> Semantic Web Use Cases and Case Studies

Case Study: Semantic tags

Vuk Milicic, Faviki, Serbia

December 2008

General Description

Faviki is a social bookmarking tool that allows users to annotate the contents of web pages by Wikipedia concepts. Using Wikipedia as a source of a universal controlled vocabulary, it provides so-called ‘semantic tags’ which are standardized and computer-interpretable. In this way, Faviki is able to solve some common problems related to classic ‘folksonomy’ tags, in particular: polysemy, synonymy, different lexical forms, and lack of a commonly agreed meaning of terms. In a wider perspective, Faviki aims to speed up the transition from Web 2.0 to the Semantic Web.

The Problem

Tags represent keywords used to describe a web resource. They have been introduced by Del.icio.us to a wider audience in 2003, and soon after they became an extremely popular way of organizing and sharing large collections of data in online social communities. Compared to previous organization methods (categories/folders), tags are simple to add, flexible, and can designate membership in more than one category at the same time. When used collectively they become a powerful categorization tool.

However, tags have some considerable disadvantages. They do not provide information about their meaning, the semantics of the words tags may be composed of. Polysemy (the same word can refer to different concepts), synonymy (the same concept can be pointed out using different words), different lexical forms (different noun forms, different verb conjugation, acronyms, different languages), misspelling errors are some problems that arise when using tags.

For example, the tag ‘orange’ might refer to the fruit or the color, and the different tags ‘movie’ and ‘film’ can be used to describe the same concept. This lack of semantic distinction leads to inappropriate connections between items, making them hard to search and browse through. Tags are essentially a full text search mechanism, not graph linkages that machines can understand. Simply put, the same reason that makes tags difficult to process also makes them so popular - they are just random words.

The Linked Open Data effort provides sets of referencable, semantically interlinked resources with defined meaning. However, motivating people to use these resources for publishing structured data is still a great challenge, as applications that would allow them to do that are still overly complex and not user-friendly.

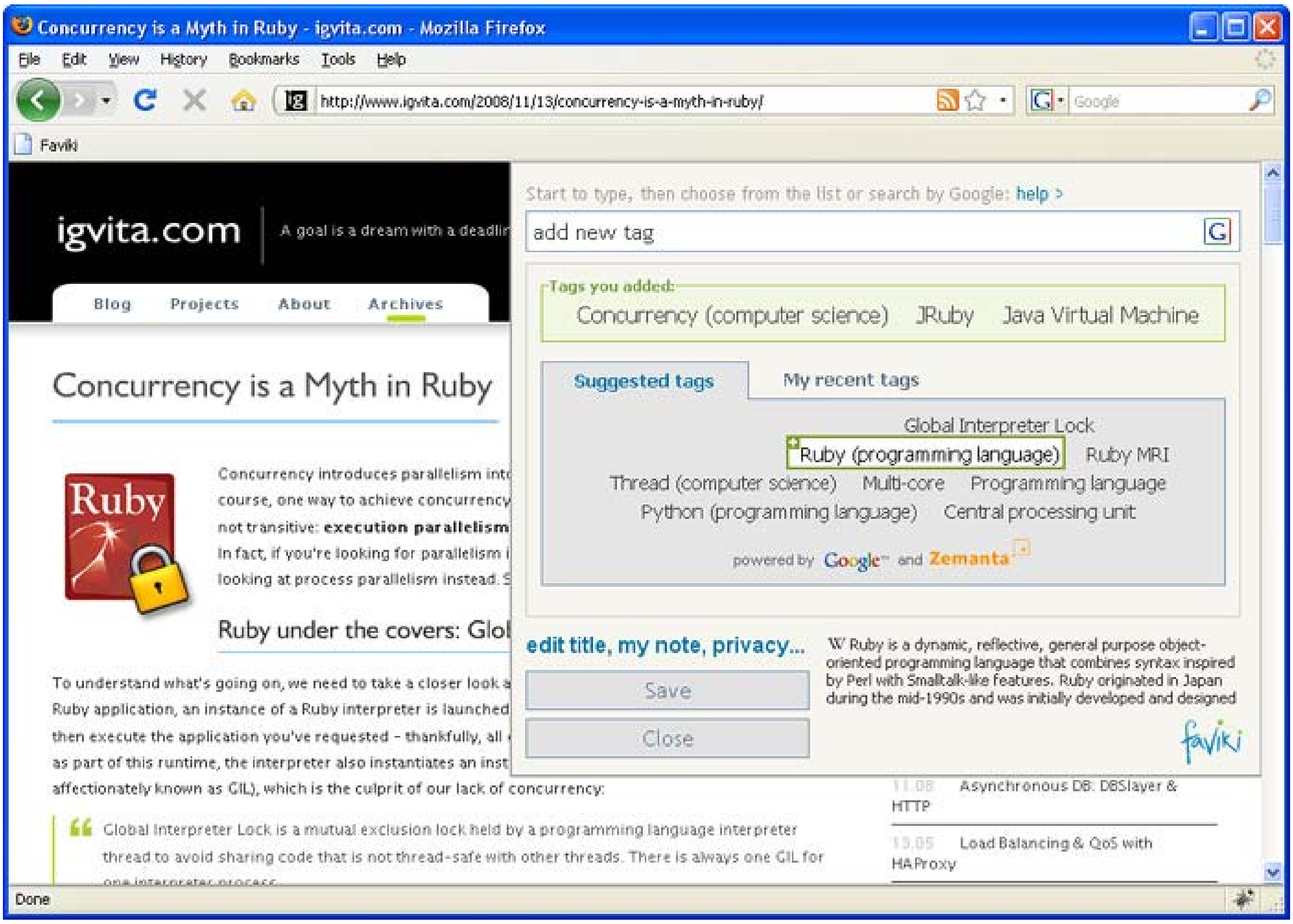
On the one hand there are flexible tags with little structure or semantics widely accepted as a way of organizing information on the Web, and, on the other hand, there exist highly formalized and complex semantic technologies and standards struggling to find their way into the mainstream.

