Seed, a Natural Language Interface to Knowledge Bases

Bahaa Eldesouky(✉), Heiko Maus, Sven Schwarz, and Andreas Dengel

Knowledge Management Department, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany
{bahaa.eldesouky,heiko.maus,sven.schwarz,andreas.dengel}@dfki.de

Abstract. The World Wide Web has been rapidly developing in the last decade. In recent years, the Semantic Web has gained a lot of traction. It is a vision of the Web where data is understandable by machines as well as humans. Developments in the Semantic Web made way for the creation of massive knowledge bases containing a wealth of structured information. However, allowing end-users to interact with and benefit from these knowledge bases remains a challenge.

In this paper, we present *Seed*, an extensible knowledge-supported natural language text composition tool, which provides a user-friendly way of interacting with complex knowledge systems. It is integrable not only with public knowledge bases on the Semantic Web, but also with private knowledge bases used in personal or enterprise contexts.

By means of a long-term formative user-study and a short-term user evaluation of a sizable population of test subjects, we show that *Seed* was successfully used in exploring, modifying and creating the content of complex knowledge bases. We show it enables end-users do so with nearly no domain knowledge while hiding the complexity of the underlying knowledge representation.

Keywords: Usability · Semantic Web · Natural language · Knowledge bases

1 Introduction and Related Work

The World Wide Web has seen rapid developments in the last decade. It is steadily transforming into the Semantic Web [6], a Web where data is understandable and consumable by machines as well as humans. Although the vision of the Semantic Web is making progress towards its realization, a gap between non-expert end-users and the content of the Semantic Web still exists. Tools for interacting with structured information on the Web remain directed almost entirely at highly trained individuals [5].

The progress of the Semantic Web vision can be observed in the field of modeling and structuring data, where huge knowledge bases such as DBPedia [7] and Freebase [8] contain millions of concepts and billions of facts about them. These huge knowledge bases comprise a web of Linked Open Data (LOD) [4].

© Springer International Publishing Switzerland 2015
S. Yamamoto (Ed.): HIMI 2015, Part I, LNCS 9172, pp. 280–290, 2015.
DOI: 10.1007/978-3-319-20612-7_27

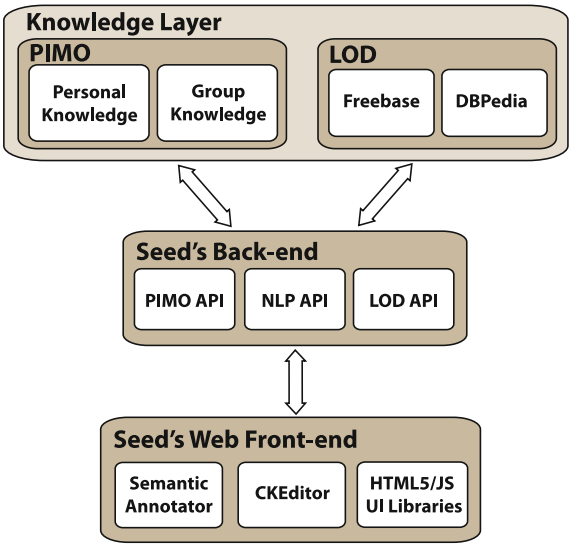

Fig. 1: *Seed* architecture diagram

In addition to public knowledge repositories, there are private ones, which focus on individual or group knowledge [15] (e.g. corporate knowledge repositories).

Development of user interfaces for non-experts that allow for user-friendly consumption and interaction with the existing knowledge on the Semantic Web is essential. In this paper, we present *Seed*, short for semantic editor. It is an extensible knowledge-supported web-based natural language text composition tool. We point out the structure of *Seed*, explain how it builds state-of-the-art developments in the fields of NLP, LOD and Semantic Web technologies to provide a user-friendly way of interacting with complex knowledge systems. By means of a long-term formative user-study and a short-term user evaluation of a sizable population of test subjects, we show that the use of *Seed* in exploring, modifying and creating semantic content reduces prerequisite domain knowledge and hides the complexity of the underlying knowledge representation.

