



1st version of RDF schema for medical content labels

Distribution: Public

MedIEQ

Quality Labeling of Medical Web content using
Multilingual Information Extraction

National Centre for Scientific Research "Demokritos"
Teknillinen Korkeakoulu – Helsinki University of Technology
Universidad Nacional de Educacion a Distancia
Col.legi Oficial de Metges de Barcelona
Zentralstelle der deutschen Ärzteschaft zur Qualitätssicherung in der
Medizin
Vysoka Skola Ekonomicka V Praze
I-Sieve Technologies Ltd

2005107 **D4.1**

October 2006

Project ref. no.	2005107
Project acronym	MedIEQ
Project full title	<i>Quality Labeling of Medical Web content using Multilingual Information Extraction</i>

Security (distribution level)	<i>Public</i>
Contractual date of delivery	<i>30 June 2006</i>
Actual date of delivery	<i>17 October 2006</i>
Deliverable number	<i>D4.1</i>
Deliverable name	<i>1st version of RDF schemas for medical content labels</i>
Type	<i>Report</i>
Status & version	<i>Final</i>
Number of pages	<i>31</i>
WP contributing to the deliverable	<i>WP4</i>
WP / Task responsible	<i>WMA</i>
Other contributors	<i>AQuMed, NCSR</i>
Author(s)	<i>Miquel Angel Mayer (WMA) Dagmar Villarroel Gonzales (AQuMed) Pantelis Nasikas, Vangelis Karkaletsis, Dan Brickley (NCSR)</i>
EC Project Officer	<i>Artur Furtado</i>
Keywords	<i>Content labelling, labelling scheme, health content labels, RDF scheme, RDF vocabularies</i>
Abstract (for dissemination)	<i>This deliverable presents our work for the development of the 1st version of the MedIEQ labelling scheme and, more specifically, the first set of labelling criteria issued, the corresponding RDF vocabulary, and use cases for different forms of health-related web content where the scheme can be exploited. A description is also given of an initial version of the user interface for the labelling expert developed to support the generation, maintenance and access of labels according to the proposed scheme.</i>

Table of contents

Executive Summary.....	5
1. Introduction	6
2. Content labeling	7
3. Describing resources with labels	9
4. How we are going to build the schema.....	14
4.1 RDF Vocabularies used for creating MedIEQ labels.	15
4.2 Label Structure	17
Label Metadata	17
Label Restrictions	17
Label Properties.....	17
5. Use Cases	18
6. Example Labels and Data access using SPARQL.....	20
6.1 Label	20
6.2 SPARQL Queries for extracting label data	22
6.3 RSS feed.....	23
7. Concluding remarks.....	25
Appendix I.....	26
Appendix II.....	31

Executive Summary

MedIEQ aims to advance current medical quality labelling technology capitalizing on the results of previous work on quality labelling and content analysis. The implementation of this objective will be based on the realisation of the following more specific objectives:

- Develop a scheme for the quality labelling of health-related web content and provide the tools supporting the creation, maintenance and access of labelling data according to this scheme;
- Specify a methodology for the content analysis of health-related web resources according to the MedIEQ scheme and develop the tools that will implement it;
- Specify a methodology and develop the tools for the creation and maintenance of the multilingual resources that will support content analysis in health-related web resources;
- Develop a prototype labelling system and demonstrate it in seven different languages and two labelling applications (third party accreditation, classification).

This deliverable presents our work for the development of the 1st version of the MedIEQ labelling scheme and, more specifically, the first set of labelling criteria issued, the corresponding RDF vocabulary, and use cases for different forms of health-related web content where the scheme can be exploited. Examples of labels are also given along with example queries to exploit the labels' content.

It must also be noted that an initial version of the user interface for the labelling expert has also been developed to support the generation, maintenance and access of labels according to the proposed scheme.

1. Introduction

According to the Technical Annex (Annex I “Description of the action”), the work in WP4 aims to achieve the following objectives:

- Extend the general vocabulary of the QUATRO project in order to create a schema for medical quality labelling, exploiting the labelling criteria established in previous relevant projects and labelling initiatives.
- Specify a methodology for exploiting the RDF labelling data in two applications that correspond to the two examined labelling mechanisms (third party rating and classification).

This deliverable presents our work for the development of the 1st version of the MedIEQ labelling scheme. More specifically, after the description of the problem of content labelling for health-related web resources (section 2), the deliverable presents:

- the approaches used so far for describing health web resources with labels, the specific approaches upon which MedIEQ work will be based, and the first set of labelling criteria issued based on the criteria currently used by the participating labelling agencies WMA and AQuMed, the eEurope criteria guidelines and a label agency of reference as HONCode (section 3),
- the initial version of an RDF vocabulary which represents the 1st set of labelling criteria, based on the scheme produced in the QUATRO project; existing RDF vocabularies are reused under this initial version of the MedIEQ vocabulary (section 4) ,
- use cases for different forms of health-related web content where the MedIEQ labelling scheme can be exploited (section 5)
- examples of labels based on the MedIEQ scheme along with example queries to exploit the labels’ content (section 6).

An initial version of the user interface for the labelling expert has also been developed to support the generation, maintenance and access of labels according to the proposed scheme (a relevant description is given in Appendix II). This belongs to the user interface for the integrated MedIEQ system that is currently being developed in the context of WP8 “System integration”.

