AQUA, a system assisting labelling experts assess health web resources

Konstantinos Stamatakis[1], Konstantinos Chandrinos[2], Vangelis Karkaletsis[1] Miquel Angel Mayer[3], Dagmar Villarroel Gonzales[4], Martin Labsky[5], Enrique Amigo[6], Matti Pöllä[7]

[1]National Center for Scientific Research (NCSR) "Demokritos", Institute of Informatics & Telecommunication, Athens, Greece, {kstam, vangelis}@iit.demokritos.gr

[2]I-sieve Technologies ltd., Athens, Greece, chandrinos@i-sieve.com

[3]Web Médica Acreditada (WMA), Barcelona, Spain, mmayer.wma @comb.es

[4]Agency for Quality in Medicine (AQuMed), Berlin, Germany, villarroelgonzales@azq.de

[5]University of Economics, Prague, Czech Republic, labsky@seznam.cz

[6]Universidad Nacional de Educacion a Distancia, Madrid, Spain, enrique@lsi.uned.es

[7]Helsinki University of Technology (TKK), Espoo, Finland, matti.polla@tkk.fi

Users visiting health related web sites would be served best if they knew whether these sites meet a minimum level of quality standards. However manually labelling health resources is a tedious task. Based upon state-of-the-art technology in the areas of semantic web, content analysis and labelling, the MedIEQ project integrates existing technologies and tests them in a novel application: AQUA, a system aiming to automate parts of the labelling process in health-related web content. AQUA provides tools that enable the creation of machine readable labels, tools that crawl the web to locate unlabelled health web resources, suggest labels for them according to pre-defined labelling criteria and monitor them. This paper describes the current status in the area of health information labelling and explains step-by-step how AQUA paves the way towards the automation of the labelling process.

Keywords

automated content assessment, content labels, semantic web technologies

1. Introduction

Various players, such as government institutions, consumer and scientific organisations, patient associations, health provider institutions, commercial organizations and also individuals, having different motivations and backgrounds, offer health information through the internet. Inevitably, the quality of the provided information is very variable and difficult to assess.

At the same time, on the patients' side, searching for health information on the Web has become a popular activity [1]. The importance of internet access in decisions affecting health can be gleaned from several surveys and studies published in the last years: eighty percent of American internet users, or some 113 million adults, have searched for information on at least one major health topic according to a survey [2]; at the same time more than four out of

ten health information seekers say the material they find affects their decisions about their own health [2]. However, health information on the Internet may be misleading or misinterpreted, compromising health behaviours and health outcomes, or resulting to inappropriate requests for clinical interventions [3].

Organisations around the world are working on establishing standards of quality in the accreditation of health-related web content [4, 5, 6, 7]. However the establishment of codes of conduct or ethics is not enough in the health domain. Self-adherence to such codes is nothing more than a claim or a pledge with little enforceability. The establishment of rating mechanisms being a necessity, various rating initiatives appeared in the recent years. These initiatives focus on credible characterization of the content of health related web sites; a number of them is acting as third party accreditation authorities issuing accreditation trustmarks or seals [8, 9, 10, 11] once the content conforms to specific criteria, while the rest maintain portals where health web sites are organised and characterised against certain criteria [12, 13, 14]. Both approaches (seal issuing or characterization and thematic organization) could be defined by the more general concept "content labelling" and the organizations behind them can be considered "labelling authorities" or "labelling agencies".

The major bottleneck of content characterization or content labelling, as applied nowadays, is the lack of standardization and automation in most parts of the labelling process. Few disjointed approaches exist [15, 16] addressing though the assessment workflow partially. The labelling agencies, in order to preserve and enhance their authority in the health content assessment environment, should continuously review and control an everyday increasing number of health web sites, a process requiring a huge amount of human effort as most work is currently manual. Existing rating approaches must be equipped with technologies that enable the automation of the rating process, such as information extraction techniques that allow the continuous monitoring of labelled web sites alerting the labelling agency in case some changes occur against the labelling criteria, or web content collection techniques that allow the retrieval of new unlabelled web sites, their characterisation and addition in a health thematic portal.

