SEKI@home, a Generic Approach for Crowdsourcing Knowledge Extraction from Arbitrary Web Pages

(Semantic Web Challenge 2012—Open Track Submission)

Thomas Steiner[1] and Stefan Mirea[2]

[1] Universitat Politècnica de Catalunya – Department LSI, Barcelona, Spain
tsteiner@lsi.upc.edu
[2] Computer Science, Jacobs University Bremen, Germany
s.mirea@jacobs-university.de

Abstract. With *SEKI@home*, which stands for *Search for Embedded Knowledge Items*, we propose a generic, browser extension-based approach for crowdsourcing the task of knowledge extraction from arbitrary Web pages. As people with the extension installed browse a targeted Web page, the extension sends extracted knowledge items according to the customizable extraction rules to a centralized, optionally publicly accessible triple store. Thereby, simply by browsing the Web as usual, participants in the knowledge extraction task can help make previously locked-in knowledge openly accessible, *e.g.*, via the standard SPARQL protocol. We have implemented and made available a prototype browser extension, which, after customization and adaptation, can serve as the basis for future knowledge extraction tasks.

1 Introduction

The term *crowdsourcing* was first coined by Jeff Howe in an article in the magazine Wired [2]. It is a *portmanteau* of "crowd" and "outsourcing". Howe writes: *"The new pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R&D"*. The difference to outsourcing is that the crowd is undefined by design. We suggest crowdsourcing for the described task of extracting knowledge from arbitrary Web pages for two reasons: *(i)* the entirety of the search space, *i.e.*, the complete set of all targeted Web pages, is often not known beforehand, *e.g.*, in the case of a video-sharing portal, where there is no list of all available videos; and *(ii)* even if there was such a list, it would not be practicable (nor typically allowed by the terms and conditions of the website owners) to crawl it.

In this paper, we describe and provide a prototype implementation of an approach titled *SEKI@home* and based on crowdsourcing via a browser extension, to make closed knowledge bases programmatically and openly accessible. A prototype browser extension, licensed under the *Apache 2.0* license, is available at `https://github.com/tomayac/seki-at-home/tree/master/extension`.

The proposed approach can be tested by installing the Chrome extension available at `http://bit.ly/SEKIatHome`, clicking the extension icon reveals latest extraction statistics (for legal reasons, not all data is accessible at the moment).

2 Related Work

Wrappers around Web services or Web pages have been used in the past to lift data from the original source to a meaningful, machine-readable RDF level. Examples are the Google Art wrapper by Guéret [1], which lifts the data from the Google Art project [5], or the now discontinued SlideShare wrapper[1] by the same author. Such wrappers typically work by mimicking the URI scheme of the site they are wrapping. Adapting parts of the URL of the original resource to that of the wrapper provides access to the desired data. Wrappers do not offer SPARQL endpoints, as their data gets computed on-the-fly.

With *SEKI@home*, we offer a related, however, still different in the detail, approach to lift and make machine-readably accessible knowledge from arbitrary Web pages. Via crowdsourcing we can distribute the heavy burden of crawling the whole search space on many shoulders. Finally, by storing the extracted knowledge items centrally in a triple store, our approach allows for openly accessing the data via the standard SPARQL protocol.

3 Methodology

Browser Extensions: We have implemented our prototype browser extension for the Google Chrome browser. Chrome extensions are small software programs that users can install to enrich their browsing experience. Via so-called *content scripts*, extensions can inject and modify the contents of Web pages.

Web Scraping: Web scraping is a technique to extract data from Web pages. This happens typically via CSS selectors [3] that allow for directly addressing elements in the DOM (Document Object Model) tree. An exemplary query selector is `.description` (all elements with class name "description"), which, via the JavaScript command `document.querySelector` returns the description of an item on an imaginary Web page.

Lifting the Extracted Knowledge Items: Web pages typically get generated based on structured databases. However, oftentimes this initial structure gets lost when the database contents get transformed into HTML. In order to make extracted knowledge items meaningful again, we need to lift it. We propose JSON-LD [6], a JSON representation format for expressing directed graphs; mixing both Linked Data and non-Linked Data in a single document. JSON-LD allows for adding meaning by simply including or referencing a so-called (data) context. The syntax is designed to not disturb already deployed systems running on JSON, but to provide a smooth upgrade path from JSON to JSON-LD. Following up on the example from before, we could define `exKnowledgeExtraction:Desc`, which might map to `dbpprop:shortDescription` from DBpedia.

[1] `http://linkeddata.few.vu.nl/slideshare/`

Maintaining Provenance Data: The facts extracted via the *SEKI@home* approach are derived from existing third-party knowledge bases. A derivation is a transformation of an entity into another, a construction of an entity into another, or an update of an entity, resulting in a new one. In consequence, it is considered good form to acknowledge the original data source, which we do by default via the property `prov:wasDerivedFrom` from the PROV Ontology [4] for each extracted knowledge item.

4 Evaluation

Our approach *SEKI@home* was successfully evaluated in [8]. The present paper can be seen as a generalized, abstract version of the referenced paper. In summary, the approach turned out to be *efficient* in the sense that knowledge was successfully extracted, *unobtrusive* in the sense that the extension worked unnoticed in the background, and *scalable* in the sense that the extension worked fine with people using it massively in parallel.

