Shortipedia

Aggregating and Curating Semantic Web Data

Denny Vrandečić[1,2], Varun Ratnakar[2], Markus Krötzsch[3], Yolanda Gil[2]

[1]Institute AIFB, KIT Karlsruhe Institute of Technology, Karlsruhe, Germany
denny.vrandecic@kit.edu
[2]ISI, USC University of Southern California, Marina del Rey, CA, USA
{varunr|gil}@isi.edu
[3]Computing Laboratory, University of Oxford, Oxford, UK
markus.kroetzsch@comlab.ox.ac.uk

Abstract. Shortipedia is a Web-based knowledge repository, pulling together a growing number of sources in order to provide a comprehensive, diversified view on entities of interest. Contributors to Shortipedia can easily add claims to the knowledge base, provide sources for their claims, and find links to knowledge already available on the Semantic Web.

Keywords: semantic wikis, data integration, semantic infrastructure technology, technology for collaboration

1 Introduction

The goal of Shortipedia is to develop a reference Web service that combines automatic aggregation of facts on the Semantic Web and manual curation by volunteer users. Since different sources can take different views and contain inconsistent information, central to the design of Shortipedia is that any assertion is in principle possibly true and must be included in the repository together with its provenance. Instead of taking assertions as true factual knowledge, we interpret them as assertions about an entity that are true within the context of their provenance record.

Key features of Shortipedia are: it 1) assigns a unique identifier to entities of interest using DBpedia as the basis for identity and the original Wikipedia as the basis for consensus on entities of interest, 2) shores those identifiers in a consensus-built resource, 3) aggregates facts about those entities from the Web of Data and other available data sources, 4) embraces the diversity of views on any given fact and documents their provenance, and 5) supports the curation of its contents by volunteers.

Shortipedia is built on top of a number of Web services and frameworks in order to provide a comprehensive and rewarding workflow for contributors to the project, and thus ultimately to create a reference website for all kind of facts.

Shortipedia is still under development and we will release it publicly before ISWC 2010 and will be able to report about contributor activities at that time. Whereas DBpedia [4] already shows the high potential of providing a nucleus for

the Web of Data, an editable extension of such a nucleus could bring together the advantages of Wikipedia and DBpedia, building on both services and enriching them.

Shortipedia is a knowledge repository formed by:

- a core of entities, based on Wikipedia articles
- mappings of those entities to entities in the Web of Data
- aggregated facts from Web of Data sources
- provenance records of all assertions
- direct contributions and corrections from users

This results in a system that provides

- a knowledge repository containing an aggregated and curated view on Semantic Web data
- a data query and browsing interface, based on Linked Open Data
- a neutral and unbiased reference site that includes alternative points of view

This system description first gives an overview of the architecture and some related approaches. In Section 4 we note some of the design decisions we have made, followed by a summary of lessons learned in Section 5. Section 6 summarizes how the submission meets the requirements of the Semantic Web Challenge call.

Due to space constraints, this system description does not address the question of user motivation and advantages. For that, we refer to [9], describing a very similar system. We also do not describe the workflow or how the system is actually used. For this we refer to the site, available at `http://shortipedia.org`

2 Architecture

Shortipedia is an application of MediaWiki, a highly extensible, open source wiki software [2]. MediaWiki is widely used both on the Web, most famously on Wikipedia [1], as well as in corporate intranets for a huge variety of tasks. In order to support these tasks, more than 1,200 extensions are being provided for download from the MediaWiki homepage. For Shortipedia, we used a number of existing extensions and developed a new one from scratch that added the missing functionality. We will describe both in this section.

The major extension used in Shortipedia is Semantic MediaWiki, which allows to add metadata to wiki pages and to query the collected metadata of the wiki [9]. Thus the wiki can also be regarded as a light-weight, schemaless database editable through the wiki user interface. So instead of having just simple, human-readable pages about persons, cities, companies, music bands, etc., the system allows the community to also add metadata describing these persons in a way that is both accessible to the human reader and the machine. This enables to write tools that can access this collaboratively maintained knowledge base and repurpose it.

