K-CAP 2003), October 26, 2003. Sanibel, Florida, USA, 2003.

[15] Domingue, J., Dzbor, M., Motta. E. *Collaborative Semantic Web Browsing with Magpie*. In Proc. of the 1st European Semantic Web Symposium (ESWS), May 2004.

[16] Tighe, K., Johnston, A. *Using Named Entities as a basis for sharing associative trails between Semantic Desktops.* 1st International Semantic Desktop Workshop (ISWC) November 2005.

[17] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna F. *Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art* Journal of Web Semantics: Science, Services and Agents on the World Wide Web (4): 14-28. 2006.