

# Quality Labelling of Medical Web Content

Miquel Angel Mayer\*, Vangelis Karkaletsis\*\*, Phil Archer\*\*\*, Pau Ruiz\*,  
Konstantinos Stamatakis\*\*, Angela Leis\*

\*Web Médica Acreditada (WMA), Barcelona, Spain, {mmayer.wma, pau.wma,  
mleis.wma}@comb.es

\*\*National Center for Scientific Research (NCSR) “Demokritos”, Institute of Informatics &  
Telecommunication, Athens, Greece, {vangelis, kstam}@iit.demokritos.gr

\*\*\*Internet Content Rating Association (ICRA), Brighton, UK, parcher@icra.org

## **Abstract**

As the number of medical web sites in various languages increases, it is increasingly necessary to establish specific criteria and control measures that give consumers some guarantee that the health web sites they are visiting, meet a minimum level of quality standards. Further, that the professionals offering the information are suitably qualified.. The paper presents briefly the current mechanisms for labelling medical web content and introduces the work done in the EC-funded project Quatro. This has defined a vocabulary for quality labels and a schema to deliver them in a machine-processable format. . In addition, the paper proposes the development of a labelling platform that will assist the work of medical labelling agencies in automating, up to a certain level, the retrieval of unlabelled medical web sites and their labelling, and the monitoring of labelled web sites as to whether they are still satisfying the criteria.

## **1. Introduction**

The number of health information web sites and online services is increasing day by day. It is known that the quality of these web sites is very variable and difficult to assess; we can find web sites published by government institutions, consumer and scientific organisations, patients associations, personal sites, health provider institutions, commercial sites, etc. On the other hand, patients continue to find new ways of reaching health information and their physicians [1] and more than four out of ten health information seekers say the material they find affect their decisions about their health itself [2]. Thus the choice of appropriate evaluation criteria as well as the development of tools to support the labelling process (retrieval of unlabelled web sites, monitoring of labelled web sites) are both crucial and challenging.

Organisations around the world are working on establishing standards of quality in the accreditation of health-related web content [3, 4, 5, 6]. However the establishment of codes of conduct or ethics is not enough in the medical domain where the quality of information delivered from medical web sites may affect the health of the citizens. Self-adherence to such codes is nothing more than a claim or a pledge with little enforceability. It is necessary to establish rating mechanisms, either by third party accreditation [7, 8, 9], or by creating portals where medical web sites are organised and characterised against certain labelling criteria [10, 11].

In order for these mechanisms to be successful, they 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 crawling and spidering techniques

that allow the retrieval of new unlabelled web sites, their characterisation and addition in a medical thematic portal.

In Section 2 of the paper we give examples of medical quality labelling criteria and outline the labelling processes followed by both rating mechanisms. Section 3 presents the on-going work in the EC-funded project Quatro [12] for the definition of a common vocabulary of quality labelling criteria, the development of a machine processable labelling schema and the development of tools that exploit in practice such a schema. Section 4 proposes the development of a labelling platform which provides tools that can automate the task of medical quality labelling. Finally section 5 concludes presenting some major remarks.

## **2. Existing Criteria and Processes for Labelling Medical Web Sites**

Labelling criteria have already been established through various initiatives. We will use as an example the criteria adopted by the medical labelling initiative Web Médica Acreditada (WMA) in Spain and Latin America of the Medical Association of Barcelona [13]. The first level of these criteria is presented in Table 1.

Identification
Content
Confidentiality
Control and validation
Advertising and Founding
Virtual Consultation
Non compliance

**Table 1: WMA labelling criteria**

For instance, the “Identification” criterion concerns the provision of information such as the site ownership, contact information, professionals involved in case the site offers consultation services. In addition, the “Content” criterion enforces the provision of information on the updates made to the site, the authors of the medical resources, references to bibliography, etc.

So far there are two major mechanisms in medical quality labelling. The first one is based on third party rating where the web site is assessed by a labelling agency and, if the criteria are met, a label is added to the web site. This is the model used by, among others, WMA. The second type of labelling mechanism examines medical web sites in specific thematic areas, characterizes them against certain criteria, filters some of them based on their characterization, and organizes the rest into web directories to facilitate access by health information consumers. This is the approach used by, for example, the Agency for Quality in Medicine (AQuMed) [14]

## **3. Quatro vocabulary and labelling schema**

Quatro is an on-going EC-funded project which aims to provide a common vocabulary and machine processable schema for quality labelling, making it possible for the many existing labelling schemes to be brought together through a single, coherent approach without affecting the individual scheme’s criteria or independence. The project has already published its vocabulary which is divided into four categories:

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

- Criteria for labelling 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.

The complete vocabulary is available on the Quatro project web site both as a plain text document and as an RDF schema [15]. Labelling 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 labelling criteria. For example, if a labelling 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 labelling operator alerted that the site needs re-checking.

On both counts the use of a common vocabulary offers commercial advantages to labelling operators by increasing the value of the labels for content providers and end-users.

