Semantic Enrichment by Non-Experts: Usability of Manual Annotation Tools

Annika Hinze[1], Ralf Heese[2], Markus Luczak-Rösch[2], and Adrian Paschke[2]

[1] University of Waikato hinze@waikato.ac.nz
[2] Freie Universität Berlin {heese,luczak,paschke}@inf.fu-berlin.de

Abstract. Most of the semantic content available has been generated automatically by using annotation services for existing content. 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 an end-user study to evaluate its acceptance by and usability for non-expert users. This paper presents the results of this user study and discusses the lessons learned about both the semantic enrichment process and our methodology of exposing non-experts to semantic enrichment.

1 Introduction

Current internet search engines typically match words syntactically; semantic analysis is not supported. The Semantic Web envisions a network of semantically-enriched content containing links to explicit, formal semantics. So far, most of the semantic content available has been generated automatically by either wrapping existing data silos or by using annotation services for existing content. We believe, however, that the success of the Semantic Web depends on reaching a critical mass of users creating and consuming semantic content. This would require tools that hide the complexity of semantic technologies and match the compelling simplicity of Web 2.0 applications: light-weight, easy-to-use, and easy-to-understand. Very little research has been done on supporting non-expert end-users in the creation of semantically-enriched content.

We studied the process of manual creation of semantic enrichments by non-experts. For this, non-expert users were observed interacting with an example annotation system. We used *loomp*, an authoring system for semantic web content [1]. In its initial design phase, *loomp* received positive feedback from non-expert users (e.g., journalists and publishing houses). Their feedback revealed great interest in adding "metadata" to content but also some difficulty in understanding the underlying principles of semantic annotations. This motivates research to derive guidelines for the design of adequate annotator tools for non-experts and to gain insight into non-experts' understanding of semantic annotations. To explore the experience of non-experts in operating a semantic annotation tool, we therefore conducted a user study of the *loomp* annotator component. This paper reports on the results of this user study and discusses its implications for the semantic web community. We will argue that manual semantic annotations need specialized task experts (instead of domain experts) and we note a lack of clearly defined use cases and accepted user-centred quality measures for semantic applications.

The remainder of this paper is structured as follows: In Section 2 we explore related work in evaluating the usability of semantic web tools. Section 3 introduces *loomp* and the *One Click Annotator*. The methodology and setup of our study is explained in Section 4. Section 5 reports on the results of the study, while Section 6 discusses the implications of the study results. Section 7 summarizes the contributions of our study and draws conclusions for semantic web tool design.

2 Related work

Here we discuss research related to the *loomp OCA* and its evaluation.

Annotation tools. Annotation can be done either automatically or manually (or in combination). Automatic annotation tools are typically evaluated only for precision and recall of the resulting annotations [2,3,4]. Most manual annotation tools have never been evaluated for their usability; many are no longer under active development [5,6,7]. We classify current manual systems into commenting tools [8,9,10], web-annotation tools [11,12], wiki-based systems [13,14,15], and content composition systems [1,16], digital library tools [17,18,19], and linguistic text analysis [20].

Naturally, due to their different application fields, the tools encounter different challenges in interaction design and usability (e.g. wiki tools require users to master an annotation-specific syntax and to cope with many technical terms). However, we believe the most significant challenge for user interaction design is defined by the conceptual level of semantic annotation. That is, the annotation process is conceptually different if tools support simple free-text annotation (e.g.,[8,9,10]), offer a shared vocabulary of concepts (e.g., [6,7,11]), use a local semantic identity (by thesaurus or ontology, e.g., [5,13,15]), or use shared semantic identity (e.g., by linked ontologies referencing with a linked data server such as DBpedia, e.g., [2,3,18]). The development of most annotation tools has a strong focus on providing novel functionality. For the manual annotation tools, usability was typically a factor in the interface development. However, end-user evaluations of interface and user interaction are very rare.

End-user experience of annotations. Few system evaluations have considered end-user experiences. Handschuh carried out a tool evaluation with user involvement; however, participants were used merely to provide annotation data that was then analysed for inter-annotator correlation [7,21]. Furthermore, the researchers expressed disappointment about the low quality of annotations. Feedback on the participants' experience and their mental model of the process were not sought. Bayerl et al. [22] stresses the importance of creating schemas and systems that are manageable for human annotators. They developed a method for systematic schema development and evaluation of manual annotations that involves the repeated annotation of data by a group of coders. Erdmann et al. [23] performed studies on manual and semi-automatic annotation involving users. They describe their participants as "more or less" able to annotate webpages. However, the majority of issues identified were of a syntactic nature that could easily be remedied by tool support. Work on rhetorical-level linguistic analysis of scientific texts is closely related to semantic annotation [20]. Teufel performed user studies in which she looked for stability (same classification over time by same annotator) and reproducibility (same classification by different annotators; similar to Handschuh's inter-annotator correlation). Similar to [22], she found that complex schemas may lead to lower quality annotations, and subsequently simplified the predefined set of common concepts that was used in the evaluation (from 23 to 7 concepts). Teufel assumed that annotators would be task-trained and familiar with the domain. We discuss observations of these studies relating to semantic understanding in Section 6.

Benchmarking. A number of researchers have discussed methodologies for comparing annotation tools using benchmarks [24,25]. Maynard developed a set of evaluation criteria for performance as well as for usability, accessibility and inter-operability [24,26]. However, usability here refers to practical aspects such as ease of installation and online help, and does not contain concepts of interaction design, user acceptance and effectiveness. Schraefel and Karger [27] identify ontology-based annotation as one of the key concepts of SW technologies and defined a set of quality criteria. One of these is usability, which for them covers ease-to-learn, ease-of-use and efficiency. Uren et al. [28] developed a survey of annotation tools in which "user-centered collaborative design" was one of the requirements. However, they mainly explore the ease-of-use of tool integration into existing workflows. They furthermore assumed that annotation would be created by "knowledge workers." Most benchmarks focus on (user-independent) performance measures; usability concepts are seldom included and rarely evaluated. Castro [25] observes that in the semantic web area, technology evaluation is seldom carried out even though a number of evaluation and benchmark frameworks exist.