2. Content labeling

Internet has become an important mass medium for consumers seeking health information and health care services on line. It is known that the quality of these web sites is very variable and difficult to assess. On the other hand, patients continue to find new ways of reaching health information and more than four out of ten health information seekers say the material they find affect their decisions about their health. However, it is difficult for health information consumers, to assess by themselves the quality of the information because they are not always familiar with the medical domains and vocabularies. Although there are divergent opinions about the need for accreditation of health Web sites and adoption by Internet users, different organizations around the world are working on establishing standards of quality in the accreditation of health-related web content and developing tools to rate them.

We could describe two main approaches in order to rate and improve the quality of health related content websites. The first one is based in establishing a set of quality criteria, quality guidelines or even user guidelines about the best practice with no active review of the web sites. There are several and key representatives of this approach. For instance, the recommendations of the American Medical Association¹ or the eHealth Code of Ethics². It should also be mentioned the recommendations of the European Commission in the document “eEurope 2002: Quality Criteria for Health related Websites”³ that was the result of an important effort from different organizations, professional and consumer institutions and label agencies with the aim “to draw up a commonly agreed set of simple quality criteria on which Members States, as well as public and private bodies, may draw in the development of quality initiatives for health related websites.”

Another kind of tool under this first approach is the User guidelines: these are developed to support the user to asses the quality of the websites. They consist usually in a questionnaire about quality characteristics. The most known user guidelines are DISCERN⁴ and NETSCORING⁵.

On the other hand, the second main approach is that organizations and label agencies, strictly speaking, are rating actively the websites through two mechanisms:

1. Filtering portals: the web pages are classified according to predetermined criteria and organized in groups in order to facilitate a quick access to quality reviewed information. Examples of this mechanism are the following: “Catalog and Index of French-speaking Medical Sites” (CISMEF)⁶, “Organising Medical Networked Information - The UK Gateway to reliable health information” (OMNI)⁷ and “Agency for Quality in Medicine” (AQUMED)⁸.

¹ <http://www.ama-assn.org/>

² <http://www.hi-ethics.org/>

³ http://europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf

⁴ <http://discern.org.uk/>

⁵ <http://www.chu-rouen.fr/netscoring/netscoringeng.html>

⁶ <http://www.cismef.org/>

⁷ <http://omni.ac.uk/>

⁸ <http://www.azq.de/>

2. Third party accreditation: an organization evaluates actively the quality of the website according to a set of criteria. Compliance with those criteria is showed with a logo or trust mark on the homepage. HON Code of the Health on the Net Foundation⁹, URAC Accreditation Program¹⁰, Web Mèdica Acreditada¹¹ are the most lively known quality seals.

In MedIEQ we focus in the second approach. The main problem that the mechanisms of the 2nd approach face is the need for a continuous review and control of the accredited or classified web sites that means a huge amount of human effort. WMA, as third-party accreditation system, for instance, periodically reviews manually the accredited web sites to renew the quality label. On the other hand, in AQuMED, as filtering and rating system, website directories are periodically updated due to the addition of new sites and changes in the characterization of the already visited ones.

In order for the medical quality labeling and filtering mechanisms to be successful, they must be equipped with semantic web technologies that enable the creation of machine-processable labels as well as the automation of the labeling process. Such technologies may involve information extraction techniques that allow the continuous monitoring of labeled web sites alerting the labeling agency in case some changes occur against the labeling criteria, or web crawling techniques that allow the retrieval of new unlabelled web sites, their characterization and addition in a medical thematic portal.

⁹ <http://www.hon.ch/>

¹⁰ <http://www.urac.org/>

¹¹ <http://wma.comb.es/>

3. Describing resources with labels

The quality of medical information in Internet is a feature that attracts the interest of the scientific community, particularly with regard to standardization of the description of health related resources. Web site quality evaluation has been proved to be not trivial task, as many methodologies and methods with a wide variation have appeared to date.

The World Wide Web Consortium developed a set of technical standards called PICS (Platform for Internet Content Selection)¹². PICS was originally developed to support applications for filtering out pornography and other offensive material, to protect children. PICS technology is not extended and it has showed poor deployment levels. HIDDEL¹³ (Health Information Disclosure, Description and Evaluation Language) was one of the first efforts in this area. HIDDEL evolved from medPICS, a basic rating vocabulary (rating system) for medical information conforming to the Platform for Internet Content Selection (PICS). This language HIDDEL is a common vocabulary to describe and evaluate health information on the internet. It was developed within the EU- Project MedCERTAIN (MedPICS Certification and Rating of Trustworthy Health Information on the net)¹⁴ and further refined during MedCIRCLE project (Collaboration for Internet Rating, Certification, Labelling and Evaluation of Health Information on the World-Wide-Web)¹⁵. The goal of the development and implementation of HIDDEL vocabulary was to allow portals to make the results of their quality evaluations accessible as XML/RDF. The Extensible Markup Language (XML) is a W3C-recommended general-purpose markup language capable of describing many different kinds of data. Furthermore, the Dublin Core metadata element set is a standard for cross-domain information resource description and it provides a simple and standardised set of conventions for describing things online in ways that make them easier to find. Dublin Core is widely used to describe digital materials such as video, sound, image, text and composite media like web pages. Implementations of Dublin Core are typically XML and Resource Description Framework based.

