Health Informatics Journal

Article

Health Informatics Journal
17(2) 95–115

DOI: 10.1177/1460458211405004
jhi.sagepub.com
SAGE

Applying Semantic Web technologies to improve the retrieval, credibility and use of health-related web resources

Miguel A. Mayer
Web Mèdica Acreditada, Medical Association of Barcelona (COMB), Spain
Research Programme on Biomedical Informatics (GRIB), IMIM-Universitat Pompeu Fabra, Spain

Pythagoras Karampiperis, Antonis Kukurikos, Vangelis Karkaletsis and Kostas Stamatakis
National Center for Scientific Research "Demokritos" (NCSR), Greece

Dagmar Villarroel
Agency for Quality in Medicine (AQuMed), Germany

Angela Leis
Web Mèdica Acreditada, Medical Association of Barcelona (COMB), Spain

Abstract

The number of health-related websites is increasing day-by-day; however, their quality is variable and difficult to assess. Various "trust marks" and filtering portals have been created in order to assist consumers in retrieving quality medical information. Consumers are using search engines as the main tool to get health information; however, the major problem is that the meaning of the web content is not machine-readable in the sense that computers cannot understand words and sentences as humans can. In addition, trust marks are invisible to search engines, thus limiting their usefulness in practice. During the last five years there have been different attempts to use Semantic Web tools to label health-related web resources to help internet users identify trustworthy resources. This paper discusses how Semantic Web technologies can be applied in practice to generate machine-readable labels and display their content, as well as to empower end-users by providing them with the infrastructure for expressing and sharing their opinions on the quality of health-related web resources.

Corresponding author:
Miguel A. Mayer, Medical Association of Barcelona (COMB), Spain and Research Programme on Biomedical Informatics (GRIB), IMIM-Universitat Pompeu Fabra, Spain
Email: mmayer.wma@comb.cat

Keywords
health, labelling, quality, Semantic Web, metadata

Introduction

The number of health-related websites is increasing day-by-day[1]; however, their quality is variable and difficult to assess. One source of difficulty is the plethora of organizations producing web content, ranging from government institutions, consumers, scientific organizations, commercial companies and patients' associations, for example.[2,3]

Different studies show that more than 60% of health information seekers began their last online health inquiry using a search engine, whereas less than 30% used a health-related website portal.[4–7] It is difficult for patients and the general public to assess the quality of information as they are not always familiar with the medical domains and terminology.[2,8] To support patients and consumers in retrieving suitable information sources, a number of quality labelling initiatives have been developed across Europe.[9–11]

One approach is the "trust mark" method, where a third-party agency ascertains on a regular basis whether the quality of the information of the website is acceptable or not. Another approach is when a third-party authority selects, or filters, websites for the public to use.[12] In each case, a website is awarded the right to display the quality label or trust mark after going through a review process. Users may click on a logo to see data supplied by the labelling scheme operator.

The aforementioned approach follows a linear model of content annotation that places a quality label at one end of the chain and the end-user at the other. In reality, the nature of the web has shifted significantly from such a model. Users have become content producers and can express and exchange directly, or indirectly, information on any topic of interest. Moreover, a form of quality labelling called social or collaborative tagging has emerged, which is a means for characterizing online resources with different criteria. The openness and generality of the collaborative tagging has resulted in widespread support for the method. Several web-based social networks, such as del.icio.us (http://delicious.com/), RawSugar (http://www.rawsugar.com/), Flickr (http://www.flickr.com/) and Last.fm (http://www.last.fm/) have adopted tagging functionalities.

Thus, an ideal solution for accurately describing health-related web content is to combine the wisdom of experts (medical organizations, labelling authorities) with the wisdom of crowds (collaborating end-users).[13] Figure 1 summarizes the roles and desired activities of each stakeholder in the Medical Content Labelling process.

Domain experts act as providers of labeling information by certifying web resources using criteria that they find suitable and adequate. Communities of users can also create labeling information by expressing their opinions, using their own criteria, which can be aggregated to produce an overall description for the corresponding web resources. Finally, the end-user, who can be a patient, a consumer or any interested web user, should be able to exploit the information produced by domain experts and web communities in order to search and efficiently retrieve quality medical content.

Such approaches raise several issues. From the perspective of the labelling authority, the major concerns are related to the discovery of content relative to their domain of interest, classifying the content, and, finally, creating the label using their criteria and associating it with the web resource. An important aspect of the labelling procedure is the need for easy and constant monitoring of the labelled web resources, as their contents could be modified or completely altered at any time.

End-users, on the other hand, need to have easy and meaningful access to the labelling information and also have a way of ensuring that the label is valid and verified. Furthermore, and taking

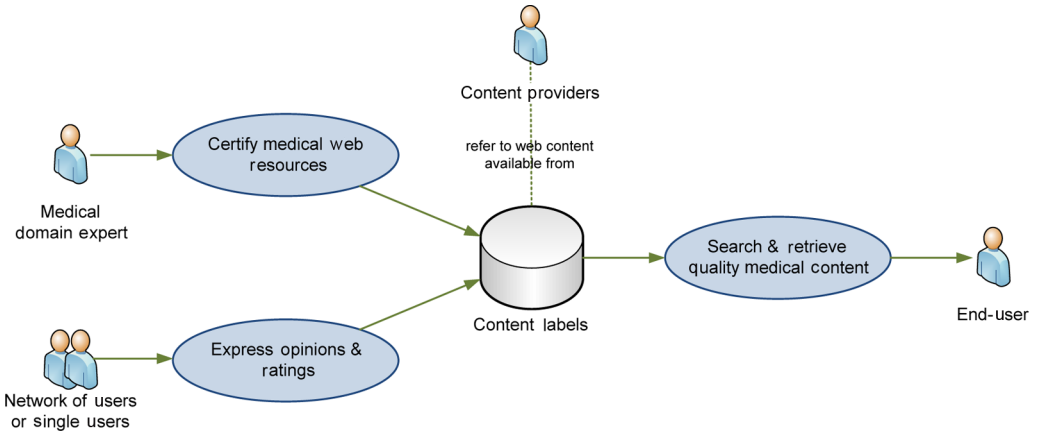

Figure 1. Stakeholders in medical content labeling

into account the current trends regarding use of the web, users should have the opportunity to express their own opinions by creating their own labels as individuals or as part of a community or organization, and also to comment on the accuracy of other labels published by other users, other communities or labelling agencies.

The Semantic Web and its associated technologies established the ground for facilitating the content discovery process and automating the labelling and monitoring procedures. In addition, the Semantic Web, in principle, empowers end-users to access the content they seek and is suitable for their needs, as well as to publish their opinions in an interchangeable, machine-processable manner. Thus, Semantic Web technologies have the potential to provide new opportunities in the field of content labelling of online health information.[14,15]

Ten years ago, the World Wide Web Consortium (W3C) developed a standard metadata language called 'Platform for Internet Content Selection' (PICS). The most common uses of PICS labels have been in filtering applications that block access to web resources based on labels associated with those resources, for example to filter pornography and other offensive material for child protection (http://www.w3.org/PICS). PICS was a system for associating metadata (PICS "labels") with internet content and provided a mechanism enabling independent groups to develop metadata vocabularies without naming conflicts. However, PICS did not implement any of the subsequent developments in web technology, such as the Extensible Markup Language (XML), and has not been widely adopted.

Subsequently, the Resource Description Framework (RDF), developed by W3C, provided a model for representing more general metadata than PICS, with more expressive power and using XML syntax (http://www.w3.org/RDF).

