Efficient High-Level Semantic Enrichment of Undocumented Enterprise Data

Markus Schröder

[1] Smart Data & Knowledge Services Dept., DFKI GmbH, Kaiserslautern, Germany
[2] Computer Science Dept., TU Kaiserslautern, Germany
markus.schroeder@dfki.de

Abstract. In absence of a data management strategy, undocumented enterprise data piles up and becomes increasingly difficult for companies to use to its full potential. As a solution, we propose the enrichment of such data with meaning, or more precisely, the interlinking of data content with high-level semantic concepts. In contrast to low-level data lifting and mid-level information extraction, we would like to reach a high level of knowledge conceptualization. Currently, this can only be achieved if human experts are integrated into the enrichment process. Since human expertise is costly and limited, our methodology is designed to be as efficient as possible. That includes quantifying enrichment levels as well as assessing efficiency of gathering and exploiting user feedback. This paper proposes research on how semantic enrichment of undocumented enterprise data with humans in the loop can be conducted. We already got promising preliminary results from several projects in which we enriched various enterprise data.

Keywords: semantic enrichment · knowledge graph building · enterprise data · human in the loop

1 Introduction and Motivation

With the accompanying digitalization we are in the midst of the "data everywhere" age. Since the advantages of electronic data processing have long been recognized, companies are increasingly digitizing their processes as well as collecting and processing data in various data pools. However, if necessary data maintenance is lost in hectic everyday work, naturally grown undocumented data piles up. That is why we frequently encounter arbitrarily structured, diverse, heterogeneous and distributed enterprise data sets accumulated over several years.

Therefore, for companies it becomes increasingly difficult to discover and make use of their data. Especially messy data is an obvious obstacle in performing complex data mining analyses which are used to gain new insights from data [35]. This observation also conforms with the statement of data scientists, who regard data preparation as an integral part of their work [10]. Moreover, it hinders employees to efficiently work with the data content in day-to-day business. Here, possible solutions could be workload reducing software tools embedded in the work environment to support employees' daily work [24]. However, such efforts require pre-processed, structured and organized data.

Thus, we propose as a solution the semantic enrichment [6] of such data by meaning. Instead of stopping at a low-level conversion or mid-level information extraction (IE), our plan is to augment data with high-level concepts. While data lifting usually converts the data structure to a knowledge representation and IE further extracts structured data from unstructured one, semantic enrichment in addition precisely annotates data with meaningful concepts obtained by experts. For example, the table with header `|Car-ID|EOP|` can be easily converted to RDF [37] and car configurations hidden in the car's ID can be extracted with rules, however, domain knowledge is needed to reveal that EOP means *End Of Production*. The knowledge to understand the data is hidden in the mindsets of domain experts working daily with their usual data assets. In companies they are usually limited in number and have only restricted time available. Therefore, we plan to integrate their feedback as efficiently as possible into the enrichment process. Only then we can achieve that high-level concepts in the employees' minds are heeded in the leveraging process.

As a concrete example, think of a manufacturing company building products in process pipelines. Its messy dataspace is full of arbitrarily structured data produced by many employees without any consultation: product databases, planning spreadsheets, XML exports, mails and shared drives. This should be semantically enriched to ease the application of data mining tasks across all those data assets. Initially, different information extraction methods are applied to these data pools by a knowledge engineer. Several domain experts are invited to give feedback on extracted results. Iterating between those two procedures, a semantic graph of high-level concepts emerges. Finally, a data analyst is able to use semantic services to exploit gained structured data for various mining tasks.

The remainder of this paper is structured as follows: Section 2 lists the state of the art with related work followed by the problem and consequential research questions in Section 3. Section 4 outlines the planned research methodology while Section 5 sketches a brief evaluation plan. Preliminary results are described in Section 6 followed by the conclusion (Section 7).

2 State of the Art

This research builds on state of the art and related work from several domains.

Semantic Technologies. The foundation of this PhD is formed by the research area Semantic Web [2]. Commonly, ontologies are used to create a formal, explicit specification of a shared conceptualization [34]. Knowledge bases, usually modelled with the Resource Description Framework (RDF) [37] as sets of facts, interlink resources using URIs and are processed by semantic technologies [15]. If a knowledge base is represented as a graph, we speak of a knowledge graph [26]. In corporations, such semantic networks of domain knowledge are known as *enterprise* knowledge graphs [12,11]. The non-trivial question however is, how to efficiently construct a sophisticated knowledge graph originating from human knowledge, expertise and diverse enterprise data.

