MedIEQ – Quality Labelling of Medical Web Content Using Multilingual Information Extraction

Miquel Angel MAYER [a,1], Vangelis KARKALETSIS [b], Kostas STAMATAKIS [b], Angela LEIS[a], Dagmar VILLARROEL [c], Christian THOMECZEK [c], Martin LABSKÝ [d], Fernando LÓPEZ-OSTENERO [e] and Timo HONKELA [f]
[a] *Web Médica Acreditada (WMA) of the Medical Association of Barcelona (COMB), Spain*
[b] *National Centre for Scientific Research "Demokritos (NCSR)", Greece*
[c] *Agency for Quality of Medicine (AQuMED), Germany*
[d] *University of Economics in Prague (UEP), Czech Republic*
[e] *Universidad Nacional a Distancia (UNED), Spain*
[f] *Helsinky University of Technology (HUT), Finland*

Abstract. Quality of Internet health information is essential because it has the potential to benefit or harm a large number of people and it is therefore essential to provide consumers with some tools to aid them in assessing the nature of the information they are accessing and how they should use it without jeopardizing their relationship with their doctor. Organizations around the world are working on establishing standards of quality in the accreditation of health-related web content. For the full success of these initiatives, they must be equipped with technologies that enable the automation of the rating process and allow the continuous monitoring of labeled web sites alerting the labeling agency. In this paper we describe the European project MedIEQ that integrates the efforts of relevant organizations on medical quality labelling, multilingual information retrieval and extraction and semantic resources, from six different European countries (Spain, Germany, Greece, Finland, Czech Republic and Switzerland). The main objectives of MedIEQ are: first, to develop a scheme for the quality labelling of medical web content and provide the tools supporting the creation, maintenance and access of labelling data according to this scheme and second, to specify a methodology for the content analysis of medical web sites according to the MedIEQ scheme and develop the tools that will implement it.

Keywords. Semantic web, medical information, quality labelling, web content analysis

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.[1] On the other hand, patients continue to find new ways of reaching health information and their physicians and more than four out of ten health information seekers say the material they find affect their decisions about their health itself.[2,3] Health information consumers, such as the patients and

[1] Corresponding Author: Colegio Oficial de Médicos de Barcelona, Pg Bosanova 47, 08017 Barcelona, Spain; E-mail: mmayer.whma@comb.es.

the general public, cannot assess themselves of the good quality of the information because of they are not always familiar with the medical domain and vocabulary.[4]

Although there are divergent opinions about the need for accreditation of health Web sites and adoption by Internet users, [5] different organizations around the world are working on establishing standards of quality in the accreditation of health-related web content.[1, 6-12]

The European Council in 2000 supported an initiative within eEurope 2002 to develop a core set of Quality Criteria for Health Related Websites. The specific aim was to draw up a commonly agreed set of simple quality criteria on which Member States, as well as public and private bodies, may draw in the development of quality initiatives for health related websites. These criteria should be applied in addition to relevant Community law. As a result, a core set of quality criteria was established. The criteria may be used as a basis in the development of user guides, voluntary codes of conduct, trust marks, accreditation systems, or any other initiative adopted by relevant parties, at European, national, regional or organisational level. By using a common set of criteria as a starting point, such initiatives can develop in a focused manner across the European Union. [13]

There are three 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, in terms of certain labelling criteria, and is asked to make some changes to get the accreditation label which then it is added onto the web site. The second one 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. The third mechanism is based on self-adherence to some codes of conduct or ethics that is nothing more than a claim or a pledge with little enforceability. [14]

On the other hand, the current Web is based on HTML (hypertext mark-up language), which specifies how to layout a web page for human readers. HTML as such cannot be exploited by information retrieval techniques to improve results, which thus to rely on the words that form the content of the page. This “current web” must evolve in the next years, from an human-understandable information, to a global knowledge repository, where the information should be machine-readable and directly processed by computers, enabling the use of advanced knowledge management technologies.[15] This change is based on Semantic Web technologies. The Semantic Web is "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" based in metadata. [16] We can think of it as being an efficient way of representing data on the World Wide Web, or as a globally linked database.[17] These metadata can be expressed in different ways as the Resource Description Framework (RDF) language. RDF, developed under the auspices of the World Wide Web Consortium (W3C), [18] is the standard language for representing information about resources in the World Wide Web. It is particularly intended for representing metadata about Web resources, such as the title, author, and modification date of a Web page, copyright and licensing information about a Web document, or the availability schedule for some shared resource.[19] RDF defines a simple, yet powerful model for describing resources.

