A topic hierarchy on the web*

Valentin Zacharias
ontoprise© GmbH
Amalienbadstr. 36 (Raumfabrik 29)
76227 Karlsruhe
zacharias@ontoprise.de

Abstract

We present the architecture and interface of a metadata registry that is very simple to integrate by content and application providers. It takes its inspiration from currently successful metadata architectures and aims to be an evolutionary change to the web – using long established standards where possible.

1 Motivation

Unitracc[1] is a internet based e-learning system for the area of canalization. It also contains tools that help the public authorities manage and monitor underground infrastructure. The system already contains a large number of information units, especially digital versions of the two standard textbooks about canalization. The information stored in the system is annotated using topics organized in a topic hierarchy.
Currently we are trying to enrich the system with third party content, in the beginning we expect this to be mostly information about tools and machines supplied by the vendors. Third party content can either be directly entered into the unitracc system or found by a crawler from sites known to have information about canalization. In both cases we would like people to adopt our topic hierarchy, making it easier to integrate the information into the system. In order to archive this goal we need to make the metadata registry as open and easy to use as possible. In this context it is unlikely that content providers will download an OWL or FLogic Ontology to learn about the topics and their relations. It is also unlikely that they will be willing to spend large amounts of time learning about RDF, OWL and how to embed metadata in web pages. In addition we want to allow other information providers to profit from the existence of annotated data in the easiest way possible.

2 Inspiration

We did a survey of existing topic annotation approaches to find those that actually work in a heterogeneous environment. The ones we found to be the most successful where the various kinds of tags used in the blogging and social software community.
The probably most successful format is the tag format pioneered by Technorati[2]. The number of blogs that publish information according to this standard is estimated to be in the millions[Gibson,2005]. An important tag like “Life 8” is used to annotate almost 15000 blog entries. In this architecture, there is no explicit schema, everybody is free to use any tag she chooses – a concept known as folksonomy. There are two ways to annotate an information unit with tags, through RSS files or by using the “reltag microformat”.
RSS is a family of XML file formats describing the current content of a webpage. Some RSS formats allow to include a <category> or a <dc:subject> tag, giving the subject of an information unit.
The reltag microformat consist of just an html link with the attribute rel=”tag”. Usually such a link points to an uri of the form http://www.technorati.com/tag/[tagname], so for example including
Life8
annotates this webpage with the tag Life8. It is important to note, that the tag URI can be used to access information about the tag – giving a list of information units recently annotated with this tag, related tags and links that often appear in these information units. Using tag URIs that can actually be accessed is also used to a great extend by other social software like del.icio.us[3] or flickr[4]. Both of these services allow to augment the URI in order to get different information, for example http://del.icio.us/tag/[tagname] returns a list of links annotated with a certain tag and http://del.icio.us/rss/tag/[tagname] returns recent changes to this list as RSS file.

* The authors acknowledge support by the German Federal Ministry of Education and Research under the ksi_underground project. The expressed content is the view of the authors and not necessarily the view of the project as a whole.

[1] http://www.unitracc.de

[2] http://www.technorati.de

[3] http://del.icio.us

[4] http://www.flickr.com

2 Implementation

Starting from the above described observations we are currently building a metadata registry with the central goal of simplicity for content providers.

2.1 Getting information from the registry

All URI for the topics are of the format http://unitracc.de/topic/[topicID]. These URIs are also the starting points when interacting with the registry. Calling the URI returns a HTML page describing the topic, its name, related topics, super- and subtopics. HTTP Get parameters can be used to augment the registries response:

- "format", with possible values html, xml allows to get versions suited for human or computer processing.
- "docs", with a integer value ask the registry to return a number of documents annotated with the topic. "from" can be used to give an index where the list begins.
- "news", returns an RSS file with recent changed to this topic and new additions. This parameter cannot be combined with the "format" or "docs".
- "sub", true or false. If documents annotated with subtopics of the current one should be returned, or if changes and additions to subtopics of the current one should be returned.

For example

http:// unitracc.de/topic/leak_test?format=xml&docs=20&from=15&sub=false, returns the documents 15-35 from the list of documents annotated with topic leak_test. The list is returned in a simple, self-explanatory xml format, documents annotated with subtopics of leak_test are not returned. Documents are sorted by the time they where added to the index, newest come first.

Calling the base url http://unitracc.de/topic/ returns a list of all topics, again formatted in html or xml and with the possibility to get the most recent changes as RSS.

2.2 Annotating informations

We plan to regularly crawl web pages that have registered with www.unitracc.de. On these pages we recognize annotations that adhere to the above described reltag microformat. So including Leak Test in a web page, annotates it with the topic leak test.
We can also parse RSS files and recognize the <category> and <dc:subject> tags. So including a category http://unitracc.de/topic/leak_test in a RSS file annotates the appropriate item. Unlike technorati or its competitors we do not interpret a category "leak_test" as http://unitracc.de/topic/leak_test.

2.3 Posting informations

We will allow sites that annotate their content with the unitracc topics to register with our site and will crawl these sites periodically. We also plan two mechanisms to alert the metadata registry to the presence of new content.
The first possibility is a HTTP POST to the base uri http://http://unitracc.de/topic/. A parameter "document" specifies the url of the changed document that is then crawled and for which any topics are added to the index. To prevent SPAM the HTTP user agent of a post request must be a key that is obtained by registering with unitracc[5].
The second possibility is to make such a POST request on the URI of a topic; this stores the information that a particular document has the topic on which the POST request was called. So making a POST request on URL http://unitracc.de/topic/leak_test with the parameter document = "www.leaktestingspec.com" adds the information that "www.leaktestingspec.com" is about leak testing, even if the page has no annotations on it. This of course makes it a trivial exercise to create an HTML form to add annotations to the registry or even Bookmarklets to do this right out of the browser.

2.4 HTTP Status codes

Where possible the system makes use of HTTP status codes to give additional information about a request in a well established and machine understandable format. Besides the obvious 401 (Unauthorized) for failed authentification and 404 (Not Found) for not existing topics, that is 301 (Moved Permanently) to inform the user that a topic id has been replaced by some other topic id and 410 (Gone) to indicate that a topic has been deleted without replacement.

3 Conclusions

We presented the architecture of a metadata registry that is not yet fully implemented nor deployed; we expect to finish the implementation within weeks.
The fact that it was possible to give a fairly comprehensive introduction to the whole interface to this metadata registry in less than two pages makes us optimistic that it may indeed by simple enough to be adopted by some content providers.

References

[Gibson, 2005] Bud Gibson. *Imitation is the surest sign of success.* http://thecommunityengine.com/home/archives/2005/07/imitation_is_th.html, 2005.

[5] To further simplify things and allow for easy testing out of every browser, we allow to simulate such post request by using the GET with parameters action="post" and user_agent="…"