Quality Labeling of Web Content: The Quatro approach

Vangelis Karkaletsis
NCSR "Demokritos"
P. Grigoriou & Neapoleos str.
15310 Ag. Paraskevi Attikis, Greece
+30 210 6503197
vangelis@iit.demokritos.gr

Andrea Perego
Università degli Studi di Milano
via Comelico 39/41
I-20135 Milano MI, Italy
+39 02503 16273
perego@dico.unimi.it

Phil Archer
Internet Content Rating Association
22 Old Steine, Brighton, East Sussex,
BN1 1EL United Kingdom
+44 (0)1473 434770
parcher@icra.org

Kostas Stamatakis
NCSR "Demokritos"
P. Grigoriou & Neapoleos str.
15310 Ag. Paraskevi Attikis, Greece
+30 210 6503215
kstam@iit.demokritos.gr

Pantelis Nasikas
NCSR "Demokritos"
P. Grigoriou & Neapoleos str.
15310 Ag. Paraskevi Attikis, Greece
+30 210 6503197
pnas@iit.demokritos.gr

David Rose
Coolwave Limited
4 -6 Greenfield House, Storrington, Nr
Pulborough, West Sussex, UK
+44 (0)870 7127000
david@coolwave.co.uk

ABSTRACT
QUATRO is an on-going EC-funded project which aims to provide a common vocabulary and machine readable schema for quality labeling of Web content, as well as ways to automatically show the contents of the label(s) found in a Web resource, and functionalities for checking the validity of these labels. The paper presents the QUATRO processes for label validation and user notification, and outlines the architecture of QUATRO system.

Categories and Subject Descriptors
H.3.5 Online Information Services: *Web-based services*

General Terms
Management, Reliability, Experimentation, Verification.

Keywords
Quality labeling, web content analysis, RDF schemas

1. INTRODUCTION
QUATRO is an on-going EC-funded project which aims to provide a common vocabulary and machine readable schema for quality labeling of web content, making it possible for the many existing labeling schemes to be brought together through a single, coherent approach without affecting the individual scheme's criteria or independence [1].

QUATRO's work on providing a platform for machine-understandable quality labels, also called trustmarks, is part of a much greater activity around the world, that of Semantic Web [2]. Three QUATRO partners, ERCIM, as European host for W3C, and ICRA and NCSR, as W3C members, are active participants in this activity. RDF, the Resource Description Framework [3], is the key technology behind the Semantic Web, providing a means of expressing data on the web in a structured way that can be processed by machines. It allows a machine to recognize that, for example, 5 blogs are commenting on the same web site, that 3 people have the same site in their (online) bookmarks (favorites) and that it gets a 4.5 rating on a recommender system.

QUATRO adds to the picture in two ways: by providing a way in which any number of web resources can easily share the same description; by providing a common vocabulary that can be used by labeling authorities. As a result, machines will be able to recognize that a site mentioned in a blog that gets a 4.5 star rating on a recommender system and is in 3 friends' online bookmarks also has a label. By basing the labels on RDF, QUATRO is effectively promoting the addition of data on the web that a wide variety of other applications can use to build trust in a given resource.

At the time of writing this paper, the details of the QUATRO vocabulary have been finalized and the complete vocabulary is available on the QUATRO site and elsewhere, both as a plain text document and an RDF schema [4]. It will be available for free usage by Labeling Authorities (LAs) as they see fit. The project's 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.

LAs 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.

Work is now underway to develop applications to make use of the machine-readable labels:

- An application for checking the validity of machine-readable labels found in web resources. A label's validity is checked against the corresponding information found in the LA's database. Furthermore, QUATRO also enables, for some cases, the checking of label's validity against the content of the web resource. The application is implemented as a proxy server, named QUAPRO.
- A browser extension, named ViQ**,** which enables the visual interpretation of label found in the web resource requested by the user, according to QUAPRO results. A user is therefore able to see that a site has a label and be notified on the label's validity and content.
- A wrapper for search engines' results, named LADI, which indicates the presence of label(s) on the web sites listed. This will be available for inspection by clicking an icon adjacent to the relevant result. As in the case of ViQ, label validation and user notification will be performed by QUAPRO.

This paper briefly presents the QUATRO processes for label validation and user notification (Section 2), the QUATRO architecture and the main functionalities of the components of the system implementing this architecture (Section 3).

2. Label validation and User notification

Before displaying the content of a label identified in a web resource, it is necessary to examine whether the label is a valid one against either the Labeling Authority's (LA) database or the content of the web resource. For this purpose, QUATRO employs two validation processes.

