SEKI@home, or Crowdsourcing an Open Knowledge Graph

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. In May 2012, the Web search engine Google has introduced the so-called Knowledge Graph, a graph that understands real-world entities and their relationships to one another. It currently contains more than 500 million objects, as well as more than 3.5 billion facts about and relationships between these different objects. Soon after its announcement, people started to ask for a programmatic method to access the data in the Knowledge Graph, however, as of today, Google does not provide one. With *SEKI@home*, which stands for *Search for Embedded Knowledge Items*, we propose a browser extension-based approach to crowdsource the task of populating a data store to build an Open Knowledge Graph. As people with the extension installed search on Google.com, the extension sends extracted anonymous Knowledge Graph facts from Search Engine Results Pages (SERPs) to a centralized, publicly accessible triple store, and thus over time creates a SPARQL-queryable Open Knowledge Graph. We have implemented and made available a prototype browser extension tailored to the Google Knowledge Graph, however, note that the concept of *SEKI@home* is generalizable for other knowledge bases.

1 Introduction

1.1 The Google Knowledge Graph

With the introduction of the Knowledge Graph, the search engine Google has made a significant paradigm shift towards *"things, not strings"* [7], as a post on the official Google blog states. Entities covered by the Knowledge Graph include landmarks, celebrities, cities, sports teams, buildings, movies, celestial objects, works of art, and more. The Knowledge Graph enhances Google search in three main ways: by disambiguation of search queries, by search log-based summarization of key facts, and by explorative search suggestions. This triggered demand for a method to access the facts stored in the Knowledge Graph programmatically [6]. At time of writing, however, no such programmatic method is available.

⋆ Full disclosure: T. Steiner is also a Google employee, S. Mirea a Google intern.

1.2 On Crowdsourcing

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 facts from SERPs with Knowledge Graph results for two reasons: *(i)* there is no publicly available list of the 500 million objects [7] in the Knowledge Graph, and *(ii)* even if there was such a list, it would not be practicable (nor allowed by the terms and conditions of Google) to crawl it.

1.3 Search Results as Social Media

Kaplan and Haenlein have defined social media as *"a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content"* [4]. We argue that search results are social media as well, especially in the case of Google with its tight integration of Google+, a feature called *Search plus Your World* [8].

1.4 Contributions and Paper Structure

In this position paper, we describe and provide a prototype implementation of an approach, tentatively titled *SEKI@home* and based on crowdsourcing via a browser extension, to make closed knowledge bases programmatically and openly accessible. We demonstrate its applicability with the Google Knowledge Graph. The extension can be added to the Google Chrome browser by navigating to `http://bit.ly/SEKIatHome`, the Open Knowledge Graph SPARQL endpoint can be tested at `http://openknowledgegraph.org/sparql`[1].

The remainder of this paper is structured as follows. In Section 2, we highlight related work for the field of extracting data from websites with RDF wrappers. In Section 3, we describe the *SEKI@home* approach in detail. We provide a short evaluation in Section 4. The paper ends with an outlook on future work in Section 5 and a conclusion in Section 6.

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 [9], or the now discontinued SlideShare wrapper[2] 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.

[1] The SPARQL endpoint and the extension were active from Aug. 11 to Sep. 6, 2012.
[2] `http://linkeddata.few.vu.nl/slideshare/`

With *SEKI@home*, we offer a related, however, still different in the detail, approach to lift and make machine-readably accessible closed knowledge bases like the Knowledge Graph. The entirety of the knowledge base being unknown, via crowdsourcing we can distribute the heavy burden of crawling the whole Knowledge Graph on many shoulders. Finally, by storing the extracted facts centrally in a triple store, our approach allows for openly accessing the data via the standard SPARQL protocol.

3 Methodology

3.1 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. We have implemented an extension that gets activated when a user uses Google to search the Web.

3.2 Web Scraping

Web scraping is a technique to extract data from Web pages. We use CSS selectors [3] to retrieve page content from SERPs that have an associated real-world entity in the Knowledge Graph. An exemplary query selector is `.kno-desc` (all elements with class name "kno-desc"), which via the JavaScript command `document.querySelector` returns the description of a Knowledge Graph entity.

3.3 Lifting the Extracted Knowledge Graph Data

Albeit the claim of the Knowledge Graph is *"things, not strings"* [7], what gets displayed to search engine users are strings, as can be seen in a screenshot available at `http://twitpic.com/ahqqls/full`. In order to make this data meaningful again, we need to lift it. We use JSON-LD [10], 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.