Following the definition of PICS and the establishment of RDF as a W3C recommendation, the MedPICS (Certification and Rating of Trustworthy Health Information on the Net) project, funded under the European Union's (EU) "Action Plan for safer use of the Internet", developed a standard vocabulary (expressed as PICS/RDF/XML) called medPICS (platform for internet content selection in medicine), an application of PICS.[16] MedPICS was created for use by information providers, users and third-parties to describe and disclose properties of e-health services.

The Health Information Disclosure, Description and Evaluation Language (HIDDEL) evolved from MedPICS and was developed within the EU MedCERTAIN project, with the goal of offering

a standard vocabulary to describe health websites.[17] The EU MedCIRCLE (Collaboration for Internet rating, certification, labelling and evaluation of health information on the World Wide Web) project[18] further developed and refined this vocabulary. HIDDEL was very extensive, detailed and based on RDF format, but at that time RDF was in the first stages of development, which hindered wide use of HIDDEL.

The Quality Assurance and Content Description (QUATRO) project applied Semantic Web technologies to trust-mark schemes and quality labels. Within QUATRO, an analysis of different standardized metadata vocabularies was carried out, including Dublin Core Metadata (http://www.dublincore.org), HIDDEL[17] and PICS. Work in QUATRO led to the RDF Content Label Schema (RDF-CL) and the corresponding W3C Incubator group (http://www.w3.org/2004/12/q/doc/content-labels-schema20050704.htm). Further development led to the creation of a W3C Working Group and the subsequent development of the W3C Protocol for Web Description Resources (POWDER). POWDER (http://www.w3.org/2007/powder/) is a general-purpose content- and quality-labelling protocol based on Semantic Web technologies. POWDER documents are expressed in XML syntax and contain attribution information regarding the publisher of the specific document and Description Resources (DR) blocks, where the actual description and its scope are defined.

In this article, we present different metadata and Semantic Web technologies, as well as tools targeted at the labelling agencies and end-users that exploit them, developed under two European projects: the Quality Labelling of medical Web content using multilingual information extraction (MedIEQ) project (http://www.medieq.org) and the Content Labels for User Empowerment (QUATRO Plus) project (http://www.quatro-project.org) – a continuation of the QUATRO project – with the aim of improving the retrieval, credibility and use of trustworthy websites. These technologies are exploited by two third-party medical labelling authorities, the medical quality certification program Web Mèdica Acreditada (WMA; http://wma.comb.es) in Spain, and the medical filtering portal Patienten-information.de (http://www.aezq.de) in Germany.

Following descriptions of the various MedIEQ and QUATRO Plus tools, we define a complete framework that aims to help certification and filtering organizations to label health-related web resources using machine-processable descriptions, update and maintain their labels, as well as publishing them, making them visible and usable for the general public, and giving users the opportunity to contribute to the evaluation of web resources.

Current approaches to health-related web content labelling

Two major approaches currently exist for the labeling of health-related Web resources: first, *filtering portals* (organizing resources in health topics and providing opinions from specialists on their content); and second, *third-party certification* (issuing certification trustmarks or seals once the content conforms to certain principles). In general, and in both approaches, the labeling process comprises three tasks that are carried out entirely, or partially, by most labeling agencies:

1. *Identification* of new web resources: this could happen either by active web searching or on the request of the content provider, i.e. the website responsible actively requests for the review in order to get a certification seal.
2. *Labeling* of web resources: this could be done for the purpose of awarding a certification seal or in order to classify and index web resources in a filtering portal.
3. *Re-reviewing* or *monitoring* labeled web resources: this step is necessary to identify changes or updates in resources, as well as broken links, and to verify if a resource still deserves to be awarded its label.

This is the general case; any particular labeling agency can integrate additional steps that may be necessary in its work. The two labeling agencies participating in MedIEQ, Agency for Quality in Medicine (AQuMed; http://www.aezq.de) and Web Mèdica Acreditada (WMA; http://wma.comb.es), represent the two approaches mentioned above: AQuMed maintains a filtering portal, while WMA acts as a third-party certification agency.

The indexing and labeling process in AQuMed consists of five steps:

1. *Inclusion of a new resource*. There are two ways through which a new resource can be identified for indexing in AQuMed database. The first way is through an internet search and the second is through a direct request from the content provider. Websites are selected according to general criteria: content, form and presentation should be serious; authorship, sponsorship and creation/update date should be clear; and only websites without commercial interest should be indexed.
2. *Website classification*. Previously unlabelled websites are classified into four groups: treatment information, background information, medical associations/scientific organizations and self-help/counseling organizations. Only sites with treatment information proceed to the next step.
3. *Evaluation*. Sites with treatment information are evaluated using the DISCERN (http://www.discern.org.uk/) and Check-In instruments. DISCERN is a well-known user guidance instrument and Check-In was developed by AQuMed in collaboration with the "Patient Forum" of the German Medical Association. Check-In is based on DISCERN and the AGREE (http://www.agreecollaboration.org/instrument/) instrument for critical evaluation of medical guidelines.
4. *Confirmation*. The database administrator has to confirm the result of the evaluation. It can be modified, erased, or simply confirmed.
5. *Feedback to the information provider*. AQuMed sends an e-mail with the result of the evaluation in the case of sites with treatment information and with the information about the admission into the AQuMed database in the case of the other categories.

Other valid categories for AQuMed are: i) background information; ii) medical associations or scientific organizations; and iii) self-help organizations.

AQuMed's database is periodically populated through new internet searches and is regularly examined for broken links. The evaluated resources are also periodically re-reviewed in order to identify changes against the criteria or other updates.

The complete WMA certification process consists of the following four steps:

1. The person in charge of a website sends a (voluntary) request to WMA in order to initiate the process. Using the online application form, the interested party provides certain information to WMA and has the chance to auto-check the WMA criteria based on the Code of Conduct and the Ethical Code.
2. The WMA Standing Committee assesses the website based on the WMA criteria (medical authorship, updating, web accessibility, rules in virtual consultation, etc.), and issues recommendations.
3. WMA sends a report to the person in charge who implements the recommendations.
4. When the recommendations have been implemented, it is possible to obtain the seal of approval. In such a case, WMA sends an HTML seal code to be posted on the accredited website. In addition, WMA includes the site's name and URL to the index of accredited websites and an RDF content label is generated.

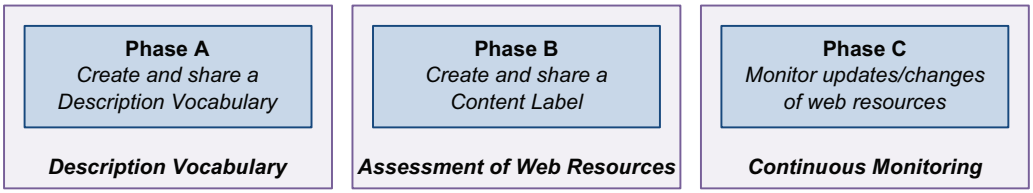

Figure 2. Content Label lifecycle

Based on the above, a typical Content Label lifecycle consists of the following key steps: (a) the creation of a description vocabulary through identification of the key attributes describing web resources; (b) the assessment of web resources based on the agreed vocabulary; and, (c) continuous monitoring to certify the quality of the resource(s) over time.