One of the case studies in Quatro concerns the labelling of medical web sites through the involvement of the WMA labelling operator.

### ***3. Proposal for a labelling platform***

The processes of continuous review and control of medical web sites and locating new unlabelled medical web sites are absolutely essential to assure the quality of health knowledge disseminated through the Web. We propose the development of a labelling platform that enables the development of labelling systems. These systems will assist the work of labelling experts, thus increasing the number of labelled medical sites and improving their monitoring. The architecture of such a labelling system is depicted in Figure 1. We will exemplify the platform functionalities by presenting the use of the labelling systems by the two rating mechanisms presented in Section 2 (WMA, AQuMed).

In the case of WMA, the application of the platform tools concerns the constant monitoring of already labelled medical web sites comparing newly extracted information from the site pages against the data stored in the labelling operator database. Taking into account the steps of the WMA labelling process these will be supported by the labelling systems in the following ways:

- Every time a new request arrives to WMA, the labelling system is invoked in order to collect an initial set of data from the corresponding web site. The type of data collected (they will vary according to the request type) will be stored in a separate database in order to be used by the WMA standing committee.
- After the site owner informs WMA that any committee recommendations have been implemented, the labelling system is invoked to examine the corresponding updates. The system outcome is again stored in order to be used by the labelling experts in WMA, who will decide whether the specific site will be labelled or not.

- After the site gets the WMA label, the system will be invoked periodically to examine whether any changes occurred, in terms of the labelling criteria. Depending on the change, the system can alert WMA, thus facilitating the review process.

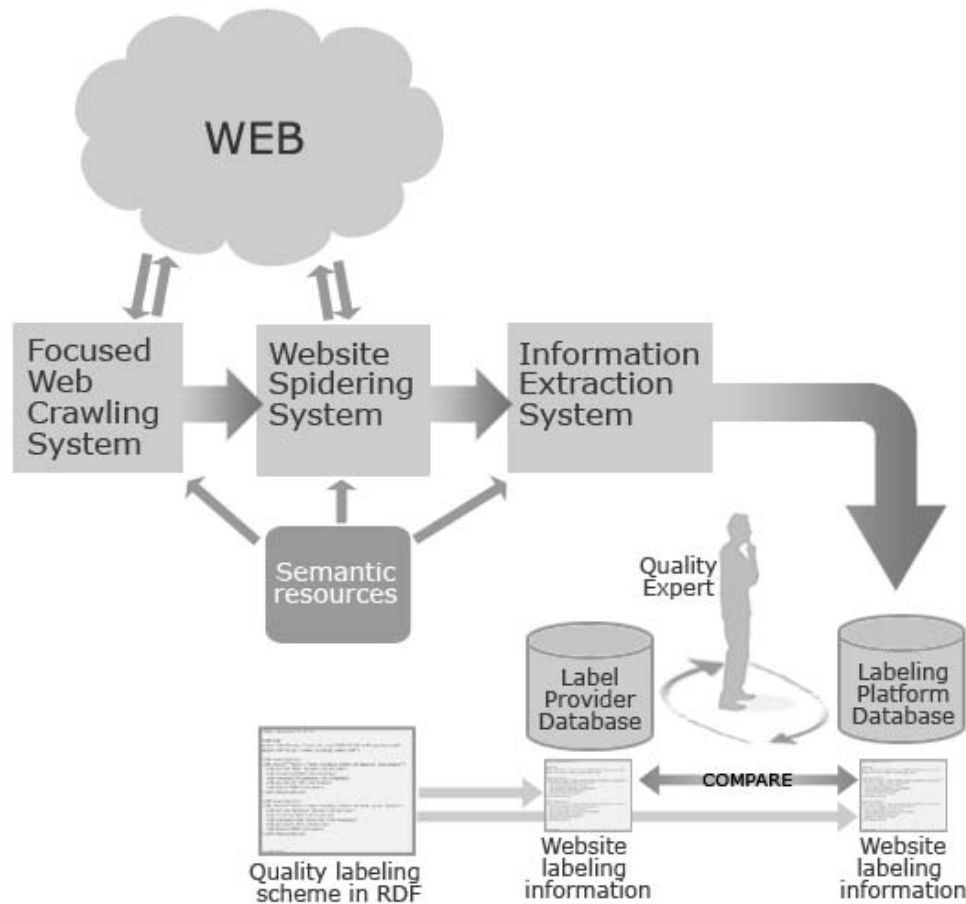


Figure 1: Architecture of the labelling systems

In order to operate as described above, the labelling system must involve components for the following tasks (see Figure 1):

- **Spidering:** Each Web page visited is evaluated, in order to decide whether it is really relevant to the topic (that is the labelling criteria), and its hyperlinks are scored in order to decide whether they are likely to lead to useful pages. Thus, a score-sorted queue of hyperlinks is constructed, which guides the retrieval of new pages. The spidering tool consists of three components: site navigation, page classification and link scoring [16].
- **Information extraction:** The pages retrieved by the spidering component are processed in order to locate and extract useful facts, that is, facts relevant to the labelling criteria. For instance, in a contact page, we are looking for entities such as organization name, person name, medical specialty, an e-mail address, etc. Based on the entities retrieved, certain key phrases, the page layout, we locate the part of the page that contains the information we are looking for. This is a well known web information extraction task, which requires the combination of technology on web wrappers and language technology [16].
- **Data storage:** The extracted information is stored in a data base according to the specification of the medical quality labelling schema.