Another recent approach is the QUATRO¹⁶ project (Quality Assurance and Content Description Project). The general goal of the project is to support the Internet user by searching information of good quality in internet. QUATRO aims to provide the means for making a quality label machine-readable. As a part of the project a common (core) vocabulary for describing internet resources was developed, it can be used and adapted for quality labelling and trust mark schemes around the world. This vocabulary is not specific for the medical area, but one of the case studies concerned medical websites, it was represented by Web Mèdica Acreditada (WMA).

QUATRO adds to the picture in two ways:

¹² <http://www.w3.org/PICS/>

¹³ <http://www.medcircle.info/metadata/hiddel.php?lanxid=7187678f6b98936f717b63c525b9d97d>

¹⁴ <http://www.medcertain.org/>

¹⁵ <http://www.medcircle.org/>

¹⁶ <http://www.quatro-project.org/>

- 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.
- 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.

The complete vocabulary is available on the QUATRO project web site both as a plain text document and as an RDF schema. Labeling 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:

- A label that is machine readable and uses common descriptors will be interpreted more easily by semantic web tools than one that uses purely proprietary elements.
- A common set of elements makes it possible to apply content analysis techniques in order to automate up to some point the difficult task of ensuring that an accredited site continues to meet the labeling criteria. For example, if a labeling scheme 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 labeling operator alerted that the site needs rechecking.

The project's vocabulary is divided into four categories:

- General Criteria, such as whether the labeled site uses clear language that is fit for purpose, includes a privacy statement, data protection contact point etc.
- Criteria for labeling to ensure accuracy of information such as the content provider's credentials and appropriate disclosure of funding.
- Criteria for labeling 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.

On both counts the use of a common vocabulary offers commercial advantages to labeling operators, by increasing the value of the labels for content providers and end-users. One of the case studies in Quatro concerns the labeling of medical web sites through the involvement of the WMA labeling operator. 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.

MedIEQ continues the work of previous projects in the area of medical quality labeling (MedCERTAIN, MedCIRCLE and WRAPIN) and quality labeling standards and platform developed by QUATRO. MedCERTAIN and MedCIRCLE established a third-party rating systems to select high quality information medical websites on the Internet. WRAPIN (Worldwide online Reliable Advice to Patient and Individuals)¹⁷ is another project that its main objective was to make available a tool to determine information quality by automatically checking a document against matching sources from databases of known quality. A reference database of trustworthy sources includes MEDLINE, Clinical Trials, FDA Drug Information, UROFrance, OESO and HON's databases. The QUATRO project (Quality Assurance and Content Description) is a platform that applies semantic web technologies to trust mark schemes and quality labels.

MedIEQ aims to continue and build upon the work of these and related projects. We plan to create a schema specific for medical quality labelling by extending the QUATRO vocabulary. It will be made based on the criteria that are currently used by the participating labeling agencies WMA and AQuMed, the eEurope criteria guidelines and a label agency of reference as HONCode.

In the first step the criteria of both labelling agencies were examined in order to identify similarities and discrepancies (see Appendix I for a detailed comparison):

- Web Mèdica Acreditada (WMA) criteria are based on its Code of Conduct for health websites created by the Medical Association of Barcelona.
- AQuMed uses the criteria from Check-In - Instrument. This instrument was developed in AQuMed in collaboration with the Patient Forum of the German Medical Association. It is based on the criteria of the user guide DISCERN and on the AGREE instrument for critical evaluation of medical guidelines.

A further step was to compare WMA's and AQuMed's criteria with one of the most known labelling schemas, HON Code of Health on the Net Foundation and with the recommendations of the European Union "eEurope 2002: Quality Criteria for Health related Websites". The purpose of this comparison was to identify criteria that are not included either in WMA's or in AQuMed's schemes and that should be considered in a common set of criteria that could be the goal set of quality criteria useful and applicable for any medical label agencies.

Based on this analysis we proposed the first version of a common set of criteria to label and describe medical information in Internet (see Table 1). It consists of eleven criteria and will be expanded and refined in the course of the project.

¹⁷ <http://www.wrapin.org/>

Criteria	Definition	Comment
1. Title	The name given to the resource	
2. Resource URI	The Uniform Resource Identifier	A URI can be classified as a locator or a name or both. A Uniform Resource Locator (URL) is a URI that, identifies a resource obtainable via HTTP
3. Responsible / Author	An entity or person responsible for creating the content of the resource	
4. Contact details of the responsible / author	Additional information useful to contact the website responsible	a. Email address b. Postal address c. Telephone number
5. Last update	Date of the last review of the resource	
6. Topic / Keywords	Some keyword (s) which describe the content of the resource	For instance using the Unified Medical Language System (UMLS)
7. Resource language(s)	Language(s) of the content of the resource.	
8. Target audience	Group for whom the resource is addressed	Considered groups: a. Professional b. Non – professional 1. Adult 2. Children
9. Advertisement	Presence of any kind of advertising.	
10. Quality seal or third party program	Presence of a quality seal from a third party program or compliance with a code of conduct	Considered seals: a. WMA b. HON Code c. pWMC d. URAC e. Health Truste f. Afgis
11. Virtual consultation:	Presence of any kind of virtual consultation service using discussion forum, chat or e-mail.	

Table 1: Set of criteria for labelling of medical websites – first version.

For the next version of this set of criteria, we will consider on the one hand a specification of some of the above named criteria as a description of the limits of the different kinds of virtual consultation. On the other hand we will consider new criteria with regard to confidentiality and personal data management, information sources, sponsorship, editorial independence, accessibility, etc.