Thus, the design process of Content Labels consists of the following key phases (as depicted in Figure 2):

- Phase A: Create a Description Vocabulary with common terms that can be identified by labelling experts/organizations, and can be represented in a common and interoperable format, as well as share this Description Vocabulary within a Community of Labelling Organizations, so as to facilitate common ways of describing web resources.
- Phase B: Create Content Labels, using agreed vocabularies and representing them using a common and interoperable format.
- Phase C: Continuously monitor web resources, so as to grant the continuous validity of the corresponding content labels.

Supporting the work of labeling experts

Taking into account the WMA and AQuMed approaches, as well as the phases of the content label lifecycle, the AQUA system[19] was designed to support the main tasks in the labeling process, as shown in Figure 3:

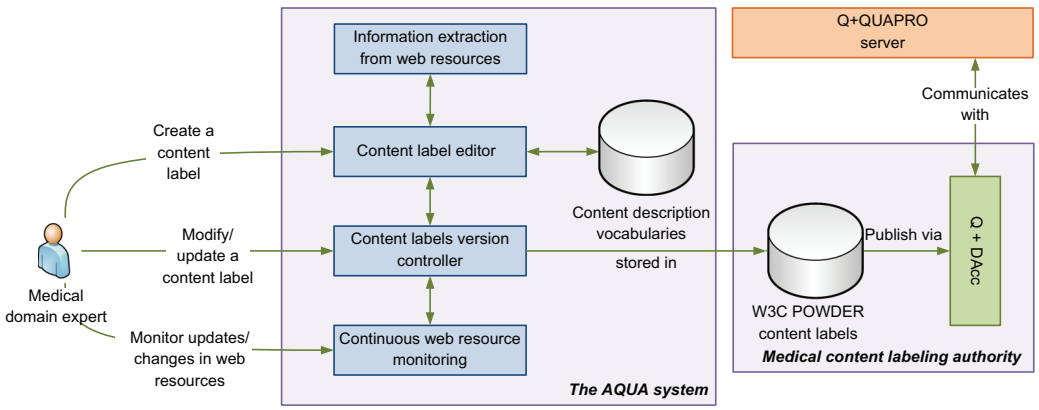

Figure 3. Key Functionalities of the AQUA system

1. Identification of unlabelled resources having health-related content.
2. Visit and review of the identified resources.
3. Generation of content labels for the reviewed resources.
4. Monitoring the labeled resources.

The AQUA system was developed within the MedIEQ project, which provides the infrastructure and the means to organize and support various aspects of the daily work of labeling experts. Compared to other approaches that partially address the assessment process,[20,21] the AQUA system is an integrated solution.

The steps towards this objective are as follows.

Step 1

Creating machine-readable labels by:

- adopting the use of the W3C Protocol for Web Description Resources (POWDER);
- creating a vocabulary of criteria, consolidating on existing ones from various Labeling Agencies (this vocabulary is used in the POWDER labels);
- developing a label management environment allowing experts to generate, update and compare content labels.

Step 2

Automating parts of the labeling process by:

- helping in the identification of unlabelled resources;
- extracting from these resources information relative to specific criteria;
- generating content labels from the extracted information;
- facilitating the monitoring of already labeled resources.

Step 3

Putting everything together; AQUA is implemented as a large-scale, enterprise-level, Web application having the following three tiers:

- the user tier, including user interfaces for the labeling expert and the system administrator;
- the application tier where all applications run;
- the storage tier consisting of the MedIEQ file repository and the MedIEQ database.

AQUA addresses a complex task. However, various design and implementation decisions helped MedIEQ partners keep AQUA extensible and easy to maintain. The main characteristics of its implementation include: a) open architecture; b) accepted standards adopted in its design and deployment; c) character of large-scale, enterprise-level web application; and, d) internationalization support.

AQUA incorporates several subsystems (see the application level in Figure 4) and functionalities for the labeling expert (for a detailed description see Karkaletsis et al.[22]). The Web Content Collection (WCC) component identifies, classifies and collects online content relative to the criteria proposed by the labeling agencies participating in the project.

The *Information Extraction Toolkit* (IET) analyzes the content collected by WCC and fills the corresponding attributes of the content labels. The *Label Management* (LAM) component

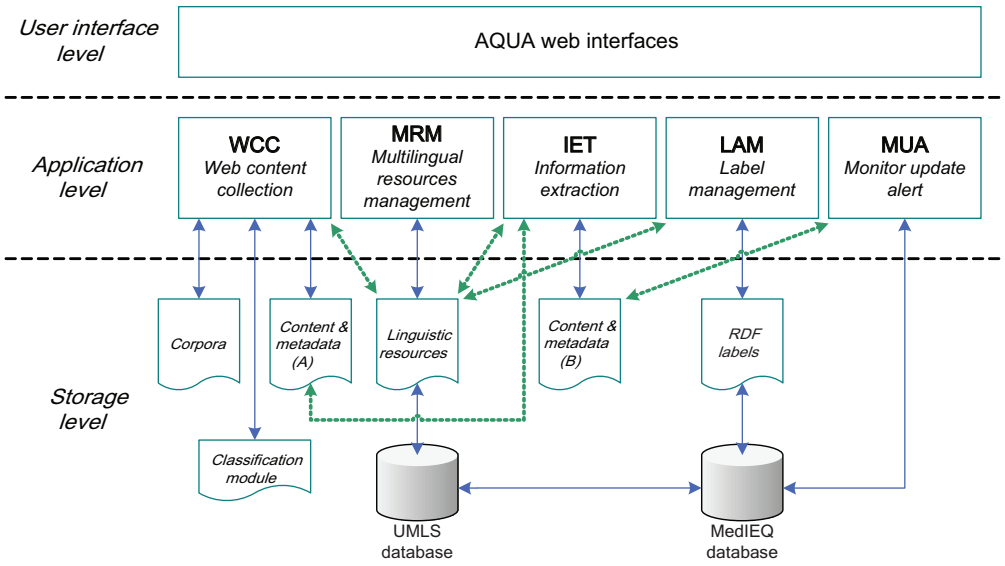

Figure 4. Architecture of the AQUA system

generates, validates, modifies and compares the content labels. The *Multilingual Resources Management* (MRM) subsystem gives access to health-related multilingual resources. In MRM we use the following vocabularies: MeSH (http://www.nlm.nih.gov/mesh/), SNOMED (http://www.ihtsdo.org/snomed-ct/), ICD10 (http://www.who.int/classifications/icd/en/) and WHO (http://www.who.int/en/). One of MedIEQ goals was to support MeSH for all project languages, namely: English, German, Spanish, Finnish, Czech, Greek and Catalan. However, this was not easily accomplished in some languages for which official MeSH translations do not exist, Greek being one of them. To this end, for the Greek language, MedIEQ contacted the Greek project IATROLEXI (http://www.iatrolexi.gr/), which has adopted the UMLS semantic network, including a version of MeSH, along with other biomedical terms created. Input from such resources is needed in specific parts of the WCC, IET and LAM toolkits. Finally, the *Monitor-Update-Alert* (MUA) tool handles auxiliary, but important, jobs like the configuration of monitoring tasks, database updates, or the alerts to labeling experts when important differences occur during the monitoring of existing content labels.