The first one concerns the label's integrity, independently from the content of the web resource. A label is generated by the corresponding LA at some point in time, and represents the content of the web resource at that time. It is possible that the provider of the web resource's content has changed the label's content without informing the LA. The validation mechanism must enable the checking of the label's content against the corresponding content stored in the LA's database, in order to ensure the label's integrity. This does not mean that a label that satisfies the integrity constraint is actually valid, since the content of the web resource may have changed. On the other hand, we cannot be completely sure that a label which does not satisfy our integrity constraints is necessarily invalid.

That's why examining a label's integrity must be supported, whenever this is possible, by an additional comparison of the label's content against the actual resource content. This constitutes the second QUATRO validation process. It is difficult to automate this validation check since it involves the use of advanced content analysis techniques. In the context of QUATRO, we use the content analyzer FilterX [5] in one of the case studies.

The criteria according to which a label should be considered valid/invalid may vary depending on the specific labeling scheme. We distinguish two different scenarios.

In the first scenario, the labels are stored at the LA's site. In such a case, labels cannot be modified directly by the web resources' content providers, and thus their integrity is granted. That is, in this case, we can only examine whether the resource's content has been modified and if the updated content is not in-line with the label's content.

In the second scenario, labels are stored at the labeled resource site. Since such labels are not under the control of the LA, they can be easily modified by the resources' content providers. In order to verify their validity, QUATRO needs to be able to verify a) whether the label stored at the labeled resource site is the same of the one that has been generated by the LA (integrity control) and b) whether the label has not expired (date control). The former may be enforced by a hash-matching while the latter by a date-comparison mechanism.

More precisely, concerning integrity control, whenever a label is generated, the LA hashes the label and the produced hash is stored in the LA database. Whenever a label is located inside a web resource, QUATRO hashes it and asks the LA to verify whether this hash matches with the hash of the label stored in the LA's database. In addition, for every label generated by the LA, a label expiry date parameter is set, which means that the label is valid until that specific date. Therefore, QUATRO gets from the LA this valid-until date in order to check the label validity.

Finally, as noted before, whenever a content analyzer is available, QUATRO can perform an additional check examining the content of the web resource against the label's content.

Thus, three different policies can be enforced for label's validation: labels' integrity, labels' expiry date, and content analysis (meaning the semantic equivalence between the actual resource content and the description provided by the label).

Note that it may be also the case that the label cannot be validated. For instance, the LA database may be down, the hosting server may be off-line, the QUATRO's proxy (QUAPRO) may be unavailable. In such cases we can simply say that the validity of the label *cannot be verified*. This applies even to the case when a content analyzer is not able to decide whether a label is valid or not. Thus we have the following possible results when evaluating labels: *valid*, *invalid*, and *cannot be verified*;

As it concerns user notification, this is performed in order to inform users whether a resource is labeled or not. Yet, when labels are invalid, the description they provide is useless. Thus, we can devise two different strategies for considering a resource as labeled:

- when valid labels are associated with it,
- when labels are associated with it, independently from their validity.

QUATRO adopts the latter strategy, since it aims at informing users about the characteristics of the requested resources, not at blocking inappropriate contents. In addition, QUATRO validation policies allow the verification of labels' validity against the LA's database in all cases, but, as it concerns the validation of the label's content against the resource's content, this can only be done when a content analyzer is available for the specific case. Thus, QUATRO's approach allows the user to access the content of a label, even though it is not valid. After being notified whether a label is valid or not, users can display the contents of any available label. It is up to them to decide whether they will trust it or not.

Label notification may then return one of the following results:

- The requested resource is unlabelled: The end user is informed that no label is available for the requested resource.

- The requested resource is labeled: The end user is informed that labels are present, and he/she is notified whether they are valid, invalid, or they cannot be evaluated.

Further work on the label validation scheme will include, incorporating XML Digital Signatures. In this scenario an LA does not need to provide an online database with labels and hashes as a web service, just a way to locate its public key (e.g. as RDF/A metadata on its website). The label file will contain the digital signature of the hash. The hash will be generated as before, and we will generate the digital signature from it, rather than from the label itself, due to performance reasons. So, once the labeling authority creates the label and the hash, and signs it with a digital signature from a private key that it (the LA) keeps secret , a user agent program can easily verify the integrity of the hash (and thus the label) if he uses the public key. One drawback in this validation scheme would be that it might take too much time to decrypt the digital signature with the public key in order to get back the original hash key , but we are working on it.

QUATRO Architecture

Figure 1 depicts the four applications participating in the QUATRO quality labels validation and notification tasks (ViQ, LADI, QUAPRO and FilterX). QUAPRO is the central server-based application which receives requests from the two end-user applications (ViQ, LADI), identifies quality labels, evaluates them and replies accordingly. A Data Access interface (DAcc), placed before an LA's database, handles the communication between QUAPRO and the database.