Considering the above, the EC-funded project Quality Labelling of Medical Web content using Multilingual Information Extraction (MedIEQ) [17] is developing AQUA. AQUA stands for Assisting Quality Assessment. Its technology is expected to have a significant impact on content labelling, assisting the work of labelling experts, increasing the number of labelled health sites across Europe and their effective monitoring, and thus improving the quality health knowledge disseminated through the Web.

Section 2 describes the current situation in labelling of online health information and outlines how AQUA can serve the different approaches and needs of labelling authorities. Section 3 presents the 1st set of test-bed criteria for AQUA. Section 4 provides a step-by-step analysis of our actions towards the partial automation of the labelling process. Our concluding remarks and future steps are given in section 5.

2. Current Situation in Labelling Health Information

As outlined above, two major approaches currently exist concerning the labelling of health information in the internet: a) filtering portals (organizing resources in health topics and providing opinions from specialists on their content) and b) third party accreditation (issuing accreditation trustmarks or seals once the content conforms to certain principles). In general, and in both approaches, the labelling process comprises three tasks that are followed entirely or partially by most labelling agencies:

- Identification of new web resources: this could happen either by active web searching or by voluntary application from the information provider, i.e. the web site responsible asks actively for a review, usually in order to get an accreditation seal.

- Labelling of the web resources: this could be done with the purpose of awarding an accreditation seal or in order to classify and index the web resources in a filtering portal.
- Re-reviewing or monitoring the labelled web resources: this step is necessary to identify changes or updates in the resources as well as broken links and to verify if a resource still deserves to be awarded an accreditation seal.

This is the general case; eventually, any particular agency can integrate additional steps which may be necessary in its work.

The two labelling agencies participating in MedIEQ, Agency for Quality in Medicine (AQuMed) [12] and Web Mèdica Acreditada (WMA) [9] represent both the approaches mentioned above: AQuMed maintains a filtering portal while WMA acts as a third party accreditation agency.

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

1. New medical patient information: there are two ways through which a new resource can be identified for indexing in AQuMed database. The first one is through internet search and the second one is through a direct request from the information provider. The web sites are selected according to general criteria: content, form and presentation should be serious, authorship, sponsorship and creation/update date should be clear, and only web sites without commercial interest should be indexed.
2. Web site classification: previously unlabelled web sites are classified in four groups: treatment information, background information, medical associations/scientific organisations and self-help/counselling organisations. Only the sites with treatment information go to the next step.
3. Evaluation: Sites with treatment information are evaluated with DISCERN [18] and Check-In [19] 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 [20] 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. Feed back to the information provider: AQuMed sends an e-mail with the result of the evaluation in case of sites with treatment information and with the information about the admission in AQuMed database in case of the other categories.

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

A complete accreditation process in WMA 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 to express acceptance of these recommendations;
2. The WMA Standing Committee assesses the web site based on the WMA criteria and issues recommendations;
3. WMA sends a report to the person in charge who implements the recommendations;
4. When the recommendations are 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 web sites and an RDF file is generated.

Taking into account WMA and AQuMed approaches, AQUA was designed to support the main tasks in their labelling processes, more specifically:

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 labelled resources.

Other attempts to automatically assess health information in the internet exist but address partially the assessment process. Automated quality assessment procedure (AQA) [15] ranks depression websites according merely to their evidence-based quality. Automatic indicator detection tool (AIDT), presented in a recent study [16], is suggested as a complementary instrument for the assessment of health information quality. AIDT is evaluated upon the automatic detection of pre-defined indicators that correspond to a number of technical quality criteria. However, AIDT focuses on extraction techniques and does not address the assessment process as a whole.

3. MedIEQ Labelling Criteria

As regards the labelling criteria, a first set of 11 criteria, to examine our methodology and test our tools, was decided, by the labelling authorities participating in the project consortium. This set of criteria (which will soon expand to include additional quality aspects[1]) is shown in table 1.

For each of these 11 criteria, AQUA aims to identify and extract relevant information and propose this information to the expert (i.e. automatically provide information, otherwise manually searched for). Then, the expert can accept or modify AQUA suggestions and generate a quality label on-the-fly.