Figure 5 presents an example of a POWDER label created by an expert from WMA using AQUA, for the Web resource www.hipocampo.org. The POWDER document initially defines the vocabularies that will be used as attributes of the root *powder* element. Then, information about the label's creation is included in the required *attribution* element. The creator of the label is defined by *issuedby*, while *issued* defines the date when the label was created, and *validfrom* and *validuntil* determine the period of validity for the label. The *dr* element is used to declare the set of resources that are described and provide the actual description. The web resources to be labeled are defined inside the *iriset* element. In our case, the set of web resources contains every resource that have www.hipocampo.org as the host part of their IRI. The *descriptorset* element then provides the description that applies to every resource in the *iriset* by using the appropriate vocabularies.

```
<powder xmlns="http://www.w3.org/2007/05/powder#"
        xmlns:wma="http://wma.comb.es/powder/vocabulary"
        xmlns:quatro="http://purl.org/quatro/elements/1.0/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">

        <attribution>                                                    } LABEL CREATOR
            <issuedby src="http://wma.comb.es"/>
            <issued>2007-07-06T00:00:00</issued>
            <validfrom>2009-02-18T00:00:00</validfrom>
            <validuntil>2010-02-18T00:00:00</validuntil>
        </attribution>

        <dr>
            <iriset>                                                     } WEB RESOURCE
                <includehosts>www.hipocampo.org</includehosts>
            </iriset>

            <descriptorset>                                              } LABEL
                <wma:wmacode>275</wma:wmacode>
                <dc:identifier>http://www.hipocampo.org</dc:identifier>
                <dc:language>es</dc:language>
                <wma:otherseals>1</wma:otherseals>
                <wma:generalupdate>1</wma:generalupdate>
                <wma:hasemail>1</wma:hasemail>
                <wma:emailok>1</wma:emailok>
                <wma:healthprof>1</wma:healthprof>
                <wma:authors>1</wma:authors>
                <wma:dateinfo>1</wma:dateinfo>
                <wma:healthsource>1</wma:healthsource>
                <wma:Webstruct>1</wma:Webstruct>
                <wma:intlinks>1</wma:intlinks>
                <wma:extlinks>1</wma:extlinks>
                <wma:scicont>1</wma:scicont>
                <wma:advsci>1</wma:advsci>
                <wma:confi>1</wma:confi>
            </descriptorset>
        </dr>
</powder>
```

Figure 5. Example of a POWDER Label generated by a Medical Domain Expert

End-user exploitation of medical Content Labels

The ultimate goal of the creation and publication of Content Labels is to provide assistance to end-users who want to find reliable web resources. More specifically, end-users must be equipped with tools that enable:

- searching for certified content on a subject;
- viewing of the labels – created by Medical domain experts – associated with certified resources;
- expression of their own opinion on the content of the resource and even on the content labels;
- sharing of their opinions with other members of communities they belong to.

The Content Labels for User Empowerment project (QUATRO Plus), part of the European Safer Internet *Plus* programme, aimed to develop the appropriate tools in order to provide a complete solution for the aforementioned issues. In detail, QUATRO Plus developed tools that are able to:

1. verify and publish to the web the labels created by medical domain experts;
2. build meta-search engines that associate the returned search results with their corresponding labels, if such exist;

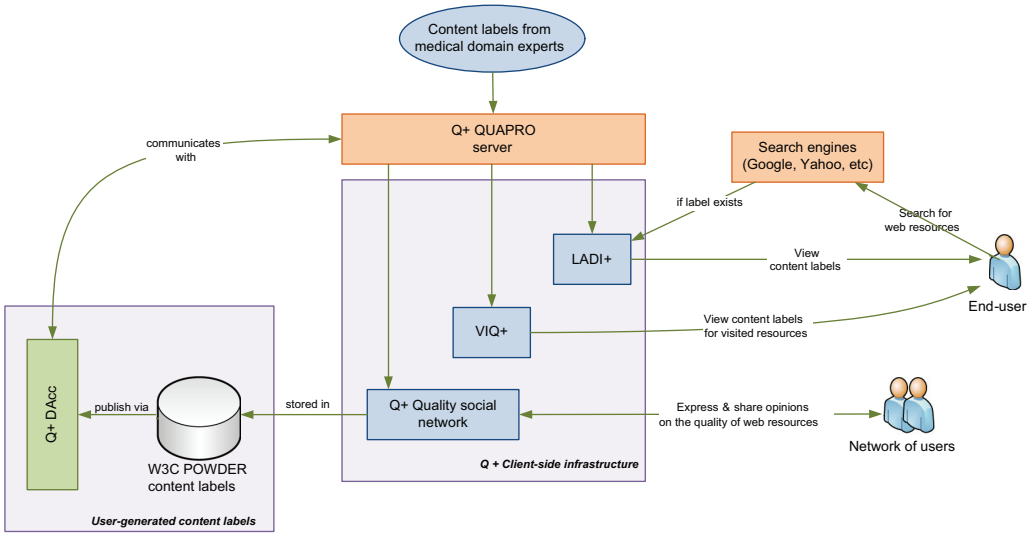

Figure 6. The QUATRO Plus infrastructure

3. view the content of the labels associated with a web resource, either from the search engine result page or through a web browser when the end-user is visiting the resource;
4. organize communities of specific interest (social networks) and give their members the functionalities required to create their own labels or ratings to existing labels, in a uniform way;
5. aggregate and publish the labels/ratings created by the members of an active user community.

Figure 6 summarizes the various QUATRO Plus components and their interaction.

Search results returned by popular engines are augmented with content certification information via the LADI+ search engine wrapper. The presence of labels is also indicated by the ViQ+ application when visiting the web resource associated with the label. Both LADI+ and ViQ+ provide visualization of the labels' content.

Finally, user empowerment is achieved by a social networking application, the Quality Social Network (QSN). The QSN includes all the basic social networking functionalities, as well as the mechanisms that allow the creation of content labels by each user and production of an aggregate label for the entire community.

The QUAPRO+ proxy server is the module responsible for accessing the label repositories of medical domain experts and those of the various user communities, verifying the validity of labels and returning their content to the client application that requested the label.

Searching for labeled web content

LADI+ (http://www.quatro-project.org/tools) is a client application that gives end-users an indication of the existence of content labels inside the web resources listed in search engine results and allows them to see more detailed information about those labels. LADI+ performs calls to the QUAPRO+ service to verify the validity of a label and to provide a summary and further details of the label's content. The availability of such information is obviously useful if the end-user wishes to know about the content of a website, its authoritativeness and its reliability before visiting the site itself.

Figure 7. Google search results appended with the LADI+ shield icon

LADI+ works on top of the Google and Yahoo! search engines. It examines the list of results returned by the selected search engine. Where one or more content labels do exist, the result is marked with a shield icon (Figure 7).

Visualise content labels

Both of the LADI+ and ViQ+ tools visualise content labels.

When using LADI+ to see the labelling information for a result marked with a shield icon, the users can click on it and a new, "pop-up" frame will appear where the label information is rendered in a human-readable form.

ViQ+ (http://www.quatro-project.org/tools) is a client application responsible for two main tasks:

1. To notify users whether a visited web resource is associated with content labels or not.
2. To display to users the content of the labels associated with the resource.

ViQ+ can be installed on any of the major browsers. As happens with LADI+, ViQ+ relies on QUAPRO+ for confirming label validity and retrieving label contents. It re-renders the content of the page the user is visiting and annotates it with a bar indicating the presence or absence of content certification labels and an active icon (Figure 8). When the icon is clicked, the labelling information is presented to the user (Figure 9).