Knowledge Graph Construction. Data lifting is a technique that converts well structured data to an RDF representation using mapping definitions (e.g. using R2RML[3] and GRDDL[4], to name prominent ones). While it can convert structured data to RDF, they have several limitations on unstructured or semi-structured data despite some enhancements that were recently made [5]. Usually, they require knowledge engineers who are familiar with the data sources, mapping languages and target ontologies. Current state-of-the-art enterprise knowledge graph construction approaches assume well (semi-)structured data (CSV, relational DB, XML, JSON, etc.) with well-known schemata [28,26,22]. They follow a kind of extract-transform-load (ETL) pipeline which converts data into RDF once the mapping is configured. However, in our scenarios, we can neither assume solely structured data, nor knowledge about schemata, nor existence of suitable ontologies in advance. A deficiency of such approaches is the necessity of comprehensive prior knowledge and rigid processing to perform the construction. We envision an agile workflow which would permit yet unknown parameters (schemata, ontologies) and intermediate feedback loops with users.

Information Extraction on Unstructured Data. We suppose that arbitrarily structured (i.e. not especially well structured) enterprise data is mostly generated by people writing texts such as labels or descriptions. To automatically gain structured information from unstructured and also semi-structured data, various information extraction (IE) methods in the field of NLP have been researched for several years [23]. This includes procedures that find entities in texts (named entity recognition), determine relevant terms from the domain (term extraction) and disambiguate these to knowledge base identifiers (entity linking). They can be used to discover initial suggestions for relevant terms, concepts, instances, or links found in the data. During the dissertation, suitability assessments of several state-of-the-art approaches will be conducted.

However, not all methods are applicable without adaptions, as they make certain assumptions about the input, for example, the text's nature. In companies we often encounter, besides usual documents, short ungrammatical text snippets in their data assets (e.g., file and folder names [3], database schema labels [27] and semi-structured data in general). Particularly, short texts contain no regular grammar, have only few statistical signals and are rather ambiguous [17]. To address text snippets human behaviour could be detected and exploited, because people tend to label elements in their own way but in a repetitive manner. For example, imagine a person always using an underscore (_) to separate tokens, while date information formatted as `YYYYMMDD` is appended at the end. Semantic labelling approaches [27] could annotate such information on a character level:

[3] `http://www.w3.org/TR/r2rml`

[4] `http://www.w3.org/TR/grddl`

an example would be the annotation in `travel_20190602.txt` which is linked to the concepts *Date* and *ESWC*. Because usually such labelled data does not exist right from the start the labelling approaches should be unsupervised. If the text additionally shows no clear separation between tokens, approaches like automatic identifier splitting [9] have to be performed up front.

Meta-Data Management on Data Lakes. Aforementioned approaches having particular input assumptions are not sufficient for the huge data diversity in companies. A company's dataspace [14] typically contains not only relational databases and well structured files, but also unstructured (short) texts and weakly semi-structured file formats. There is a new trend to gather raw data in a data lake [20] and integrate it step by step in a pay-as-you-go fashion [18]. Should this fail, the data lake turns more and more into a data swamp: its content becomes difficult to understand and to discover. New research directions tackle this challenge with sophisticated meta-data management [4,13,33]. However, in order to reach a high-level enrichment, we often need to capture and explicitly model the meaning hidden in the data content on a more fine-grained level. This requires to directly descend deeper into its content and annotate meaning on a character level. That way we can more precisely acquire feedback from the user.

Human in the Loop. Regarding involvement of human experts, recent research enables users to semi-automatically populate an ontology by user-defined conceptualisations [7]. They automatically align them with an ontology and let users add new instances or concepts as well as split or merge them. In the context of IE, interactive information extraction [21] allows users to verify and correct extraction results. However, there is still a considerable amount of manual labour to be performed by the user. That is why other research focuses on reducing human annotation effort, for instance on named entity recognition [36] and extraction [8]. Those approaches present users selected sentences containing most likely correctly found entities. We would like to further reduce their effort by showing automatically generated summaries and quickly graspable visualizations. For instance, we could summarize entities by their suggested types. Knowledge acquisition [25] investigates methods to extract expertise from experts. We can apply proven direct methods in order to enrich data with the user's expertise. Promising methods like questionnaire or interview are integrable into the semantic enrichment process.