Table 1 First set of criteria examined in MedIEQ.

	Criterion	Description
1	Resource title	A title clearly identifying the resource
2	Resource URI	No redirects, no masked domains
3	Resource last update	Must be present and clearly stated
4	Resource language(s)	The language(s) of the resource
5	Resource topic/keywords	The medical topic(s) of the resource's content
6	Target audience	The intended audience of the resource
7	Resource responsible name(s)	Must be present and clearly stated
8	Resource responsible contact details	Must be present and clearly stated
9	Virtual consultation (VC)	Is there any VC service available?
10	Advertising	Is advertising present and clearly separated from editorial content?
11	Other seal	Are there other seals or trustmarks assigned to the resource?

4. The AQUA Solution

Compared to other approaches that address partially the assessment process [15, 16], the AQUA system is an integrated solution. AQUA aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labelling experts by making them computer assisted. The steps towards this objective are the following:

Step 1: Creating machine readable labels; this is achieved by:

[1] The final set of criteria will be announced through the project website: http://www.medieq.org

- Adopting the use of the RDF model [21] for producing machine readable content labels; at the current stage, the RDF-CL model [22] is used (CL stands for Content Labels).
- Creating a vocabulary of criteria, consolidating on existing ones from various Labelling Agencies; this vocabulary is used in the machine readable RDF labels.
- Developing a label management environment allowing experts to generate, update and compare content labels.

Step 2: Automating parts of the labelling 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 labelled 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 the user interfaces for the labelling expert and the system administrator.
- The application tier, where all applications run.
- The storage tier, including the MedIEQ file repository and the MedIEQ database.

Details on the above steps are provided in 4.1, 4.2, and 4.3 that follow.

4.1 Creating Machine Readable Labels

4.1.1 Making Labels Machine Readable

To make content labels machine readable the use of the RDF model is adopted. At the current stage, the RDF-CL model is used. The RDF-CL model is issued by the EC-funded project Quality Assistance and Content Description (QUATRO) [23]; it is going to be refined to a new model (and this will be completed already before the end of the MedIEQ project) by the recently initiated W3C Protocol for Web Description Resources (POWDER) working group [24].

4.1.2 Proposing a Vocabulary of Criteria

The final vocabulary to be proposed by MedIEQ for use in these RDF labels will exploit the criteria[2] that are currently used by the participating labelling agencies WMA and AQuMed, the eEurope criteria guidelines for health related websites [6] and a labelling agency of reference such as Health On the Net Foundation (HON) [8]. To express these criteria, existing vocabularies are re-used where possible. Our aim is not to suggest the MedIEQ vocabulary as "the only one to use" but to show instead the value of machine readable labels. The emphasis is on the technology and not on the terms (criteria) included in the vocabulary. The terms selected capture important aspects of health related content and form the case study for MedIEQ technology partners. The two labelling authorities in the project (WMA, AQuMed) will evaluate the impact of MedIEQ's technology on their work, examining its usefulness in practice.

4.1.3 Developing a Label Management Environment

MedIEQ develops a label management interface and tools, called LAM, allowing experts to generate, update and compare content labels. LAM is also integrated within AQUA. We note that from within the LAM user interface someone is able to a) generate new RDF labels from information automatically extracted by other AQUA tools, b) manually fill the relevant fields

[2] In section 3, Table 1 we only see the first set of criteria. The final set of criteria will be announced through the project website: http://www.medieq.org

and generate new RDF labels, c) edit and update existing RDF labels, and d) compare RDF labels between them.

The user interface to generate/edit a label is a web form (see Figure 1) with input boxes, single and multiple select boxes, links and buttons. It is split in 4 distinct areas. The top three reflect the three logical parts of an RDF-CL-based content label: they are filled either with information automatically extracted by AQUA tools or manually by the user himself.

The first part of the label generation/editing form asks from the user to fill the input boxes with label metadata. The fields to be filled hold data for identifying who created the label, when the label was created and altered and who is accountable for using certain RDF vocabularies.