Up to now a standard RDF schema for medical web sites does not exist. MedIEQ will put forward a specific medical metadata vocabulary, making use of the experience in previous projects in this area, the EC Quality Criteria for Health Related Websites, the RDF Content labels schema that is developed in QUATRO, and other standardized vocabularies as the Dublin Core Metadata Initiative and FOAF project. Furthermore, previous initiatives didn't use web content collection and extraction technologies that enable the automation of the rating process, such as information extraction techniques that allow the continuous monitoring of labeled web sites alerting the labeling agencies in case some changes occur against the labeling criteria, alerting experts the sites content is updated against the quality criteria, thus facilitating the work of medical quality labeling agencies.

4. How we are going to build the schema

The labelling criteria for describing medical resources found on the web, that have been mentioned and analysed in section 3, need to be organized and structured in a way that will fit all the MedIEQ project architecture. Building upon the experience from the Quatro project and the workings (finished and prospect) of W3C Web Content Labelling Incubator¹⁸ and forthcoming Working Group we decided to use the RDF¹⁹ language and more specifically RDF-CL²⁰ model to create the vocabulary for the medical labels.

MedIEQ project partner NCSR is actively supporting the launch of W3C working group on Web Content Labels. This working group will continue the work already done in the Incubator group on WCL and hopefully deliver a detailed W3C recommendation on Content labels. So far it has been decided that Web Content Labels will be created using RDF . What is left to be decided is whether to follow, refine or drop RDF-CL as a model. In any case MedIEQ will adapt to the WCL-WG recommendation.

An RDF vocabulary is a set of terms organized in Classes and Properties. Classes and Properties can be constrained to certain ranges and domains and thus along with the entailment we can create a semantic model. RDF-CL creates one for Content Labels. This model has Classes and Properties to specify resources and labels for them. It can constrain the use of labels for certain parts of a resource so as to have a more fine grained description of the resource under description. It accomplishes this using a set of rules and by defining relations among the rules and and the labels. The labels properties are implemented as binary properties. RDF-CL also defines Classes and Properties for specifying label metadata. These metadata give out information and credentials for label creation and accountability. For example a third party organization can create its content labels based on RDF-CL model and add its domain specific properties. That organization would hold accountability for the labels it would create.

Project QUATRO partners have already developed four RDF vocabularies for content labeling based on RDF-CL. Each vocabulary contains Classes and Properties for labelling certain types of content. For example, the Quatro²¹ vocabulary focuses on terms for e-business, validity of information etc., whereas the WMA vocabulary on medical content.

What we aim is to reuse existing RDF vocabularies and create new Properties and Classes under a MedIEQ vocabulary only when our needs are not covered. An example could be to use existing vocabularies for label metadata such as label creator and add new terms/properties for certain types or properties of medical content.

¹⁸ <http://www.w3.org/2005/Incubator/wcl/>

¹⁹ <http://www.w3.org/RDF/>

²⁰ <http://www.w3.org/2004/12/q/doc/content-labels-schema20050704.htm>

²¹ <http://www.quatro-project.org/vocabulary/1.0/>

4.1 RDF Vocabularies used for creating MedIEQ labels.

The criteria defined in section 3 already exist as properties in the RDF vocabularies shown in the table below. The left column shows the vocabulary name and the right the URI namespace that this vocabulary has been defined and all its terms can be found.

rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
dc	http://purl.org/dc/elements/1.1/
dcterms	http://purl.org/dc/terms/
quatro	http://purl.org/quatro/elements/1.0/#
wma	http://wma.comb.es/rdf/vocabularyv01#
foaf	http://xmlns.com/foaf/0.1/#
label	http://www.w3.org/2004/12/q/contentlabel#

Table 2: RDF Vocabularies for Medieq labels

Some details about the vocabularies:

- foaf : Has classes and properties for describing people, the links between them and the things they create and do.
- rdf : Provides a number built-in types and properties for representing groups of resources and RDF statements, and capabilities for representing XML fragments as property values.
- rdfs : Used for specifying rdf vocabularies.
- dc : General metadata about content such as creation ,author etc.
- dcterms : Refinements and further encodings to dc.
- quatro : To ensure accuracy of information ,compliance with rules and legislation for e-business , operating the trust mark scheme.
- wma : Related to medical content.
- label : Allows groups of resources to be linked to a common classification , description and also rules and operators to combine groups with descriptions.

The criteria as defined in table 1 of section 3 will be mapped to the RDF properties from the vocabularies mentioned above as seen in the following table. The first column contains the criterion in plain language, the second the RDF term and the third the value it will be accepting.

Criteria	RDF Vocabulary Term	Value
1. Title	dc:title	String
2. Resource URI	label:hostRestriction	URI
3. Responsible / Author	foaf:name	String
4. Contact details of the responsible / author	dc:creator foaf:organization foaf:mbox foaf:homepage	Email address Postal address Telephone number
5. Last update	dcterms:modified	date
6. Topic / Keywords	dc:MESH	For instance using the Unified Medical Language System (UMLS)
7. Resource language(s)	dc:language	String for language and language code
8. Target audience	wma:target	Considered groups: Professional Non – professional Adult Children
9. Advertisement	quatro:ac	boolean
10. Quality seal or third party program	wma: otherseals	Considered seals: WMA HON Code pWMC URAC Health Truste Afgis
11. Virtual consultation:	wma:virtcons	boolean

Table 3: RDF terms for labelling criteria

4.2 Label Structure

The label will be structured in three main parts :

1. Label Metadata
2. Label Restrictions
3. Label Properties

Label Metadata

It will be referring to the file that stores several labels and their restrictions. It will be holding information about who and when created the labels ,when they were last modified , details about the label authority itself ,the vocabulary it is responsible for. Terms will be used from *dc* , *dcterms*, *foaf* and *label* vocabularies.

