User-Defined Semantic Enrichment of Full-Text Documents: Experiences and Lessons Learned

Annika Hinze[1], Ralf Heese[2], Alexa Schlegel[2], and Markus Luczak-Rösch[2]

[1] University of Waikato, Department of Computer Science
hinze@cs.waikato.ac.nz
[2] Freie Universität Berlin, Department of Computer Science
{heese,alexa.schlegel,markus.luczak-roesch}@inf.fu-berlin.de

Abstract. Semantic annotation of digital documents is typically done at meta-data level. However, for fine-grained access semantic enrichment of text elements or passages is needed. Automatic annotation is not of sufficient quality to enable focused search and retrieval: either too many or too few terms are semantically annotated. User-defined semantic enrichment allows for a more targeted approach. We developed a tool for semantic annotation of digital documents and conducted a number of studies to evaluate its acceptance by and usability for non-expert users. This paper discusses the lessons learned about both the semantic enrichment process and our methodology of exposing non-experts to semantic enrichment.

1 Introduction

Semantic technologies are of increasing importance in digital library (DL) research and practice [10]. Semantic enhancements have been used both at data level and at service level. At data level, FRBR provides an ontological scheme for bibliographic records [4] that increases the expressiveness of retrieval in library catalogues by incorporating information about user tasks. At service level, semantic enrichment has been used to give access to heterogeneous digital libraries and to support collaboration between location-based services and digital libraries. Supporting semantically-enriched services requires DL systems to handle different semantic models [8]. Semantic models and annotations at data and service level are typically defined by domain experts (*i.e.*, librarians or DL designers). Both types of models refer to the *conceptual aspect* of the data, *i.e.*, they annotate the meta-data of documents or document classes. So far, very little support is given for annotating the full-text body of documents.

Even though DL systems support full-text search, semantic enrichment is typically restricted to bibliographic data. Few approaches aim to enrich the full-text of DL documents; to the best of our knowledge, those approaches all use automatic text annotation methods [1,16,17]. Automatic annotations can deal with the large text corpora of Digital Libraries. However, selective annotations for domain-specific context and disambiguation of homonyms are challenging and require complex sentence analysis. Automatic tools provide excellent recall but poor precision. Furthermore, even though automatic tools are well developed for English language texts [16,17,5], other languages are poorly supported.

Our research therefore focuses on an alternative approach: we aim to support readers in manually enriching full-texts. We developed *loomp* – a tool to create user-defined

P. Zaphiris et al. (Eds.): TPDL 2012, LNCS 7489, pp. 209–214, 2012.

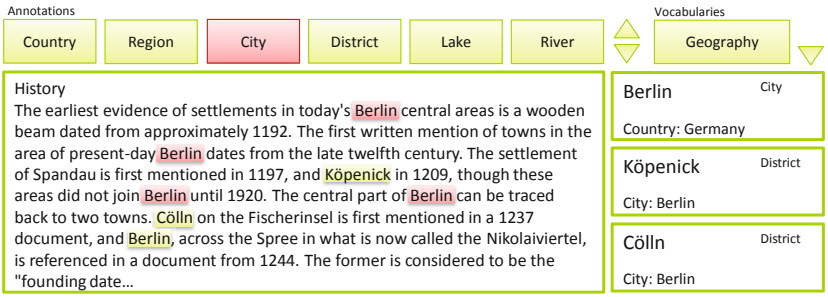

Fig. 1. *loomp* interface (stylized screen-shot for clarity)

semantic annotation of full-texts [12]. Although other tools exist for semantically annotating texts manually [2,15,9], those are typically extensions of wiki environments and require considerable technical knowledge. Moreover, the processes of manually enriching texts have not been evaluated to date. In this paper we report on our experiences and the lessons learned from observing how readers (*i.e.*, non-experts) create those semantic annotations. Annotation tools for non-experts are essential for creating a large body of high-quality annotations (*e.g.*, via crowd-sourcing) as required for the Semantic Web.

The paper is structured as follows: Sect. 2 briefly introduces *loomp*. Sects. 3 and 4 describe the setup and execution of our two user studies for annotating full-texts. Sect. 5 presents our lessons learned, and Sect. 6 the implications for digital libraries.