To the best of our knowledge there is no system yet that deeply integrates experts in the enrichment process like the one we envision.

3 Problem Statement and Contributions

This PhD will be guided by the following main research question:
How can an efficient high-level semantic enrichment of undocumented enterprise data be conducted?

The question focuses specifically on arbitrarily structured data occurring in companies. Such data is rather undocumented and thus difficult to comprehend and to process. In contrast to research with mass data on the web, we focus on semantic enrichment on an enterprise level. Compared with the internet, enterprise data is more limited in quantity while enrichment results are expected by companies to be of high quality. Thus, statistical signals (e.g. based on frequency) will not reveal relevant data points as expected from approaches which use the Internet's mass data. Conversely, enterprises usually have a manageable domain and participants with low variety of possible information. An alternative to statistical signals is human expertise directly suggesting relevant data points. Since experts are costly, we need to make the process efficient. Only with the human component can we reach the intended high semantic level.

Subsequently, the main research question is further divided into subquestions.

RQ 1 *What state-of-the-art approaches can be utilized and how can they be adapted to gain precise results on limited diverse enterprise data?*

By the application of current approaches, we expect an initial mid-level enrichment. However, the special data situation in enterprises will limit the selection of already existing methods. Some supervised approaches can be excluded in advance because they require a considerable amount of labelled data to provide an acceptable outcome (e.g. deep learning). Others have too strong assumptions about the data's nature that do not match with data found in corporations (e.g. short ungrammatical texts). Especially very special data structures (e.g. file trees, spreadsheets, PDFs) require procedures that exploit the unique nature of their contents. Therefore, during this PhD the suitability of several state-of-the-art approaches with regard to concrete data situations will be evaluated. If necessary they will be adapted to the special nature of enterprise data. For instance, unsupervised terminology extraction could be extended with additional metrics reflecting the term's occurrence in the data (e.g. folder hierarchy depths).

Sophisticated algorithms can achieve a considerable enrichment level automatically. However, only with domain experts integrated in this process we will reach the intended high level. That is why we ask the following research question:

RQ 2 *How can we efficiently integrate human experts in the process to achieve our envisioned high level of enrichment?*

In contrast to crowdsourcing, we only have a limited number of experts in the company who can give feedback to enriched results. In addition, employees have only restricted time available for feedback loops. Because of these constraints we need time-saving human-in-the-loop methods that allow to give targeted feedback desirably on a large number of data points. Therefore, the research question also includes the design of efficient graphical user interfaces together with suitable interaction patterns. This involves giving feedback to already organized, thus quickly graspable statements without great effort, for example, using short mouse movements on clustered elements. Every human input, no matter how small, could immediately contribute to the enrichment.

The questions so far emphasized a “high-level” as well as an “efficient” enrichment. We aim to estimate these parameters as well as possible during the process at any time. Assessing the status quo will benefit in choosing the appropriate enrichment algorithms and design of feedback loops. Hence, the next research questions are concerned with such assessments.

RQ 3 *How can we quantify the enrichment level of algorithms?*

The level of enrichment approximates how comprehensive the gained knowledge about enterprise data is. This can vary greatly, depending on current circumstances and field of application. While a high coverage of enterprise data is desired, at the same time, the enrichment results should be as precise as possible. In order to objectively judge the results of algorithms, we will develop semantic enrichment measures. For this, a good starting point provide ontology evaluation metrics [16] like accuracy, completeness, consistency, clarity, etc. However, we have to further adjust them to also reflect data dependent aspects, like for instance file coverage. Our metrics are applied to the algorithms' results to quantify what enrichment level they can reach. This allows to assess and compare state-of-the-art procedures in terms of their suitability in semantic enrichment.

Last, we intent to quantify human effort.

RQ 4 *What are appropriate measures to assess the efficiency of gathering and exploiting user feedback?*

In order to make human in the loop approaches comparable, we will quantify their efficiency in collecting feedback. One dimension is the number of verifications divided by the time the expert needed. This can be combined with the expertise, willingness and available time of domain experts. Moreover, we are interested to quantify how well various feedback types help in the enrichment process. Typical ones include direct methods found in knowledge acquisition [25] like interviews moderated by the knowledge engineer, (auto-generated) questionnaires, user observation, protocol analysis and drawing closed curves. They will be adapted for (digital) enterprise data, for example, a questionnaire directly refers to the data items where feedback is needed.