The second part lets the user constrain the application of the label to certain hosts by explicitly declaring the host URIs or by adding regular expressions that properly identify them. Multiple hosts can be defined. Regular expressions for more fine-grained addressing can be defined as well. These definitions can be combined via Union and Intersection and thus create rules that link different parts of a web resource with different labels.

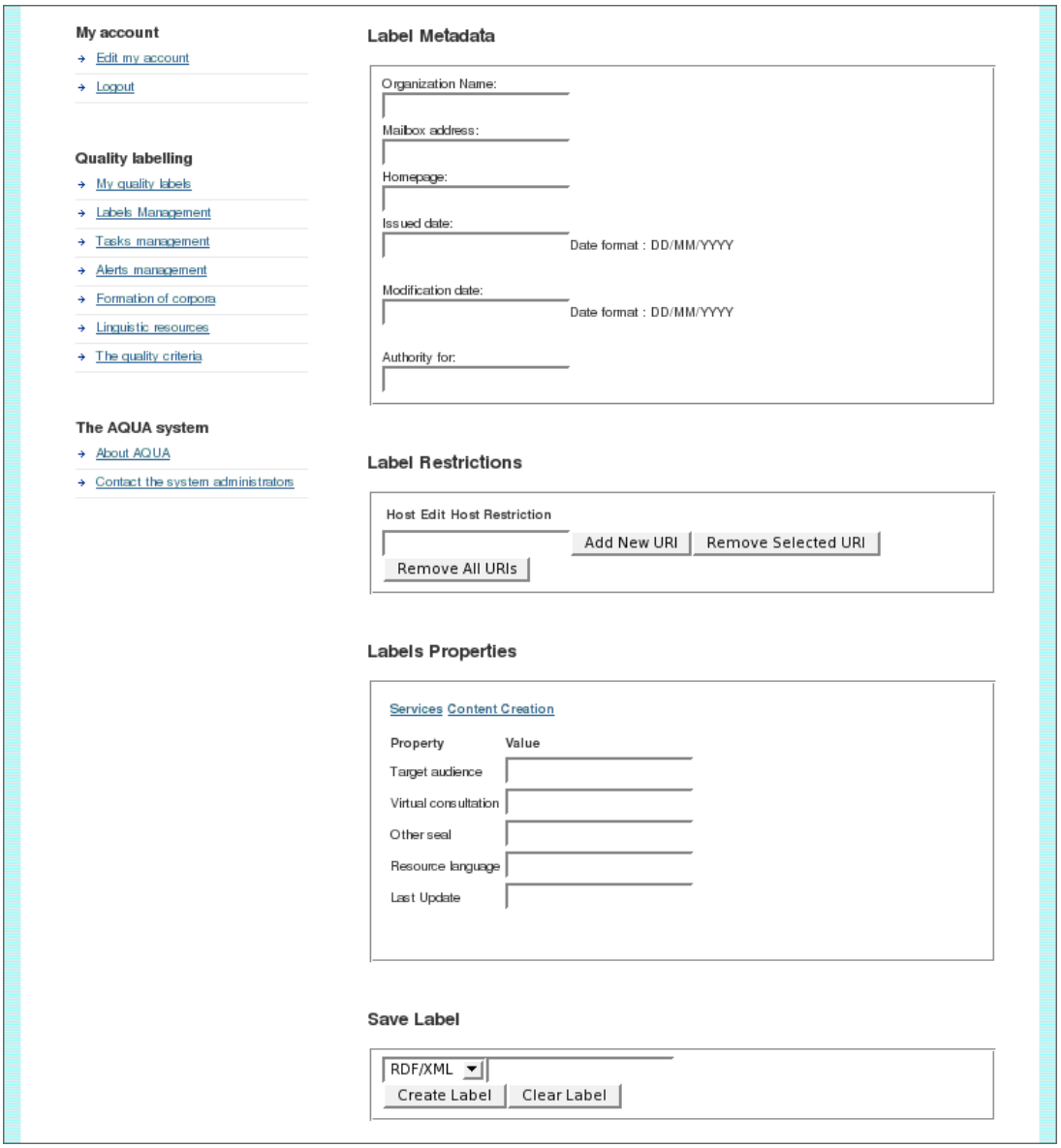

Figure 1 The AQUA label management environment (LAM).

