

The Assisting Quality Assessment (AQUA) system

P. Karampiperis^a, V. Karkaletsis^a

^a*Software & Knowledge Engineering Lab, Institute of Informatics and Telecommunications,
National Center for Scientific Research (NCSR) "Demokritos",
P. Grigoriou & Neapoleos, 15310 Aghia Paraskevi Attikis, Greece
{pythk, vangelis}@iit.demokritos.gr*

Extended Abstract

The ever increasing amount of available health-related information on the web creates excellent conditions for self-education of patients and better education of physicians, but entails substantial risks if such information is trusted irrespectively of low competence or even bad intentions of its authors. This is why medical web resources certification (also called 'quality labeling') by renowned authorities is of high importance. In this respect, the labeling process could benefit from the employment of web mining and information extraction techniques, in combination with methods of web-based information management developed within the semantic web initiative. Achieving such a synergy is the central issue in the EC-funded project MedIEQ "Quality Labelling of Medical Web content using Multilingual Information Extraction"¹.

The AQUA (Assisting QUality Assessment)² system, developed within the MedIEQ project, aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts. MedIEQ enables, through AQUA, the issue of machine-readable labels. Such labels enable the label data to be processed so that it can be displayed as text by a web browser or as logos displayed next to search results so that end users can see immediately which web sites have been reviewed by labelling authorities before the sites are visited.

Two major approaches currently exist concerning the labeling 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 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:

- *Identification* of new web resources: this could happen either by active web searching or on the request of the information provider.
- *Labeling* of the web resources: this could be done with the purpose of awarding a certification seal or in order to classify and index the web resources in a filtering portal.
- *Re-reviewing* or *monitoring* the labeled web resources: this step is necessary to identify changes or updates in the resources and to verify if a resource still deserves to be awarded the certification seal.

This is the general case; eventually, any particular agency can integrate additional steps which may be necessary in its work. The two labeling agencies participating in MedIEQ, Agency for Quality in Medicine – AQuMed (<http://www.aeqzq.de>) and Web Mèdica Acreditada - WMA

¹ <http://www.medieq.org>

² <http://www.medieq.org/aqua/welcome.seam>

(<http://wma.comb.es>), represent the two approaches mentioned above: AQuMed maintains a filtering portal while WMA acts as a third-party certification agency.

Taking into account WMA and AQuMed approaches, the AQUA tool (Stamatakis et. al., 2007) was designed to support the main tasks in their labeling 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 labeled resources.

Compared to other approaches that partially address the assessment process (Griffiths et. al., 2005; Wang & Liu, 2006), 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 labeling experts by making them computer-assisted. The steps towards this objective are the following:

Step 1: Creating machine readable labels by:

- Adopting the use of the RDF model (W3C, 2004) for producing machine-readable content labels; at the current stage, the RDF-CL model (W3C, 2005) is used. In the final version of AQUA, another model called POWDER, introduced by the recently initiated W3C Protocol for Web Description Resources (POWDER) working group (W3C, 2007), will be supported.
- Creating a vocabulary of criteria, consolidating on existing ones from various Labeling 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 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 the 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.

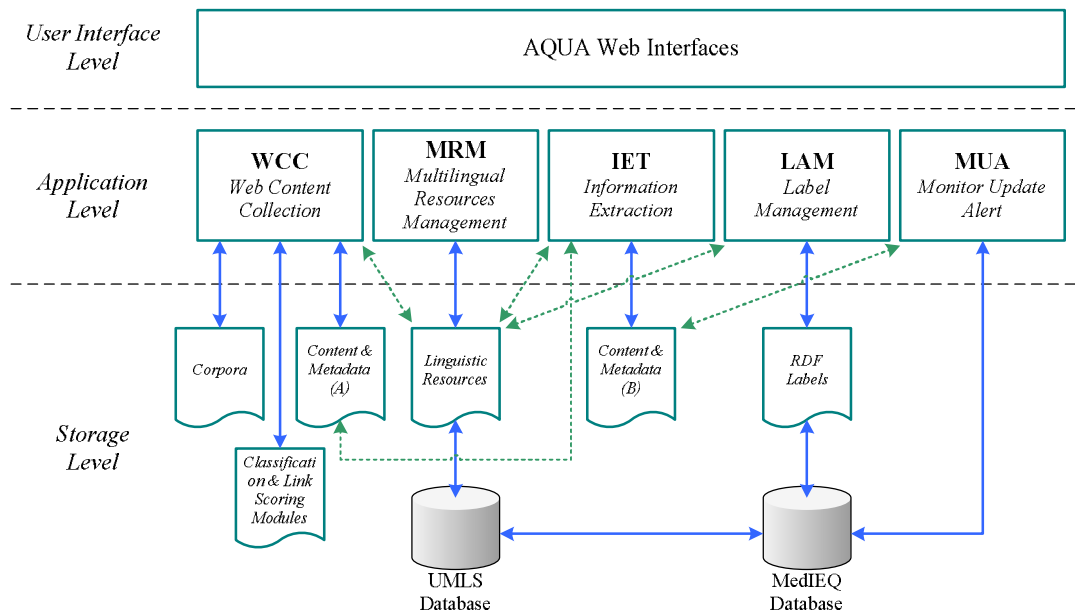


Figure 1. Architecture of the AQUA system.

AQUA incorporates several subsystems (see the application level in Figure 1) and functionalities for the labeling expert. 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) analyses the web content collected by WCC and extracts attributes for MedIEQ-compatible content labels. The *Label Management* (LAM) component generates, validates, modifies and compares the content labels based on the schema proposed by MedIEQ. The *Multilingual Resources Management* (MRM) 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 tasks like the configuration of monitoring tasks, the MedIEQ database updates, or the alerts to labeling experts when important differences occur during the monitoring of existing content labels.

The first version of AQUA prototype, made operational in autumn 2007, addresses the certification of new resources and covers two languages (English and Spanish). The full version of the system, to be delivered in autumn 2008 will also enable monitoring of already labeled resources and will cover 7 languages in total.

References

- 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.
- Stamatakis K, Chandrinou K, Karkaletsis V, Mayer M.A, Gonzales D.V, Labsky D.V, Amigó E, Pöllä M. AQUA, a system assisting labelling experts assess health web resources. 12th Intern. Symposium for Health Information Management Research (iSHIMR 2007), Sheffield, UK, 18-20 July, (2007), 75-84.
- W3C, Resource Description Framework (RDF), 2004. <http://www.w3.org/TR/rdf-schema/>
- W3C, RDF-Content Labels (RDF-CL), 2005. <http://www.w3.org/2004/12/q/doc/content-labels-schema.htm>
- W3C, Protocol for Web Description Resources (POWDER), 2007. <http://www.w3.org/2007/powder/>
- Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the Web. *Int J. Med Inform.* 2006 May 31.