Summarising, the main contribution is an innovative semi-automatic method to lift enterprise data to a high semantic level by efficiently integrating human in the loop. Suitable procedures will be identified to initially enrich enterprise data. We expect new insights on how to involve experts in the process the best way. The outcome of the PhD will also provide a base for further research in the field of enterprise knowledge graph construction and bootstrapping semantic services (e.g. Semantic Desktop).

4 Research Methodology and Approach

In our research institute, we work closely with industry customers, for whom we develop and apply innovative AI solutions. During the application we usually face different data sources which have to be processed for the actual project's objective. We expect that our developed knowledge services will provide even better results, once we enrich the data with domain knowledge. For such cases, our proposed methodology enables an efficient enrichment in which answers to the stated research questions are provided.

Regarding RQ 1, we would like to mutually compare state-of-the-art approaches and, if necessary, extend them appropriately. By making their enrichment level quantifiable, we verify the usability of individual procedures. Bottom-up, we semantically annotate raw data in order to form a semantic graph originating from the data. Top-down, we heed the conceptualization of the employees by involving them into the process at an early stage. In doing so, we consider two aspects simultaneously: First, information extraction is not solely data-driven, since domain experts directly give feedback to their outcome. Second, the experts' conceptualisation is not modelled in isolation, because concepts are immediately linked to the corresponding data item. By this, we expect to get better results than with existing methods.

Per use case, we will have contact with several domain experts who can be consulted for user studies. In this event, various user interfaces, feedback types and interaction patterns will be designed and tested systematically. We plan to develop graphical interfaces presenting enrichment suggestions in easily comprehensible arrangements. Feedback types will vary from simple questionnaires to costly interviews with a knowledge engineer. Interaction patterns include various mouse gestures (e.g. drag & drop) and keyboard shortcuts. Iteratively, we contribute to the question how to integrate human experts efficiently (RQ 2).

In order to quantify the enrichment level in RQ 3, many dimensions must be considered. First, we will create a classification system to make the current data situation in the enterprise (more) clear. This includes an evaluation of the data variety, availability and quality. In addition, the desired quality of the resulting knowledge graph and its usage by other systems is also taken into account.

5 Evaluation Plan

Our plan is divided into two parts: First, suitability assessments of several state-of-the-art approaches will be conducted. We plan to perform a couple of data-driven evaluations with labelled datasets. This will reveal potential improvements of tested methods on special enterprise data. Various adaptions are planned to be implemented and tested. Proven approaches will be collected in a framework.

Second, our plan is to compare various user interfaces which are designed to gather expert feedback. For that we will conduct several user studies, mainly with university students, but possibly also with domain experts of our different projects. At the same time, we will investigate per use case whether our enrichment results are consistent with the domain experts' conceptual view.

6 Preliminary Results

Regarding the main research question, we already enriched enterprise data in several projects. In the research project PRO-OPT[5], we have gathered first experiences with the construction of a semantic data dictionary in the automotive domain. The insights have been used for a similar industrial project which had the objective to construct a knowledge graph from a given data lake. In another project, we tackled the challenge of enriching several spreadsheets containing manufacturing concepts. Currently, we lift a file tree taken from a shared drive. Regarding the first research question, we tried several unsupervised terminology extraction approaches on tokenized file names, in order to find domain-relevant concepts. However, preliminary findings show that well known approaches like CValue [1] produce insufficient results. Thus, we plan to include certain file system features in the terminology extraction process. Similarly, for our special requirements, we have already enhanced named entity recognition to be tolerant of inflections and for real-time applications [19]. This approach uses language information together with ontologies to also recognize word variations induced by inflection. Our evaluation on Wikipedia shows that we recall considerably more named entities than the baselines.