Finally, the problem is — how to add meaning to the tags, or, from the other perspective, how to bring Web 2.0 practices to semantic technologies and make them more flexible and easier to use.

The Application

Faviki is a social bookmarking tool which allows users to use common, predefined tags to annotate web pages. Using the Wikipedia collection of articles as a source of a universal controlled vocabulary, it provides so-called ‘semantic tags’. In this way, web pages saved by users are not just described by random words but connected to uniquely defined concepts.

Bookmarks in Faviki are saved through a simple bookmarklet interface (Figure 1). Here, users can add tags and edit information about bookmarks, such as a title of the web page, users notes, a quote from the web page, or the privacy options.


Concurrency is a Myth in Ruby - igvita.com - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.igvita.com/2008/11/13/concurrency-is-a-myth-in-ruby/
Google
Faviki
igvita.com
A goal is a dream with a deadlin
Blog Projects About Archives
Ruby
Start to type, then choose from the list or search by Google: help >
add new tag
Suggested tags
My recent tags
powered by Google and Zemanta
edit title, my note, privacy...
Save
Close
faviki
11.08 Asynchronous DB: DBSlayer & HTTP
13.05 Load Balancing & QoS with HAProxy
Done

Tags you added: Concurrency (computer science) JRuby Java Virtual Machine

Suggested tags: Global Interpreter Lock, Ruby (programming language), Ruby MRI, Thread (computer science), Multi-core, Programming language, Python (programming language), Central processing unit

Concurrency is a Myth in Ruby

Concurrency introduces parallelism int course, one way to achieve concurrency not transitive: **execution parallelism** In fact, if you're looking for parallelism i looking at process parallelism instead. S

Ruby under the covers: Glol

To understand what's going on, we need to take a closer look a Ruby application, an instance of a Ruby interpreter is launched then execute the application you've requested – thankfully, all as part of this runtime, the interpreter also instantiates an inst affectionately known as GIL), which is the culprit of our lack of concurrency:

> Global Interpreter Lock is a mutual exclusion lock held by a programming language interpreter thread to avoid sharing code that is not thread-safe with other threads. There is always one GIL for one interpreter process.