Label Restrictions

This sections will be holding mappings of labels to hosts and more detailed rules for explicitly linking parts of websites of resources to labels. Also elements for combining these rules will be available. The terms come from the *label* vocabulary.

Label Properties

Label properties will be expressed using the criteria from table 3. Each different label for a resource will have its own set of properties.

5. Use Cases

To give a better understanding of Medieq system's intended use we provide a set of use cases. These use cases aim describe common internet usage scenarios for discovering medical related content and demonstrate how a labelling scheme would be of great value to internet users.

Use case 1: Trustmark Scheme operator to content portal

The Example Medical Trustmark Scheme reviews online articles/resources about health, providing a trustmark for those that meet a set of published criteria. The scheme operator wishes to make its trustmark available as machine readable code as well as a graphic so that content aggregators, search engines and end-user tools can recognize and process them in some way.

The trustmark operator maintains a database of sites it has approved and makes this available in two ways:

First, the labelled site includes a link to the database. This can be achieved in a variety of ways such as an XHTML Link tag, an HTTP Response Header or even a digital watermark in an image. A user agent visiting the site detects and follows the link to the trustmark scheme's database from which it can extract the description of the particular site in real time.

Secondly, the scheme operator makes the full database available in a single file for download and processing offline.

Since the actual data comes directly from the trustmark scheme operator, it is not open to corruption by the online trader and can therefore be considered trustworthy to a large degree. To reduce the risk of spoofing, however, the data is digitally signed.

This use case fits well with WMA.

Use case 2: Website to end-user

A local physician and IT enthusiast, makes her medical related materials available through her personal website. She adds metadata to her material that describes the subject areas it covers and identification information for authors. In order to gain wider trust in her work she submits her site for review by her local medical accreditation authority and a trustmark scheme. Both reviewers offer her a digitally signed, machine-readable version of their trustmark that she can add to her site. She merges these into a single pool of metadata to which she adds content descriptors from a recognized vocabulary. She adds her own digital signature to the metadata. The set of digital signatures allow user-agents to identify the origin of the various assertions made. As in use case 2, links from the content itself point to this metadata.

Since the metadata is on the website itself, user agents are unlikely to take the assertions made in the metadata at face value. Unlike the trustmark operator, the local

authority does not operate a web service that can support the label, it does, however, digitally sign its labels and publishes its public key on its website. This can be used to verify that it is indeed the local education authority that issued the relevant data in the label.

Separately, a user-agent can interrogate the trustmark operator's database in real time to check whether the physician is authorized to make the assertions relevant to their namespace. Furthermore, the use of a recognized vocabulary for the content description means that a content analyser trained to work with that vocabulary can give a probabilistic assessment of the accuracy of the relevant data.

Taken together, these multiple sources of data can provide confidence in the quality of the content and the local authority trustmark which is not directly testable. The multiple data sources may be further supported by recognising that her work is cited in many online bookmarks, blog entries and postings to medical-related message boards.

In this case the local medical accreditation authority can be AQUAMED.

Use Case 3: Rich Metadata for RSS/ATOM

A Labelling Authority has a website that offers reviews of articles for everyday health topics (nutrition , etc) for childrens, teenagers and young people . The site is summarized in both RSS and ATOM feeds. Most of the articles reviewed have an ratings from different labelling organizations (each one expert on certain field). However, the Labelling Authority (L.A.) includes reviews of some articles rated from independent experts as well , which are declared at the item level and override the channel level metadata.

Separately, another Authority's web service combines the L.A.'s and other review feeds to provide alternative reviews of the articles by transforming the ATOM feeds into RDF and creating an aggregate view using SPARQL queries.

WMA already produces rss feeds with new labelled websites.

6. Example Labels and Data access using SPARQL

6.1 Label

What follows is an RDF/XML serialised version of an RDF label. Namespace declarations lines for the RDF vocabularies used start with the `xmlns` string.

The next part of the label, encapsulated within the `rdf:Description` tag holds the label metadata. It reveals who the label creator is when it has been first issued, when modified and finally a URL for the namespace that this authority is responsible for.

What follows is the basic host restrictions within the `label:Ruleset`. All the hosts that the default label is associated with, are listed under the `label:Hosts`.

Closing this part we find the `label:hasDefaultLabel` pointing to another resource, through a relative URI. That resource is the third part of our labels file and holds the properties for the default label.