Regarding integrating human experts (RQ 2), some demos have been published that motivate how domain experts can add or interact with enriched data. It has been shown that spreadsheets enable various kinds of users to easily create semantic data [32,31]. Together with the concept of deep linking [30] they can also semantically annotate semi-structured file contents. For example, via browsing a user obtains a deep link referring to a presentation slide's title and uses the RDF spreadsheet editor to easily make statements about it, like <*this title*, is about, Semantic Web>. We also demonstrated how to intuitively query a semantic graph without knowing about SPARQL [29]. Those approaches are intended to allow users query and enter semantic data in a more familiar way. Yet, these tools are not designed to collect user feedback on the system's enrichment decisions.

7 Conclusion

This paper proposes a PhD topic to investigate an efficient methodology for generating high-level semantic enrichments from undocumented enterprise data. We separated the main research into four partial questions: (1) the utilization of suitable state-of-the-art approaches, (2) the integration of human experts, (3) the quantification of enrichment levels and (4) the efficiency assessment of gathering and exploiting user feedback. Preliminary results show a wide range of potential applications and further research directions. This will impact how enterprise data is transferred into a usable form (again) using semantic technologies. In addition, it will raise awareness of the important involvement of humans during the enrichment process.

[5] `http://www.pro-opt.org/`

Acknowledgements Parts of this work have been funded by the German Federal Ministry of Economic Affairs and Energy in the project PRO-OPT (01MD15004D) and by the German Federal Ministry of Food and Agriculture in the project SDSD (2815708615). I thank my doctoral supervisor Prof. Dr. Andreas Dengel and my colleagues Christian Jilek, Dr. Heiko Maus, Dr. Sven Schwarz, Dr. Jörn Hees and Dr. Ansgar Bernardi for their helpful discussions, comments and feedback.

References

1. Ananiadou, S.: A methodology for automatic term recognition. In: COLING 1994 Vol. 2: The 15th Int'l Conf. on Computational Linguistics. pp. 1034–1038 (1994)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5), 34–43 (2001)
3. Bouquet, P., Serafini, L., Zanobini, S., Sceffer, S.: Bootstrapping semantics on the web: Meaning elicitation from schemas. pp. 505–512. WWW '06 (2006)
4. Brackenbury, W., Liu, R., Mondal, M., Elmore, A.J., Ur, B., Chard, K., Franklin, M.J.: Draining the data swamp: A similarity-based approach. In: Proc. of the Workshop on Human-In-the-Loop Data Analytics. HILDA'18, ACM (2018)
5. Chortaras, A., Stamou, G.: D2RML: Integrating heterogeneous data and web services into custom RDF graphs. In: Proc. of the LDOW. vol. 2073. CEUR (2018)
6. Clarke, M., Harley, P.: How smart is your content? Using semantic enrichment to improve your user experience and your bottom line. Science Editor **37**(2) (2014)
7. Clarkson, K., Gentile, A.L., Gruhl, D., Ristoski, P., Terdiman, J., Welch, S.: User-centric ontology population. In: The Semantic Web: 15th Int'l Conf., ESWC 2018. pp. 112–127. Springer (2018)
8. Culotta, A., McCallum, A.: Reducing labeling effort for structured prediction tasks. In: AAAI. vol. 5, pp. 746–751 (2005)
9. Enslen, E., Hill, E., Pollock, L., Vijay-Shanker, K.: Mining source code to automatically split identifiers for software analysis. 2009 6th IEEE Int'l Working Conf. on Mining Software Repositories pp. 71–80 (2009)
10. Figure Eight Inc.: Data scientist report 2018 (2018), `https://www.figure-eight.com/figure-eight-2018-data-scientist-report/`, accessed: Feb. 1st, 2019
11. Galkin, M., Auer, S., Scerri, S.: Enterprise knowledge graphs : A backbone of linked enterprise data. In: 2016 IEEE/WIC/ACM Int'l Conf. on Web Intelligence (2016)
12. Galkin, M., Auer, S., Vidal, M.E., Scerri, S.: Enterprise knowledge graphs: A semantic approach for knowledge management in the next generation of enterprise information systems. In: Proc. of the 19th Int'l Conf. on Enterprise Information Systems (ICEIS). vol. 2, pp. 88–98. SciTePress (2017)
13. Hai, R., Geisler, S., Quix, C.: Constance: An intelligent data lake system. In: Proc. of the 2016 ACM SIGMOD Int'l Conf. on Management of Data. ACM (2016)
14. Halevy, A.Y., Franklin, M.J., Maier, D.: From databases to dataspaces: A new abstraction for information management. ACM Sigmod Record **34**, 27–33 (2005)
15. Hitzler, P., Krotzsch, M., Rudolph, S.: Foundations of semantic web technologies. Chapman and Hall/CRC (2009)
16. Hlomani, H., Stacey, D.: Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. Semantic Web Journal **1**(5), 1–11 (2014)
17. Hua, W., Wang, Z., Wang, H., Zheng, K., Zhou, X.: Short text understanding through lexical-semantic analysis. In: 2015 IEEE 31st Int'l Conf. on Data Engineering. pp. 495–506 (2015)

