

Assisting Quality Assessment (AQUA) – a system based on semantic web and information extraction technologies to support medical quality labelling agencies

D. Villarroel¹, M.A. Mayer², A. Leis², V. Karkaletsis³, K. Stamatakis³, V. Metsis³, P. Nasikas³,
M.Labsky⁴, M. Ruzicka⁴, V. Svátek⁴, F. López-Ostenero⁵, V. Peinado⁵, E.A. Cabrera⁵,
T. Honkela⁶, M. Pölla⁶, K. Chandrinos⁷

¹Agency for Quality in Medicine (AQuMed), Germany

²Web Mèdica Acreditada of the Medical Association of Barcelona (WMA), Spain

³National Center for Scientific Research “Demokritos” (NCSR), Greece

⁴University of Economics in Prague (UEP), Czech Republic

⁵Universidad Nacional a Distancia (UNED), Spain

⁶Helsinki University of Technology (TKK), Finland

⁷I- sieve Technologies ltd., Greece

Abstract

The EU funded project MedIEQ (Quality Labelling of Medical Web content Using Multilingual Information Extraction) aims to pave the way to an automation of the quality appraisal of health related information on the Internet.

This paper presents AQUA (Assisting Quality Assessment), the assisting system for labelling agencies of medical web sites. This system, based on semantic web and information extraction technologies, aims to support the work of the labelling expert in three different aspects: 1. it searches the Web looking for unlabeled websites, 2. it reviews the health web resources according to a set of machine-readable criteria creating a RDF/XML label, 3. finally, it automatically monitors the web resources in order to identify modifications against the criteria and it alerts the expert about it.

AQUA is expected to reduce labelling agencies' workloads, thus increasing the number of quality-rated web pages as well as their effective monitoring.

1. Introduction

The Internet has become one of the principal sources of health related information [1-5]. Consumers access online health information for different reasons: searching directly for health information, participating in support groups, and consulting with health professionals [6]. The acquired information influences the decision the patients take; 70% said the online information affects their decision about how to treat an illness, and 50% said it led them to ask a doctor a new question or to get a second opinion [7]. However the unequal and often indeterminate

quality of online health information has raised concern. For these reasons many organisations around the world strive towards the creation of quality standards.

Although the approaches differ, they all have a similar goal: to guarantee that the searched information satisfies minimal quality criteria. The first approach is to establish codes of conduct. This implies that the information provider voluntarily follows the criteria, without being controlled. A more active approach is the third party accreditation of web resources. A web site is rated against a defined set of criteria, and if it satisfies them, it obtains a quality seal [8-11]. Another approach is the filtering portals, which classify and organize the web sites according to certain criteria, in order to make this information more accessible for the user [12-15].

In the last two cases, the third party accreditation and the filtering portals, the rating and classification work is done principally “by hand”, resulting in a high workload and requiring a lot of resources.

On the other hand, the Semantic Web offers new opportunities on the field of quality labelling of online health information [16]. Using metadata, it is possible to create a computer readable label, telling the user e.g. who is behind the web site, who is the sponsor, which is the target audience, when the information was created, etc [17,18]. These labels can be “read” by a computer/software, which would provide the user with more specific information according to his/her requirements. The metadata can be expressed in the Resource Description Framework language (RDF) [19], which was developed within the World Wide Web Consortium [20]. RDF is a standard vocabulary designed to represent information on the Web.

This paper presents the work developed within the EC co-funded project Quality Labelling of Medical Web Content using Multilingual Information Extraction (MedIEQ) [21]. We present a system that supports the work of the labelling expert, called AQUA (Assisting Quality Assessment).

It builds on the experiences of other groups, who also worked with semantic web as a tool for the quality labelling of health-related information [10,22,23] and with information extraction technologies [25].