The applications mentioned above have to exchange messages since QUAPRO needs information from all the parties involved (ViQ/LADI, LA's database, content analyzer) to assess the labels' validity. The Simple Object Access Protocol (SOAP), a W3C recommendation [6], is used for this purpose. An XML schema has been devised that must be followed by any application that wants to use the services provided by QUAPRO. This enables, for instance, to employ another content analysis tool, or add another labeling authority. SOAP has been selected because it uses http (in our case) as its transfer protocol, and therefore no special configuration is required from the end user when installing the ViQ plug-in.

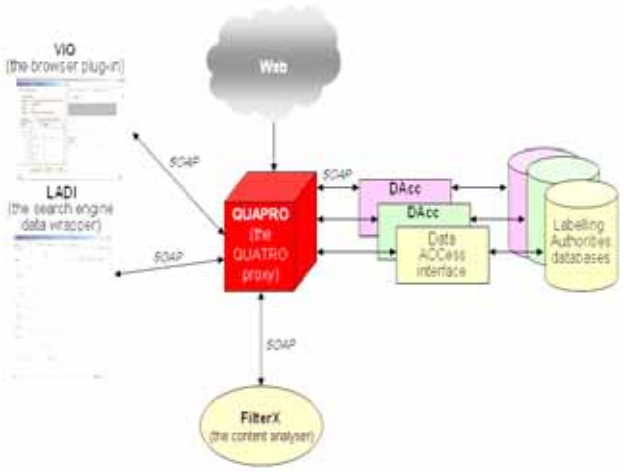

Figure 1. QUATRO architecture

The next sub-sections provide more information on the functionalities of QUATRO components.

2.1 ViQ

The Metadata Visualizer (ViQ) is a client application in charge of two main tasks:

- to notify users whether a requested Web resource is associated with content labels or not;
- to display to the users the contents of the labels associated with Web resources.

ViQ is being developed as a browser extension for the three most popular Web browsers (i.e., MS Internet Explorer, Mozilla Firefox and Opera), providing a toolbar (the ViQ Toolbar), a status bar icon, and an additional item in the browser main menu. Users are notified of the presence/absence of labels by specific icons. If labels are available, the user can display their contents.

ViQ relies on QUAPRO for verifying labels' validity. Moreover, QUAPRO will be in charge of returning the information needed by ViQ to display the label summary and details. More precisely, whenever a Web resource is requested by the user, ViQ performs the following steps:

- if QUAPRO says that labels are absent, the user is notified that no labels are available for the requested resource;:
- otherwise, ViQ notifies that labels are present, and it displays the lists of available labels, marked with an icon denoting their validity status (valid, invalid, and "cannot be verified" – see Figure 2).

Figure 2. ViQ browser extension

2.2 LADI

The Search Engine Wrapper LADI is a server application that gives users an indication of the existence of a label or labels inside the web resources listed in search engine results and then allows them to see more detailed information about those labels. As with ViQ, LADI calls on QUAPRO to provide label summary and details and to verify the validity of labels. Where ViQ provides information about resources that have already been

visited, LADI will provide the same or similar information before a resource is visited. LADI's task is therefore quite different in that it must check with QUAPRO for each of, say, ten results per page of search results that are viewed per user search. It must then provide the indicators and a method for viewing the information within the browser as part of the search result listing returned to the user.

So, LADI will:

- Provide a web search form initially.
- Accept a search term from the user and, using the appropriate API, perform a server-to-server request to the appropriate search engine (Google, Yahoo! in QUATRO case studies).
- For each of the resources returned by the search engine(s), make a server-to-server request to QUAPRO to check for the existence of a label or labels and to obtain the information about those labels.
- Produce the HTML for the search results to be returned to the user, merging the results obtained from the chosen search engine with any relevant information from QUAPRO.

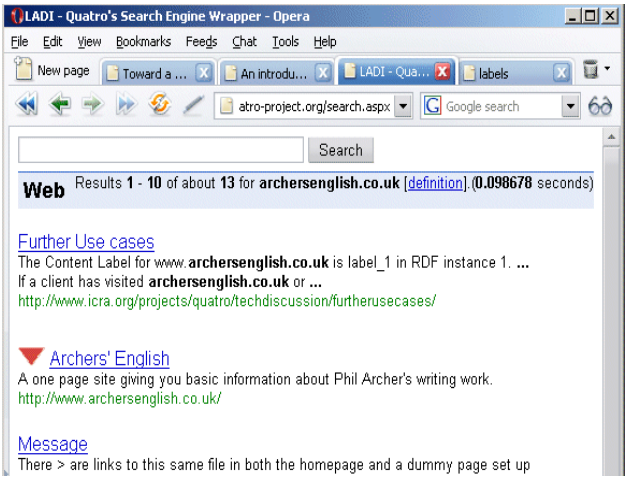

Figure 3. LADI-annotated search results

2.3 QUAPRO

QUAPRO is a server-based application that processes requests from both ViQ and LADI. In order to decide on a quality label's validity, QUAPRO can perform 3 different types of controls: date control, hash control, content analysis control. The first two checks are used to decide on label's validity against the LA's database, whereas the third check examines the label's validity against the content of the corresponding resource. In case all three checks are used, a composition of the verdicts gives the final validity value for the label (valid, invalid, "cannot be verified").