The third part is where the label properties are assigned values. The label properties are the actual descriptors of a web resource, mapping the labelling criteria. A set of label descriptors can be linked with a set of host restrictions defined in the second part. Related properties are grouped to help the user in filling them.

Once the user has filled the label metadata, restrictions and properties, he can save the label. There is a notification field that informs the user if the label already exists in the system and its changes are tracked from the AQUA version control system. In this case the user can save the label as a revision of an existing label. If this is a new label, the user just selects to save the label. In both cases, the user has the option to download an RDF/XML serialized form of the label. This serialized label can be assigned to the web resource by the site webmaster.

4.2 Automating Parts of the Labelling Process

4.2.1 Locating Unlabelled Web Resources

The AQUA crawling mechanism is part of the Web content collection environment (WCC). Its AQUA interface is shown in Figure 2. The Crawler searches the Web for health related content, which doesn't already have a content label (at least not a label found in MedIEQ records). It is a meta-search-engine, exploiting results returned from known search engines and directory listings from known Web directories. All collected URLs from all sources are merged and filtered and a pre-final URLs list is returned. The merging / filtering process: a) removes possible duplicates, b) ignores sub-paths of URLs already in list, and c) removes URLs having already a content label (the Crawler consults the MedIEQ database for this).

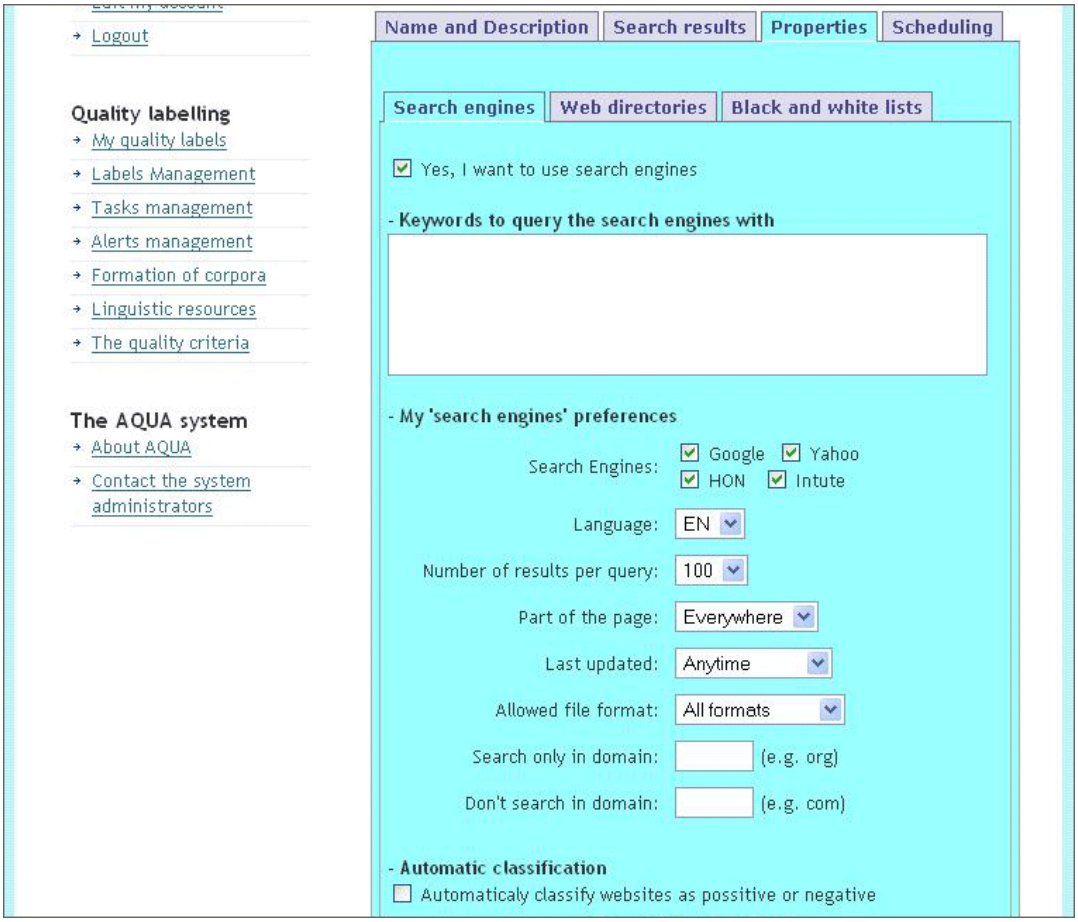

Figure 2 Configuring the MedIEQ Crawler from the AQUA interface.