Note how we link the different parts through of the file using implicit and explicit URI notation thus creating a graph model, serialised in XML.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:quatro="http://purl.org/quatro/elements/1.0/#"
  xmlns:wma="http://wma.comb.es/rdf/vocabularyv01#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/#"
  xmlns:label="http://www.w3.org/2004/12/q/contentlabel#">

  <rdf:Description rdf:about="">
    <dc:creator>
      <foaf:organization>
        <foaf:name xml:lang="en">MedIEQ</foaf:name>
        <foaf:mbox rdf:resource="mailto:pnas@iit.demokritos.gr" />
        <foaf:homepage rdf:resource="http://www.medieq.org" />
      </foaf:organization>
    </dc:creator>
    <dcterms:issued>2006-5-18</dcterms:issued>
    <dcterms:modified>2006-05-29</dcterms:modified>
    <label:authorityFor>http://www.medieq.org</label:authorityFor>
  </rdf:Description>

  <label:Ruleset rdf:ID="Ruleset">
    <label:hasHostRestrictions>
```

```
<label:Hosts>
  <label:hostRestriction>www.medieq.org</label:hostRestriction>
</label:Hosts>
</label:hasHostRestrictions>
<label:hasDefaultLabel rdf:resource="#label_1"/>
</label:Ruleset>

<label:ContentLabel rdf:ID="label_1">
  <rdfs:comment>Label for all/most of website</rdfs:comment>
  <dc:title>The Resource's Title</dc:title>
  <dcterms:modified>2006-05-29</dcterms:modified>
  <dc:creator>
    <foaf:organization>
      <foaf:name xml:lang="en">An Author</foaf:name>
      <foaf:mbox rdf:resource="pnas@iit.demokritos.gr" />
      <foaf:homepage rdf:resource="http://url.to.the.resource.com" />
    </foaf:organization>
  </dc:creator>
  <!-- content publisher -->
  <dc:publisher>MedIEQ</dc:publisher>
  <!-- content language -->
  <dc:language xml:lang="en-US">English</dc:language>
  <dc:MESH>
    <rdf:Bag>
      <rdf:li>Quality of Health Care</rdf:li>
    </rdf:Bag>
  </dc:MESH>
  <!--Target group for this publication -->
  <wma:target>unk</wma:target>
  <!-- Adevrtising Present -->
  <quatro:ac>1</quatro:ac>
  <!--Is the site a subscriber of any other seal quality or third-party
  program-->
  <wma:otherseals>0</wma:otherseals>
  <!--There is a service of virtual consultation for health user-->
  <wma:virtcons>1</wma:virtcons>
</label:ContentLabel>
</rdf:RDF>
```

6.2 SPARQL Queries for extracting label data

Five SPARQL queries that extract information from certain parts of that label , follow.

The first query returns a human readable version of an rdf element's vocabulary for the Catalan language

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?x ?label ?comment ?t
WHERE {
    ?x rdfs:label ?label .
    ?x rdfs:comment ?comment .
    ?x rdf:type ?t .
    FILTER ( lang(?label) = "ca-es" ) .
}
```

The following query returns the properties :

1. Has Advertisement
2. Content Language
3. Target Audience
4. Other seals presence
5. Existence of virtual consultation

for the default label "label_1".

```
PREFIX label: <http://www.w3.org/2004/12/q/contentlabel#>
PREFIX wma: <http://wma.comb.es/rdf/vocabularyv01#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX quatro: <http://purl.org/quatro/elements/1.0/#>
SELECT *
{
    ?g dc:creator ?z .
    FILTER REGEX( str(?g) , \"label_1\")
    ?g a label:ContentLabel ;
        quatro:ac ?has_advert ;
        dc:language ?language ;
        wma:target ?target ;
        wma:otherseals ?otherseals ;
        wma:virtcons ?virtcons }
```

Next query returns label metadata and identification information for the creator and a namespace for his/her labelling scheme. Label metadata for our labelling scheme come from the foaf vocabulary.

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX label: <http://www.w3.org/2004/12/q/contentlabel#>
SELECT ?y ?z
WHERE {
  ?x label:authorityFor "http://www.medieq.org" .
  ?x dc:creator ?o .
  ?o ?y ?z }
```

This query returns the default label for a web resource.

```
PREFIX label: <http://www.w3.org/2004/12/q/contentlabel#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?host ?hostrest ?label
WHERE {
  ?x rdf:type label:Ruleset .
  OPTIONAL { ?x label:hasHostRestrictions ?host . }
  OPTIONAL { ?x label:hasDefaultLabel ?label . }
  OPTIONAL { ?host label:hostRestriction ?hostrest . } . }
```

This query returns the rules for identifying other labels for a web resource. The rules are perl5 regular expressions.

```
PREFIX label: <http://www.w3.org/2004/12/q/contentlabel#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?has_URI ?label ?rule ?u
WHERE { ?x rdf:type label:Ruleset .
  OPTIONAL { ?x label:rules ?rule . }
  OPTIONAL { ?rule rdf:Description ?u . }
  OPTIONAL { ?u label:hasLabel ?label . }
  OPTIONAL { ?u label:hasURI ?has_URI . } . }
```

6.3 RSS feed

We give an example rss feed with links to RDF content labels for each feed item. The format is rss 1.0 which is RDF based and this helps us manipulating the feed with a SPARQL query as said in use case 3. An option can be combination and filtering of several feeds based on their contents and their labels. This feature is under development yet so consider this as a first example and nothing more.

This example feed is adapted from WMA's rss feed . It lists two items ,for our case two newly labelled websites. Each item has an empty description element, a title and

an `rdfs:seeAlso` element that points to the file that holds the RDF labels for that resource. The RDF link is imaginary and only serves the purpose of providing a first example.