2 *loomp* Annotation Tool

loomp is an authoring platform for creating, managing, and accessing semantically enriched content.[1] Similar to content management systems that allow non-experts (*i.e.*, people unfamiliar with HTML) to create websites, *loomp* supports non-experts (*i.e.*, people unfamiliar with semantic technologies) in creating semantic annotations. It can be used as a stand-alone tool or as a manual correction of annotations created by automatic tools which recognize named entities in analysed texts and add semantic annotations. Automatic annotation alone is not sufficient for scenarios requiring concise annotations of high quality (*e.g.*, where precision is more important than recall).

To support non-experts, *loomp* was designed to resemble current word processors. In a process similar to assigning formatting (*e.g.*, heading 1) to text passages, *loomp* users can select vocabularies and assign annotations. Figure 1 shows the *loomp* UI with its key elements of text pane, annotation toolbar, and annotation sidebar. The text pane contains (part of) the full-text with highlighted annotations. *loomp*'s interface offers references to concepts in different ontologies (shown as annotations and vocabularies in the annotation toolbar), highlights annotated text passages and shows existing annotations (*e.g.*, 'Berlin' list in sidebar in Fig. 1). We explored alternatives to highlighting annotated text passages in a simplified user interface (discussed in Sect. 3) and observed the readers' understanding of the annotation process using the full *loomp* interface (Sect. 4).

[1] interactive *loomp* software online at `demo.loomp.org`

3 Annotation Process

As the tool was developed for use by non-exerts (*i.e.*, readers of a DL), effective feedback about the process of creating semantic annotations is particularly important. Users need to be able to easily recognise which terms and phrases have already been annotated. Identification of different annotations (*i.e.*, ontological concepts) and clear distinction of elements in overlapping annotations are important. Analysing systems for (non-semantic) annotation support,[2] we identified four characteristics for visual feedback: (1) highlighting atoms and annotations, (2) position of annotations, (3) handling of overlapping atoms, and (4) connecting atoms and annotations.

For *loomp*, we explored the two alternatives of bar layout and border layout, implementing all four characteristics. In the bar layout, each atom within the text is indicated by a vertical bar in the left margin (Fig. 2, left). The colour of the bar reflects the annotation concept. The bars are ordered by length and order in the text. Atoms in the text are highlighted by a mouse-over of the corresponding bar. The border layout highlights annotations by enclosing an atom in a coloured frame (Fig. 2, right). Both layouts allow for many-to-many relationships between atoms and annotations. We observed 12 non-expert participants interacting with both interfaces (some starting with the bar and others with the border layout). During a learning phase, participants familiarized themselves with *loomp* using a short practice text. During the application phase, they had to execute a number of annotation tasks on a longer text.

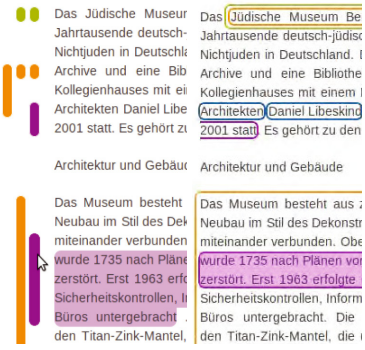

Fig. 2. Bar layout and border layout

4 Annotation Concept

We executed a second user study to evaluate *loomp*'s suitability for non-experts, with particular attention on how these users experience and apply the concept of semantic annotation. Here we focussed on user interaction and understanding (not interface design issues). Even though *loomp* is fully operational, we used a paper prototype to allow for greater flexibility in reacting to user activities and to elicit richer feedback. Its design allowed us to react easily to unexpected user behaviour and to make small changes to the user interface on the fly. It was prepared by printing the framework of the *loomp* UI and outlines of interaction elements; alternatives and pull-down menus were simulated by folding the paper into concertinas. Labels on interaction elements were handwritten so they could be changed dynamically. In the paper version, all UI components of *loomp* are present: the text pane, the annotation toolbar (consisting of annotation concepts and vocabularies), the annotation sidebar, and a resource selector.