HCI challenges in the Semantic Web. A series of workshops on Semantic Web HCI identified areas for research contribution, one of which is the capture of semantically-rich metadata without burdening the user [26]. Karger suggests hiding the complexity of the Semantic Web by developing tools that look like existing applications and to develop better interfaces to bring the semantic web forward "before AI is ready" [29]. Jameson addresses a number of concerns of the SW community about user involvement and stresses the value of both positive and negative results of user studies [30].

Ontology engineering. The development of ontologies faces similar challenges to that of the semantic annotation of texts: It is a complex task that often needs (manual) user input [31,32,33]. However, ontology engineering is typically executed by experts in semantic technologies and is not necessarily suitable for end-users. However, Duineveld at al. [31] report that often the bottleneck in building ontologies still lies in the social process rather than in the technology. User-oriented evaluation focuses predominantly on syntactic problems (e.g., how much knowledge of the representation language is required), but not on conceptual questions such as the user's mental models of the system.

Summary. Even though aspects of HCI and user involvement have been identified as important aspects for Semantic Web technologies, typical benchmarks and evaluation strategies do not contain complex user aspects. Few studies involving end-users have been executed in the context of semantic annotations. In particular, manual annotation tools have so far not been systematically evaluated for appropriate interaction design and semantic understanding. System evaluations that incorporated human participants did not seek their feedback on interaction issues nor did they evaluate the participants' mental models of the system interaction. So far, issues of understanding of semantic annotations by (non-expert) users have not been studied in a systematic manner.

3 Semantic Content Enrichment using *loomp*

loomp is a light-weight authoring platform for creating, managing, and accessing semantically enriched content. Similarly to content management systems allowing people unfamiliar with HTML to manage the content of websites, *loomp* allows people unfamiliar with semantic technologies to manage semantically enriched content. *loomp*'s *One Click Annotator* enables users to add semantic annotations to texts. We first discuss user-related design considerations for the *loomp OCA*, and then briefly give an introduction into the concepts of the *OCA*.

3.1 User-oriented Design Considerations

A key goal for *loomp OCA* was to hide the complexity of semantic technologies (cf. [29]) but nevertheless allow for the creation of meaningful and consistent annotations. We identified following key requirements for non-expert users.

Established interaction patterns. Karger argues that there is an advantage of making Semantic Web applications look like existing applications and to use familiar interaction paradigms. [29,27]. A similar argument has been made in the area of personal information management, where new tools are more successful if they are able to extend the functionality of existing applications rather than introducing an entirely new way of doing things [34]. For the *loomp One Click Annotator*, we therefore adopt well-known interaction procedures of widely used software (such as text highlighting and formatting in MS Word$^{\text{TM}}$).

Simple vocabularies. It has been shown that complex thesauri and category structures are disadvantageous for quality annotations [22,20]. For a given task, users may only require a small part of a vocabulary that is modeled in a large ontology with deep hierarchical structures. Thus, *loomp OCA* only offers an appropriate subset of terms and provides support in choosing the right annotation.

Contextual semantic identity. The RDF data model differs in subtle ways from the cognitive models that humans create for the content of a text. In RDF, resources are assigned URIs for unique identification and disambiguation of concepts and instances. Although people may recognize URLs as addresses of websites, they are not used to identifying real-world entities by URL/URIs and are typically not familiar with the concept of namespaces. Instead, humans use labels to refer to entities (e.g., "baker") and disambiguate meaning by the textual context (e.g., as reference to the person Baker or the profession baker). The annotator has to bridge this gap between objective knowledge (as encoded in the RDF data model) and subjective knowledge of human cognition [35,36]. The *loomp OCA* aims to support this process by presenting labels and contextual information for identification of semantic identity.

Focus on the user's task. Handschuh and Staab observed that semantic authoring and semantic annotations have to go hand in hand [37]. As a consequence, we integrated the *loomp OCA* toolbar for creating semantic annotations seamlessly into the *loomp* editor, so that the user can add annotations without being distracted from their primary task.

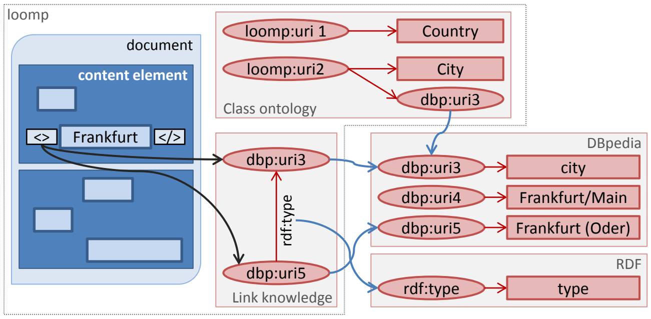


Fig. 1. *loomp OCA*: Conceptual data model

3.2 *loomp OCA* Conceptual Design

loomp's domain model consists of content elements, ontological concepts and annotations that create links between them, as well as encapsulating documents. Each document consists of a sequence of content elements (see Fig. 1), where a content element can belong to several documents. We use DBpedia as the base ontology to identify annotation concepts (classes) and instances.