Research on enabling end-users to interact with the Semantic Web in a friendly way is increasingly gaining interest. Examples include SemCards [17], which provides an intermediate ontological representational level that allows end-users to create rich semantic networks for their information sphere. OntoAnnotate [16] is an ontology-based annotation environment for web pages based on RDF [13] and RDFschema [9]. RDFaCE [11] provides an easy way for adding RDF-based annotations to text. RDFauthor [18] bases on making arbitrary XHTML views with integrated RDFa annotations editable. In [12] a WYSIWYG tool called OntosFeeder is proposed which annotates text for the news/journalism domain. In [3], Epiphany is used to get RDFa enhanced versions of their articles linking to underlying Linked Data models.

2 Seed Architecture and Implementation

As depicted in Fig. 1, Seed consists of three main logical parts: a knowledge layer, a back-end and a front-end. All three parts are loosely coupled and communicate via standard Web APIs.

2.1 Knowledge Layer

This logical part embodies the knowledge integrated with Seed. We distinguish between two scopes of knowledge: a personal and a world scope. With the personal knowledge scope, we refer to the knowledge model of the user(s) which contains things relevant to an individual or a group. Things present in the personal knowledge scope may not be relevant for many outside of the group. With the world knowledge scope, we refer to common knowledge publicly accessible on the Semantic Web such as from LOD sources (e.g. Freebase and DBPedia). It contains things common to a large group of people such as celebrities, companies, countries ... etc. This distinction gives priority to user knowledge and complements missing information using public knowledge.

PIMO. Personal/group knowledge refers to structured information about concepts from the point of view of an individual user or a specific group of users. The PIMO [15] is a personal and group knowledge base reflecting the mental models of the users with concepts such as persons, projects, tasks, topics, events and resources such as emails, files, webpages, notes, pictures. PIMO knowledge of the author and possibly that of the group to which the author may belong is integrable in *Seed*

LOD. General common knowledge on the other hand, refers to structured information available from public knowledge repositories such as DBPedia, Freebase and other LOD sources. The integration of common knowledge is important for complementing the user's knowledge. It also helps expand the context of knowledge in a text authored by the user.

2.2 Back-End

This component, physically realized on the server-side, consists of multiple APIs responsible for:

- Communication with integrated personal and public knowledge sources.
- Analyzing content authored by the user and enriching with information retrieved from knowledge sources.

NLP API. This component is implemented as a Java service which builds on two state-of-the-art NLP toolkits; Stanford CoreNLP [14] and Apache OpenNLP [1]. It can perform major NLP tasks such as named entity recognition, coreference resolution and relation extraction. The NLP API currently supports two languages; English and German.

LOD API. This component of the back-end communicates in real-time with live LOD sources to extract information about concepts mentioned in the text. This component can work independently or in combination with the NLP API. Following are example tasks where this component is involved:

- Entity disambiguation: By an entity we refer to a thing (e.g. person, city, event, organization ... etc.) that is mentioned in the text and is of interest to the user.
- Relation extraction: Finding relations between arbitrary entities.

PIMO API. This component interacts with personal and group knowledge from PIMO on behalf of the user. The interaction scenarios include but are not limited to:

- Identifying and disambiguating entities or suggesting related ones.
- Extracting information about recognized entities from the users's personal or group PIMO.
- Finding relations between entities mentioned in the text.
- Adding new entities from the text to the user knowledge in PIMO.

2.3 Web Front-End

The current prototype of *Seed's* front-end is meant to run in the browser. Therefore, it is written completely in HTML5 and JavaScript. However, it is also possible to embed it in graphical user interfaces (GUIs) built using other languages. The only prerequisite is the availability of an HTML capable user interface (UI) element to run the editor component.