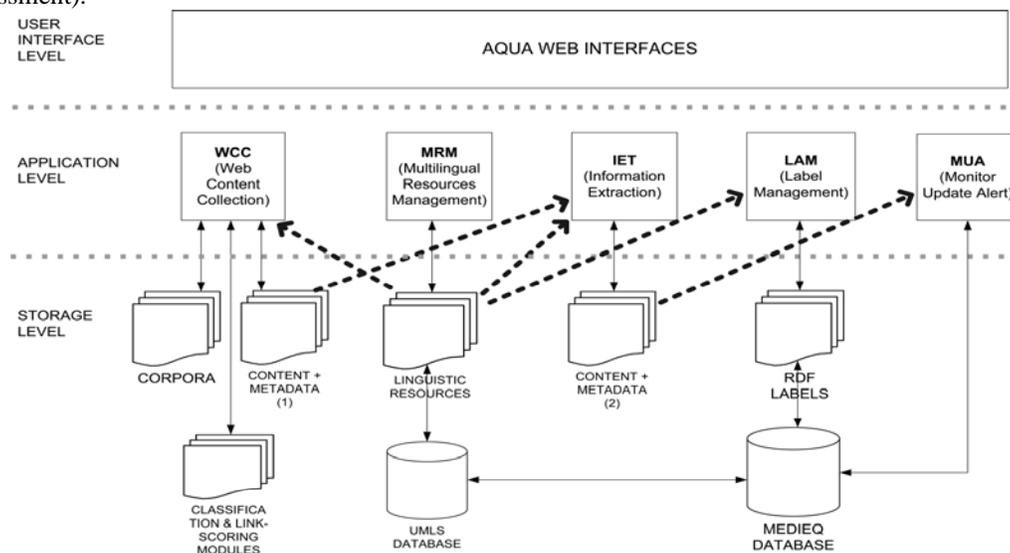


Figure 1. Architecture of AQUA system

2. Assisting Quality Assessment (AQUA)

AQUA is a web application that incorporates different toolkits interacting with each other. Figure 1 shows the complete architecture of AQUA.

AQUA will assist the expert in different processes of the quality labeling work:

1. identification of unlabeled web sites,
2. semi-automatic labelling of health web sites according to certain criteria and creation of machine readable labels.
3. monitoring and continuous re-labelling of the web sites in order to identify updates and to control the compliance with the criteria.

The first step of the labeling process where AQUA can be applied is locating unlabeled websites on the web. Thanks to a content classifier, AQUA is able to distinguish web sites with and without health information. A crawler is also part of AQUA; it searches the web for health-related web sites that do not yet have a label. The crawler is a meta-search machine that utilizes the results retrieved by other search engines (Google, Yahoo, Intute and HON). The starting points

are the user-selected keywords and/or web directories.

Besides, it is possible to prepare a corpus in order to train the classifier. In this way, the results of the next search will be more precise.

For the semi-automatic labelling of health web sites, AQUA offers a Label Management Environment (LAM), where the experts can create, update and compare the RDF labels. The RDF labels are created according to the vocabulary proposed by MedIEQ, which is based on a set of criteria selected by the participating labeling agencies, the Accreditation Program Web Mèdica Acreditada (WMA) and the Agency for Quality in Medicine (AQuMed). The selection is based on an analysis of the most familiar (or the most well known) international quality initiatives: the recommendations of the EU in the document "eEurope 2002: Quality Criteria for Health related Websites [26], the criteria used by the participating agencies (Code of Conduct of the Medical Association of Barcelona [27] and Check In Instrument [28] of AQuMed), as well as the criteria used by Health on the Net Foundation (HON) [8]. Table 1 shows the 10 criteria with their 36 attributes.

The proposed vocabulary serves as a case study to show the functionalities of AQUA and the value of machine-readable labels. The vocabulary terms can be

modified by adding or eliminating some terms according to the needs of the labelling agency.



Figure 2. Welcome page of AQUA

A RDF label, based on several metadata vocabularies (e.g. Dublin Core, FOAF, etc.), can be automatically created, i.e. with information automatically extracted by AQUA, as well as manually, i.e. the expert should search the information and fill the label.

The Information Extraction (IET) and the Web content collection (WCC) toolkits are responsible for the identification and extraction of the information that should be extracted for each criterion.

