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

Thomas Steiner[1], Stefan Mirea[2]
[1]Universitat Politècnica de Catalunya, tsteiner@lsi.upc.edu
[2]Jacobs University Bremen, Germany, s.mirea@jacobs-university.de

TL;DR

SEKI@home stands for **S**earch for **E**mbedded **K**nowledge **I**tems. It is a generic, **browser extension-based** approach for **crowdsourcing** the task of **knowledge extraction** from arbitrary Web pages. Simply by browsing the Web, participants in the knowledge extraction task can help make locked-in knowledge openly accessible, *e.g.*, via the standard SPARQL protocol.

1) Why Crowdsourcing

We suggest crowdsourcing for the 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.

(ii) Even if the search space was known, it would not be practicable to crawl it.

2) Background

Web Scraping

Technique to access data from Web pages, *e.g.*, via CSS query selectors [1].

JSON-LD

Semantic lifting of extracted knowledge items with JSON-LD [2], a JSON representation format for expressing directed graphs. JSON-LD allows for adding meaning by including or referencing a data context:

```
{
  "@context":
    "http://ex1.org/context.ld",
  "@id":
    "http://ex2.org/videos/123",
  "name": "Some Example Video",
  "length": "00:12:00.000",
  "...": "..."
}
```

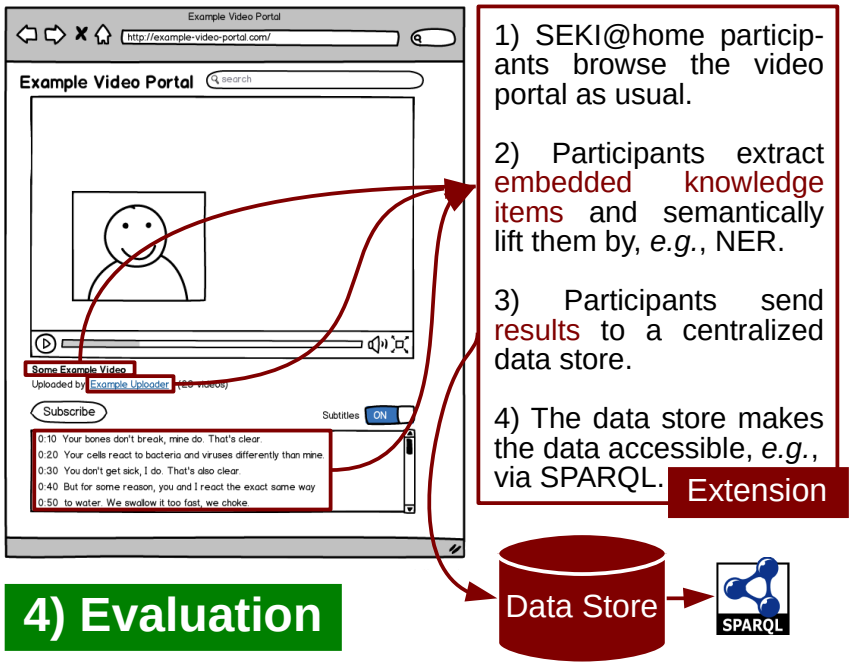
Provenance

For the derivates, give credit to the original data source via `prv:wasDerivedFrom` from the W3C PROV Ontology [3]. Handled transparently by the extension.

SEKI@home extension
It's free and open source:
http://goo.gl/EQiYE

3) Example Use Case

Crowdsource Semantic Video Annotation



1) SEKI@home participants browse the video portal as usual.

2) Participants extract embedded knowledge items and semantically lift them by, *e.g.*, NER.

3) Participants send results to a centralized data store.

4) The data store makes the data accessible, *e.g.*, via SPARQL.

Extension

4) Evaluation

It works! See http://openknowledgegraph.org.

[1] L. Hunt and A. van Kesteren. *Selectors API Level 1*. Candidate Recommendation, W3C.
[2] M. Sporny, D. Longley, et al. *JSON-LD Syntax 1.0, A Context-based JSON Serialization for Linking Data*. Working Draft, W3C.
[3] T. Lebo, S. Sahoo, D. McGuinness, et al. *PROV-O: The PROV Ontology*. Working Draft, W3C.

Partially funded by the EU FP7 I-SEARCH Project Project Reference 248296