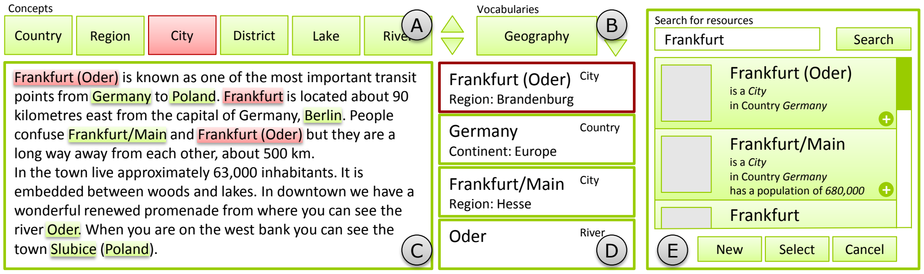
The *loomp OCA* offers a selection of class ontologies (subsets of DBpedia) to present vocabularies to the users (see Fig. 2, B and C).

A user can assign an annotation to a text atom (i.e. a part of a content element) in two steps.

1. The user marks an atom in the text field and then selects a concept from an offered vocabulary (see Fig. 2, B). For example, they mark the atom Frankfurt and select the concept City from the vocabulary Geography. Internally, the system then creates a link between the atom and the concept id, which is inserted into the content element as RDFa element (transparent to the user).
2. The system sends the text of the atom as a query to DBpedia. The labels of the resulting instances are filtered by the concept, and then presented to the user for selection. For example, the system sends Frankfurt as a query to DBpedia, where all instances containing this phrase are identified. The result set is then filtered by the concept term city. The user is presented with the resulting instance list containing references to Frankfurt/Oder and Frankfurt/Main (see Fig. 2, right). They identify the appropriate instance to be linked to the atom. Internally, the system creates a second link for the atom, linking to the instance id (here, linking to the DBpedia id of Frankfurt (Oder), see Fig. 1).

The creation of the links from the atom to both concept id and instance id allows identification of link knowledge, such as the type of the instance resource (Frankfurt (Oder) rdf:type City[3]). As a result, when linking atoms from different documents to the same semantic identifier (e.g., to the DBpedia id of Frankfurt (Oder)), a user creates conceptual cross-links between these documents. As this paper focuses on the user interaction with the *loomp OCA*, we refer for technical details of *loomp OCA* to [1,38].

[3] For example, encoded as (dbp:uri5 rdf:type dbp:uri3), using ids from Fig. 1



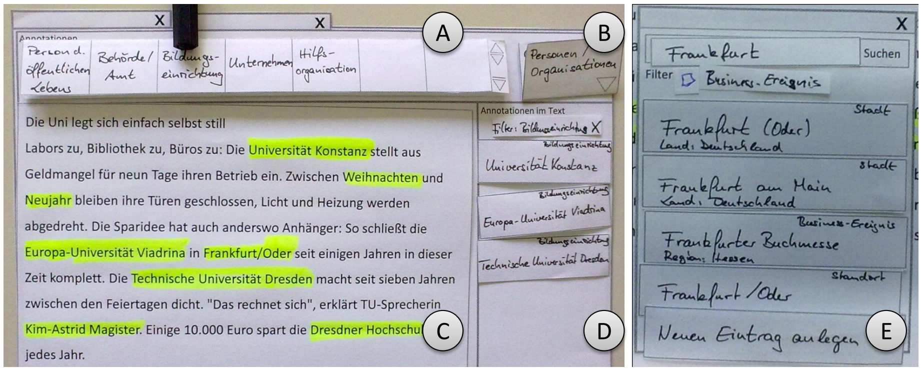
Frankfurt (Oder) is known as one of the most important transit points from Germany to Poland. Frankfurt is located about 90 kilometres east from the capital of Germany, Berlin. People confuse Frankfurt/Main and Frankfurt (Oder) but they are a long way away from each other, about 500 km.
In the town live approximately 63,000 inhabitants. It is embedded between woods and lakes. In downtown we have a wonderful renewed promenade from where you can see the river Oder. When you are on the west bank you can see the town Slubice (Poland).

Fig. 2. Interface of *loomp OCA*

4 User Study Methodology

We studied the usability of the *loomp OCA* in an end-user study. Following Degler [39], who evaluated methods for improving interaction design for the Semantic Web, we performed usability tests and interviews with real users (not student stand-ins). The aim of our study was to (1) evaluate the suitability of the tool for non-experts, and (2) explore how non-expert users experience and apply the concept of annotating texts. Even though the *loomp* system is fully operational, the user study was executed with a paper prototype. This allowed us to gather feedback from non-expert users in a non-threatening technology-free context. A paper prototype consists of mock-up paper-based versions of windows, menus and dialog boxes of a system. One of two researchers plays the role of the 'system', the other one acts as facilitator. Participants are given realistic tasks to perform by interacting directly with the prototype – they "click" by touching the prototype buttons or links and "type" by writing their data in the prototype's edit fields. The facilitator conducts the session; the participants are video-taped and notes are taken. The 'system' does not explain how the interface is supposed to work, but merely simulates what the interface would do. In this manner, one can identify which parts of the interface are self-explanatory and which parts are confusing. Because the prototype is all on paper, it can be modified very easily to fix the problems. Paper prototyping is an established usability method, which has been shown to allow greater flexibility in reacting to user activities and to elicit high quality and creative feedback as users do not feel restricted by an apparently finished software product [40]. The user study was set up as follows:

Paper prototype. Mirroring the *loomp OCA* interface, the paper prototype consisted of two windows (see Fig. 3). All UI components of the functional software described in Section 3 are present: (A) the text pane, (B) the annotation toolbar consisting of two parts, (C) the annotation sidebar, and (D) the resource user (as separate pop-up window). The framework of the user interface and outlines of interaction elements were printed on paper and cut to size; alternatives and pull-down menus were simulated by folding the paper into concertinas. All labels on interaction elements were hand-written to allow dynamic changes. The available texts to annotate were printed onto text pane templates. Designing the paper prototype in this way allowed us to react easily to unexpected user behaviour (e.g., by creating resources for unexpected annotations) and to make small changes to the user interface on the fly.



