

Quatro – a metadata platform for trustmarks

Phil Archer
Internet Content Rating Association
Tel: +44 1473 434770
parcher@icra.org

Contributions from other Quatro
project members
See www.quatro-project.org

Abstract

The Quatro project has applied semantic web technologies to trustmark schemes and quality labels. Drawing on past and original research, the project has defined a vocabulary that can be used by any trustmark scheme (TMS) and a technical platform to deliver the trustmarks in a format that can be processed by semantic web agents.

Keywords

Quatro project, Dublin Core Conference, Submission Paper, Trustmarks, Quality Labels, Group URIs, Common description

1. Introduction

Trustmark schemes have been established in many parts of the world, some are online versions of existing schemes, others have been developed specifically for the web. Two notable areas of interest for trustmarks are those designed to give consumers confidence in eCommerce operations and those that indicate that medical information has been peer reviewed. Operators of both types of TMS are among the partners in the Quatro project .

In all cases encountered, the model is essentially the same: a website is submitted for review by the TMS. If the site meets the TMS criteria it is allowed to show a logo. If a user clicks on the logo, a database is interrogated and the current record for that site is displayed, usually showing information such as the date on which the site was last reviewed. Despite the presence of a hyperlink that links to a database record, trustmarks are designed solely to be read by humans and not machines. As a result of Quatro, they will be available to both.

2. The vocabulary

A significant amount of research has been done into

trustmarks, particularly in Europe¹. Research has focussed on how trustmark schemes operate, what benefits they confer on the user and the websites carrying them etc. One such project in 20012 produced a list of criteria that any trustmark scheme would be likely to use when assessing a website. Quatro

Contributions from other Quatro project members
See www.quatro-project.org

has used that a starting point to create a generic vocabulary, available for royalty-free use by quality label and trust mark schemes around the world.

The vocabulary is divided into four categories:

- **General Criteria**, such as whether the labelled site uses clear language that is fit for purpose, includes a privacy statement, data protection contact point etc.
- **Criteria for labelling to ensure accuracy of information** such as the content provider's credentials and appropriate disclosure of funding.
- **Criteria for labelling to ensure compliance with rules and legislation for e-business** such as fair marketing practices and measures to protect children
- **Terms used in operating the trust mark scheme itself** such as the date the label was issued, when it was last reviewed and by whom.

The complete vocabulary is available on the Quatro project website both as a plain text document³ and as an RDF schema,⁴ the namespace for which we have defined as <http://purl.org/quatro/elements/1.0/>.

Trustmark schemes will, of course, continue to devise their own criteria. However, where those criteria are equivalent to those in the Quatro schema, use of common elements offers some distinct advantages.

Firstly, a trustmark that is machine readable and uses common descriptors will be interpreted more easily by semantic web tools than one that uses purely proprietary elements and a proprietary platform. If a user agent is configured to look for Trustmark A but

finds a site that is accredited by Trustmark B, at least the common elements will be recognised, even if those specific to Trustmark B are not. The incentive for content providers to gain accreditation for their material is therefore enhanced if the TMS uses at least some of the common descriptor set.

Secondly, a common set of elements makes it possible to apply machine-learning techniques to the difficult area of ensuring that an accredited site continues to meet the TMS criteria. A machine cannot tell whether an email sent to an eCommerce operator will be responded to within a given time, but it can detect that a contact route is still provided 6 months after the site was last reviewed by a human, even if the nature of the contact route changes.

For example, a site may offer a simple mailto link for contact but subsequently change this to a web form. Content analysis by machine learning will continue to recognise this as a contact route. Likewise, a document that is properly referenced is relatively easy for a machine to identify. If a TMS includes the criterion that all medical documents are properly referenced and a new medical document is added without such references, it can be detected and the TMS alerted that the site needs re-checking.

On both counts the use of a common vocabulary offers commercial advantages to trustmark scheme operators by increasing the value of the labels for content providers and end-users.

3. The Technical Platform

In its simplest form, a trustmark would be a series of elements encoded in much the same way as any other metadata. However, a trustmark will generally apply not to a single resource but to a group of resources, such as all those on a particular website. This presents a problem for RDF which is based on a single URI as a subject. An identical problem obtains for content labelling for other purposes such as child protection.

Project partners' experience of working with PICS⁵ has been informative in devising a schema for RDF Content Labels⁶. A set of documents produced under the aegis of the Quatro project and other activities in Europe and Japan gives use cases, test data and a full description of the schema⁷. Essentially the system allows for a single description to be applied to any number of resources. This can be done in two ways. Firstly a resource can be linked directly to a description using a tag such as:

```
<link rel="meta"
href="http://www.example.org/labels.rdf#label_1"
type="application/rdf+xml" />
```

The RDF instance, labels.rdf, would include a

description – a content label – with an rdf:ID of “label_1”.

However, the real power of the system comes from the second method - a simple rule set. All resources on a content management system or server can include a common link or HTTP response header that points to a single RDF instance. It is likely that this file will be under the control of the content provider's editorial department rather than a production centre. Data in the RDF instance will allow an agent to take the URI of a particular resource and apply the rules that then lead to the correct content label.

Using this method, a trustmark operator, for instance, would be able to accredit a limited portion of a website or a suite of web properties. For ICRA's child-centred labelling system⁸, it allows content providers to apply different labels to different resources on their network. Further uses quickly become apparent, such as film classification or applying a single set of management information to a large collection of resources.