QUAPRO either accepts a single URL (ViQ) or a list of URLs (LADI) and checks if they are labeled. It looks for links to labels in the HTML code of the web page or the HTTP headers when accessing a URL. If a label is found, QUAPRO proceeds by querying the label to find the label's creator and subsequently returns this information to ViQ/LADI. QUAPRO is using the SPARQL query language [7], for accessing information stored in the RDF labels, such as the label creator, the label expiry date and the URLs that this label applies to.

When QUAPRO receives a request for one of the labels found in a specific URL, it queries the label in order to find its expiry date, creates its hash and contacts the corresponding LA database (via DAcc) to assess the validity of the label. While waiting for the DAcc response, and in case a content analyzer is available (FilterX in our case), it also sends a message to it. When the responses from DAcc and the content analyzer come, QUAPRO compiles the new message to be sent to ViQ/LADI. This message contains links to unique URLs in the QUAPRO server that contain the labels in natural language so that it can be accessed if requested from ViQ/LADI.

2.4 DAcc

The labeling authorities maintain a database of the web sites that have been labeled as well as metadata about the labels such as expiration date, language, the hash key for the label. For QUAPRO, DAcc is a "black box" receiving and sending SOAP messages in conformity to the SOAP messages schema.

The DAcc application receives from QUAPRO the URL of the web site, the URL of the RDF label on the web site and the hash key generated from QUAPRO. DAcc in response returns whether the hash keys match, and the expiration date status.

2.5 FilterX

FilterX is a content analyzer which enables the intelligent blocking of obscene content accessible through browsers on the World Wide Web. FilterX is a product of i-sieve [3], a spin-off of QUATRO's partner NCSR "Demokritos". I-sieve provides FilterX to NCSR for the research purposes of the QUATRO project.

For the purposes of QUATRO, FilterX has been adapted to perform as an independent software module which will be invoked by QUAPRO to evaluate labeled Web resources and return a message compatible to QUATRO specification. So, FilterX accepts a URL sent by QUAPRO and returns a message with the results of content analysis.

3. Concluding remarks

Currently, web sites carrying quality labels such as those administered by the QUATRO partners, Internet Quality Agency and Web Mèdica Acreditada, carry a logo. Clicking the logo, results in the display of a database entry confirming the logo's validity, last review date etc. However, such labels work in isolation and are only visible to human visitors to sites. They cannot be harvested, aggregated or otherwise utilised by machines.

QUATRO offers a substantial improvement to the current situation. First, project members have worked to create a flexible platform that encodes the labels. Secondly, it offers a vocabulary that encompasses the common elements of a wide variety of labeling schemes. The two together have the potential to make many different quality labels highly interoperable. It must be noted that Segala [8] is using the system to encode its certification scheme for web accessibility. RDF content labels are also examined in a W3C's Incubator Activity [9] which is feeding directly into the Mobile Web Initiative's development of a mobileOK trustmark [10].

Furthermore, QUATRO provides the means for users navigating the web with a common web browser to be notified when quality labels are present (using appropriate graphics) and, if they are, whether they are valid or not. The two end-user applications, ViQ and LADI, currently under development, serve this purpose.

4. Acknowledgments

This research was partially funded by the EC through the SIAP project QUATRO (Quality Assurance and Content Description). QUATRO involves the following partners: Pira International (Coordinator), Internet Content Rating Association, Internet Quality Agency, Web Mèdica Acreditada, NCSR "Demokritos", University of Milan, Coolwave, ECP.NL, ERCIM.

5. References

[1] http://www.quatro-project.org

[2] http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

[3] http://www.w3.org/RDF/

[4] http://purl.oclc.org/quatro/elements/1.0/

[5] http://www.i-sieve.com

[6] http://www.w3.org/TR/soap

[7] http://www.w3.org/TR/rdf-sparql-query/

[8] http://www.segala.com

[9] http://www.w3.org/2005/Incubator/wcl/wcl-charter-20060208.html

[10] http://www.w3.org/Mobile/

[11] http://www.w3.org/TR/xmldsig-core/