Case Study: Semantic Content Description to improve discovery

Kevin Smith, Vodafone Group R & D

August 2007

General Description

Vodafone Group R & D have designed and implemented a content description vocabulary using W3C Semantic Web technology standards. This has shown a significant improvement in discovery and subsequent download of content described with such metadata.

The Problem

Vodafone is the world's largest international mobile communications group with operations across five continents. In 2001 Vodafone Group designed and built a mobile web portal ("Vodafone live!") which aimed to allow subscribers using a wide variety of mobile devices to discover, download and consume content from the portal. This portal was subsequently deployed across the Vodafone operating companies in Europe and elsewhere. Two problems inherent in supporting such a large global footprint were:

- the wide variety of handsets to support
- managing content from different sources

Supporting a wide variety of handsets

One goal of Vodafone live! was to allow the portal content to be accessible from a huge range of mobile handsets. In 2001 there was a great disparity between the capabilities of mobile Web browsers; with support for one or more of WML, HTML, cHTML, XHTML Basic and XHTML-MP. Whilst the ability to deliver a common user experience via multiple markup languages was provided through a device independent markup language and an XSLT transformation layer, a more pressing problem was support for content downloads. These include ringtones, games and "wallpaper". Unlike Web pages, content downloads typically involve a different binary for each mobile client (or range of clients). As such the discovery and consumption of such content downloads involves extra steps to ensure that the correct binary be delivered to the relevant client.

Managing content from multiple sources

Vodafone live! brokers content from multiple content providers. This content is typically retrieved from a wide variety of content management systems. Such systems typically feature proprietary methods and vocabularies to describe their content. Since these are generally not interoperable, it becomes difficult to identify a relationship between content sourced from one provider, and that which is sourced from others. There was also not a common means to identify any restrictions on the content (such as that of a violent, explicit or sexual nature), nor a way to denote which devices or connection speeds that the content would apply to.

The Solution

Vodafone identified the Resource Description Framework (RDF) over HTTP as a suitable technology for content partners to model and transport metadata for a given URI (or list of URIs). Through binding to appropriate vocabularies, RDF enables semantic descriptions. This means that the meaning of the description itself is defined and published for reference. RDF's use of namespace binding ensures that a machine can determine the scope of the description given, and that conflicts between descriptive vocabularies are avoided.

The RDF model was published as an XML Schema. Appropriate industry vocabularies were identified and referenced in the schema:

- Dublin Core Metadata Element Set to represent Web publication metadata (creator, keywords, teaser description)
- Publishing Requirements for Industry Standard Metadata (PRISM) for additional publication constraints such as content embargo and expiration dates
- The Internet Content Ratings Association (ICRA) vocabulary to accurately describe the nature of any violent, explicit or profane or gambling content.

In addition, Vodafone created its own XML namespace in order to model metadata for which there was no industry standard. This included the device or device range for which the content was suitable, and whether the content was targetted at customers on a fast (3G) connection. The benefit of namespaces was that the same metadata term could be reused in a different context. For example, `dc:title` was at the time defined as the title which would appear as the title for the resource in a Web page. Vodafone also defined their own `vf:title` to represent the full title of a music track or game. `vf:title` was not necessarily to be displayed, but rather to provide the full title of the work as separate to the `vf:artist` (i.e. the musician). In addition further information which helps to describe (and hence locate) the content may be provided through any number of `dc:keywords`. This allows the following separation of concern:

- `<dc:title>Frank Zappa 'Peaches En Regalia' </dc:title> <:!-- the title of the track to appear on the page-->`
- `<vf:title>Peaches En Regalia</vf:title> <:!-- the title of the music track itself -->`
- `<vf:artist>Frank Zappa</vf:artist> <:!-- the musician/performer/band -->`
- `<vf:genre>Rock</vf:genre> <:!-- a genre of music -->`

- `<vf:genre>Jazz fusion</vf:genre> <:!-- a genre of music -->`
- `<dc:keyword>hot rats</dc:keyword> <:!-- associated info (the album) -->`
- `<dc:keyword>guitar solo</dc:keyword><:!-- associated info (famous guitar solo) -->`

Figure 1: Vodafone live! portal

Previously, a "full text" crawl of provider's content pages yielded a rough guess as to the contents, but with no semantic context, it was easy to mistake the nature of content and to repeat search results (for example, is "Apple" a `fruit/corporation/city/recording label/juice?`). The separation of concern and defined context in the RDF improves content discovery. Vodafone were able to associate tracks from different content providers, allowing for genre- and device-specific pages from disparate content sources. In addition there was a significant increase in content downloads resulting from search, as the search results were now more accurate and could be filtered according to the user's device or their connection speed. This ensured that users were only presented links suitable to their context, which avoided subsequent error pages.

Content providers were willing to adopt the RDF model, which could be validated via XML Schema.

How it Works

A content provider creates an RDF document based on Vodafone's published XML Schema. This vocabulary is used to describe a content item (game, music, ringtone, video, wallpaper) or a Web page. This RDF document is ingested and processed by the Vodafone portal. Vodafone can then relate the provider's content descriptions to those from other partners (typically by mapping the RDF to a relational database). The RDF predicates can be modeled as database and search form fields as appropriate.

The following is a example fragment of the content provider's RDF that they submit to Vodafone

```
<rdf:Description about="partner/ringtones.xml">
        <vsrch:ringtone serviceid="myserviceid" >
                <vsrch:provider>classicalmusiccorp</vsrch:provider>
                <dc:title>Delibes - Flower Duet from Lakme</dc:title>
                <vsrch:artist>Delibes</vsrch:artist>
                <vsrch:title>Flower Duet from Lakme</vsrch:title>
                <dc:description>As used in the British Airways advert</dc:description>
                <vsrch:genre>classical</vsrch:genre>
                <vsrch:format>SMAF</vsrch:format>
                <dc:subject>lesley garrett</dc:subject>
                <dc:subject>british airways</dc:subject>
                <dc:subject>romantic</dc:subject>
                <vddr:device>Sharp GX-10</vddr:device>
        </vsrch:ringtone>
</rdf:Description>
```

Key Benefits of Using Semantic Web Technology

- By providing semantic metadata, the content provider facilitates the discovery of their content to Vodafone users.
- Vodafone are able to infer content relationships from disparate providers, and to produce common-themed pages
- Users can search more precisely via extended search forms, which allow search based on predicates used in the RDF ("song title", etc.)
- Vodafone can filter pages and search results based on the user's context (preferences, device and connection speed)

Key Learnings of Using Semantic Web Technology

- Use an open, flexible model

- Research existing vocabularies and reuse industry metadata standards where possible.
- Provide full documentation and a schema (RDF or XML) for validation and IDE support.
- Ensure appropriate stakeholder support

Note: Vodafone created its own vocabulary for content since such a vocabulary (or ontology) mapping all possible Web content is not available. Vodafone will continue to research and continue to contribute to W3C Semantic Web activity with a view to improving Web content discovery and reuse of standard vocabularies.