Die Uni legt sich einfach selbst still

Labors zu, Bibliothek zu, Büros zu: Die Universität Konstanz stellt aus Geldmangel für neun Tage ihren Betrieb ein. Zwischen Weihnachten und Neujahr bleiben ihre Türen geschlossen, Licht und Heizung werden abgedreht. Die Sparidee hat auch anderswo Anhänger: So schließt die Europa-Universität Viadrina in Frankfurt/Oder seit einigen Jahren in dieser Zeit komplett. Die Technische Universität Dresden macht seit sieben Jahren zwischen den Feiertagen dicht. "Das rechnet sich", erklärt TU-Sprecherin Kim-Astrid Magister. Einige 10.000 Euro spart die Dresdner Hochschul jedes Jahr.

Fig. 3. Paper prototype of the *loomp OCA*

The participants received a marker pen to simulate the use of 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 (see top left of Fig. 3).

Texts and Ontology. We prepared two texts for annotation that contained only general knowledge concepts. Thus every participant was a domain expert. The first document was used as a training set; it contained a short text about Konrad Adenauer, the first chancellor of West Germany. This allowed the participants to explore the interface without any pressure to "get it right." The second, longer text was about the weather and universities being closed during the cold period. Both texts were based on news portal entities that were shortened for the study. We adapted the texts so as to explore interesting semantic problems, such as place names with similar words (Frankfurt (Oder) and Frankfurt/Main), nested concepts (Universität Konstanz) and fragmented references (Konrad Hermann Joseph Adenauer and Konrad Adenauer referring to the same person). These adaptations ensured that participants had to select the correct resources to link to an atom. We used the same underlying 'news' ontology for both texts. A subset of classes of this ontology was selected manually to provide a set of concepts tailored to the example texts (while allowing for variations). The classes were grouped into three named vocabularies: Persons & Organizations, Events, and Geography. They contained 12, 8, and 10 annotations respectively. Identical underlying ontology and annotations sets were used for learning and application phase.

Study phases. The study was performed by two researchers: the first interacted with the participants, while the second acted as system simulator (no direct interaction between participants and second researcher). The study was performed in four phases: introduction, learning phase, application phase, and guided interview. During the introduction, the aim of the project and the prototype were explained and the participant was shown the paper prototype. During learning phase and application phase, the participant interacted with the prototype. The researchers took notes and recorded the interactions.

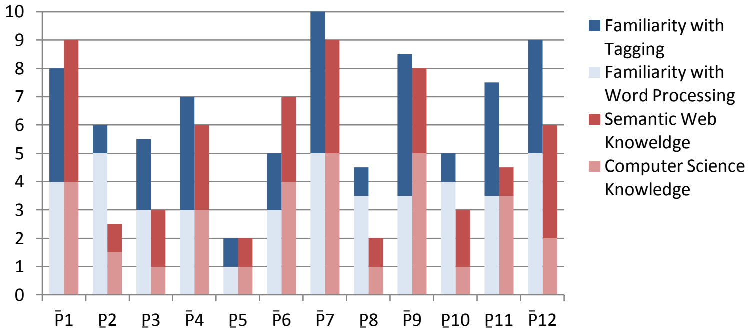


Fig. 4. Participants' self assessment

In the learning phase, the researcher explained the purpose of the application by way of a use case for semantic search and thus illustrated the need for semantic annotations. The participant received instructions on the task of annotating. Participants were given time to familiarize themselves with both the task and the interface. In the application phase, the longer text was used with the same task. The participants were encouraged to think out loud while making decisions in interaction with the prototype, instead of asking for the 'correct' procedure. The study had 12 participants (up to 1.5 hours interaction each).

5 Results

We here describe our observations of how participants interacted with the *One Click Annotator*. We differentiate between general observations about the participants (Sect. 5.1), observations related to the interaction with UI elements (Sect. 5.3) and observations related to the task of creating semantic annotations (Section 5.4). Implications for tool design and the Semantic Web community will be discussed in Section 6.

5.1 Participant demographics

As the tool is designed for non-experts, we selected 12 participants with varied backgrounds (e.g., librarians, PR managers, secretaries). We enquired about their familiarity with word processing (as a measure of computer literacy), tagging (as an annotation task), computer science and Semantic Web (as technical expertise). Participants rated their knowledge on a 5-point scale (1=no knowledge, 5=very knowledgeable). Fig. 4 shows the distribution of expertise for the 12 participants. 11 of 12 participants are computer literate (basic computer literacy is a requirement for *loomp*), six are familiar with tagging and setting annotations and thus have advanced computer literacy skills. Six participants have very little knowledge in computer science and Semantic Web; they are (technical) non-experts – the user group for which *loomp* was designed. Based on their self-assessment, we identified participants P̄1, P̄4, P̄6, P̄7, P̄9 and P̄12 as *technical* *experts* (CS+SW knowledge ≥ 6) and participants P̱2, P̱3, P̱5, P̱8, P̱10 and P̱11 as *non-experts* (CS+SW knowledge < 5). Throughout the paper, we visually indicate expertise thus: P̱x and P̄x. We observe that technical experts are also (highly) computer literate.