[2] Amongst others, `www.veeeb.com`, `atlas.ti`, `diego.com`, `http://itunes.apple.com/us/app/bible+/id332615624`

The resource selector, a separate pop-up window (see Fig. 3, not shown in Fig. 1), supports the selection of semantic identities via resource labels. For example, the atom 'Frankfurt' is annotated with the concept 'city' from vocabulary 'geography' and linked to the resource 'Frankfurt(Oder)' (see Fig. 3), which is internally referencing the resource-id 'http://dbpedia.org/resource/Frankfurt_(Oder)'. The participants used a marker pen to simulate a computer mouse (used for highlighting text in the text pane and selecting UI elements by clicking with closed pen). This simulated mouse was readily accepted by the users; some additionally invented right clicks and alternate keys. The fast changing highlighting of UI elements (indicated by a pressed button and colour change in the *loomp* software) were indicated by pen caps being placed onto the elements. The study was performed by two researchers: the first one interacted with the participants while the second one acted as system simulator. The learning phase continued until each participant felt they understood the system and then the application phase continued until they had created sufficient annotations.. We observed 12 non-expert participants interacting with the *loomp* prototype; none of the participants had taken part in the first study.

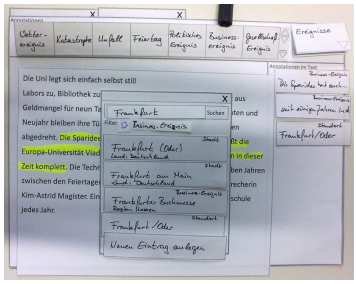

Fig. 3. *loomp* paper prototype

5 Discussion

Both studies observed readers (thus, non-expert users) creating semantic annotations on full-texts. While the study on highlighting text passages (*i.e.*, focussing on the annotation process) required only computer literacy, the second study required a much greater conceptual understanding of semantic enrichment. In this section, we discuss our insights into both the semantic annotation process and our methodology of exposing non-experts to semantic enrichment. We found that for both studies, participants had few problems interacting with the annotation system. They openly embraced the concept of semantic annotations and aimed to create complex, partially overlapping annotations. Participants felt that indicating annotations using the bar layout was better suited to longer annotations whereas the border layout was more appropriate for shorter annotations. A combination of both forms needs to be explored.

As expected, interactions with the more complex user interface of the complete *loomp* system (annotation concept study) were more challenging. We observed that participants had difficulty recognising the implications of some of the more complex features of the user interface. In the simplified interface (study on annotation process), semantic annotation only required highlighting and selection of an annotation, whereas in the full interface (study on annotation concept), annotations required highlighting, annotation selection and assignment of resource identifier.

In the simpler study, all 123 (bar layout) and 116 (border layout) annotations were correctly formed and semantically meaningful. In the more complex second study (using a shorter text), 54 annotations were correct and meaningful, 14 were incorrectly formed but semantically meaningful, and 16 were both incorrectly formed and semantically incorrect. Two participants created several semantically meaningless annotations

(P10 and P11) and two others (P2 and P5) failed completely to create meaningful annotations. From our observations of the participants' interactions with *loomp*, we conclude that not every user group can be educated to be good annotators.

We observed that several participants of the second study had difficulty keeping in mind which passages they had selected and what their intention was (*e.g.*, people often wanted to refer back to the text, the flow was interrupted). Five participants forgot the task they were given and changed their perspective from being an information provider to an information consumer (*i.e.*, wanting to query the system). Five treated the system as a knowledge base (such as wikipedia) and wanted to insert additional information (*e.g.*, create cross-references, extend the vocabulary by synonyms, and insert unit conversions of kmph to mph). Three of the 12 participants tried to create summaries of the text by selecting whole sentences. We see one reason for this observed behaviour in the novelty of the annotation task. The wikipedia model is already well established but creation of semantic annotations (or even indexing) is not a typical task for a reader. This observation holds even for people familiar with the creation of keywords (such as librarians) and tags (such as web 2.0 users). Moreover, semantic search is not widely used and readers therefore do not have examples of well-established use-cases readily available.

Another problem is the concept of semantic identity, which is difficult for non-experts to grasp. Annotation of text has been used in digital libraries [7], for text mining [5] and for shared reading of texts [14,11]. All of these provide closed worlds of annotations (no linkage to vocabulary). Simple semantic markup of text has long been used in libraries (text keywords) as well as in web 2.0 markup of string literals [3]. Full semantic annotation requires the assignment of a semantic identity (*e.g.*, *loomp* uses DBpedia [12], LDP annotates biological texts with references to the Gene Ontology [6]). In *loomp*, this identity assignment is done implicitly by the resource selector (see Fig. 3). Other manual semantic annotation tools [2,9,15] require their users to assign this identity explicitly, thus making them unsuitable for non-expert users. SWickyNotes [13] provides a complex graphical interface that targets advanced non-expert users. The FRBR equivalent of semantic identity is the concept of *work*. The concept of semantic identity was problematic in our studies as some users did not understand the implications of freely assigning new identities to similar atoms.