Each label associated with the resource is presented in a different tab. Each tab is named after the certification authority or organization that created and published the label. In the tab, each vocabulary descriptor that holds a value in the label is presented by using an appropriate lexical description. For example, the *wma:hasemail* vocabulary descriptor is presented as 'Valid e-mail address' to the end-user.

Empower users to express and share their opinions

The QSN is a social network with an added dimension – quality labels.[13] It has all the standard features of a social network, such as registration, editable user profiles, message

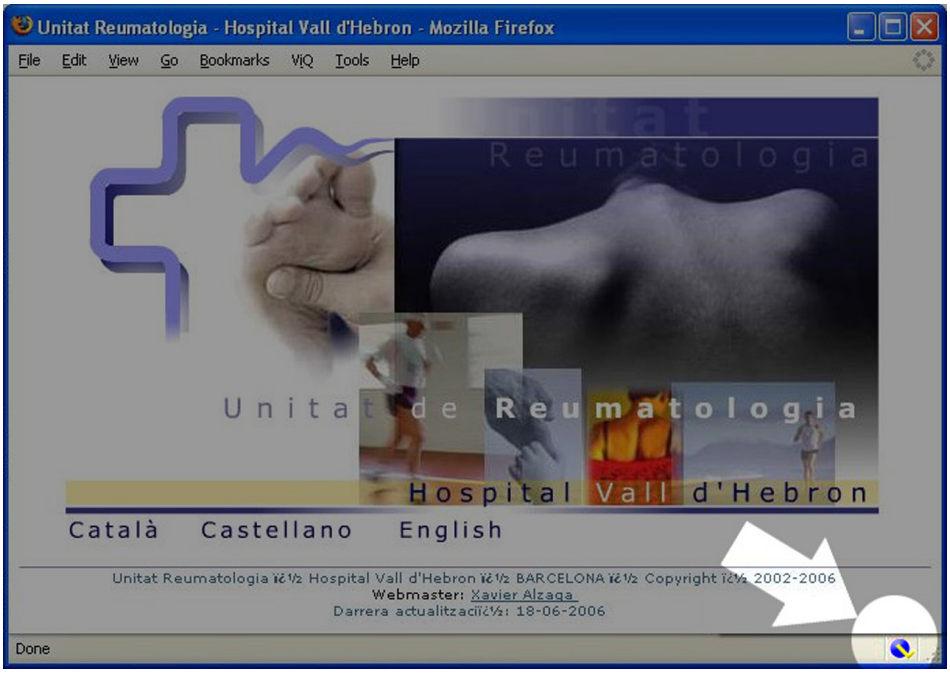

Figure 8. The blue icon (bottom-right) indicating the presence or absence of content labels

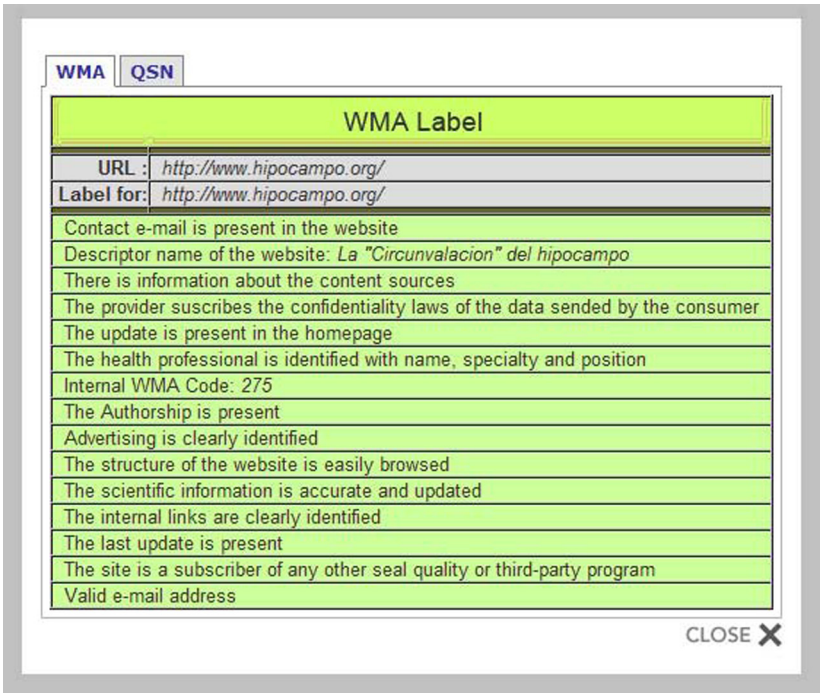

Figure 9. Labelling information presented by LADI+ or ViQ+

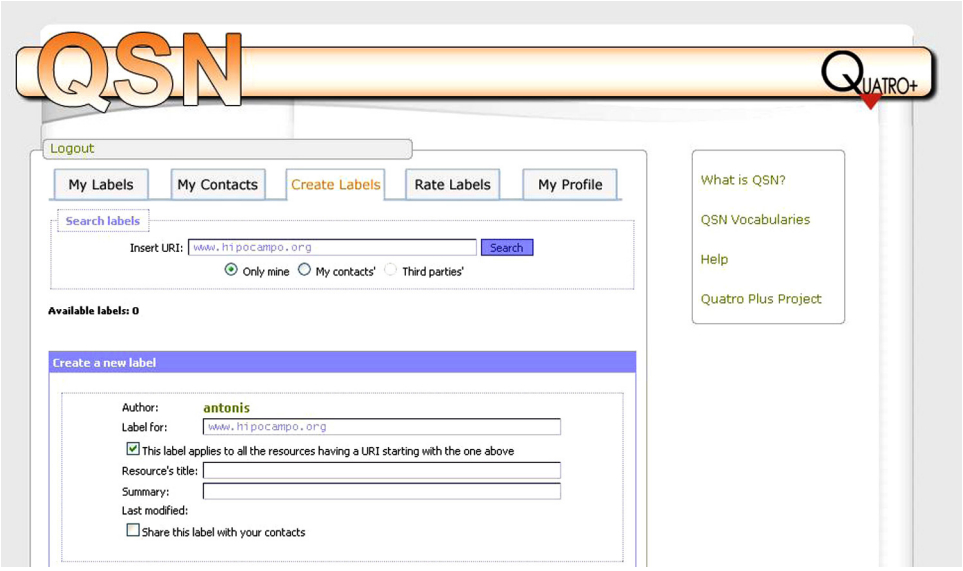

Figure 10. The 'Create Labels' section of the QSN

exchanging and establishing connections with other users. In addition, the QSN allows users to collect and store (anonymously) both user-defined labels for web resources and user-defined ratings of third-party labels. The latter can be aggregated to provide end-users with an indication of the reliability of existing labels. User-defined labels can be shared with other members of the QSN community, as can the ratings, depending on the permission levels set by the individual user.

The QSN environment is divided in distinct sections, each of them providing different functionalities to the users (Figure 10):

- 'My profile' contains the personal information submitted by the user, who can update it at any time;
- the 'My contacts' section presents a list of the members with which the user has established a relationship as well as the type of relationship. Furthermore, it offers search functionalities in order to discover other users and establish new relationships with them;
- within the 'Create Labels' section, QSN members may have access to one or more label vocabularies that they can use to create their own content labels. The user is asked to provide information regarding the web resource, such as its URI, its title and a brief summary, and then applies values to the vocabulary descriptors provided by the QSN in order to create their label. The user can also define if the created label is visible to their contacts and the rest of the QSN community.