The label schema supports three basic “types” of description:

- A content label – a class whose properties provide the description. This is the one used by the Quatro and ICRA labelling schemes.
- A classification – a class that itself provides a description such as “Suitable for persons aged 12 years and over”
- Management Information – a class whose properties would typically include the DC metadata set, Creative Commons licence etc.

An important component of the RDF Content Labels schema is the idea of defaults and overrides. An RDF instance can declare global, default descriptions that are then overridden if a rule leads to a label of the same type. In other words, one might declare a website to be published by the Example Content Production Company with unrestricted copyright as default management information. However, a different set of management information would override this in the “Madrid” section of the site were published by España Example and all rights are reserved. Classifications and Content Labels can be overridden in the same way but act independently of each other.

3.1 Example

The following code fragment exemplifies several features of the platform.

```
<label:Ruleset> <label:hostRestriction>example.org
</label:hostRestriction>
<label:hostRestriction>example.com
```

```

</label:hostRestriction>
<label:hasDefaultLabel
  rdf:resource="#label_1" />
<label:hasDefaultManagementInfo
  rdf:resource="#mgt_1" />
<label:rules rdf:parseType="Collection">

  <rdf:Description>
<label:hasURI>photography
</label:hasURI>
<label:hasLabel

rdf:resource="#label_2"/>
<label:hasManagementInfo
rdf:resource="#mgt_2" />
</rdf:Description>

<label:UnionOf>
<label:hasURI>guestbook</label:hasURI>
<label:hasURI>messages</label:hasURI>
<label:hasLabel
  rdf:resource="#label_3" /> </label:UnionOf>
</label:rules> </label:Ruleset>
<label:ContentLabel rdf:ID="label_1" />
<rdfs:label>Use of clear language fit for
purpose, Privacy statement, no nudity...
<quatro:gb>1</quatro:gb>
<quatro:gc>1</quatro:gc>
<icra:nz>1</icra:nz>
...

</label:ContentLabel>
<label:ContentLabel rdf:ID="mgt_1"> <dc:publisher
rdf:resource="http://www.example.org" />
<dc:rights>© Example Inc</dc:rights> <cc:license
rdf:resource="http://www.creativecommons.
org/licenses/example1" />

```

... The first two elements in the Ruleset define that information is available only about material on the example.org and example.com hosts. Subdomains are defined as being in scope. The default label and the default management information are then given for these hosts.

In the absence of further information, the assertions made in label_1 (which in the example includes both Quatro and ICRA elements) are true; everything on example.org and example.com is published by example.org and is copyright Example Inc.

However, if the URL in question includes the string "photography" then it is described by label_2 and has a different set of management information. (The values of label:hasURI properties are processed as Perl 5 regular expressions.)

The second rule says "if the URL includes 'guestbook' or 'messages' then use label_3." However, the

management information is not overridden so that the default publisher and copyright information still applies.

4. Relevance to Dublin Core

Although Quatro and Dublin Core are responses to very different demands made by different constituencies, there are clear areas of common interest and interoperability.

4.1 The Vocabulary

There is no direct mapping between the bulk of the Quatro vocabulary and the DC elements and terms since they serve different purposes.. However, Dublin Core metadata *is* highly relevant to the elements used by TMS operators in the administration. of their schemes dcterms:issued is used directly, quatro:lastReviewed and quatro:withdrawn are both defined as subProperties of dc:date.

4.2 The Platform

As the example in section 3.1 shows, the RDF Content Labels platform makes specific provision for management information as a separate entity from descriptions such as quality and content labels. Dublin Core elements can therefore readily be applied to groups of resources in a manner that is machine processable. Critically, management information can be applied in a manner that readily fits in with the typical workflow of large content providers.

5. Application

Quatro is approaching the end of its first year. Both the vocabulary and technical platform are already published with implementation under way by two trustmark schemes (IQUA⁹ and WMA10) and ICRA. Work has now begun to develop applications to make use of the machine-readable labels. These are:

A browser-independent helper application that will recognise semantic web data where present on websites and provide a visual interpretation. A user will therefore be able to see that a site has a trustmark whether or not the actual trustmark logo is visible to them.

A wrapper for search results that will indicate the presence of trustmarks and/or other metadata on the websites listed. This will be available for inspection by clicking an icon adjacent to the relevant result.

The applications will use common code elements to identify the labels and use relevant methods to attempt to gain trust in them. These include automated database look-up and machine-learning based content analysis. The first application sits on an end-user's computer, the second is an option for search engines.

6. Summary

The Quatro project presents a method of grouping URIs that share common descriptions. It is hoped that this will have wide interest and application in the DC community, however, the focus of the project is on bringing trustmarks (quality labels) into the semantic web. A royalty-free vocabulary has been devised for use by trustmark schemes. Use of this common basis for a variety of labelling schemes offers significant advantages to trustmark operators and endusers.

7. References

1. See, for example, http://europa.eu.int/information_society/activities/sip/docs/pdf/reports/qual_lab_bkgd.pdf
2. The UNICE -BEUC e-Confidence project. The final report, published 22/10/01 is available from www.beuc.org but is more easily found at www.quatro-project.org/unice-beuc/eConfidence.pdf
3. www.quatro-project.org/vocabulary/1.0/
4. <http://purl.org/quatro/elements/1.0/>
5. PICS – the Platform for Internet Content Selection. See www.w3.org/PICS/
6. e.g. <http://www.icra.org/press/www2004/>
7. See <http://www.w3.org/2004/12/q/doc/rdf-contentlabels.html>
8. <http://www.icra.org>
9. <http://www.iqua.net>
10. <http://wma.comb.es/>