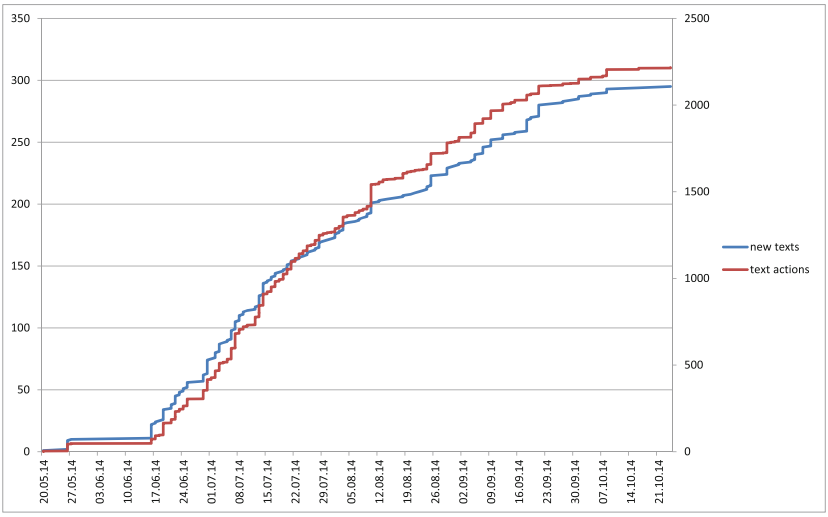

Fig. 2: Statistics about textual content created in *Seed* over a six-month period

CKEditor. At the core of the front-end, *Seed* builds upon CKEditor [2], the open source WYSIWYG HTML editor.

Semantic Annotator. The semantic annotator is a JS/HTML extension responsible for monitoring the HTML and continuously performing basic analysis of the content typed by the user. It also communicates with the back-end retrieving annotations and applying them to the text in a proactive way.

HTML/JS UI Libraries. Depending on the application scenario in which *Seed* is integrated, various HTML/JS libraries (e.g. jQuery, UI, Bootstrap ... etc.) are used to build the GUI.

2.4 Experimental Evaluation

In order to assess the value of *Seed* for end-users interacting with knowledge bases, we have performed two user-evaluations.

Long-Term Formative User-Study. This long term evaluation aimed at assessing the suitability of *Seed* as a natural language interface to PIMO. During the evaluation, we iteratively improved existing features of *Seed* or added new ones to it based on the feedback from our test subjects.

Demographics:

- Number of users: 4
- Occupation: Students of non-technical majors
- Duration: 6 months
- Languages: German, English

For the duration of six months between May and November 2014, four students extensively used a shared PIMO for personal and professional reasons. It allowed them to model things important to them such as places, events, companies, photos, files ... etc. The had the ability to share and collaborate on the information they increasingly accumulated throughout their use of PIMO. A lot of the PIMO things they modeled where composed of unstructured text. Examples included shopping lists, meeting notes, recipes ... etc.

We integrated *Seed* in Web-based and Java-based UIs of PIMO. The goal was to use it as the main interface for interacting with textual PIMO things or textual properties of many non-textual PIMO things. Before integrating *Seed*, tasks like annotating concepts, creating relations between them and adding new ones could be done only through interaction with menus, buttons and conventional UI elements. Using *Seed* in the case of textual PIMO things, users could perform the aforementioned tasks automatically or semi-automatically while composing natural language text. Users had access to *Seed* for the following purposes:

- Writing textual descriptions of existing PIMO entities, such as persons, photos, institutions, events or many other types.

– Composing free text documents, such as notes, meeting minutes, diary entries, shopping lists ... etc.

While composing text, *Seed* identified mentioned entities and suggested related ones to the users. It helped them save time and effort of manually searching for and annotating entity with related ones.

During the six months, we met the students regularly on a weekly basis. In each meeting, they reported the types of activities they undertook using PIMO for the past week and highlighted success stories and fail stories. Their iterative feedback about interaction with textual PIMO things was used to guide the development of *Seed*. Figure 2 shows a plot of the number of texts created over the period of six month by the test subjects. It also shows the number editing actions performed on those texts. The so-called text actions refer to interactions with the content of the text through *Seed*.

During the first month, *Seed* was being integrated in various GUIs of PIMO. Test subjects got introduced to it and provided preliminary feedback about their most frequent text composition needs. As seen during the first month, users created few texts and rarely interacted with them once created. They mostly interacted with PIMO through non-textual interfaces. As can be seen in Fig. 2, by the end of the first month, the students had become more familiar with *Seed* and started using it more frequently for creating new texts. They increasingly adopted it for interacting with textual PIMO things. However, during most of the second month, test subjects tended to rarely edit texts once they were created. Using their feedback we iteratively improved interaction possibilities with annotations to address the problem. Towards the end of the third month, test subjects edited documents they had created substantially more often. As seen in Fig. 2, from the fourth month onwards, multiple edits per document became more often resulting in a considerable increase over the number of documents created.