After the extraction, AQUA presents the results to the expert, together with a link to the page where the information was found. The expert has to decide if it is appropriate or not.

AQUA also supports the expert by monitoring labelled web sites. This happens with the help of another AQUA tool, Monitor-Update-Alert (MUA). It allows the expert to create monitoring tasks, so the labelled web sites will be reviewed again in a predetermined period of time. In the case the monitored web sites were updated against the criteria, MUA sends an alert e-mail to the expert.

3. Final remarks

To help health information seekers identify web sites that follow at least minimal quality standards is one of the big challenges of a labelling agency. Its work usually requires a lot of human effort, because most of the steps of the labelling process are done manually.

In this context, AQUA aims to facilitate the work of the labelling authorities by automating some of the steps

of the quality labelling process.

Criterion	Attributes
1. Resource defining information	1.1 Resource URI 1.2 Resource title 1.3 Resource last update 1.4 Resource language (s)
2. Ownership / Creatorship	2.1 Organization name 2.2 Organization type 2.3 Responsible name 2.4 Responsible title 2.5 Responsible contact details 2.6 Webmaster name 2.7 Webmaster contact details
3. Purpose / Mission	3.1 Purpose/mission of the resource 3.2 Purpose/mission of the owner 3.3 Target audience 3.4 Statement declaring limitation of the provided information
4. Topic/Keywords	4.1 Topic/Keywords (UMLS)
5. Virtual consultation (VC)	5.1 VC service available 5.2 VC responsible name 5.3 VC responsible contact details 5.4 Statement declaring limitation of the VC service
6. Funding / Advertising	6.1 Statement declaring sources of founding 6.2 Name of founding (sponsoring) organization 6.3 Statement declaring limitation of influence of sponsors on the content 6.4 Is advertising present? 6.5 Are advertisement clearly separated from editorial content? 6.6 Policy with regal of advertisement.
7. Other seal or Recommendation	7.1 Is other seal present? 7.2 Which other seal?
8. Information supporting scientific content	8.1 References, bibliography, 8.2 Publication / creation date 8.3 Last revision / modification date 8.4 Author name(s) 8.5 Author(s) contact details 8.6 Editorial policy
9. Privacy policy Confidentiality /	9.1. Explanation how personal data is handled.
10. Accessibility	10.1 Accessibility level (A,AA,AAA)

Table 1. 10 Criteria and their attributes

AQUA will be available in seven languages and will support the 10 criteria and 36 attributes. The first prototype, delivered in July 2007, supports two languages (English and Spanish) and only a selection of the criteria.

Furthermore, the creation of machine-readable labels opens new possibilities for the application of advanced knowledge management technologies (e.g. to make a search more precise).

Both participating labelling agencies will evaluate the

first prototype in order to determinate its performance, usability as well as the influence of AQUA in the daily work of a labelling expert. The results of the evaluation will be presented until the end of October 2007.

Acknowledgements

AQUA was developed in the context of the project MedIEQ (Quality Labelling of Medical Web Content using Multilingual Information Extraction). This project is co-funded by the European Union (DG-SANCO) under the Programme of community action in the field of Public Health (2003-2008).