W Ruby is a dynamic, reflective, general purpose object-oriented programming language that combines syntax inspired by Perl with Smalltalk-like features. Ruby originated in Japan during the mid-1990s and was initially developed and designed

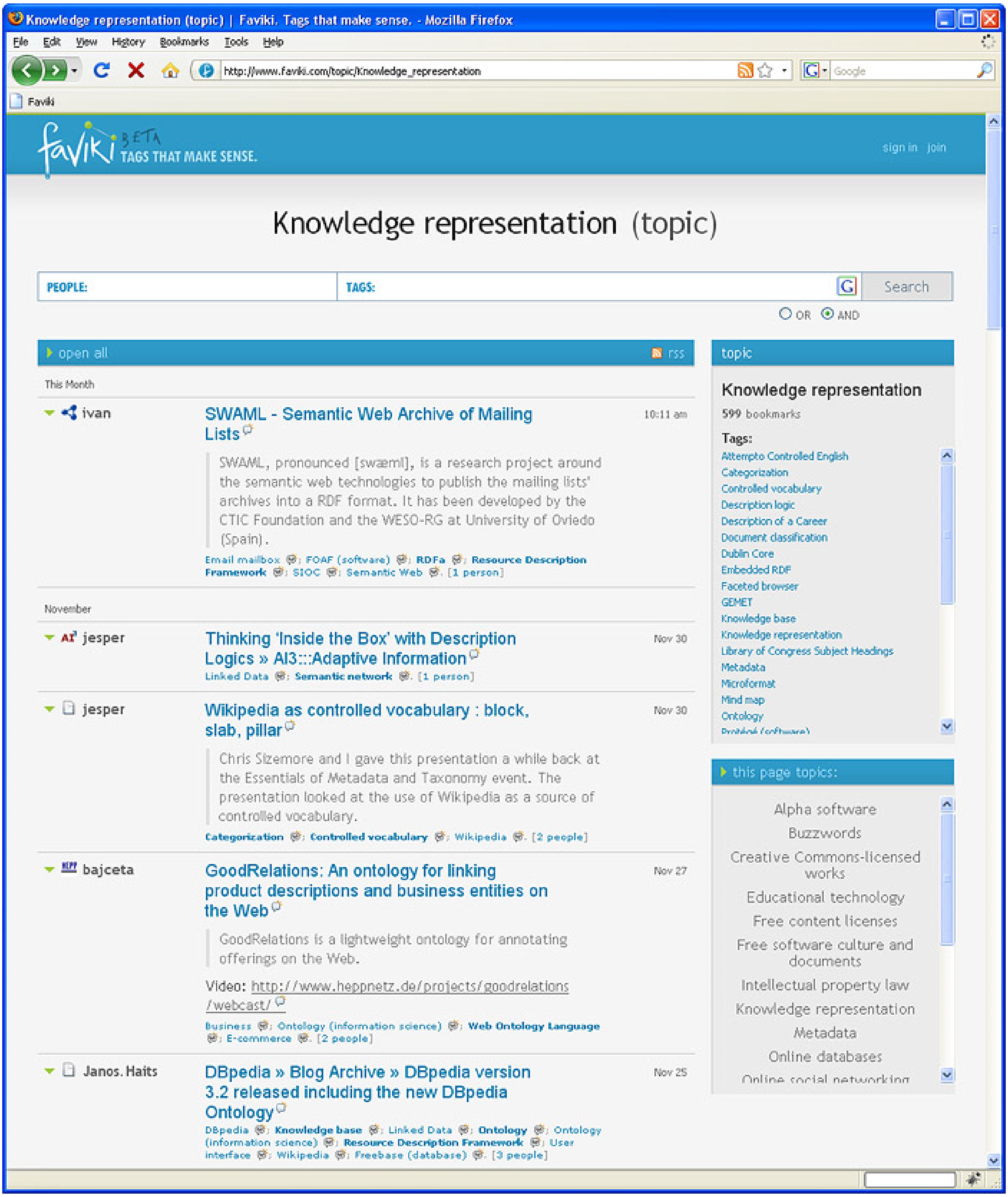
Figure 1: Faviki bookmarklet interface

Adding tags in Faviki is slightly different from classic tagging because semantic tags must correspond to some Wikipedia concept. Faviki allows users to add tags in more than one way, namely:

- by starting to type in the input field and choosing a suggested Wikipedia concept from the autocomplete list.
- by clicking on auto-suggested concepts. Auto-suggestion is accomplished by means of the Zemanta API which analyzes the content of a page and finds the most relevant Wikipedia concepts (Figure 1).
- by searching for Wikipedia concepts by means of an integrated Google search restricted to the domain en.wikipedia.org.
- by browsing through the user's tags entered for the previously bookmarked web pages.

Bookmarks can be browsed in Faviki via tag clouds or the retrieved by search. Moreover, Faviki automatically shows related semantic tags based on previously added bookmarks in the system. Faviki helps users to expand their knowledge by providing more information on tags, such as short abstracts, images, homepage URLs and links to original Wikipedia articles.

Faviki experiments with Wikipedia categories used in Wikipedia for classification of articles. In Faviki, categories allow users to track bookmarks by broader topics. It automatically classifies tags and thus it is able to 'know' that, for instance, 'RDFa', 'Web Ontology Language' and 'Controlled vocabulary' belong to the common category 'Knowledge representation', as shown in Figure 2.


Knowledge representation (topic) | Faviki. Tags that make sense. - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.faviki.com/topic/Knowledge_representation
Google
Faviki
faviki BETA
TAGS THAT MAKE SENSE.
sign in join
PEOPLE:
TAGS:
Search
○ OR ● AND
open all
rss

Knowledge representation (topic)

This Month

ivan — SWAML - Semantic Web Archive of Mailing Lists — 10:11 am

SWAML, pronounced [swæml], is a research project around the semantic web technologies to publish the mailing lists' archives into a RDF format. It has been developed by the CTIC Foundation and the WESO-RG at University of Oviedo (Spain).

Email mailbox ; FOAF (software) ; RDFa ; Resource Description Framework ; SIOC ; Semantic Web . [1 person]

November

jesper — Thinking 'Inside the Box' with Description Logics » AI3:::Adaptive Information — Nov 30

Linked Data ; Semantic network . [1 person]

jesper — Wikipedia as controlled vocabulary : block, slab, pillar — Nov 30

Chris Sizemore and I gave this presentation a while back at the Essentials of Metadata and Taxonomy event. The presentation looked at the use of Wikipedia as a source of controlled vocabulary.

Categorization ; Controlled vocabulary ; Wikipedia . [2 people]

bajceta — GoodRelations: An ontology for linking product descriptions and business entities on the Web — Nov 27

GoodRelations is a lightweight ontology for annotating offerings on the Web.

Video: http://www.heppnetz.de/projects/goodrelations/webcast/

Business ; Ontology (information science) ; Web Ontology Language ; E-commerce . [2 people]

Janos.Haits — DBpedia » Blog Archive » DBpedia version 3.2 released including the new DBpedia Ontology — Nov 25

DBpedia ; Knowledge base ; Linked Data ; Ontology ; Ontology (information science) ; Resource Description Framework ; User interface ; Wikipedia ; Freebase (database) . [3 people]

topic

Knowledge representation

599 bookmarks

Tags:

- Attempto Controlled English
- Categorization
- Controlled vocabulary
- Description logic
- Description of a Career
- Document classification
- Dublin Core
- Embedded RDF
- Faceted browser
- GEMET
- Knowledge base
- Knowledge representation
- Library of Congress Subject Headings
- Metadata
- Microformat
- Mind map
- Ontology

this page topics:

- Alpha software
- Buzzwords
- Creative Commons-licensed works
- Educational technology
- Free content licenses
- Free software culture and documents
- Intellectual property law
- Knowledge representation
- Metadata
- Online databases
- Online social networking

Figure 2: Knowledge representation topic page on Faviki

Faviki exploits another dimension of semantic tags that has not been possible before — multilingual semantic tagging. It allows users to tag in 14 different languages, keeping web resources connected to the main English version and translating tags into the users' languages. This means that, for instance, a resource that is tagged by different users in Japanese, French, and German tags, can be still found by using tags in Russian, because English is used as an universal reference.