```
<rss:channel rdf:about="http://www.w3.org/2001/sw/Overview.rss">
  <rss:title>Noticias de Web Mèdica Acreditada (WMA)</rss:title>
  <rss:link>http://wma.comb.es</rss:link>
  <rss:description>Noticias diarias para el mundo de la acreditación
    mèdica</rss:description>
  <rss:items>
    <rdf:Seq>
      <rdf:li rdf:resource="http://www.mundosanitario.es"/>
      <rdf:li rdf:resource="http://www.crecerjuntos.com.ar"/>
    </rdf:Seq>
  </rss:items>
</rss:channel>
```

```
<rss:item rdf:about="http://www.mundosanitario.es">
  <rss:title>Nueva WMA:</rss:title>
  <rdfs:seeAlso rdf:Resource="http://www.mundosanitario.es/labels.rdf"/>
  <rss:description>
</rss:description>
</rss:item>
```

```
<rss:item rdf:about="http://www.crecerjuntos.com.ar">
  <rss:title>Nueva WMA: Crecer Juntos</rss:title>
  <rdfs:seeAlso rdf:Resource="http://www.crecerjuntos.com.ar/labels.rdf"/>
  <rss:description>
</rss:description>
</rss:item>
```

7. Concluding remarks

We've given a description of our approach to Content Labelling. Our interest is towards machine readable labels that provide a foundation for semantic web applications. Cooperation with expert organizations on labelling help us model the task of Content Labelling and make it comprehensible to machines and programs to automate tasks such as inference and decision making on the interest of the average internet user.

After giving an overview of the history of content labelling and approaches for medical content more specifically, we listed a first set of criteria whose identification can be automated in web resources. Those criteria have been put under a formal context (RDF) and we give a first version of a set of guidelines to use this formal context for creating medical content labels.

This work will evolve, more criteria will be added and user tools are already under heavy development to make the user interaction with content labelling as friendly as possible and closer to the average user's world.

Appendix I

Comparison AQuMED / WMA Criteria

At the beginning of the project AQuMED and WMA had a joint meeting (March 15-16, Barcelona) where they compared their quality criteria in order to identify their differences and similarities. These are depicted in the table below. Similarities are marked in grey and differences in red. It must be noted that AQuMED decided to use Check-in instrument instead of DISCERN because it's easier to apply in MedIEQ in terms of machine readable format.

The major outcome of this comparison was the following:

- A major part of the criteria are similar and they could be unified in one proposal (a common RDF Schema).
- The main differences were: AQuMED does not consider as criteria the “medical virtual consultation” issues and WMA doesn't consider details of the “medical treatments” issues.
- The point is (it's already an open issue) if to create “two different versions” considering the differences or not.

AQuMED Criteria	WMA Criteria	
<i>CHECK - IN</i>	A. Identification	Description
Additional questions for information in Internet IN3. Is it possible to contact the author and webmaster? a. Address, telephone number and/or e-mail of author b. Address, telephone number and/or e-mail of webmaster	1.1. E-mail address	E-mail of contact in the website easily visible
	1.2. Valid e-mail address	The e-mail is a valid address
	2. Person in charge e-mail address	E-mail address in the request form
Participation of interested groups 3. Is/are the author/s mentioned? a. Name of the author b. If an institution is mentioned as author, a responsible hast to be specify 4. Is/are the qualification of the author/s mentioned? a. Short description of the professional career of the author/s b. It is indicated, where the author actually works c. It is clear and explicit describe that different authors from different fields participated in the development of the information Additional questions for information in Internet IN1. Information about the information provider an his aims? a. Address of the provider (“Impresssum”) b. Aims of the provider are describe (“About us”)	3. Person in charge	Person in charge is identified
	4. Health care professional	The health care professional is identified with name, speciality and position
	5. Website category	Type of the website in the request form

	6. Website name	Name website
Area of application and aims 1. Is it exactly describe, which is the goal of the publication? a. Is there an introduction, which describes which topics are treated? b. Is there is detailed table of content?	7. Website description	Website description: target, topics
Correctness of the development 12. Is it specified, if the publication participates in a quality initiative (Hon Code, MedCIRCLE)? a. Logo at the homepage b. Logo hast to be active	8. Another Seal of quality or third-party program	The website is a subscriber of any other seal quality or third-party program
	B. Contents	
Correctness of the development 8. Date of creation	9. The general last update	The last general update of the website is present in the homepage
Participation of interested groups 3. Is/are the author/s mentioned? a. Name of the author b. If an institution is mentioned as author, a responsible hast to be specify	10. Authorship of the Documents	Authorships of the document are present in the website
Correctness of the development 8. Date of creation	11. Date of the last update of the documents or information provided	Update of the documents
Correctness of the development 6. Does the publication base on scientific sources? a. Is there a list of sources with a reference at the end of the publication (including date of publication)? b. List of the references at the end of the publication. c. The sources are not listed, however there is an advice that the sources can be obtain from author (only if the address of the author is mentioned)	12. Sources of health contents	Sources of health contents (documents)
Design 22. Is the most important content easy to identify? (e.g. summary, important recommendations) 23. Is it the information comprehensible? a. Glossary or explanation of medical vocabulary b. Sentence structure is easy (8-10 words) c. User friendly navigation d. Tables and graphics	13. Structure of the website	The structure of the website is easily browsed
	13.1. Accessibility	The accessibility priority level based on WAI is A, AA or AAA
	14. Internal links	The internal links are clearly identified
Correctness of the development 13. Hast the publication enough additional support information or	15.1. External links	The external links are clearly identified

new sources?	15.2 External link	There is information about the website link (a description of the type of the website you can link)