Fig. 1. The Shortipedia page for Shanghai. The middle column presents the content on Shanghai from Wikipedia (including the Wikipedia infobox), the right column displays the facts known within the system (top) and entities from the Web of Data that may match to the given site (bottom, with the entity from the New York Times expanded). From `http://shortipedia.org/index.php/Shanghai`

We have not yet fully exploited the rich set of MediaWiki extensions that can further enhance the service, e.g by providing rich query capabilities with systems like Ask the wiki [8] or enabling the shared exchange and responsibility of knowledge with systems like Distributed SMW [11].

Building on this baseline, we already have a system that is known to be scalable with regards to the size of the data and the size of the community it is supposed to handle, and that provides us with all the basic functionalities like user management, rights management, an infrastructure for Restful APIs, complete and rich article history features, and a comprehensive extension framework. For Shortipedia we developed a MediaWiki extension to provide us with the missing features we need for the system. These include:

- the ability to reach out to the Semantic Web and load data about entities that is available as Linked Open Data [3]. The data is then visualized for the end user, and can then be mapped to the wiki-internal vocabulary, and the knowledge can then be copied into the wiki system, retaining the link to the original source as a reference
- a simple interface for editing facts and references, so that no wiki-syntax needs to be learned by the user. In fact, Shortipedia does not allow nor-

mal wiki-editing of the entity pages but only through the dedicated editing interface
– providing a interface where the user can seamlessly change languages

On every wiki page we display the Wikipedia article, if available. Wikipedia and DBpedia [4] have together created the largest resource of community-built consensus on the meaning of Web identifiers. For bootstrapping, we assume the mapping that DBpedia provides for Wikipedia articles, i.e. we assume that the Wikipedia article at `http://en.wikipedia.org/wiki/X` is about the entity referenced by `http://dbpedia.org/resource/X`. With this identifier, we query the sameAs.org service. sameAs.org [7] is a service that, given an URI, provides with further URIs that are considered co-referent. We also query Sindice with the name of the page. Sindice [10] is a Semantic Web search engine that, among other services, provides entities given a search term. The system is designed so that we can incorporate further sources of identity later, e.g. for special fields like proteins or music.

We incorporate the RDF parser ARC2[1] into Shortipedia, so that we can then dereference the URIs suggested by sameAs and Sindice, and provide data about them to the user upon request (applying AJAX via the jQuery library). This requires the data to be available as Linked Open Data. The system then displays the assertions gathered from the Web of Data and allows the user to map the external vocabulary to the Shortipedia vocabulary. Once mapped, the user can integrate a fact into the Shortipedia knowledge base. The system can be used at `http://shortipedia.org`.

3 Related approaches

DBpedia [4] is a project to extract structured information from Wikipedia. It provides the Web of Data with a much needed nucleus of stable, Wikipedia-based URIs. DBpedia uses a multitude of automatic and semi-automatic techniques to extract information from Wikipedia articles. Whereas Shortipedia also takes Wikipedia as its starting point for its entities (and especially for establishing identity), it is not based on an automatic extraction from Wikipedia, but rather on the completely human effort for extracting and curating every single fact. Also, DBpedia does not provide the direct editing of its content but relies on respective changes in the original source, i.e. Wikipedia, or the extraction mechanism. Recently, Ultrapedia[2] has been released to provide a wiki-based interface on top of DBpedia data, including the possibility to easily and selectively edit the source Wikipedia text and thus provide a feedback loop to the original source.

Freebase [5] is a Web-based database that allows to create and edit data entries for any entity of general interest. Whereas close in spirit, Shortipedia refrains from Freebase's approach of types and their schemas for entities, but provides a completely schemaless environment (a trait inherited by the usage of

[1] http://arc.semsol.org/
[2] http://wiking.vulcan.com/up/index.php/Main_Page

Semantic MediaWiki as the underlying implementation), and also puts a strong emphasis on retaining references for every piece of information. Freebase can, due to the usage of schemas, provide a more polished user interface and improved support for data entry.