5.2 UI Elements: Observed Interaction Patterns

We now describe participant interactions with the key elements of the *loomp OCA* with a focus on the participants' understanding of the annotation process.

i) Annotation toolbar (A+B in Fig.3). Some participants had difficulties interacting with the annotation toolbar. Some participants selected first an annotation without highlighting atoms. P̄1 had forgotten to select an atom first; P̄12 intended to use the annotation as a paint brush to assign it to several atoms. A number of participants had difficulty remembering the available concepts they had just looked at.

ii) Text pane (C). Most participants had no problems selecting atoms to assign annotations. Five of 12 participants tried to create nested annotations, which is currently not supported in *OCA*. Taking the atom Universität Konstanz as an example, they wanted to assign Educational institution to the phrase and City to the term Konstanz. Two participants (P̄4, P̱10) lost the annotation of the city because the later assignment of the larger atom Universität Konstanz overwrote the previous smaller atom Konstanz. Only P̄4 understood and corrected the mistake. P̱10 observed the loss of annotation but did not recognize the problem. In contrast, two participants allocated Educational institution before allocating City with the result that the first annotation covered only the atom Universität. Five of 12 participants wanted to assign more than one annotation to the same atom, e.g., Adenauer is a person and a politician. Participant P̄12 wanted to use the ALT-key to select all text occurrences of an atom and then select the annotation only once to link all occurrences to the same resource (e.g., all occurrences of the same city name). The same participant suggested that the system should also support the following process: Users highlight all text occurrences of entities of the same type (e.g., all cities), click on an annotation button, and choose a resource to each of these one after the other.

iii) Annotation sidebar (D). Participants used the annotation sidebar either as a simple list for compiling the created annotations, or as a tool for highlighting occurrences of created annotations (P̄1, P̄7) and for copying existing annotations to newly selected atoms (P̄9, P̱11). P̄9 also clicked on the concept (e.g., Person) of an annotation shown in the sidebar and, thus, filtered the entries of the sidebar according to that concept. Participants did not recognize the temporal order of annotations in the sidebar (Participant P̄4 suggested they be sorted alphabetically).

iv) Resource selector (E). Several participants had difficulties in understanding the resource selector. The selector recommends resources for a selected atom, which was readily accepted by the participants without reflecting on the source of the recommendations. The recommendations are created based on DBpedia search results for the atom text. As a consequence, generic atom names (e.g., university) lead to long result lists (P̄7 was so overwhelmed by the long result list that he closed the window without selecting an entry). Only five of the participants (P̄1, P̄6, P̄7, P̄9, P̱10) recognized the text field as a means for searching for resources manually and only two of them (P̄7, P̱10) understood the filter option below the search field.

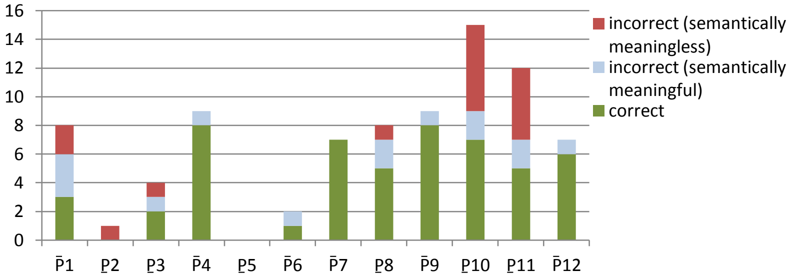


Fig. 5. Correctness of Annotations

5.3 Task: Understanding Semantic Concepts

We were interested in participants' conceptual understanding of the creation of semantic annotations. We evaluated the selected atoms and annotations, as well as the participants' reaction to the resources recommended in the resource selector.

i) Quality of Annotations. Fig. 5 shows an analysis of the annotations created by the participants (second phase only). The gold-standard annotation for the text contained eight annotations (for the given vocabulary). An annotated atom is considered semantically meaningful if it refers to a named entity, e.g., if participants allocated City to the atom Dresden within Technische Universität Dresden. Finally, annotations are considered semantically meaningless if they do not refer to a named entity, e.g., light and heating are turned off. Six of 12 participants identified at least six correct annotations and another one created 5 of 8 correct annotations. However, two participants additionally created many semantically meaningless annotations (P̲10,P̄12). P̲2 and P̲5 failed to create any meaningful annotations.

ii) Assuming system knowledge. Six participants switched from being an information provider to an information consumer in the course of the study (P̲2, P̲3, P̲5, P̲8, P̄9, P̲11). (P̲3 "Now I want to know something about his political career;" P̲2: "Now I cannot find any information"). Three of them clicked concepts without selecting any text because they expected that the system would highlight the respective named entities, e.g., highlight all mentioned cities. Five participants assumed that the system comes with background knowledge (e.g., P̲8 clicked on the term 'chancellor' and said "There should be a list of chancellors of FRG.")

iii) Selecting annotations. In the first annotation step (see Section 3.2), four participants wanted to create a summary of the text by selecting whole sentences as atoms (P̄1, P̲2, P̲3, P̄6). For example, P̄1 selected light and heating are turned off and allocated the annotation Political Event. P̄6 comented that "Somehow the whole text is an event."The selection of unsuitable atoms resulted in difficulties when selecting the correct annotation: P̲10 selected the atom library and allocated Educational Institution. She observed: "I find it difficult to find the correct annotation." She proceeded similarly with laboratory and office. Several participants aimed to annotate such classes of instances (in addition to instances), which almost always led to the unnecessary creation of new resources. We also observed difficulties when one concept is a subclass of another one (e.g., Person and Politician). As the prototype did not support overlapping annotations, almost all participants chose the more specific concept. Only P̄4 explained that he assumed the system would contain information about the relationship between the two concepts. In contrast, three participants (P̄7, P̄9, P̱10) developed a workaround and annotated one occurrence with Person and another one with Politician. Four participants had difficulty deciding on the granularity of atoms (P̄4, P̱8 P̄9, P̱10), e.g., whether to annotate the city in the atom Technische Universität Dresden.

iv) Interaction patterns. We observed different strategies for creating annotations. Two participants (P̱8, P̄12) first seemed to select a vocabulary and then scan the text for occurrences of matching atoms, e.g., to annotate all cities (P̄12: "The cities are done now."). P̄12 suggested having an "annotation painter" (similar to the format painter in office software) that allows for selecting a concept from a vocabulary and then selecting atoms. Another common strategy was to annotate an entity and search for further occurrences in the text. A few participants felt unsure whether they had created enough annotations, e.g., P̄7 commented "I would ask a colleague."