In the case of AQuMed, the application of the platform tools concerns the identification of new medical web sites, in specific thematic areas, their characterization, the filtering of some of them based on their characterization, and their organization into web directories. Taking into account the steps of the AQuMed labelling process, these will be supported by the labelling systems in the following ways:

- A focused web crawler will be trained to locate medical web sites for specific subjects..
- Every time a new web site is retrieved, the labelling system will examine it against AQuMed criteria and store the data collected in a data base separate from the data base storing the meta-data of the AQuMed web directories.
- In case the labelling system has to re-examine an already characterized web site, it checks first whether the previously collected meta-data are still valid and in case changes occurred it updates the data collected in the data base, alerting the labelling expert.
- The sites that do not meet certain criteria are filtered and their data are stored separately in order to be examined by the labelling expert who will take the final decision on adding, excluding or withdrawing a site from the directory.
- The labelling system operates periodically in order to locate new web sites or update the data on existing ones.

In order to operate as described above, the labelling system must involve components for one more task apart from the tasks described:

- Crawling: The focused crawler searches for medical web sites on specific subjects/problems [17]. The crawler may exploit for this purpose, specific subject-related web hierarchies, keywords (phrases) from subject-related ontologies, thesauri, lexica. The result is a list of medical web sites that is compared to the previously collected list as well as to AQuMed web directories in order to keep only those sites found for the first time.

#### **4. Concluding Remarks**

This work proposes the use of semantic web technologies (RDF labelling schemas, focused crawling, spidering, information extraction) to tackle the main problem of current medical quality labelling mechanisms, that is, the need for a continuous review and control of the accredited or filtered medical web sites, a process which requires a huge amount of human effort. The resulting technology is expected to have a significant impact on medical quality labelling assisting the work of labelling experts, increasing the number of labelled medical sites across Europe and their effective monitoring, and thus improving the quality health knowledge disseminated through the Web.

#### **Acknowledgements**

This research was partially funded by the EC through the SIAP project Quatro (Quality Assurance and Content Description).

#### **References**

- [1] G. Eysenbach. "Consumer health informatics". Br Med J 2000; 320(4):1713-1716.
- [2] S. Fox, L. Rainie. "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](http://www.pewinternet.org/PPF/r/26/report_display.asp))
- [3] Internet Healthcare Coalition. <http://ihealthcoalition.org>
- [4] Health Internet Ethics: Ethical Principles for Offering Internet Health Services to Consumers. (<http://www.hiethics.com/Principles/index.asp>)
- [5] European Commission. "Quality Criteria for Health Related Websites". [http://europa.eu.int/information\\_society/europe/ehealth/doc/communication\\_acte\\_en\\_fin.pdf](http://europa.eu.int/information_society/europe/ehealth/doc/communication_acte_en_fin.pdf)

- [6] M. Winker, A. Flanagan, B. Chi-Lum, J. White, K. Andrews, R. Kennett, C. DeAngelis, R. Musacchio. "American Medical Association. Guidelines for Medical Health information Sites on the Internet". (<http://www.ama-assn.org/ama/pub/category/1905.html>)
- [7] Health On the Net Foundation. <http://www.hon.ch>.
- [8] M.A. Mayer, S. Darmoni, M. Fiene, Kohler, T. Roth-Berghofer, G. Eysenbach. "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: 667-672.
- [9] URAC American Accreditation Healthcare Commission. <http://www.urac.org/>
- [10] OMNI. <http://omni.ac.uk/>
- [11] CISMeF. <http://www.chu-rouen.fr/cismef/>
- [12] Quatro. <http://www.quatro-project.org>
- [13] WMA. [http://wma.comb.es/home\\_eng.htm](http://wma.comb.es/home_eng.htm)
- [14] AQuMed. <http://www.aeqzq.de>
- [15] [www.quatro-project.org/vocabulary/1.0/](http://www.quatro-project.org/vocabulary/1.0/) , [www.quatro-project.org/rdfs/vocabulary1.0.rdf](http://www.quatro-project.org/rdfs/vocabulary1.0.rdf)
- [16] V. Karkaletsis, C.D. Spyropoulos, C. Grover, M.T. Pazienza, J. Coch, D. Souflis, "A Platform for Cross-lingual, Domain and User Adaptive Web Information Extraction" In *Proceedings of the European Conference in Artificial Intelligence (ECAI)*, pp. 725 - 729, Valencia, Spain, 2004.
- [17] K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, C. Grover, J. R. Curran, S. Dingare. "Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler," In *Proceedings of the Second International Workshop on Web Document Analysis (WDA 2003)*, pp. 75-78, Edinburgh, UK, August, 2003.