18. Jeffery, S.R., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go user feedback for dataspace systems. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. pp. 847–860 (2008)
19. Jilek, C., Schröder, M., Novik, R., Schwarz, S., Maus, H., Dengel, A.: Inflection-tolerant ontology-based named entity recognition for real-time applications. In: 2nd Conf. on Language, Data and Knowledge. vol. 70. OASIcs (2019), (in print)
20. Khine, P.P., Wang, Z.S.: Data lake: a new ideology in big data era. ITM Web Conf. **17**, 03025 (2018)
21. Kristjansson, T., Culotta, A.: Interactive information extraction with constrained conditional random fields. In: AAAI. vol. 4, pp. 412–418 (2004)
22. Li, H., Zhai, J.: Constructing investment open data of chinese listed companies based on linked data. In: Proc. of the 17th Int'l Digital Government Research Conf. on Digital Government Research. pp. 475–480. ACM (2016)
23. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web (Preprint), 1–81 (2018)
24. Maus, H., Schwarz, S., Dengel, A.: Weaving personal knowledge spaces into office applications. In: Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives, pp. 71–82. Springer (2013)
25. Olson, J.R., Rueter, H.H.: Extracting expertise from experts: Methods for knowledge acquisition. Expert systems **4**(3), 152–168 (1987)
26. Pan, J.Z., Vetere, G., Gomez-Perez, J.M., Wu, H.: Exploiting linked data and knowledge graphs in large organisations. Springer (2017)
27. Pham, M., Alse, S., Knoblock, C.A., Szekely, P.: Semantic labeling: A domain-independent approach. In: The Semantic Web – ISWC 2016. Springer (2016)
28. Rao, S.S., Nayak, A.: LinkED: A novel methodology for publishing linked enterprise data. Journal of computing and information technology **25**(3), 191–209 (2017)
29. Schröder, M., Hees, J., Bernardi, A., Ewert, D., Klotz, P., Stadtmüller, S.: Simplified SPARQL REST API: CRUD on JSON object graphs via URI paths. In: The Semantic Web: ESWC 2018 Satellite Events. pp. 40–45. Springer (2018)
30. Schröder, M., Jilek, C., Dengel, A.: Deep linking desktop resources. In: The Semantic Web: ESWC 2018 Satellite Events. pp. 202–207 (2018)
31. Schröder, M., Jilek, C., Hees, J., Hertling, S., Dengel, A.: RDF spreadsheet editor: Get (g)rid of your RDF data entry problems. In: ISWC 2017 Posters & Demonstrations and Industry Tracks. vol. 1963. CEUR (2017)
32. Schröder, M., Jilek, C., Hees, J., Hertling, S., Dengel, A.: An easy & collaborative RDF data entry method using the spreadsheet metaphor. arXiv 1804.04175 (2018)
33. Skluzacek, T.J., Kumar, R., Chard, R., Harrison, G., Beckman, P., Chard, K., Foster, I.: Skluma: An extensible metadata extraction pipeline for disorganized data. 2018 IEEE 14th Int'l Conf. on e-Science pp. 256–266 (2018)
34. Studer, R., Benjamins, V.R., Fensel, D., et al.: Knowledge engineering: principles and methods. Data and knowledge engineering **25**(1), 161–198 (1998)
35. Terrizzano, I., Schwarz, P., Roth, M., Colino, J.E.: Data wrangling: The challenging journey from the wild to the lake. 7th Biennial Conf. on Innovative Data Systems Research (CIDR'15) (2015)
36. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Accelerating the annotation of sparse named entities by dynamic sentence selection. In: BMC Bioinformatics (2008)
37. W3C: RDF 1.1 concepts and abstract syntax (2014)