Readers also do not necessarily feel bound to a particular vocabulary or do not share its understanding. For example, one participant wanted to mark some parts of the text as "political event" because she did not agree with its social implications. Thus the task of semantic annotation may be strongly bound to one's value system.

6 Conclusions

Semantic enrichment of DL full-texts (beyond FRBR markup) provides opportunities for rich and complex retrieval. However, semantic search is currently poorly supported and semantic enrichment almost completely absent. As a consequence, readers have not yet been able to form stable mental models of the markup and retrieval processes.

Using non-expert readers (*e.g.*, by crowd-sourcing) for the enrichment process is challenging. The resulting mark-up may be coloured by personal opinion and offers the opportunity to reflect diverse understandings of a text. However, because semantic enrichments are complex with potentially far-reaching consequences, testing the semantic annotation process requires clearly defined use-cases and better integration into the reader's context. The open definition of shared semantic annotations by non-expert readers may not be viable if a certain quality of annotations is required.

Semi-automatic creation of semantic enrichments is a variation of manual annotation: automatic tools create an initial markup, which is then confirmed, deleted or extended by manual annotations. This process may be best done by a single user – support for collaborative aspects needs further exploration.

In the context of smaller and well defined use-cases (*e.g.*, location markup), tools such as *loomp* are an attractive alternative to excessive markup through automatic tools. We are planning to study the use of *loomp* for creating location markup of full-texts in a mobile digital library setting with location-based access. We are currently developing educational tutorials for non-experts with the goal of raising the quality of user-defined semantic markup.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
2. Auer, S., Dietzold, S., Riechert, T.: OntoWiki – A Tool for Social, Semantic Collaboration. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 736–749. Springer, Heidelberg (2006)
3. Brickley, D., Miller, L.: FOAF vocabulary specification 0.91 (May 2007), `http://xmlns.com/foaf/spec/20070524.html`
4. Buchanan, G.: FRBR: enriching and integrating digital libraries. In: 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 260–269. ACM Press, New York (2006)
5. Cunningham, H.: Gate: general architecture for text engineering, `http://gate.ac.uk`
6. García-Castro, L.J., Giraldo, O.L., Castro, A.G.: Using the annotation ontology in semantic digital libraries. In: Polleres, A., Chen, H. (eds.) ISWC Posters&Demos, CEUR Workshop Proceedings, vol. 658 (2010), `CEUR-WS.org`
7. Gazan, R.: Social annotations in digital library collections. D-Lib. Magazine 14(11/12) (November/December 2008)
8. Hinze, A., Buchanan, G., Bainbridge, D., Witten, I.H.: Semantics in greenstone. In: Kruk, S., McDaniel, B. (eds.) Semantic Digital Libraries, pp. 163–176. Spinger (2009)
9. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 935–942. Springer, Heidelberg (2006)
10. Kruk, S.R., McDaniel, B. (eds.): Semantic Digital Libraries. Springer (2009)
11. Kruk, S.R., Woroniecki, T., Gzella, A., Dabrowski, M.: JeromeDL - a semantic digital library. In: Golbeck, J., Mika, P. (eds.) Semantic Web Challenge, CEUR Workshop Proceedings, vol. 295 (2007), `CEUR-WS.org`
12. Luczak-Rösch, M., Heese, R.: Linked data authoring for non-experts. In: Proceedings of the Linked Data on the Web Workshop (co-located to WWW 2009). LNCS (March 2009)
13. Morbidoni, C.: SWickyNotes Starting Guide. Net7 and Universita Politecnica delle Marche (April 2012), `http://www.swickynotes.org/docs/SWickyNotesStartingGuide.pdf`
14. Pearson, J., Buchanan, G.: CloudBooks: An Infrastructure for Reading on Multiple Devices. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) TPDL 2011. LNCS, vol. 6966, pp. 488–492. Springer, Heidelberg (2011)
15. Schaffert, S.: Ikewiki: A semantic wiki for collaborative knowledge management. In: WS on Semantic Technologies in Collaborative Applications (STICA 2006), Manchester, UK (June 2006)
16. Thomson Reuters Inc. Open Calais website, `http://www.opencalais.com/`
17. Zemanta Ltd. Zemanta website, `http://www.zemanta.com/`