By combining several services, namely Zemanta API, Google Language API and DBpedia, Faviki connects non-English contents to English Wikipedia concepts. The whole process from the fetching the core text of a web page to suggesting relevant tags in various languages is described in Figure 3.

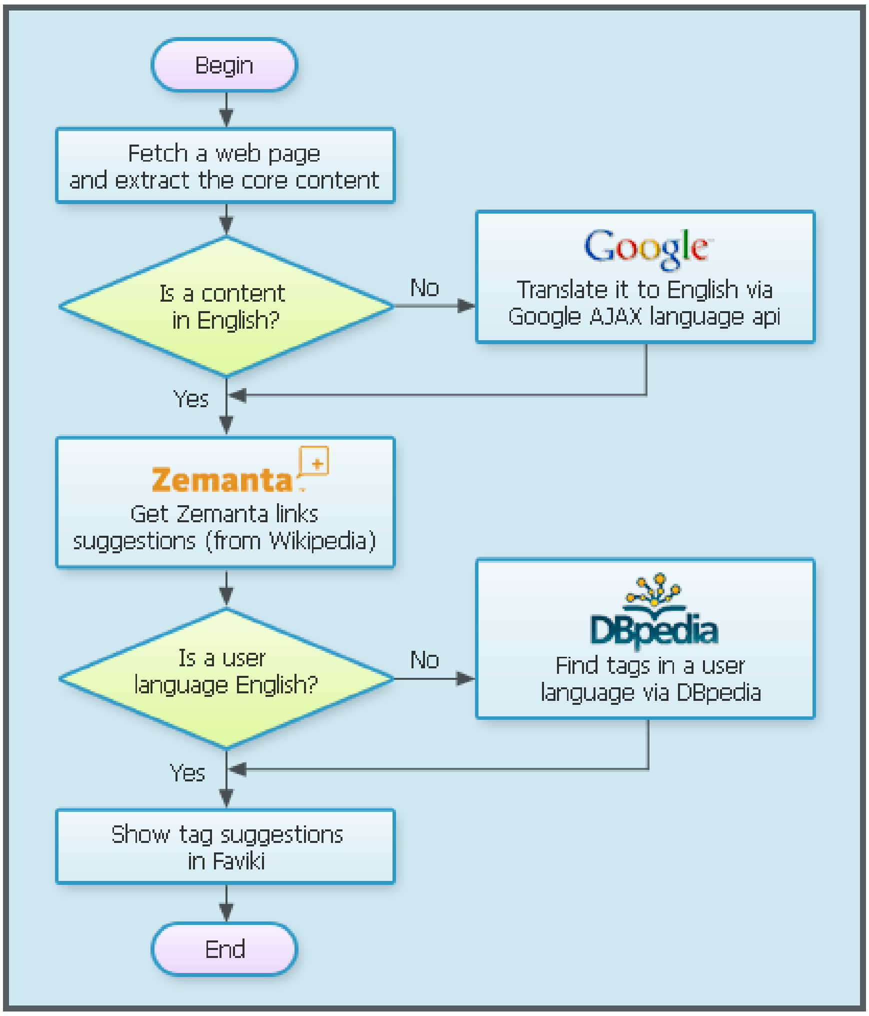

Begin
Fetch a web page and extract the core content
Is a content in English?
No
Google
Translate it to English via Google AJAX language api
Yes
Zemanta
Get Zemanta links suggestions (from Wikipedia)
Is a user language English?
No
DBpedia
Find tags in a user language via DBpedia
Yes
Show tag suggestions in Faviki
End

Figure 3: Suggesting semantic tags using Zemanta, Google Language API and DBpedia

Faviki's features go beyond those of classic social bookmarking tools thanks to the fact that it is based on a consistent system containing the same collection of tags for every user. The improved search and organization of data makes Faviki significantly more effective, especially in the long run.

The Solution

In order to be computer-interpretable, tags have to be unique, standardized, and their names have to be agreed upon. Moreover, the set of tags has to be huge. Is there a source of such tags? Fortunately, there is. The tremendous job of creating definitions of a huge number of terms has already been done and the result can be found in the 'central knowledge source of mankind' — Wikipedia.