The user-generated labels regarding the same resource can be aggregated in order to produce a label representing the opinions of the entire community. The aggregated labels can be expressed following the POWDER format and be made available to the web via the QUAPRO+ server. The aggregated QSN label contains the vocabulary descriptors selected by community users, along with the percentage of agreement within the community for each descriptor.

```
<powder xmlns="http://www.w3.org/2007/05/powder#"
        xmlns:sianpos="http://tethis.iit.demokritos.gr/vocabularies/sianpos#"
        xmlns:sianneg="http://tethis.iit.demokritos.gr/vocabularies/sianneg#">
        <attribution>
                <issuedby src="http://tethis.iit.demokritos.gr/yp/qui" />
                <issued>2009-07-20T00:00:00</issued>
                <validfrom>2009-07-20T00:00:00</validfrom>
                <validuntil>2010-07-20T00:00:00</validuntil>
        </attribution>

        <dr>
                <iriset>
                        <includehosts>hipocampo.org</includehosts>
                </iriset>

                <descriptorset>
                        <sianpos:edu>75</sianpos:edu>
                        <sianpos:cons>100</sianpos:cons>
                        <sianneg:nude>60</sianneg:nude>
                </descriptorset>
        </dr>
    </powder>
```

LABEL CREATION INFO

DESCRIBED WEB RESOURCE

QSN DESCRIPTORS /w PERCENTAGE

Figure 11. An aggregated QSN label expressed in the POWDER format

As is depicted in Figure 11, the POWDER document that is produced contains the *attribution* element with the IRI of the particular QSN community as the creator of the label and the date of the aggregation as the date when the label was issued. The *DR* element defines the resource that is described, while the actual descriptors that apply to the resource are included in the *descriptorset* element, with the percentage of agreement in the community given as the value of each descriptor. In the example, the specific community expressed the opinion that the web resources under the host www.hipocampo.org contain valuable educational content. This is indicated by including the sianpos:edu descriptor in the DR. Similarly, the community concluded that the resources contain nudity (the descriptor sianneg:nude).

Finally, in the 'Rate Labels' section, the QSN user can express their dis/agreement with existing labels by rating them. The QSN communicates with the QUAPRO+ server, in the same way as the other client applications, to retrieve labels possibly associated with a resource that the user is interested in. QSN presents the content of the label and the user can associate a rating in a specific scale to comment on the accuracy of the label. User ratings can also be expressed as POWDER labels that apply to the initial label (as this is also a web resource). Figure 12 gives an example of such a label. The POWDER document contains the required attribution which defines the creator as the specific installation of the QSN, while the *iriset* elements denotes that the description applies on the POWDER label that is being rated, using its IRI. The actual rating is the value of the *rating* descriptor. QUAPRO+ is then able to associate a label with its ratings and return the complete information to the client applications.

In the overall architecture, the QSN acts both as a data provider and a data consumer. The aggregated QSN label constructed by combining the evaluation for a resource by the network members is made available through QUAPRO+ to all the existing, and future, client applications. On the other hand, the QSN is also a client for QUAPRO+ in order to receive the labels published by the authorities and make them available for rating to its users. The aforementioned ratings can then also be (optionally) attached to the information returned by QUAPRO+ regarding the authorities' labels to which they apply.

```
<powder xmlns="http://www.w3.org/2007/05/powder#"
                xmlns:ratings="http://qsn.example.org/vocabulary/ratings#">
<attribution>
    <issuedby src="http://qsn.example.org"/>
    <issued>2009-07-05T00:00:00</issued>                    } LABEL
    <validfrom>2009-07-05T00:00:00</validfrom>              } CREATION
    <validuntil>2010-07-05T00:00:00</validuntil>            } INFO
</attribution>

<dr>
    <iriset>
           <includeresources>                                          } DESCRIBED
           http://authority.example.org/labels/powder/label.xml        } WEB RESOURCE
           </includeresources>
    </iriset>

    <descriptorset>
           <ratings:rating>2</ratings:rating>                          } RATING AS A DESCRIPTOR
    </descriptorset>
</dr>
</powder>
```

Figure 12. A POWDER label expressing the rating on an exemplary content label

Case Studies

The technologies developed have been applied in several case studies in order to prove their efficiency and impact on the labelling process, and its significance for checking, certifying and exploiting medical content. The case studies aimed to promote the standard-based labelling initiative by creating new labels, transforming existing labels to the POWDER scheme, retrieving and labelling new content in various medical fields, and using resources in various languages.

Using AQUA for semi-automatic Content Labeling

The first pilot use of the MedIEQ system (AQUA) was made by WMA in Spain and by AQUMED in Germany, as the medical partners of the project.

AQUMED used AQUA to locate unlabelled health-related web resources in English and German, and classify them according to their subject, as well as to generate POWDER conformant labels following the MedIEQ vocabulary. WMA used AQUA to transform their existing labels into the POWDER format, as well as to monitor already labeled resources.

The usage cases were directed towards assessing the efficiency and potential of the proposed framework in all the stages of the labelling process. The scope of this evaluation was the performance of AQUA in supporting the labeling process, as well as the usability of AQUA interface. The evaluation showed that by only using the links proposed by AQUA, it was possible for labeling experts to identify the right value in more than 80% of the different labeling cases. Details on the evaluation methodology and the corresponding results for supporting the work of domain experts can be found in the final public report of the MedIEQ project.[23]

Porting AQUA to new languages

A primary objective of the MedIEQ project was to offer a methodology for extending AQUA to support new languages. AQUA was originally implemented to support seven languages, namely:

English, German, Spanish, Finish, Czech, Greek and Catalan. In order to measure the effort needed to extend AQUA, work was jointly undertaken with University of Gothenburg. University of Gothenburg undertook the task of localizing in Swedish both the AQUA user interface and the internal extraction engines, using the following methodology:

- *User Interface Localization*. This is, in practice, the translation of a text message file.
- *Spider Model Training*. The training of Spider's classifiers is facilitated using a specialized tool called Corpus Formation Tool for the collection and annotation of corpus.
- *IET Model Training*. The training of information extraction models [Information Extraction Tool (IET)] is supported using the BOEMIE Annotation Tool[†], which enables the annotation of named entities and relations.
- *Topic Categorizer Configuration*. AQUA's topic categorizer employs the Automatic Ontological Concepts Extraction Tool (POKA), a general-purpose tool for automatic extraction of ontological concepts. In the current implementation of AQUA, MeSH is supported.

This experiment showed that the expected human effort for porting AQUA to a new language is less than one person for a month.

Transforming existing Content Labels to POWDER

Another primary objective of the MedIEQ project was to offer a methodology for transforming existing labels to POWDER. For testing purposes, we applied this methodology to convert the existing labels of the Health On the Net (HON) foundation to POWDER labels, following a relevant agreement with HON.

As a result, MedIEQ developed a software tool for transforming content labels stored in HON's database to W3C POWDER format. Access to HON's database was given via a specialized web service provided by HON. These labels were integrated into AQUA's label database, and a limited version of the Label Management Toolkit (LAM), requested by HON, was provided to HON to view and manage the POWDER content labels. This version is available online at: www.medieq.org:8280/aqua/seam_login.seam. Figure 13 displays a label produced by the aforementioned system. It bears all the required POWDER elements, while the descriptor set uses the terms defined by a limited version of the HON vocabulary, constructed for the use case, to assemble the description for the web resource.