Short-Term User-Study. In this evaluation, we assessed the usability of *Seed* in annotating texts with semantic information from public knowledge source, namely Freebase and DBPedia.

Demographics:

– Number of users: 115
– Language: English

Figure 4 shows statistics about the population of test subjects who participated in our evaluation experiment. As can be seen, the diversity of profiles of the participants as well as the number of participants are high enough to guarantee representative results.

Procedure: We have set up an evaluation website where test subjects, performed the experiment which proceeded as follows:

1. Registration, where participants provided demographic information about themselves.

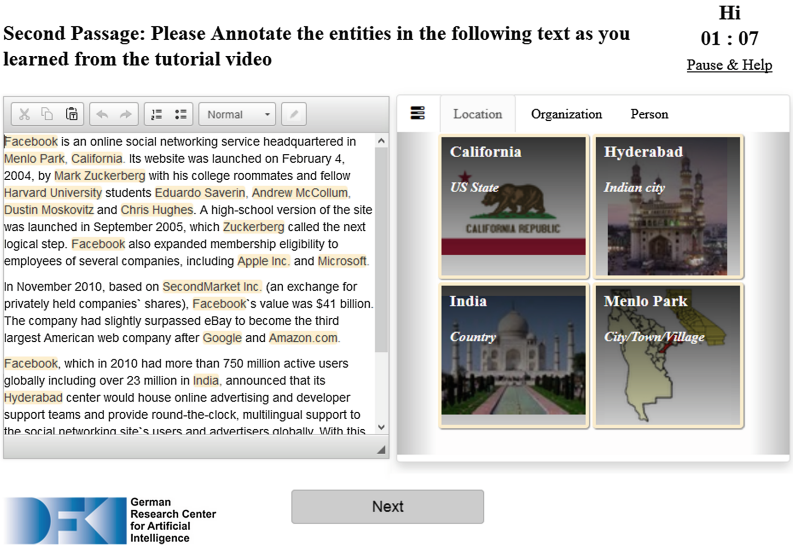

Fig. 3: Sample snapshot of the evaluation interface

2. Then, participants watched a short non-technical tutorial video [1] explaining the concept and use of *Seed*. We avoided technical aspects in the video in order to safely assume no technical domain knowledge.
3. Participants were asked to annotate 3 text passages from various domains using *Seed*. They reviewed automatic annotations by *Seed* and suggestions for annotations that they could confirm or reject.
4. Afterwards, participants were asked to type in a given text passage, which *Seed* proactively annotated as it got written.
5. Finally, participants were asked questions including but not confined to a Standard Usability Score (SUS) [10] questionnaire.

Standard Usability Score. After the end of the evaluation, users filled in a SUS questionnaire. Figure 5 shows the aggregated results of the questionnaire. Across the population of test subjects, *Seed* scored an overall mean SUS of 73.56 with standard deviation equal to 13.71. According to [10], this means *Seed* has above average usability.

Knowledge Exploration. As seen in Fig. 3, *Seed*'s front-end contains a faceted browsable view of the things mentioned in the text. It provides a user friendly way of interacting with the knowledge present in the text. The content of the faceted view is automatically extracted from PIMO, DBPedia and Freebase by *Seed*. In order to evaluate the faceted view, we asked users after the evaluation two questions whose answers are not mentioned the in text passages, but are available through the faceted view as well as through the interactive annotation

[1] Seed, the Semantic Editor - https://www.youtube.com/watch?v=CSFS4sxWm0w.

