Quality labelling of medical web content
Miquel Angel Mayer, Vangelis Karkaletsis, Phil Archer, Pau Ruiz, Konstantinos Stamatakis and Angela Leis

HEALTH INFORMATICS J 2006; 12: 81
DOI: 10.1177/1460458206061230

The online version of this article can be found at:
http://jhi.sagepub.com/cgi/content/abstract/12/1/81

Published by:
SAGE Publications
http://www.sagepublications.com

Additional services and information for Health Informatics Journal can be found at:

Email Alerts: http://jhi.sagepub.com/cgi/alerts

Subscriptions: http://jhi.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav
Quality labelling of medical web content

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

As the number of medical websites in various languages increases, it is increasingly necessary to establish specific criteria and control measures that give consumers some guarantee that the health websites they are visiting meet a minimum level of quality standards. Further, reassurance is needed that the professionals offering the information are suitably qualified. The paper briefly presents 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 websites and their labelling, and the monitoring of labelled websites as to whether they are still satisfying the criteria.

Keywords
quality labelling, semantic web technologies, web content analysis

Introduction

The number of health information websites and online services is increasing day by day. It is known that the quality of these websites is very variable and difficult to assess; we can find websites published by government institutions, consumer and scientific organizations, patients’ associations, individuals, health provider institutions, companies, 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 10 health information seekers say the material they find affects 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 websites, monitoring of labelled websites) are both crucial and challenging.

Organizations around the world are working on establishing standards of quality in the accreditation of health-related web content [3–6]. However, the establishment of codes of conduct or ethics is not enough in the medical domain where the quality of information delivered from medical websites 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 [7–9] or by creating portals where medical websites are organized and characterized 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 websites alerting the labelling agency when some changes occur against the labelling criteria, or web crawling and spidering techniques that allow the retrieval of new unlabelled websites, their characterization and addition in a medical thematic portal.

In the next section we give examples of medical quality labelling criteria and outline the labelling processes followed by both rating mechanisms. The third section presents the ongoing 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. The fourth section proposes the development of a labelling platform which provides tools that can automate the task of medical quality labelling. The conclusion summarizes expectations for these labelling advances.

Existing criteria and processes for labelling medical websites

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.

For instance, the ‘identification’ criterion concerns the provision of information, such as site ownership, contact information, and the professionals involved if 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.

Table 1 WMA labelling criteria

<table>
<thead>
<tr>
<th>Identification</th>
<th>Content</th>
<th>Confidentiality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control and validation</td>
<td>Advertising and founding</td>
<td>Virtual consultation</td>
</tr>
<tr>
<td>Non-compliance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
So far there are two major mechanisms in medical quality labelling. The first is based on third party rating where the website is assessed by a labelling agency and, if the criteria are met, a label is added to the website. This is the model used by, among others, WMA. The second type of labelling mechanism examines medical websites 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].

Quatro vocabulary and labelling schema

Quatro is an ongoing 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, and 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 website 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 rechecking.

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 websites through the involvement of the WMA labelling operator.
Proposal for a labelling platform

The processes of continuous review and control of medical websites and locating new unlabelled medical websites are absolutely essential to ensure 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 labelling systems according to the two rating mechanisms WMA and AQuMed presented earlier.

In the case of WMA, the application of the platform tools concerns the constant monitoring of already labelled medical websites, 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 website. 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.

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 and 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 database 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 websites 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 websites for specific subjects.
- Every time a new website is retrieved, the labelling system will examine it against AQuMed criteria and store the data collected in a database separate from the database storing the meta-data of the AQuMed web directories.
- When the labelling system has to re-examine an already characterized website, it checks first whether the previously collected meta-data are still valid; when changes have occurred, it updates the data collected in the database, 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 websites or update the data on existing ones.

In order to operate as described above, the labelling system must involve components for one more task in addition to those described above:

- **Crawling**. The focused crawler searches for medical websites on specific subjects/problems [17]. For this purpose the crawler may exploit specific subject-related web hierarchies, keywords (phrases) from subject-related ontologies, thesauri and lexica. The result is a list of medical websites 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.

### Concluding remarks

This work proposes the use of semantic web technologies (RDF labelling schemata, focused crawling, spidering, information extraction) to tackle the main problem of current medical quality labelling mechanisms, that is, the need for continuous review and control of the accredited or filtered medical websites, 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

Mayer et al. Quality labelling of web content

10 OMNI. http://omni.ac.uk/
12 Quatro. http://www.quatro-project.org/
13 WMA. http://wma.comb.es.
15 http://www.quatro-project.org/vocabulary/1.0

Correspondence to: Miquel Angel Mayer

Miquel Angel Mayer
Web Médica Acreditada (WMA), Barcelona, Spain
E-mail: mmayer.wma@comb.es

Pau Ruiz
Web Médica Acreditada (WMA), Barcelona, Spain
E-mail: pau.wma@comb.es

Vangelis Karkaletsis
National Center for Scientific Research (NCSR) ‘Demokritos’, Institute of Informatics & Telecommunication, Athens, Greece
E-mail: vangelis@iit.demokritos.gr

Konstantinos Stamatakis
National Center for Scientific Research (NCSR) ‘Demokritos’, Institute of Informatics & Telecommunication, Athens, Greece
E-mail: kstam@iit.demokritos.gr

Phil Archer
Internet Content Rating Association (ICRA), Brighton, UK
E-mail: parcher@icra.org

Angela Leis
Web Médica Acreditada (WMA), Barcelona, Spain
E-mail: mleis.wma@comb.es