Creating a white list for child nutritional disorders

The consortium of the QUATRO Plus project developed a methodology for creating white lists of quality web resources based on content analysis technologies developed in the MedIEQ project. The methodology can be applied as it is to any domain. The overall process is depicted in Figure 14.

The process is initiated with the definition of a keyword set by a domain expert. This set is used from a specialized crawler engine to obtain the first set of results that can amount to millions (depending on the domain).

For the second phase, the results are analyzed in order to discard resources that are not relevant to the topic of interest. These could be expired or modified resources, resources that contain the keywords in advertisements, image tags, etc. but not in their core content, and so on. The clean set of resources will generally contain thousands of URIs. A random sample from these is given to the

```
<powder xmlns="http://www.w3.org/2007/05/powder"
        xmlns:honcode = "http://www.hon.ch/RDF/">
<attribution>
    <issuedby>
        <foaf:Organization>
                <foaf:name>Health-On Net</foaf:name>
                <foaf:mbox>arnaud.gaudinat@healthonnet.org</foaf:mbox>
                <foaf:homepage rdf:resource="http://www.hon.ch"/>
        </foaf:Organization>
    </issuedby>
    <issued>2009-02-05T17:11:44</issued>
</attribution>

<dr>
    <iriset>
        <includehosts>www.myhealthscanner.com/</includehosts>
    </iriset>
    <descriptorset>
        <dc:identifier>http://www.myhealthscanner.com/</dc:identifier>
        <dc:title>My HealthScanner</dc:title>
        <dc:type>text</dc:type>
        <dc:Format>text/html</dc:Format>
        <dc:language>en</dc:language>
        <honcode:audience>
        Seniors,Adults,Adolescents,Medical Professionals,Individuals
        </honcode:audience>
        <honcode:status>compliant</honcode:status>
        <honcode:initialReview> 06 Dec 2006</honcode:initialReview>
        <honcode:lastReview>19 Feb 2008</honcode:lastReview>
        <honcode:countryLocation>Singapore</honcode:countryLocation>
        <honcode:mainContent>Medical/Health information</honcode:mainContent>
        <honcode:siteType>Commercial</honcode:siteType>
        <honcode:siteSize>Big (Database)</honcode:siteSize>
    </descriptorset>
</dr>
</powder>
```

Figure 13. A HON label expressed in POWDER

expert for evaluation. The expert categorizes them just as approved or disapproved without having to provide any further details or remarks. The classification is used for retraining the crawling module to obtain a set of web resources that have a higher possibility of being classified as approved. After the retraining, the results amount to hundreds or a few thousands. The process of random sampling, expert classification, retraining and refined searching can be repeated as many times as necessary in order to achieve the success ratio that is required from the interested party.

As a use case, we collaborated with the Greek Adolescence Health Unit (AHU – www.youth-health.gr), a non-profit organization that aims to provide guidance and help to children, adolescents and their parents. The desired white list should contain resources that provide valuable and accurate information about eating disorders and nutritional problems encountered in minors and teenagers. The initial set of keywords was as generic as possible and contained the following terms: children, adolescent, teenager, anorexia and obesity. Each step of the procedure gave the following results:

1. Initial search: 1,250,000 URIs.
2. Content analysis: 2000 URIs.
3. Set of randomly selected resources given to the expert for evaluation: 200 URIs.

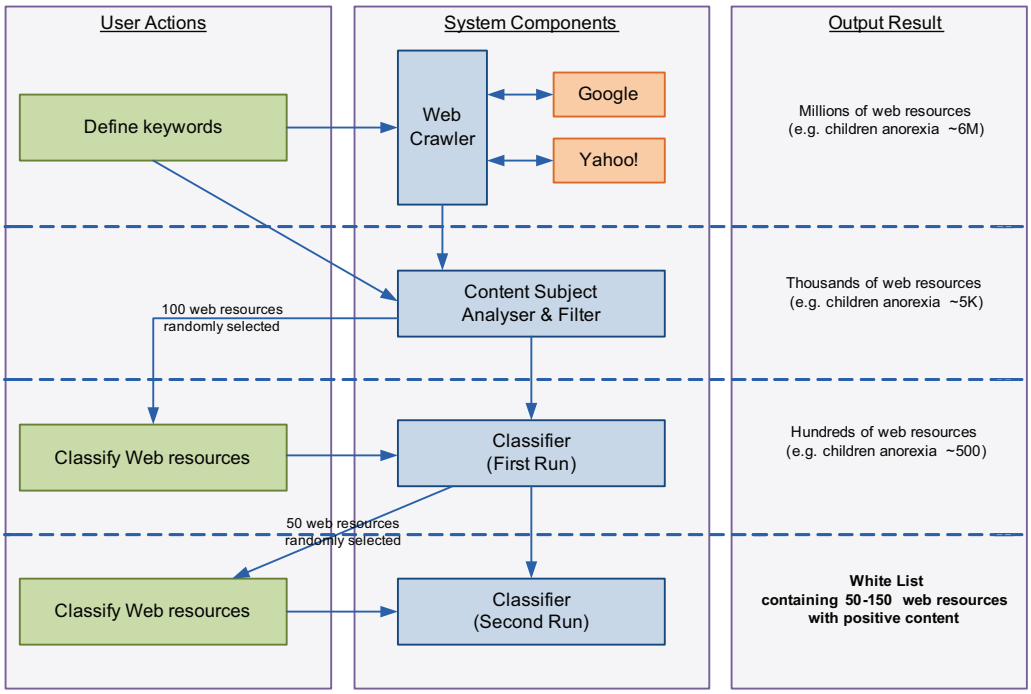

Figure 14. White list creation methodology

4. Set of resources approved by the expert: 142 URIs (71% approved).
5. Retraining and refined search: 220 URIs.
6. Second random sample of resources given to the expert: 50 URIs.
7. Approved resources: 41 (82%).
8. Second retraining and refined search: 70 resources.
9. Third classification by the expert: 65 resources (~93%).

According to the results, each re-run of the core procedure offered better results by more than 10% in comparison to the previous run, along with a significant decrease of the set that must be examined by the domain expert. The use case indicated that the desired accuracy can be obtained in a relatively short time and with little human effort.

The resources ultimately included in the white list are essentially characterized by the Adolescent Health Unit as high quality content. This approval is expressed as a POWDER label published by the organization. The vocabulary in this case contains just one descriptor that denotes the recognition of the resource as useful and valuable. Thus, if a web resource with an IRI of http://disorders.example.org is included in the white list, the corresponding POWDER label produced is presented in Figure 15. The POWDER document includes the required attribution element with information about the creator and the period of validity for the description, while the DR element defines the web resource and sets a value of 1 (i.e. true) on the single descriptor of the vocabulary.

As POWDER labels, these can be linked to QUAPRO+, in order to be published on the web and accessed by end-users via applications such as ViQ+ and LADI+.

```
<powder xmlns="http://www.w3.org/2007/05/powder#"
        xmlns:ahu="http://www.youth-health.gr/vocabulary/whitelist#">

<attribution>                                                    } LABEL CREATION INFO
    <issuedby src="http://www.youth-health.gr"/>
    <issued>2009-07-05T00:00:00</issued>
    <validfrom>2009-07-05T00:00:00</validfrom>
    <validuntil>2010-07-05T00:00:00</validuntil>
</attribution>

<dr>
    <iriset>                                                     } APPROVED WEB RESOURCE
            <includeresources>
                    http://disorders.example.org/
            </includeresources>
    </iriset>

    <descriptorset>                                              } SINGLE DESCRIPTOR
            <ahu:approved>1</ahu:approved>                         DENOTING APPROVEMENT
    </descriptorset>
</dr>
</powder>
```

Figure 15. A POWDER label for a resource included in the white list

Conclusions

The tools and platform described provide the means for content labelling of health-related web resources based on the use of machine-readable metadata descriptions, along with the access to these labels through browsers, search engines and other applications.