v) Identifying entities. In the second annotation step (see Section 3.2) of linking atom to resource, we observed problems in choosing the appropriate entry in the resource selector. P̄9 wanted to annotate the name of the river Oder in the name of the city Frankfurt/Oder. The resource selector offered two rivers, one of them within the correct region. P̄9 wondered: "Why is that, is that one river or two?" and continued by creating a new resource. However, all five participants annotating Frankfurt/Oder successfully selected Frankfurt (Oder) as resource (i.e., it was clear that both labels referred to the same city).

vi) Additional knowledge work. During the learning phase, five participants wanted to insert additional information, e.g., create cross references or even extend a vocabulary. For example, P̄1 wanted to relate Kim Astrid Magister with Technical University Dresden because she was the spokeswoman of that university. Later, while annotating the term Christmas the same participant stated: "I want to create synonyms because I can create a larger vocabulary faster." Another participant wanted to convert units, e.g., 150 kmph to mph. P̄6 was not satisfied with the available vocabularies and wanted to add his own.

5.4 Reflection: Participant feedback

We interviewed the participants about their experience in using the *loomp OCA*. Fig. 6 shows the participants' self-assessment regarding their mastery of annotations (1=no knowledge, 5=completely understood). These results are also interesting in the light of the quality of the created annotations (see Fig. 5).

We asked the participants for feedback on their *understanding* of annotations (left), ease of *creating* annotations (middle), and the ease of creating the *right* annotation (right). 9 participants found annotations easy to understand, 7 found it easy to create annotations, and 3 found it easy to create the right annotations. On average, expert participants (P̄1, P̄4, P̄6, P̄7, P̄9, P̄12) found annotations easier to understand (4.42 vs 4.0) and create (4.17 vs 3.33) than non-experts. However, both experts and non-experts found it somewhat difficult to select the right annotations (both 3.33).

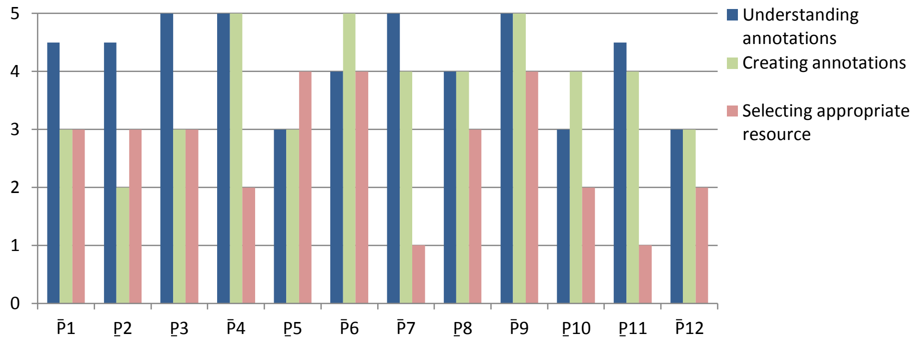


Fig. 6. Self-assessment on understanding and using annotation concepts

6 Discussion

We now discuss the insights gained from the study and draw conclusions for the design of manual annotation tools for the semantic web. We distinguish between Semantic Web 2.0 approach to annotations and the closed world of a corporate context. People responsible for creation of corporate content are typically domain experts but non-experts with respect to semantic web.

Task understanding. Some participants expected the system to be all-knowing and all-powerful. This well-known aspect from novice computer users (cf. [41]) here applied to non-experts with respect to semantic technologies. They assumed background knowledge about people mentioned in the documents, as well as complex semantic relationships between concepts in the vocabulary. This was tightly interwoven with the problem of participants' switching roles from information provider to information consumer. The task of providing conceptual, semantic information seemed so foreign to some of the participants that eventually ingrained habits of information consumers came to the fore; see Sect. 5.2/ii. This was different for participants with a strong CS/SW background; they created higher quality annotations (5.2/i) and also felt more confident about the task (5.4). However, these individuals would not be suitable as knowledge workers in a corporate environment (too highly skilled technically), but could be for an open crowd-sourcing approach.

Suitability of non-experts. Based on annotation results, we identified three groups of participants: acceptable annotators (P̄4, P̄7, P̄9, P̄12), annotators with room for improvement (P̄1, P̲8, P̲10, P̲11), and failed annotators (P̲2, P̲3, P̲5, P̄6), see Sect. 5.3/i-iii. We hypothesize that the results for the first two groups may be improved by further instructions on correct annotation. We do not believe that participants in the last group are suited as annotators. This is surprising as some were in professions that required regular creation of keywords and tagging for books (e.g., P̲5: librarian). We note that the participants in the group of acceptable annotators were all technical experts. Comparing their individual backgrounds we observed that participants of the second had more experience in working with text, in managing knowledge, and were more open to new conceptual ideas those in the third group. Furthermore, technical knowledge does not always guarantee high quality annotations (see Sect. 5.3/i). We conclude that semantic annotations cannot be done by *domain experts* (as typically assumed in the SW context) but needs *task experts*. These would be familiar not just with the domain (as all our participants were) but also with the subtleties of the annotation task.

User guidance through vocabularies. Even though the three provided vocabularies were relatively small, it was difficult for some participants to know which concepts to select and when to stop (5.2/i+5.3/iv). We concur with observations in [22,20], and observe that a reduction in size of available vocabularies and annotations helps to keep annotators focused on identifying (named) entities and will increase the quality of annotations. However, reducing the size of the vocabulary is not always a viable option and therefore documentation (SW2.0) and education (corporate) have to be explored. Dynamic online history in the annotation sidebar had mixed results (5.2/iii) and needs to be explored further.