Area of application and aims 2. Is it exactly describe, which is the goal of the publication? a. Is there an introduction that describes which topics are treated? b. Is there is detailed table of content?	16. Target audience	Who's the target audience of the website
	17. Website language	
Correctness of the development 14. Does the publication describe how the treatment/procedure works? 15. Does it describe the benefits of the treatment/procedure? a. Verbal description (with a reference) b. Statistic measures (with a reference) 16. Does it describe sufficiently the risks of the treatment/procedure? c. Verbal description (with a reference) d. Statistic measures (with a reference) 17. Is it specified, if the treatment/procedure affects daily life? a. Mention b. Description of the effects c. Personal opinion whether the description is good enough 18. Does it describe, if there are contradictory experiences (results) with the application of the treatment/procedure? 19. Is it stated, that all of the established treatments for this disease are mentioned? 20. Is it describes, what happened if no treatment is use? a. Description of the natural course of the disease	18. Scientific content	"First look" of the scientific content
	C. Confidentiality	
Additional questions for information in Internet IN2. Does the provider give information about protection and handling of personal data? a. It is explain, if and for which purpose personal data collected b. It is explain, how does with these data be dealt	19. Information about use of consumer data	General policy about the use of the personal data sent
	20. Request form	Information about handling of personal data in each request of the website

<p>Additional questions for information in Internet</p> <p>IN2. Does the provider give information about protection and handling of personal data?</p> <p>a. It is explain, if and for which purpose personal data collected</p> <p>b. It is explain, how does with these data be dealt</p>	21. Confidentiality	The provider subscribes the confidentiality laws of the data sent by the consumer
	D. Advertising and Sponsorship	
<p>Exclusion criteria of AQuMED: if the document has advertising, is not rated</p>	22. Advertising present	The website contains advertising
	23. Distinguished from the scientific content	The advertising information can be distinguished by the users easily from the scientific content
<p>Editorial independence</p> <p>21. Is the information independent and neutral?</p> <p>a. The financing is exposed, sponsor</p> <p>b. Explanation that the sponsor do not have influence on the content of the publication</p> <p>c. Neutral formulation</p>	24. Sponsor	If there is sponsorship
	25. Policy sponsor	Information about the sponsor policy
	E. Virtual Consultation	
	26.1. Virtual consultation users	There is a service of virtual consultation for health user
	26.2. Limits of this service	This service has some limits
	26.3. Identification	The virtual consultation is a healthcare service and it's compulsory the identification of the professional
	27.1. Virtual consultation professionals	There is a service of virtual consultation for professionals
	27.2. Limits of this service	This service has some limits
	27.3. Identification	It's compulsory the identification of the professional
	28.1. Chat	It can be a service moderated or not (specify)

	28.2. Limits of this service	There is warning about the limits and use of this service
	29.1. News	It can be a service moderated or not (specify)
	29.2. Limits	There is any warning about the limits and use of this service
	G. Non compliance	
	Detecting a misuse of the seal	
Participation of interested groups		
5. Do patients and/or self-help groups participate in the development of the information ? a. There is an explanation that patients were involved in the development of the information b. Experiences from patients are part of the information c. It is noted that patients/consumer read the information before its publication and they has the opportunity to make comments		
Correctness of the development		
7. Is the kind of scientific sources mentioned (evidence level)? 9. Validity of the information a. Publication date and an advice that it will be update, if new know ledges are available b. Question 10 is positive 10. Date of the next update 11. Is there an indication that the information was develop according to certain quality guidelines (e.g. DISCERN)? a. The author states that he followed quality criteria by creating the information b. The criteria are describe or there is a link to the criteria	Points 9 and 10 could be optional in WMA	
Additional questions for information in Internet		
IN4. Is the access to the information without limitations? a. Without password and free b. Logo IN5. Can the information be continuous printed? a. PDF b. Link	Although WMA consider if there is a information with / without limitations, take into account that AQuMED rates documents and not the whole website	

Appendix II

Label Management Tool (LAM)

A label management tool is under development and a first version had already been released at <http://medieq.iit.demokritos.gr:8080/medieq-lam/create.seam>.

This tool is web based and in its final version it will provide an interface for label creation, editing, comparison, validation and persistence. It will also give the option to add new labelling vocabularies and build forms from the administration interface.

We work to make a user friendly interface and hide from the user the complexities of the RDF-CL label model. We are using forms, links and buttons laid out in such a manner that will be consistent with the workflow and logical model of labelling that an expert might have in his/her mind. There are dynamic lists, tables to input and edit data as well as buttons to clear and delete fields (e.g. the case of adding host restrictions). Labels properties will be grouped so as not to have long and tedious to fill forms. After creating the label the user can select the persistence method and then save it to the server and/or download it locally to attach it to any web resource he/she wishes.

Each user will have his/her own workspace where she will be able to edit and update old labels, to upload new label files, to validate and persist in the system's database. For RDF label validation we will be using and extending the ICS-Forth's VRP²² validator.

The LAM tool will be accepting input not only from humans but also from other modules in the MedIEQ system as well. Information Extraction systems and classifiers will be extracting labelling criteria from requested resources and then fill forms and create labels semi-automatically. Work on integrating all these tools is underway.

²²<http://139.91.183.30:9090/RDF/VRP/>