The crawling process becomes even more focused with the aid of a content classifier, trained to distinguish health from non-health content. This classification component visits every URL from the merged / filtered pre-final URL list and checks its contents, thus filtering out some more entries.

The current version of the AQUA Crawler queries Google and Yahoo! search engines (with terms proposed by the user) and explores Web directories (proposed again by the user). By using merely general purpose search engines, the Crawler inevitably inherits their misses. Therefore, aiming to further enhance our Crawler we also include two specialized to the health domain searching mechanisms, one provided by HON [8] and another by Intute's Health and Life Sciences branch [13] (see Figure 2).

4.2.2 Extracting Information Relative to Criteria

MedIEQ continues and builds upon the work of previous projects in the area of information extraction (IE) [25, 26, 27, 28]. The AQUA IE toolkit (IET) employs a set of components responsible for the extraction of the parts of information found in each document and the integration of these parts in a set of useful instances. IET is being designed as a generic information extraction toolkit which should make it easy to incorporate any changes or additions to the utilised labelling schemes. In this way, IET could also be used for IE using third-party labelling schemes and within different domains. The IET components implement documented APIs to ensure they can be invoked by components from other AQUA toolkits or even by software other than AQUA.

4.2.3 Generating Content Labels from the Extracted Information

To allow labelling experts generate, update and compare RDF content labels AQUA incorporates LAM, the label management environment described in 4.1.3.

4.2.4 Monitoring of Already Labelled Resources

Another part of AQUA, called MUA (from Monitor-Update-Alert) handles things like the configuration of monitoring tasks, the necessary MedIEQ repository updates and the alerts to labelling experts when important differences occur during monitoring of previously labelled sites. MUA complements the content collection and extraction toolkits, WCC and IET respectively. MUA enables a user to create and configure monitoring tasks and to configure the rules for the maintenance of the MedIEQ database (both automatic and manual updating mechanisms are provided); finally, MUA provides a configurable mechanism alerting the user in case the content of the monitored resources is updated against the quality criteria.

4.3 Putting everything together: the AQUA system

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) large-scale, enterprise-level web application, and d) internationalization support.

AQUA incorporates several subsystems (see application level in Figure 3) and functionalities for the labelling expert. Web Content Collection (WCC) identifies, classifies and collects online content relative to the criteria (proposed by the labelling agencies participating in the project). Information Extraction Toolkit (IET) analyses the web content collected by WCC and extracts attributes for MedIEQ compatible content labels. Label Management (LAM) generates, validates, modifies, compares content labels based on the schema proposed by MedIEQ. Multilingual Resources Management (MRM) subsystem gives access to health-related multilingual resources; input from such resources is needed in specific parts of the WCC, IET and LAM toolkits. Finally, Monitor-Update-Alert (MUA) handles auxiliary but important jobs, like the configuration of monitoring tasks, the MedIEQ database updates, the

alerts to labelling experts when important differences occur during monitoring existing content labels. Figure 1 shows all the possible data flows in AQUA (dashed arrows): a) From WCC to IET: pages collected by WCC, once undergone a first level extraction by WCC (extraction of metadata 1), are then forwarded to IET for further processing (extraction of metadata 2); b) From IET to MUA: MUA takes all metadata collected by both WCC and IET and updates the MedIEQ database; c) From MRM to WCC, IET, LAM: custom vocabularies generated by the MedIEQ users through MRM interface, can be accessed from other toolkits (WCC, IET, LAM), where the user may need them.