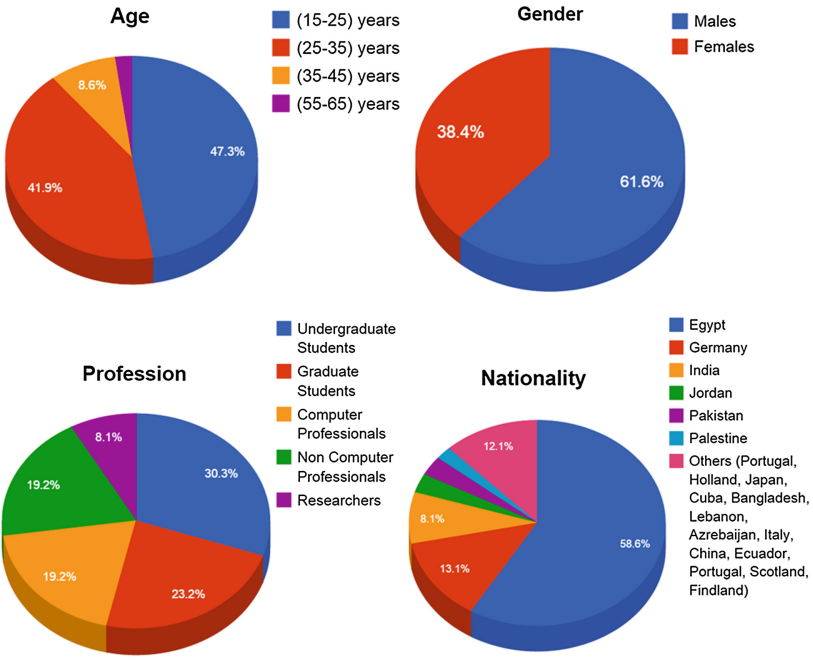

Fig. 4: Age, gender, background and nationalities of participants

information pane which shows up when users inspect the annotations in the text. Our hypothesis was that users would discover the information in the UI easily. The results of users answers were as follows:

- For the first question, 94.9 % of the participants managed to find the correct answer.
- For the second question, 51.5 % of the participants managed to find the correct answer.

To avoid the participants looking up the answer elsewhere outside of *Seed*, we explicitly asked them how they found it. For those who could answer at least one question, 93.9 % did so using *Seed*'s UI elements. This shows how *Seed* managed to relieve the majority of the users from the manual effort of browsing the knowledge base to search for the answers.

Learnability. To quantitatively evaluate how fast users learned to use *Seed*, we measured the time required to annotate each of the text passages. After cleaning up the data and normalizing it with respect to length, we got the values shown in Fig. 6. We can see an overall trend of decrease in the time required for interaction with the text using *Seed*. Although the number of the text passages was limited due to practical reasons, it can be seen that the time required to read and annotate passages decreased with progress in the experiment. It is worth noting that the number of annotations increased with the progress in the experiment,

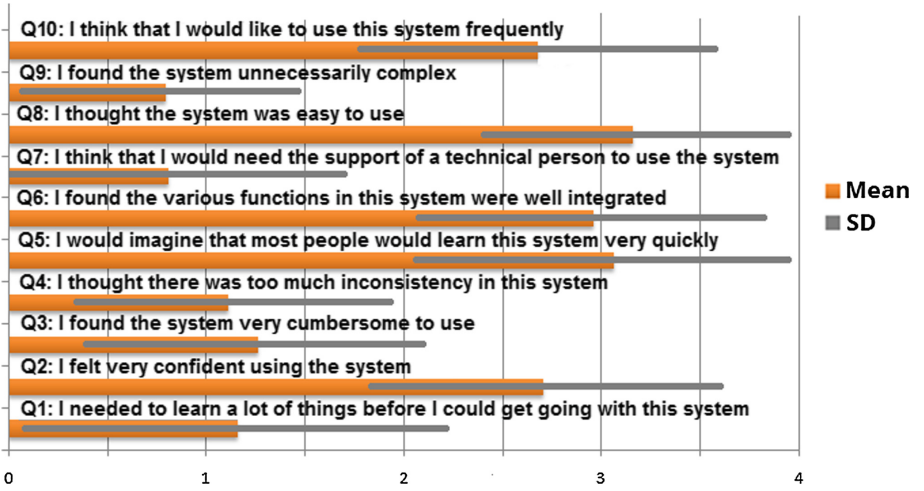

Fig. 5: Aggregated results of the SUS questionnaire. (Strongly Disagree: 0, Disagree: 1, Neutral: 2, Agree: 3, Strongly Agree 4)

