The Policy Argument, the Commercial Imperative and the Wow Factor

Phil Archer

Internet Content Rating Association, 22 Old Steine, Brighton, BN1 1EL, United Kingdom
parcher@icra.org

Abstract. The semantic web is a tnuly exciting development and has the potential to revolutionise the way people access online data. This is obvious to those involved with its development. However, conveying that potential and sharing the enthusiasm with policy makers can be very difficult. The semantic web must be repackaged as a commercial sales tools.

1 Introduction

In the early days of the internet, metadata such as keywords, descriptions and content labels were commonplace. It is ironic that at a time when the need for metadata is ever more apparent, and when RDF and OWL are recognised as being mature standards, the trivial amounts of metadata provided on the majority of websites is of such poor quality as to be largely useless.

Having developed the architecture, semantic web practitioners must now think in terms of commercial imperatives rather than cool ideas if the hard work is to bear fruit for the benefit of content providers, users and, importantly from ICRA's point of view, children.

2 Metadata on the Web Today

ICRA and its predecessor organisation, RSACi, have been around for a long time promoting a simple idea: that content providers add a label to their material that describes it in terms that reflect the concerns of parents. The Platform for Internet Content Selection, PICS[1], was one of the first W3C Recommendations, was built into Internet Explorer 3.0 and the RSACi website was once the 37th most visited site on the net.

Times have changed. Today, encouraging people to add structured metadata kind to their online material is an uphill struggle. Webmasters do generally include a title and may add keywords in the belief, rightly or wrongly, that it will improve their search engine visibility. Beyond that, however, there is little agreement on how to proceed. To take some high profile examples: on 10th August 2005 the metadata on the Microsoft homepage included an RSACi PICS label plus:

<meta name="KEYWORDS" …
<meta name="DESCRIPTION" …
<meta name="MS.LOCALE"…
<meta name="CATEGORY" …

On the same day, the Yahoo! homepage had an ICRA PICS label and no other meta or link tags. cnn.com had a variety of link tags, including RSS feeds, but no <meta> tags beyond a character set definition. In contrast, government websites, such as 10downingstreet.gov.uk, will often provide a plethora of metadata using Dublin Core and a self-defined classification system that aids their own content management.

The only metadata that is commonly provided on web pages is a title. Beyond that, the working assumption is that if you want metadata for a resource you're going to have to be ready invoke some sort of AI content analyser or screen scraper to extract it for you. For example, the SIMILE project[2] uses JavaScript and/or XSLT screen scrapers to render highly structured websites into RDF. Job listings, apartments for rent and so on can easily be converted into an RDF format but only because the HTML itself follows a predictable pattern.

Given enough ground truth data, AI analysis can classify documents and give their result in RDF if asked. For example, The Software and Knowledge Engineering Laboratory at NCSR, Athens[3], trained their FilterX classification system using roughly 2000 web pages each of three types: pornography, glamour (e.g. magazines like FHM) and non-sexual material. As a result, the system can describe any web page using one of three pre-written ICRA labels. This is a less-than ideal solution since the ICRA vocabulary can yield a total of 35×10^{12} possible labels!

If the vision of the semantic web as a system in which computers and people are better able to work in cooperation is to succeed, the need for a more consistent approach seems clear.

3 Overcoming Scepticism

ICRA has now stopped using PICS labelling and is promoting the addition of RDF descriptions to websites. Our primary goal is the protection of children but to achieve this, we need to get a lot of people to add the right sort of metadata to their content. We're working with trustmark scheme operators to get their "quality seals" expressed in RDF[4] and are excited about ideas like Mobile OK[5].

We really are asking every webmaster in the world to add a little bit of RDF to their content. This should be as natural and normal as using CSS; the addition of RDF should be as much a part of "Create a Website in 21 Days" guides as instructions for creating tables. It's not impossible. For different reasons, different organisations understand the potential benefits of detailed, machine-processable metadata on the web.

Many more people, however, are deeply sceptical. There are two essential obstacles to the wide-scale provision of metadata.

First, as is well understood, is the question of trust. ICRA moved from PICS to RDF in the belief that semantic web technologies have the potential to crack that nut by means that will be familiar to the semantic web/OWL community. For example, by

the end of 2005 it is likely that ICRA will have set up a system whereby a site's label can be cross-referenced with a database populated by a network of volunteers run on the Open Directory Project model[6]. Furthermore, the AI module, FilterX, mentioned above is being built into a proxy system so that the analysed result can be compared with the content provider's own label to see if the latter is *likely* to be accurate.

Second is the cost-benefit balance. From a content provider's point of view, is it worth the time and effort involved to systematically add metadata? Take an institution like the Natural History Museum in London, a world-class repository of information. The metadata on the pages concerned with a recent exhibition called Face to Face[7] was:

```
<meta content="Percussion Rhythmyx" name="generator"/>
<title>Face to Face - Photography by James Mollison</title>
<meta content="text/html; charset=UTF-8" http-equiv="content-type"/>
<meta content="description here" name="description"/>
<meta content="keywords here" name="keywords"/>
```

It is incumbent on the semantic web community to identify clear benefits of the technology to content providers as well as end users.

4 Targeting the Message

The semantic web community comprises enthusiasts for the power of structured data and, importantly, the inferences that can be drawn from it. The potential is something those immersed in the subject can readily feel. Convincing senior executives, policy makers, lawyers and accountants of the advantages of the semantic web is not trivial. They will generally have a different set of questions and criteria when deciding whether a project is a good idea or not, such as:

1. If the metadata states that the author is John Smith and it turns out to be John Doe, what is my legal liability?
2. Will it take more than, say, 4 clicks to install, because if so, it's too complicated for the average user.
3. What will be the increase in customer satisfaction that can be attributed directly to the work done, preferably within the current reporting period?
4. If it's that good, how come everyone isn't doing it already?
5. What does Google think about this?
6. What does Microsoft say about this?

URIs, vocabularies, schemas, ontologies and inference engines don't come into the discussion.

5 Messages That Work

There are two aspects of the semantic web that are highly attractive to senior executives:

1. The ability for consumers to be contributors.
2. The ability to sell additional goods and services related to what the users have already shown themselves to be prepared to pay for.

Creating content is expensive. Persuading your customers to pay you to publish a few kilobytes of content they've created … has obvious advantages!

At a basic level, marketing is about identifying the characteristics of your customers and then finding as many more people with the same characteristics as you can. This surely is a job for the semantic web.

6 Making the Case

If the semantic web is to reach its full potential we need to make a compelling case to the ISPs and the mobile network operators that semantic web technology can increase use of their service. Content providers should want to add metadata because they will make a greater return on their creative effort if they do. Software manufacturers should see how much better their products can be if they make use of the available data.

ICRA's position is clear: that child protection can and should be part of this process.

So I come to my basic question - if RDF and OWL data were ubiquitous on the web what could we do with it? You can link shared bookmarks and smart recommender systems, you can notice that a blog talks about something and so on, but is this a scalable real-world scenario available to average consumers or an academic exercise?

Would the end-user experience be so enhanced that it's worth the content provider's time to add the data?

In the month following ICRA's switch from PICS to RDF, around 1,000 webmasters successfully added RDF ICRA labels to their sites. We'd like it to be 10 or 100 times that figure. Maybe there are Experiences and Directions in the OWL community that can help.

---

[1] http://www.w3.org/PICS/
[2] http://simile.mit.edu/
[3] http://www.iit.demokritos.gr/skel/
[4] http://www.quatro-project.org/
[5] http://www.w3.org/2005/MWI/
[6] http://www.dmoz.org
[7] http://www.nhm.ac.uk/face-to-face/