Semantic identity. All participants were able to select the correct resource from the list if the entries contained enough information to identify the entity they had in mind. Problems arose when participants were unable to disambiguate recommended entities. Only four technical experts and one non-expert recognized the search field and the filter in the Resource selector (5.2/iv). No correlation was detected between the understanding of the resource selector and the correctness of annotations (5.3/i+v). However, selection of atoms dominated the annotation process. If in doubt about atoms, participants created new resource ids. It highlights the importance of educating annotators with the conceptual model of semantic identity and its difference to tagging. We believe non-expert users need targeted teaching material with appropriate use cases highlighting the benefits of annotation-based applications.

Interaction Patterns aligning to Mental model. The paper prototype resembled the look and feel of MS Word (following [29]) to allow easy recognition of known interaction patterns (5.2/i+ii). However, we found that the participants' mental model of how the system works had strong references to Internet-based interactions, even though the interface did not contain any typical Web elements (5.3/ii+vi). P̲11 mentioned wikipedia search as a related concept. One reason for this mismatch may be the participants' (conceptual or practical) familiarity with web-based tagging. Thus it needs to be explored which references and expectations non-expert users bring to semantic web systems.

Corporate vs. Public Semantic Web. The usage context of annotators in a corporate or public setting may differ significantly (e.g., editing new text vs. annotation of existing text). Clear *use cases* may help explore these user expectations and usage contexts. Furthermore, questions of annotation quality are expected to be answered differently within each of these contexts (cf. 5.2/i). Corporate semantic web annotations would be expected to follow a pre-defined semantics with quality measures of inter-annotator correlation [7] and stability & reproducibility [20]. However, in the public Semantic Web 2.0 sector, variation in annotation may not only be permissible (5.2/iii) but sought as an opportunity to reflect multiple perspectives on a source (e.g., supporting the needs of vision-impaired people [4]).

Annotation Nesting. A number of participants wanted to create nested annotations (5.2/ii+5.3/iii). *loomp OCA* does not currently support nested annotations as it uses XHMTL+RDFa to include annotations into the texts. Overlapping annotations cannot easily be represented in XHTML as they they would result in ill-formed XML. Visualization of nested annotations is also challenging [42]. We therefore focused on the annotation process in a paper prototype according to *loomp OCA* and did not allow for nested annotations. However, the observations of our study clearly indicate a need for the support of nested annotations.

7 Conclusions and future work

In this paper, we described a user study to evaluate the conceptual understanding of semantic annotations by non-expert users. All participants of the user study were computer-literate domain experts, six of the 12 were non-experts with respect to semantic technologies.

The results of our user study indicated that not every domain expert is a good annotator due to difficulties in the understanding of the conceptual model of semantic annotations. Though some participants had familiarity with providing content and metadata (e.g., from their occupational background), many fell back into the role of content consumers and expected the editor to provide information. Because very few use cases and applications for non-experts require the creation of semantic annotations, we assume these participants were insufficiently accustomed to this task. Even though most participants readily understood the process of creating annotations, we observed a number of challenges: granularity of atoms (e.g., sentence, phrase, word), well-formed atoms (i.e referring to named entities), annotating both instances and concepts, complexity of vocabulary, and the tendency to create new resources even though an appropriate resource exists. Some participants wanted to create a summary or synonyms instead of annotations; they felt unsure if an annotation was useful or whether they had finished. We see the reasons for these difficulties predominantly in the lack of conceptual understanding, a lack of easy-to-understand use cases and in deficits in the interaction design.

Although the study used the graphical interface of the *loomp One Click Annotator*, our results can be transferred to other editors for manually or semi-automatically annotating contents by non-experts:

Task experts: Current literature distinguishes between technical experts and domain experts. Based on our study observations, we introduce the new concept of *task experts*. Task experts are domain experts who conceptually understand the task of annotating texts and have insight into the characteristics of semantic annotations (e.g., semantic identity).

Need for use cases: We note a lack of use cases illustrating the process of annotating texts and demonstrating the benefits of semantic annotations. Use cases may need to be customized to corporate or public semantic web context.

Semantic identity: For high quality annotations, users need help in selecting appropriate resources for linking. The recommendation algorithm therefore plays an important role, and needs to be supported by an appropriate interface representation of recommended resources to users. In particular, these need to take into account that users have difficulties distinguishing between instances and classes of instances.

User evaluation methodology: We noted a lack of commonly accepted quality measures for manual semantic annotation. Furthermore, there is a lack of clearly defined methodologies for evaluating the user aspects of semantic web applications.

We currently investigate user interaction with a variety of software prototypes for semantic annotation [42] as well as their implications for digital document repositories [43]. For future research, we plan to conduct further user studies on semantic annotation in more specific usage contexts, such as combined editing and annotating in a corporate setting. Furthermore, *loomp* users may first invoke an automatic annotation service and then revise the generated annotations manually (e.g., to resolve overlapping or ambiguous annotations). A similar approach of a-posteriori annotation is supported by RDFaCE [11]. We plan to evaluate the influence of such pre-existing annotation sets on the subsequent manual annotation.

We plan to additionally investigate alternative user interfaces for selecting annotations and resources. We are interested in the impact of clearly defined use cases with practical relevance and accompanying teaching material on the quality of annotations defined by non-experts.