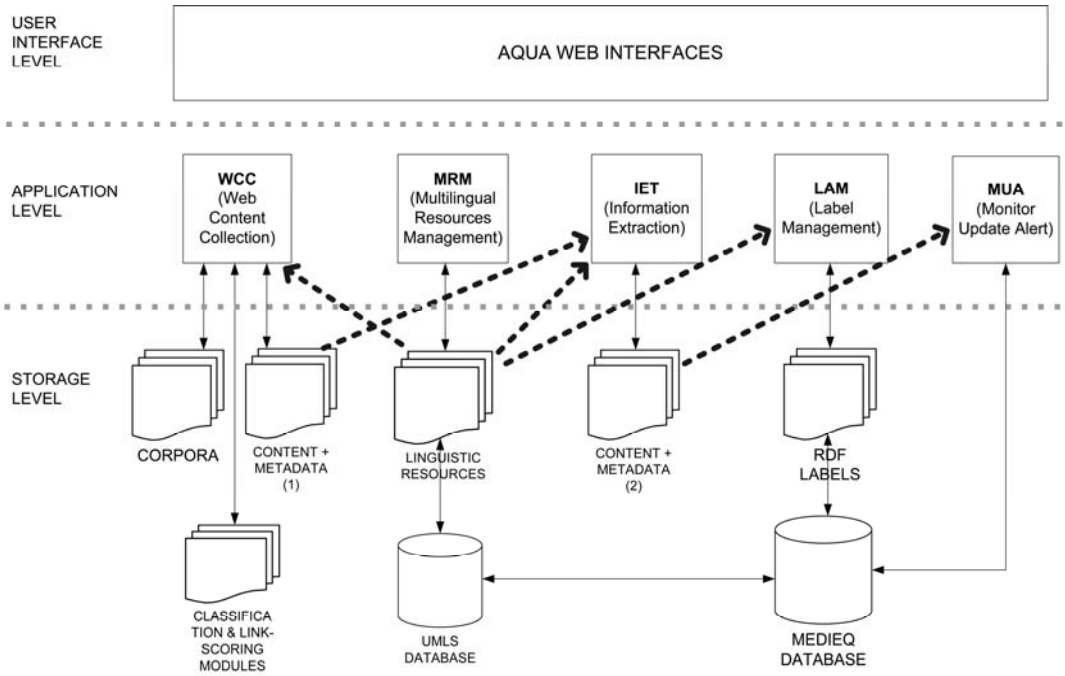

Figure 3 Architecture of the AQUA system.

5. Concluding Remarks

Assessing health information published in the internet is a complex task as it has to examine the conjunction of a number of different aspects. Various initiatives around the world have attempted to codify these aspects into criteria, principles, codes of conduct, etc. Health specialists review online health resources and label them, either by issuing accreditation trustmarks or by including them in a thematic health portal. However this work can be proven quite tedious even for experienced users. Additionally, as it currently relies on manual effort, it is very time consuming: the labelling authorities participating in MedIEQ project claim that an average expert can effectively review no more than five health resources per week. Instruments to assist certain parts of the work exist; however they focus on specific problems and none of them addresses the assessment process as a whole.

In this context, MedIEQ introduces the standardization of the content labels in the health domain by making them machine readable, while AQUA provides the infrastructure and tools to create and exploit these machine readable labels. Additionally, MedIEQ proposes a vocabulary of criteria and AQUA incorporates tools that automatically identify and extract information relative to these criteria.

AQUA's first version is scheduled to be delivered by the end of June 2007. It will cover the first set of eleven labelling criteria in two languages, English and Spanish. It is going to be tested in everyday labelling practice by experts from both labelling agencies participating in

MedIEQ, between July and September 2007 according to the project's evaluation strategy. Evaluation results will guide future enhancements in the tools performance and interface usability.

Acknowledgements

The work described in this paper is undertaken in the context of the EC-funded (DG-SANCO) project MedIEQ (Quality Labelling of Medical Web Content using Multilingual Information Extraction. Project, website: http://www.medieq.org/).