Not surprisingly, both Freebase and DBpedia are frequent sources suggested for mappings in Shortipedia, and thus data from both efforts are reused inside the project. In the same vein, all data in Shortipedia is available via Semantic Web standards, and thus in return reusable by the other systems. Wherever users improve the knowledge base, all parties win.

Also, using external data in SMW is not a new idea [6]. Instead of allowing the tight integration for querying that many of these other solutions provide, Shortipedia rather aims at integrating the external data inside of Shortipedia, and then to allow to query this integrated data set. Shortipedia extends the capabilities of current SMWs by being able to provide references for facts, and by directly accessing the Web of Data.

4 Design decisions

In this section we describe some key decisions on the design of Shortipedia and our rationales. We acknowledge that these decisions are not imperative, but could have well been taken differently (resulting in a different system).

Wikipedia and DBpedia as identity providers. In order to circumnavigate the core problem of how to establish identity on the Semantic Web, we build up on the work already provided by these two established sources.

Consistency is not required. We deliberately do not require the data to be consistent, but rather just to be referenced well. This is also consistent with the wiki-like idea of trying to accept anything, and then let it be repaired later. We do expect though, that the community will come up with gardening approaches that will check for certain consistency rules, but the system itself does not require or enforce such a consistency (with one exception, see below).

Triples are immutable. We do not allow for facts to be changed. Instead, contributors can add new facts and provide better sources, or they can delete facts (note that we retain a history of all edits to easily repair vandalism).

Don't trust the triples. Every fact can be given a reference, and references can be used to decide which facts to trust. Initially we wanted to add a rating system for triples (e.g. users would be able to agree / disagree / like the fact that Paris is the capital of France, etc.), but we dropped that idea.

sameAs is different. Stating co-reference for two URIs is not regarded as merely another assertion in the system, but rather is completely materialized – meaning that, once an assertion is added to the system it does not keep track of the original URIs used to describe the assertion. Deleting the mapping to the external URI will thus not delete the assertions added while the URI was mapped. We do not expect this to become a problem since 1) facts still need to be added manually even though two URIs have been mapped, and 2) often a specific external source uses a specific URI for an entity, and since the source

information is retained the URI can often be reconstructed. In order to enable this, the system also enforces that the mapping property is inverse functional, i.e. that the same URI cannot be mapped to two different URIs in Shortipedia.

First map, then add. Just adding facts from the Web of Data without linking them to internal entities first is not possible. First the property and the value (if not a literal) have to be mapped to URIs internal to Shortipedia.

The browser does the work. In order to provide a performant and responsive interface we had moved a lot of the processing to the browser. Even though this may lead to some browsers or devices not being able to access the site, we think that such a UI is necessary to engage a broader community. We do plan an alternative, simple browse-only interface in the close future, that cannot be used to edit Shortipedia.

5 Lessons learned

Since Shortipedia has not officially launched yet, our experiences are limited. We will note here the few problems we stumbled upon unexpectedly during the initial development of Shortipedia.

Data is noisy or hard-to-understand. The recent explosive growth of the linked open data cloud gave us hope to find huge amounts of interesting and useful data out there – and indeed, we found huge amounts of data. But many of these are hard to understand and some datasets, especially automatically extracted ones, sport an unfavorable amount of noise. These challenges point to interesting future work.

What's in a name? Many RDF resources on the Web only carry few or distributed labeling annotations, so that it raises the question of how to display those entities to the end user. Whereas entities, when dereferenced, do often have some labeling annotation for themselves, they do not carry such annotations for the other entities mentioned in the RDF document. A possible solution is to dereference all mentioned entities, but this leads to the next issue.

The Semantic Web is slow. Dereferencing many entities, i.e. loading several hundred RDF documents just to create a single page, would be unacceptably slow. We had very soon to start to balance the number of calls and the additional information we expected from the calls, as well as move to more calls initiated by user actions.

6 Appendix

This appendix summarizes how the evaluation requirements are met.

6.1 Minimal requirements

End-user application. The application is an end-user application as it is 1) useful as a reference page, and 2) editable by end-users.