5 Future Work

In the future, we want to apply *SEKI@home* to videos from video-sharing portals like YouTube or Vimeo, which can be semantically enriched, as we have shown in [7] for the case of YouTube. We plan to apply *SEKI@home* to semantic video enrichment by splitting the computational heavy annotation task, and store the extracted facts centrally in a triple store to allow for open SPARQL access. In [9], we have proposed the creation of a comments archive of things people said about real-world entities on social networks like Twitter, Facebook, and Google+, which we also plan to realize via *SEKI@home*.

6 Conclusion

In this paper, we have shown a generalizable approach to first open up closed knowledge bases by means of crowdsourcing, and then make the extracted facts universally and openly accessible. The extracted facts can be accessed via the standard SPARQL protocol. Granted that provenance of the extracted data is handled appropriately, we hope to have contributed a useful socially enabled chain link to the Linked Data world.

Acknowledgments

T. Steiner is partially supported by the EC under Grant No. 248296 FP7 (I-SEARCH).

References

1. C. Guéret. “GoogleArt — Semantic Data Wrapper (Technical Update)”, SemanticWeb.com, Mar. 2011. `http://semanticweb.com/googleart-semantic-data-wrapper-technical-update_b18726`.
2. J. Howe. The Rise of Crowdsourcing. *Wired*, 14(6), June 2006. `http://www.wired.com/wired/archive/14.06/crowds.html`.
3. L. Hunt and A. van Kesteren. Selectors API Level 1. Candidate Recommendation, W3C, June 2012. `http://www.w3.org/TR/selectors-api/`.
4. T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology. Working Draft, W3C, July 2012. `http://www.w3.org/TR/prov-o/`.
5. A. Sood. “Explore museums and great works of art in the Google Art Project”, Google Blog, Feb. 2011. `http://googleblog.blogspot.com/2011/02/explore-museums-and-great-works-of-art.html`.
6. M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and M. Birbeck. JSON-LD Syntax 1.0, A Context-based JSON Serialization for Linking Data. Working Draft, W3C, July 2012. `http://www.w3.org/TR/json-ld-syntax/`.
7. T. Steiner. SemWebVid – Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois. In *Proceedings of the ISWC 2010 Posters & Demonstrations Track*, Nov. 2010.
8. T. Steiner and S. Mirea. SEKI@home, or Crowdsourcing an Open Knowledge Graph. In *Proceedings of the First International Workshop on Knowledge Extraction and Consolidation from Social Media (KECSM2012)*, Boston, USA, Nov. 2012.
9. T. Steiner, R. Verborgh, R. Troncy, J. Gabarro, and R. V. de Walle. Adding Realtime Coverage to the Google Knowledge Graph. In *Proceedings of the ISWC 2012 Posters & Demonstrations Track*, Nov. 2012.

Annex: Challenge Criteria—Open Track Submission

Minimal Requirements

The proposed approach is available to end-users in the form of a freely installable browser extension, both in binary form for an exemplary knowledge extraction task, and in form of source code for custom experiments. People can decide to donate their computers' idle time to a knowledge extraction task of their choice, or even create their own knowledge extraction tasks. By design, data plays a central role in the proposed approach and the information sources used are under diverse ownership or control and highly heterogeneous (syntactically, structurally, and semantically), and during our evaluation of the approach [8], it has been shown that substantial quantities of real world data could be processed. We propose to represent meaning via Semantic Web technologies, namely by using JSON-LD [6]. The approach extracts unstructured data from arbitrary Web pages, lifts them to a semantically meaningful level, and thereby enables previously locked-in knowledge items to be queried using standard Semantic Web technologies such as the SPARQL protocol.

Additional Desirable Features

The application provides an attractive and functional Web interface (for human users) in the form of the popup window in the browser extension. This user interface makes the actual knowledge extraction process more graspable in the form of accumulated statistics (see the screenshot available at `http://bit.ly/SEKIatHome`). We have shown the scalability of the approach during our evaluation described in [8]. While the approach does not use all data that is currently published on the Semantic Web, in contrary, it adds even more data to the Semantic Web from previously locked-in Web pages. As detailed in [8], rigorous evaluations have taken place that demonstrate the benefits of semantic technologies and validate the results obtained. Our approach is one of the few browser extensions that uses semantic technologies, and that, apart from information retrieval via SPARQL, also copes with knowledge extraction and semantic lifting. Knowledge extraction from arbitrary Web pages can have commercial value, especially in the fields of opinion mining, or even accessibility, with the described future work project of semantically enriching videos. While the approach is currently implemented for the Google Chrome browser, through open-sourcing the source code of the extension at `https://github.com/tomayac/seki-at-home/tree/master/extension`, we open the door widely for the code to be ported over to other browser platforms like Firefox, Internet Explorer, Safari, and Opera. For the case of Firefox for Android, there is even the possibility to port the code to mobile versions of the Firefox browser[2]. We believe in the huge potential of the *SEKI@home* approach, as all we do is stand on the shoulders of giants[3].

[2] `https://developer.mozilla.org/en-US/docs/Extensions/Firefox_on_Android`

[3] `http://setiathome.berkeley.edu/`