We have modeled the plaintext Knowledge Graph terms (or predicates) like "Born", "Full name", "Height", "Spouse", etc. in an informal Knowledge Graph ontology under the namespace `okg` (for Open Knowledge Graph) with spaces converted to underscores. This ontology has already been partially mapped to common Linked Data vocabularies. One example is `okg:Description`, which directly maps to `dbpprop:shortDescription` from DBpedia. Similar to the unknown list of objects in the Knowledge Graph (see Subsection 1.2), there is no known list of Knowledge Graph terms, which makes a complete mapping impossible. We have collected roughly 380 Knowledge Graph terms at time of writing, however, mapping them to other Linked Data vocabularies will be a permanent work in progress. As an example, Listing 1 shows the lifted, meaningful JSON-LD as returned by the extension.

```
{
  "@id": "http://openknowledgegraph.org/data/H4sIAAAAA[...]",
  "@context": {
    "Name": "http://xmlns.com/foaf/0.1/name",
    "Topic_Of": {
      "@id": "http://xmlns.com/foaf/0.1/isPrimaryTopicOf",
      "type": "@id"
    },
    "Derived_From": {
      "@id": "http://www.w3.org/ns/prov#wasDerivedFrom",
      "type": "@id"
    },
    "Fact": "http://openknowledegraph.org/ontology/Fact",
    "Query": "http://openknowledegraph.org/ontology/Query",
    "Full_name": "http://xmlns.com/foaf/0.1/givenName",
    "Height": "http://dbpedia.org/ontology/height",
    "Spouse": "http://dbpedia.org/ontology/spouse"
  },
  "Derived_From": "http://www.google.com/insidesearch/↩
      features/search/knowledge.html",
  "Topic_Of": "http://en.wikipedia.org/wiki/Chuck_Norris",
  "Name": "Chuck Norris",
  "Fact": ["Chuck Norris can cut thru a knife w/ butter."],
  "Full_name": ["Carlos Ray Norris"],
  "Height": ["5' 10\""],
  "Spouse": [
    {
      "@id": "http://openknowledgegraph.org/data/H4sIA[...]",
      "Query": "gena o'kelley",
      "Name": "Gena O'Kelley"
    },
    {
      "@id": "http://openknowledgegraph.org/data/H4sIA[...]",
      "Query": "dianne holechek",
      "Name": "Dianne Holechek"
    }
  ]
}
```

Listing 1. Subset of the meaningful JSON-LD from the Chuck Norris Knowledge Graph data. The mapping of the Knowledge Graph terms can be seen in the @context.

3.4 Maintaining Provenance Data

The facts extracted via the *SEKI@home* approach are derived from existing third-party knowledge bases, like the Knowledge Graph. 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 source, *i.e.*, the Knowledge Graph, which we have done via the property `prov:wasDerivedFrom` from the PROV Ontology [5] for each entity.

4 Evaluation

4.1 Ease of Use

At time of writing, we have evaluated the *SEKI@home* approach for the criterium *ease of use* with a number of 15 users with medium to advanced computer and programming skills who had installed a pre-release version of the browser extension and who simply browsed the Google Knowledge Graph by following links, starting from the URL `https://www.google.com/search?q=chuck+norris`, which triggers Knowledge Graph results. One of our design goals when we imagined *SEKI@home* was to make it as unobtrusive as possible. We asked the extension users to install the extension and tell us if they noticed any difference at all when using Google. None of them noticed any difference, while actually in the background the extension was sending back extracted Knowledge Graph facts to the RDF triple store at full pace.

4.2 Data Statistics

On average, the number of 31 triples gets added to the triple store per SERP with Knowledge Graph result. Knowledge Graph results vary in their level of detail. We have calculated an average number of about 5 Knowledge Graph terms (or predicates) per SERP with Knowledge Graph result. While some Knowledge Graph values (or objects) are plaintext strings like the value "Carlos Ray Norris" for `okg:Full_name`, others are references to other Knowledge Graph entities, like a value for `okg:Movies_and_TV_shows`. The relation of reference values to plaintext values is about 1.5, which means the Knowledge Graph is well interconnected.

4.3 Quantitative Evaluation

In its short lifetime from August 11 to September 6, 2012, the extension users have collected exactly 2,850,510 RDF triples. In that period, all in all 39 users had the extension installed in production.

5 Future Work

A concrete next step for the current application of our approach to the Knowledge Graph is to provide a more comprehensive mapping of Knowledge Graph terms to other Linked Data vocabularies, a task whose difficulty was outlined in Subsection 3.3. At time of writing, we have applied the *SEKI@home* approach to a concrete knowledge base, namely the Knowledge Graph. In the future, we want to apply *SEKI@home* to similar closed knowledge bases. Videos from video portals like YouTube or Vimeo can be semantically enriched, as we have shown in [11] 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 [12], 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 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. As an example knowledge base, we have used the Google Knowledge Graph. The extracted facts can be accessed via the standard SPARQL protocol from the Google-independent Open Knowledge Graph website (`http://openknowledgegraph.org/sparql`). Just like knowledge bases evolve over time, the Knowledge Graph in concrete, the facts extracted via the *SEKI@home* approach as well mirror those changes eventually. 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. A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68, Jan. 2010.
5. 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/`.
6. Questioner on Quora.com. "Is there a Google Knowledge Graph API (or another third party API) to get semantic topic suggestions for a text query?", May 2012. `http://bit.ly/Is-there-a-Google-Knowledge-Graph-API`.
7. A. Singhal. "Introducing the Knowledge Graph: things, not strings", Google Blog, May 2012. `http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`.
8. A. Singhal. "Search, plus Your World", Google Blog, Jan. 2012. `http://googleblog.blogspot.com/2012/01/search-plus-your-world.html`.
9. 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`.
10. 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/`.
11. 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.
12. 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*. (accepted for publication).