MedIEQ tools allow the manual or automatic assigning of standardized metadata to describe web resources, i.e. the creation of content labels, whereas QuatroPlus tools enable end-users to exploit these content labels, as well as create their own and share them within their communities. The machine-readable labels can also be exploited by different types of agents and applications, serving the purposes of policy enforcement, filtering, etc.

These tools offer a substantial improvement to the current situation for various reasons. A flexible platform that encodes the labels is created. In addition, a vocabulary is offered that encompasses the common elements of a wide variety of labelling schemes. The system supports the experts in monitoring the labelled resources, allowing the expert to create monitoring tasks. The end-user is empowered by having easy access to labelling information and being able to express their own opinion and dis/agreement. This results in greater precision and flexibility on describing web resources and promotes the idea of resource evaluation. The creation of machine-readable labels leads to new possibilities for the application of Semantic Web technologies (e.g. to make a search more precise). In addition, the development and use of well-defined standards provides the potential to make different labels highly interoperable, thus facilitating the development of generic tools and the presentation of a description that refers to different perspectives (e.g. medical, ethical or technological).

However, there are still several issues to be tackled. Medical web content publishers should be motivated to enrich their content with machine-readable labels. Such motives could be, for example, the improvement of ranking in popular general-purpose search engines (e.g. Google, Yahoo!, etc.) when a specific resource is accompanied with a label. For *Medical Web Content Certification*

Authorities, common accreditation policies for authenticating labeling authorities to certify health-related web content, at National and European level, need to be applied. Finally, as end-users' annotation of health-related web resources with opinions and comments has the potential to provide a useful additional source of information to both end-users and labeling authorities, the continued development of such tools is important.

Acknowledgments

The work presented in this paper was supported by the EC-funded project MedIEQ - Quality Labeling of Medical Web content using Multilingual Information Extraction (www.medieq.org), under the DG-SANCO Programme "Public Health", as well as the EC-funded project QUATRO Plus - Content Labels for User Empowerment (www.quatro-project.org), under the DG-INFSO Programme "Safer Internet Plus".

Notes

†. The BOEMIE annotation tool was developed for the purposes of the BOEMIE EC-funded project (www.boemie.org).

References

1. Ferguson T and Frydman G. The first generation of e-patients. *BMJ* 2004; 328: 1148–1149.
2. Mayer MA, Leis A, Sarrias R and Ruíz P. Web Médica Acreditada Guidelines: reliability and quality of health information on Spanish-language Websites. In: Engelbrecht R et al. (eds) *Connecting Medical Informatics and Bioinformatics*. Proceedings of the Medical Informatics Europe 2005; Vol. I (1), pp.1287–1292.
3. Peng Z and Logan R. Content Quality, Usability, Affective Evaluation, and Overall Satisfaction of Online Health Information. *Paper presented at the annual meeting of the International Communication Association*. New York, 25 May 2009.
4. Jansen B and Pooch U. A review of Web searching studies and a framework for futures research. *JASIST* 2001; 52 (3): 235–246.
5. Fox S. *Online Health Search 2006: Most internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find*. Pew Internet & American Project, 29 October 2006.
6. Pew Research Center. *Internet health resources*. Washington: Pew Internet and American Life Project, 2003, pp. 20–25, accessed October 2009).
7. Halkias D, Harkiolakis N, Thurman P and Caracatsanis S. Internet use for health-related purposes among Greek consumers. *Telemedicine and e-Health* 2008; 14 (3): 255–260.
8. Soualmia LF, Darmoni SJ, Douyère M and Thirion B. Modelisation of Consumer Health Information in a Quality-Controled gateway. In: Baud R et al. (eds) *The New Navigators: from Professionals to Patients*. IOS Press. Proceedings of the Medical Informatics Europe 2003, pp. 701–706.
9. Nabarette H, Romaneix F, Boyer C, Darmoni SJ, Rémy PL and Caniard E. Certification des sites dédiés à la santé en France. *Presse Med*, August 2009.
10. Wilson P. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the Internet. *BMJ* 2002; 324 (7337): 598–602.
11. Health on the Net. Analysis of 9th HON Survey of Health and Medical Internet users. Winter 2004-2005. Available online at: http://www.hon.ch/Survey/Survey2005/res.html (accessed November 2009).
12. Rice RE and Katz JE (eds). *The Internet and Health communication. Experiences and Expectations*. London: SAGE Publications, Inc. 2001.
13. Archer P, Ferrari E, Karkaletsis V, Konstantopoulos S, Koukourikos A and Perego A. QUATRO Plus: Quality you can trust? *Proceedings: Workshop on Trust and Privacy on the Social and Semantic Web (SPOT) 2009*, 2009.

14. Berners-Lee T, Hendler J and Lassilla O. The semantic Web. *Sci Am* 2001; 284 (5): 34–43.
15. Giustini D. Web 3.0 and medicine. *BMJ* 2007; 335 (7633): 1273–1274.
16. Eysenbach G, Yihune G, Lampe K, Cross P and Brickley D. MedCERTAIN - MedPICS certification and rating of trustworthy and assessed health information on the net. *Proceedings AMIA Symposium* 2000; 230–234.
17. Eysenbach G, Kohler C, Yihune G, Lampe K, Cross P and Brickley D. A metadata vocabulary for self- and third-party labeling of health Web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL). *JAMIA* (Suppl) 2001: 169–173.
18. Mayer MA, Darmoni SJ, Fiene M, Kohler, Roth-Berghofer TR and Eysenbach G. MedCIRCLE: Collaboration for Internet rating, certification, labelling and evaluation of health information on the World-Wide-Web. In: Baud R et al. (eds) *The New Navigators: from Professionals to Patients*. IOS Press: Proceedings of the Medical Informatics Europe 2003, pp. 667–672.
19. Stamatakis K, Chandrinos K, Karkaletsis V, Mayer MA, Gonzales DV, Labsky DV, et al. AQUA, a system assisting labelling experts assess health Web resources. In: *Proceedings of the 12th International Symposium for Health Information Management Research* (iSHIMR 2007), Sheffield, 18-20 July 2007, pp. 75–84.
20. Griffiths KM, Tang TT, Hawking D and Christensen H. Automated assessment of the quality of depression Web sites. *J Med Internet Res* 2005 30; 7 (5): e59.
21. Wang Y and Liu Z. Automatic detecting indicators for quality of health information on the Web. *Int J. Med Inform* 2007; 76 (8): 575–582.
22. Karkaletsis V, Stamatakis K, Karampiperis P, Labský M, Ruzicka M, Svátek V, et al. Management of Medical Web site Quality Labels via Web Mining. In: Berka P, Rauch J, Abdelkader Zighed D (eds) *Data Mining and Medical Knowledge Management: Cases and Applications*. USA: IGI Global Inc, 2009, pp. 206–226.
23. Karkaletsis V, Karampiperis P, Stamatakis K, Spyropoulos CD, Artikis A and Charou E. MedIEQ Final Report. Available online at: http://www.medieq.org/system/files/WP1+-+Deliverable+Public+D3.3.pdf, (accessed November 2009).