References

1 Lenhart A, Horrigan J, Rainie L, Allen K, Boyce A, Madden M, et al. The ever-shifting Internet population. a new look at Internet access and the digital divide. Washington, DC: Pew Internet & American Life Project. 2003 Apr 16. http://www.pewinternet.org/reports/pdfs/PIP_Shifting_Net_Pop_Report.pdf 2007 April.
2 Fox S, Rainie L. The online health care revolution: how the web helps Americans take better care of themselves. http://www.pewinternet.org/PPF/r/26/report_display.asp 2007 April.
3 Eysenbach G, Köhler C. Does the internet harm health? Database of adverse events related to the internet has been set up. BMJ 2002 Jan 26;324(7331); p. 239.
4 Internet Healthcare Coalition. http://ihealthcoalition.org.
5 Health Internet Ethics: Ethical Principles for Offering Internet Health Services to Consumers. (http://www.hiethics.com/Principles/index.asp 2007 January).
6 European Commission. "Quality Criteria for Health Related Websites". http://europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf 2007 April.
7 Winker M, Flanagin A, Chi-Lum B, White J, Andrews K, Kennett R, DeAngelis C, Musacchio R. American Medical Association. Guidelines for Medical Health information Sites on the Internet. (http://www.ama-assn.org/ama/pub/category/1905.html 2007 April).
8 Health On the Net Foundation. http://www.hon.ch 2007 April
9 Web Mèdica Acreditada (WMA). http://wma.comb.es 2007 April
10 Mayer MA, Darmoni SJ, Fiene M, Kohler C, Roth-Berghofer TR, Eysenbach G. MedCIRCLE: Collaboration for Internet rating, certification, labelling and evaluation of health information on the World-Wide-Web. In The New Navigators: from Professionals to Patients. R. Baud et al. (Eds). IOS Press. Proc MIE 2003; p. 667-72.
MedCIRCLE project website: http://www.medcircle.org/ 2007 April.
11 American Accreditation Healthcare Commission (URAC). http://www.urac.org/ 2007 April.
12 Agency for Quality in Medicine (AQuMed). http://www.aezq.de or http://www.patienten-information.de 2007 April.
13 Intute, Health & Life Sciences. http://www.intute.ac.uk/healthandlifesciences/ 2007 April.
14 Catalogue et Index des Sites Médicaux Francophones (CISMeF). http://www.cismef.org/ 2007 April.
15 Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. J Med Internet Res. 2005 Dec 30;7(5):e59.
16 Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the Web. Int J Med Inform. 2006 May 31.
17 Quality Labelling of Medical Web content using Multilingual Information Extraction (MedIEQ). Project website: http://www.medieq.org/ 2007 April.
18 DISCERN: Quality criteria for consumer health information. http://www.discern.org.uk/ 2007 April.
19 http://www.patienten-information.de/content/informationsqualitaet/informationsqualitaet/images/check_in.pdf 2007 April.
20 Appraisal of Guidelines Research and Evaluation (AGREE). AGREE instrument: http://www.agreecollaboration.org/instrument/ 2007 April.
21 Resource Description Framework (RDF). http://www.w3.org/TR/rdf-schema/ 2007 April.
22 RDF-Content Labels (RDF-CL). http://www.w3.org/2004/12/q/doc/content-labels-schema.htm 2007 April.

23 Quality Assurance and Content Description **(**QUATRO). http://www.quatro-project.org/ 2007 April.
24 Protocol for Web Description Resources **(**POWDER). http://www.w3.org/2007/powder/ 2007 April.
25 Cross-lingual multi-agent retail comparison (Crossmarc). http://www.iit.demokritos.gr/skel/crossmarc/ 2007 April.
26 Reusable Architecture for Intelligent Brokering Of Web information access (Rainbow). http://rainbow.vse.cz/descr.html 2007 April.
27 Karkaletsis V, Spyropoulos CD, Grover C, Pazienza MT, Coch J, Souflis D. A Platform for Cross-lingual, Domain and User Adaptive Web Information Extraction. In Proceedings of the European Conference in Artificial Intelligence (ECAI); 2004; Valencia, Spain; p. 725-9.
28 Labsky M, Svatek V. Information Extraction with Presentation Ontologies. In: ESWC'06 Workhshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro; 2006 June.