References

1. Luczak-Rösch, M., Heese, R.: Linked data authoring for non-experts. In: Linked Data on the Web Workshop, World Wide Web Conference. (April 2009)
2. Thomson Reuters Inc.: Open Calais website. http://www.opencalais.com/
3. Zemanta Ltd.: Zemanta. http://www.zemanta.com/ (2012)
4. Yesilada, Y., Bechhofer, S., Horan, B.: Cohse: Dynamic linking of web resources. Technical Report TR-2007-167, Sun Microsystems (August 2007)
5. Kalyanpur, A., Hendler, J., Parsia, B., Golbeck, J.: SMORE - semantic markup, ontology, and RDF editor. Technical Report ADA447989, Maryland University (2006)
6. McDowell, L., et al.: Mangrove: Enticing Ordinary People onto the Semantic Web via Instant Gratification. In: ISWC. Volume 2870. (2003) 754–770
7. Handschuh, S., Staab, S.: Cream: Creating metadata for the semantic web. Computer Networks **42**(5) (2003) 579–598
8. Booktate: Booktate website. http://booktate.com/ (2012)
9. Textensor Limited: A.nnotate. http://a.nnotate.com (2012)
10. Olive Tree Bible Software, Inc: Bible+. available online at http://itunes.apple.com/ (2012)
11. Khalili, A., Auer, S.: The RDFa content editor - from WYSIWYG to WYSIWYM. http://svn.aksw.org/papers/2011/ISWC_RDFaEditor/public.pdf (2011)
12. Morbidoni, C.: SWickyNotes Starting Guide. Net7 and Universita Politecnica delle Marche, http://www.swickynotes.org/docs/SWickyNotesStartingGuide.pdf. (April 2012)
13. Dello, K., Simperl, E.P.B., Tolksdorf, R.: Creating and using semantic web information with makna. In: First Workshop on Semantic Wikis - From Wiki to Semantics. (2006)
14. Auer, S., Dietzold, S., Riechert, T.: OntoWiki - A Tool for Social, Semantic Collaboration. In: ISWC. Volume 4273 of LNCS., Springer (2006) 736–749
15. Schaffert, S.: Ikewiki: A semantic wiki for collaborative knowledge management. In: Workshop on Semantic Technologies in Collaborative Applications, Manchester, UK (June 2006)
16. Hinze, A., Voisard, A., Buchanan, G.: Tip: Personalizing information delivery in a tourist information system. J. of IT & Tourism **11**(3) (2009) 247–264
17. Kruk, S.R., Woroniecki, T., Gzella, A., Dabrowski, M.: JeromeDL - a semantic digital library. In: Semantic Web Challenge. (2007)
18. faviki: faviki - tags that make sense. http://www.faviki.com
19. Passant, A.: MOAT-project. http://moat-project.org
20. Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, University of Edinburgh (1999)

21. Handschuh, S.: Creating Ontology-based Metadata by Annotation for the Semantic Web. Ph.d. thesis (dr. rer. pol.), University of Karlsruhe (TH) (2005)
22. Bayerl, P.S., Lüngen, H., Gut, U., Paul, K.I.: Methodology for reliable schema development and evaluation of manual annotations. In: Workshop on Knowledge Markup and Semantic Annotation. (2003) 17–23
23. Erdmann, M., Maedche, A., Schnurr, H., Staab, S.: From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. Group **6**(i) (2000) pp. 79–91
24. Maynard, D., Dasiopoulou, S., et al.: D1.2.2.1.3 Benchmarking of annotation tools. Technical report, Knowledge Web Project (2007)
25. Castro, R.: Benchmarking Semantic Web technology. Studies on the Semantic Web. IOS Press (2010)
26. Degler, D., Henninger, S., Battle, L.: Semantic Web HCI: discussing research implications. In: Extended abstracts on Human factors in computing systems, ACM (2007) 1909–1912
27. Schraefel, M., Karger, D.: The pathetic fallacy of rdf. In: International Workshop on the Semantic Web and User Interaction (SWUI) 2006. (2006)
28. Uren, V., Cimiano, P., et al.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semant. **4**(1) (January 2006) 14–28
29. Karger, D.: Unference: Ui (not ai) as key to the semantic web. http://swui.semanticweb.org/swui06/panel/Karger.ppt (2006) Panel on Interaction Design Grand Challenges and the Semantic Web at Semantic Web User Interaction Workshop.
30. Jameson, A.: Usability and the semantic web. In: European Semantic Web Conference, Berlin, Heidelberg, Springer-Verlag (2006) 3 m18.
31. Duineveld, A.J., Stoter, R., et al.: Wondertools? a comparative study of ontological engineering tools. Journal of Human-Computer Studies **52**(6) (June 2000) 1111–1133(23)
32. Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., Antoniou, G.: Ontology change: Classification and survey. Knowl. Eng. Rev. **23**(2) (June 2008) 117–152
33. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems **16**(2) (March 2001) 72–79
34. Jones, W., Karger, D., Bergman, O., Franklin, M., Pratt, W., Bates, M.: Towards a Unification & Integration of PIM support. Technical report, University of Washington (2005)
35. Aimé, X., Furst, F., Kuntz, P., Trichet, F.: Conceptual and lexical prototypicality gradients dedicated to ontology personalisation. In: OTM 2008, Springer-Verlag (2008) 1423–1439
36. Hogben, G., Wilikens, M., Vakalis, I.: On the ontology of digital identification. In: OTM Workshops. Volume 2889 of Lecture Notes in Computer Science., Springer (2003) 579–593
37. Handschuh, S., Staab, S.: Authoring and annotation of web pages in cream. In: WWW. (2002) 462–473
38. Heese, R., Luczak-Rösch, M., Paschke, A., Oldakowski, R., Streibel, O.: One click annotation. In: Workshop on Scripting and Development for the Semantic Web. (May 2010)
39. Degler, D.: Design 10:5:2 for semantic applications. In: Semantic Technology Conference. (2011) online at http://www.designforsemanticweb.com/.
40. Snyder, C.: Paper prototyping: The fast and easy way to design and refine user interfaces. Morgan Kaufmann Pub (2003)
41. IBM: User expectations. http://www-01.ibm.com/software/ucd/initial/expectations.html (2012)
42. Schlegel, A., Heese, R., Hinze, A.: Visualisation of semantic enrichment. In: Interaction and Visualisation in the Data Web, Workshop at Informatik'2012. (2012)
43. Hinze, A., Heese, R., Schlegel, A., Luczak-Rösch, M.: User-defined semantic enrichment of full-text documents: Experiences and lessons learned. In: Theory and Practice of Digital Libraries. (2012)