References

- [1] Cooper CP, Mallon KP, Leadbetter S, Pollack LA, Peipins LA. Cancer Internet search activity on a major search engine, United States 2001-2003. *J Med Internet Res* 2005;7:e36
- [2] Eysenbach G, Köhler C. What is the prevalence of Health-related Searches on the World Wide Web? Qualitative and Quantitative Analysis of Search Engine Queries on the Internet. *Proc AMIA Annu Fall Symp*; 2003;225-229T
- [3] Phillipov G, Phillips PJ. Frequency of health-related search terms on the Internet. *JAMA* 2003;290:2258-2259
- [4] Bansil P, Keenan NL, Zlot AI, Gilliland JC. Health-related information on the Web: results from the HealthStyles Survey, 2002-2003. *Prev Chronic Dis* 2006;A36
- [5] Giustini D. How Google is changing medicine. *BMJ* 2005;331:1487-1488
- [6] Cline RJW, Haynes KM. Consumer health information seeking on the internet: the state of the art. *Health Educ Res*; 2001;16:671-692
- [7] 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_displays.asp Access July 2007
- [8] Health on the Net Foundation (HON) <http://www.hon.ch>
- [9] Mayer MA, Leis A, Sarrias R, Ruiz P. Web Médica Acreditada Guidelines: Reliability and Quality of Health information on Spanish-language. *Proceedings of the 19th International Congress of the European Federation for Medical Informatics* 2005; I:1287-1292
- [10] Kohler C, Darmoni SD, Mayer MA, Roth-Berghofer T, Fiene M, Eysenbach G. MedCIRCLE – The Collaboration for Internet Rating, Certification, Labeling, and Evaluation of Helath Information. *Technology and Health Care, Special Issue: Quality e-Health. Technol Health Care* 2002;10:515
- [11] American Accreditation Healthcare Commission (URAC) <http://www.urac.org>
- [12] Catalogue et Index des Sites Médicaux Francophones (CISMef) <http://www.chu-rouen.fr/cismef/>
- [13] Agency for Quality in Medicine (AquMed) <http://www.patienten-information.de>
- [14] Intute, Health & Lifescience <http://www.intute.ac.uk/healthandlifesciences/>
- [15] New York Online Access to Health (NOAH) <http://www.noah-health.org/>
- [16] Semantic Web by W3C <http://www.w3.org/2001/sw/> Access July 2007
- [17] Eysenbach G, Köhler C, Yihune G, Lampe K, Cross P, Brickley D. A metadata vocabulary for self- and third-party labelling of health web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL). *Proc AMIA Symp* 2001;169-173
- [18] Thirion B, Loosli G, Douyère M, Darmoni SJ. Metadata element set in a quality-controlled subject gateway: a step to health semantic Web. *Stud Health Technol Inform* 2003; 95:707-712
- [19] Resource Description Framework (RDF) <http://www.w3.org/TR/rdf-schema/> Access July 2007
- [20] World Wide Web Consortium (W3C) <http://www.w3.org/> Access July 2007
- [21] Mayer MA, Karkaletsis V, Stamatakis K, Leis A, Villarroel D, Thomeczek C et al. MedIEQ – Quality Labeling of Medical Web Content Using Multilingual Information Extraction. *Proc ICMCC Event* 2006; 183-190
- [22] Karkaletsis V, Stamatakis K, Metsis V, Redoumi D. Helath-related Web Content: quality labelling mechanisms and the MedIEQ approach. *Proc 4th International conference on Information Communication Technologies in Health (ICICTH 2006*
- [23] Gaudinant A. & Boyer C. WRAPIN (Worldwide online Reliable Advice to Patients and Individuals). *MEDNET 2003. The 8th Annual World Congress on the Internet and Medicine, Geneva.*
- [24] Mayer MA, Karkaletsis V, Archer R, Ruiz P, Stamatakis K, Leis A. Quality Labeling of Medical Web Content. *Health Informatics J* 2006; 12:81-87
- [25] Karkaletsis V, Spyropoulos CD, Grover C, Paziienza MT, Coch J, Souflis D. A Platform for Crosslingual, Domain and User Adaptative Web Information Extraction. *Proc European Conference in Artificial Intelligence (ECAI) 2004; 725-729*
- [26] Commission of the European Community. eEurope 2002: Quality Criteria for Health related Websites http://europa.eu.int/information_society/europe/ehealth/doc/communication_acte_en_fin.pdf
- [27] Web Médica Acreditada - Code of Conduct of the Medical Association of Barcelona <http://wma.comb.es/eng/codi.htm>
- [28] Agency for Quality in Medicine - Check In Instrument - http://www.patienten-information.de/content/informationsqualitaet/download/heck_in.pdf

Address for correspondence:
 Dagmar Villarroel
 Agency for Quality in Medicine
 Wegelystraße 3, 10623 Berlin (Germany)
 E-mail address: villarroelgonzales@azq.de