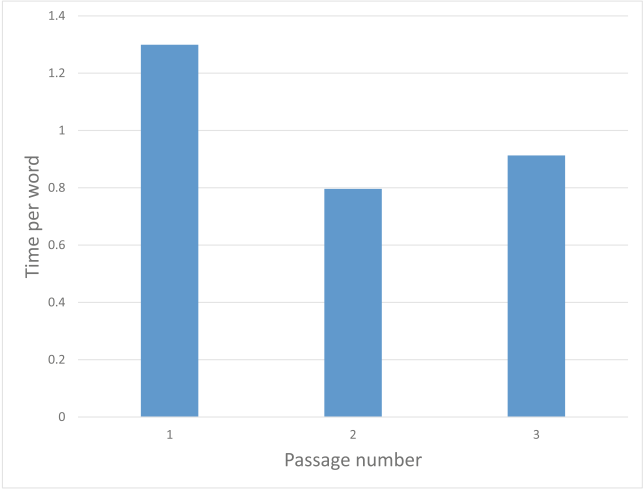

Fig. 6: Mean annotation time in seconds for all participants

which implies increased user familiarity with the system. This can also explain the slight increase in the time of the third passage 3 in comparison to the second passage.

3 Conclusion

In this paper, we presented *Seed*, a user-friendly natural language interface to complex knowledge bases. We explained its architecture and pointed out how

it aims to enable user-friendly interaction with knowledge bases through natural language text composition. We put the system to a long-term formative user-study and a short-term usability evaluation involving a large group of test subjects. The results of the experiments show how *Seed* enabled non-expert end-users to interact with personal as well as common public knowledge bases.

For future work we plan to add more interaction possibilities with the text beyond annotation and knowledge browsing. We also plan to investigate other aspects related to collaborative interaction with the knowledge through the text.

Acknowledgments. The work presented was partially funded by the European Commission in the context of the FP7 ICT project ForgetIT (under grant no: 600826).

References

1. Apache opennlp, October 2014. URL: http://opennlp.apache.org/index.html
2. CKEditor Website, October 2014. URL: http://ckeditor.com
3. Adrian, B., Hees, J., Herman, I., Sintek, M., Dengel, A.: Epiphany: adaptable RDFa generation linking the web of documents to the web of data. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS, vol. 6317, pp. 178–192. Springer, Heidelberg (2010)
4. Auer, S., Bryl, V., Tramp, S.: Linked Open Data-Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project. LNCS, vol. 8661. Springer, Heidelberg (2014)
5. Benson, E., Karger, D.R.: End-users publishing structured information on the web: an observational study of what, why, and how. In: Proceedings of the 32nd annual ACM Conference on Human Factors in Computing Systems, pp. 1265–1274. ACM (2014)
6. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
7. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semant: Sci. Serv. Agents World Wide Web **7**(3), 154–165 (2009)
8. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM (2008)
9. Brickley, D., Guha, R.V.: Resource description framework (rdf) schema specification 1.0: W3c candidate recommendation, 27 March 2000 (2000)
10. Brooke, J.: SUS: a quick and dirty usability scale. In: Jordan, P.W., Thomas, A., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry. Taylor and Francis, London (1996)
11. Khalili, A., Auer, S., Hladky, D.: The RDFa content editor - from WYSIWYG to WYSIWYM. In: 2012 IEEE 36th Annual Computer Software and Applications Conference (COMPSAC), pp. 531–540. IEEE (2012)
12. Klebeck, A., Hellmann, S., Ehrlich, C., Auer, S.: OntosFeeder – a versatile semantic context provider for web content authoring. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 456–460. Springer, Heidelberg (2011)

13. Lassila, O., Swick, R.R., et al.: Resource description framework (rdf) model and syntax specification (1998)
14. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford coreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
15. Sauermann, L., van Elst, L., Dengel, A.: PIMO - a framework for representing personal information models. In: Pellegrini, T., Schaffert, S. (eds.) I-SEMANTICS Conference, pp. 270–277. J.UCS, Know-Center, Graz, 5–7 September 2007
16. Staab, S., Maedche, A., Handschuh, S.: Creating metadata for the semantic web - an annotation environment and the human factor. In: Institute AIFB (2000)
17. Thórisson, K.R., Spivack, N., Wissner, J.M.: Semcards: a new representation for realizing the semantic web. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 425–436. Springer, Heidelberg (2009)
18. Tramp, S., Heino, N., Auer, S., Frischmuth, P.: RDFauthor: employing RDFa for collaborative knowledge engineering. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS, vol. 6317, pp. 90–104. Springer, Heidelberg (2010)