Besides displaying articles in a standardized (semi-structured) way, Wikipedia also has a standardized way of uniquely naming titles and the URLs of associated articles which have been created and are constantly perfected by thousands of contributors.

Polysemy is solved in Wikipedia by adding additional information directly into names. 'Orange (colour)' and 'Orange (fruit)' both have a clear meaning now. In addition, take an often used example of a word with plenty of meanings — 'java'. The article 'Java' represents an island in Indonesia, 'Java (programming language)' is a popular programming language, 'Java coffee' is a coffee produced on the island of the same name, 'Java (cigarette)' refers to a brand of Russian cigarettes, etc. Note that an article with the name consisting solely of the term in question (in this case - 'Java') tends to be the most representative one, if there is any.

Synonymy, on the other hand, is solved in Wikipedia by the system of redirects. Try to type 'http://en.wikipedia.org/wiki/Movie' into the address bar. You will be automatically redirected to the article named 'Film'.

Thanks to DBpedia, it is possible to use the information from Wikipedia in an elegant way. DBpedia is a project that extracts structured information from Wikipedia and makes it accessible on the Web in the form of RDF triples, under the GNU Free Documentation License. As Wikipedia contains articles about many general-purpose concepts, DBpedia can also be seen as a huge ontology that assigns URIs to a large number of concepts. This knowledge base can serve as the universal controlled vocabulary we are looking for.

Put into context, tags referring to DBpedia are not just words anymore, they act as objects with properties specified further by literals or typed links to other objects. In DBpedia there are some properties common to all tags, such as: an abstract, a picture, labels in multiple languages, and several properties dealing with classification.

For example, if we look at the DBpedia page for ‘Keith Richards’, we can learn some additional properties about him such as his year of birth, his type of voice, the genre of music he plays, as well as his connections to other tags, such as that he is: born in ‘Dartford’, a current member of the band ‘The Rolling Stones’, plays ‘Fender Telecaster’ and ‘Gibson Les Paul’, and has occupations of ‘Music producer’, ‘Musician’, and ‘Songwriter’.

Of course, the creation of a rich ontology out of Wikipedia data defining all these properties in a consistent fashion is a great challenge. The current state is not perfect, but this data is much more useful compared to classic non-semantic tags. DBpedia tags have great potential for supporting information integration and for enhancing the ‘intelligence’ of the Web.

Faviki currently uses 5.6 million DBpedia concepts — 2.7 million English titles and 2.9 million titles from other 13 languages. Some of DBpedia datasets that are used frequently, such as titles and links to categories, are imported into Faviki. Other data, like additional information about concepts (multilanguage descripitons, images, homepages) is accessed via simple SPARQL queries. To ensure that Faviki is up-to-date with Wikipedia, it is periodically synchronized with new DBpedia releases. Moreover, It uses different APIs for finding Wikipedia articles in order to help users find most appropriate and most recent concepts.

Key Benefits of Semantic Technology

- enabling a tag standardization by adding a meaning;
- connecting Web 2.0 and the Semantic Web with semantic tags;
- connecting Wikipedia and social bookmarking;
- re-using RDF data available on the Web;
- facilitating information integration and knowledge discovery
- facilitating communication between different services;
- enabling auto-classification and multilingual semantic tagging;

Conclusion

Web 2.0 showed that it is possible to have successful systems based on decentralized creation and collaboration of big online communities. Despite all its disadvantages, tagging emerged as a (good enough) way to integrate and organize the data. Semantic tags, as an intersection point of the two worlds, have the potential to enable much faster evolution of the Web by providing a solid foundation from which the Semantic Web can grow soundly.

We believe that Faviki’s approach demonstrates a number of benefits compared to classic social bookmarking systems, thanks to the fact that semantic tags are just parts of a much more powerful system. By providing consistency, additional useful information about tags, auto-organization using Wikipedia categories, and the ability to ‘understand’ tags in different languages, it just scratches the surface of the possibilities that continuously emerge.