Information sources. The system is able to pull in any information source published as Linked Open Data, to map them, and to integrate and use such information sources for reference. The data is thus under diverse ownership and highly heterogeneous.

Data meaning. The meaning of the data plays a central role for mapping external data sources into the system. The meaning is represented using RDF and mapped via the OWL sameAs, equivalentProperty or equivalentClass property, respectively. The query language in Semantic MediaWiki is based on OWL2 EL.

6.2 Additional desirable features

User interface. The user interface was designed to be simple and to allow easy editing and enriching of the data. We have worked hard in order to provide a clutter-free interface, but that was extremely challenging given the desired functionality.

Scalability. Semantic MediaWiki has been shown to work with millions of pages. We will closely watch how the scalability of Shortipedia turns out, but previous experience with the platform makes us optimistic.

Novelty. The combination of a schemaless Web-based knowledge repository with an explicit reliance on references and a multi-lingual interface is novel.

Functionality. The major functionality of Shortipedia is as a reference work and as a massive community-built data integration platform.

Commercial applicability. We are planning to create a number of MediaWiki extensions from the novel functionality developed for Shortipedia, like integrating LOD data, simple fact editing, etc.

Contextual information. All claims are provided with their sources. Thus consumers of the Shortipedia knowledge-base can decide on the trustworthiness of the sources and restrict the set of claims they take want to use.

Accuracy. All facts in Shortipedia have been added manually and curated by a community. Every single assertion can be explicitly provided with a reference.

Multilinguality. Support for multiple languages is built into the system. The facts are language independent and can be displayed in all languages where the language labels are provided. We aim to support all 200+ languages of the Wikimedia projects.

Range of devices. The system is tested to run on all modern browsers, and was also tested on a number of mobile devices. The editing interface relies heavily on browser-side Javascript execution, and an alternative toned-down interface is currently being planned.

Acknowledgements

Work presented in this paper has been funded by the EU IST FP7 projects RENDER and ACTIVE and the National Science Foundation (NSF) under Grant number IIS-0948429. We acknowledge (and appreciate) the MediaWiki and Semantic MediaWiki community for their continuous support for the SMW platform and being the incubator for many of the ideas presented in this work.

References

1. Phoebe Ayers, Charles Matthews, and Ben Yates. *How Wikipedia works*. No Starch Press, San Francisco, CA, October 2008.
2. Daniel J. Barret. *MediaWiki*. O'Reilly, 2008.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, 2009. Special Issue on Linked Data.
4. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.
5. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, New York, NY, USA, 2008. ACM.
6. Basil Ell. Integration of external data in semantic wikis. Master's thesis, Hochschule Mannheim, 2009.
7. Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *Linked Data on the Web (LDOW2009)*, Madrid, Spain, April 2009.
8. Peter Haase, Daniel M. Herzig, Mark Musen, and Duc Thanh Tran. Semantic wiki search. In *6th Annual European Semantic Web Conference, ESWC2009, Heraklion, Crete, Greece*, volume 5554 of *LNCS*, pages 445–460. Springer Verlag, Juni 2009.
9. Markus Krötzsch, Denny Vrandečić, Max Völkel, Heiko Haller, and Rudi Studer. Semantic wikipedia. *Journal of Web Semantics*, 5:251–261, September 2007.
10. Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovannia Tummarello. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3(1), 2008.
11. Charbel Rahhal, Hala Skaf-Molli, Pascal Molli, and Stéphane Weiss. Multi-synchronous collaborative semantic wikis. In Gottfried Vossen, Darrell D. E. Long, and Jeffrey Xu Yu, editors, *Proceedings of the International Conference on Web Information Systems Engineering (Wise 2009)*, volume 5802 of *LNCS*, Poznan, Poland, October 2009.
12. Denny Vrandečić. Towards automatic content quality checks in semantic wikis. In *Social Semantic Web: Where Web 2.0 Meets Web 3.0*, AAAI Spring Symposium 2009, Stanford, CA, March 2009. Springer.