Thus the choice of appropriate evaluation criteria as well as the development of tools to support the labelling process (retrieval of unlabeled web sites, monitoring of labeled web sites) are both crucial and challenging. [20]

1. MedIEQ project

MedIEQ [21] continues the work of previous projects in the area of medical quality labeling (MedCERTAIN,[22] MedCIRCLE [10] and WRAPIN [23]) and quality labelling standards (QUATRO [24]). MedCERTAIN (MedPICS Certification and Rating of Trustworthy Health Information on the net) and MedCIRCLE (Collaboration for Internet Rating, Certification, Labelling and Evaluation of Health Information on the World-Wide-Web) were some projects that established a third-party rating systems to select high quality information medical websites on the Internet. These systems used a metadata language (HIDDEL: Health Information

Disclosure, Description and Evaluation Language) which allows expression of descriptive and evaluative annotations in RDF. [25] WRAPIN (Worldwide online Reliable Advice to Patient and Individuals) was another project that its main objective was to make available a tool to determine information quality by automatically checking a document against matching sources from databases of known quality. The QUATRO project (Quality Assurance and Content Description) is a platform that applies semantic web technologies to trust mark schemes and quality labels. [26]
The overall objective of MedIEQ is to advance current medical quality labelling technology, drawing on past and original research in the area. The implementation of this objective will be based on the realisation of the following more specific objectives:

1. Develop a scheme for the quality labelling of medical web content and provide the tools supporting the creation, maintenance and access of labelling data according to this scheme;
2. Specify a methodology for the content analysis of medical web sites according to the MedIEQ scheme and develop the tools that will implement it;
3. Specify a methodology and develop the tools for the creation and maintenance of the multilingual resources that will support content analysis in medical web sites;
4. Integrate the above technologies into a prototype labelling system implemented using an open architecture;
5. Demonstrate the resulting prototype in 7 different languages and two labelling applications (third party accreditation and classification).

2. MediEQ tools

MedIEQ will examine two major mechanisms in medical quality labelling which are currently being used by the medical quality labelling agencies participating in the project: Web Médica Acreditada (WMA) and the Agency for Quality in Medicine (AQuMED). WMA is based on third party rating and grants the websites a quality seal. AQuMED filters quality websites and organizes them into directories.

In the case of WMA, the accreditation process is as follows: [1]
1. The person in charge of a website sends a (voluntary) request to the WMA website in order to begin the process. Using the online application form, the person in charge provides certain information for the WMA and auto-checks the WMA criteria (based on the Code of Conduct and the Ethical Code) (step/level1) to express acceptance of these recommendations;
2. The Standing Committee assesses the website based on the WMA criteria (step/level 2/ medical expert);
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 and WMA sends an html seal code to be posted on the accredited website, as well as adding its name and URL to the index of accredited websites.

In the case of AQuMED the medical websites are selected according to some criteria of Health on the Net Code [5] and classified in four categories: treatment information, background information, self-help and counselling organisations and finally medical organizations. After that, the treatment information are evaluated according to DISCERN guidelines [26] and with CHECK-IN instrument (http://www.patienteninformation.de/content/informationsqualitaet/download/check _in.pdf). Patients have access to this information through the website http://www.patienten-information.de/.

MedIEQ aims 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 that requires a huge amount of human effort. To achieve this, MedIEQ integrates

the efforts of relevant organizations on medical quality labelling, multilingual information retrieval and extraction mechanisms and semantic resources from six different European countries (Spain, Germany, Greece, Finland, Czech Republic and Switzerland).

The labelling system must involve components for the following tasks (see Figure 1):

- Crawling: crawl the Web to locate interesting web sites.
- 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 [28].
- 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 [28].
- Data storage: The extracted information is stored in a database according to the specification of the medical quality labeling schema.

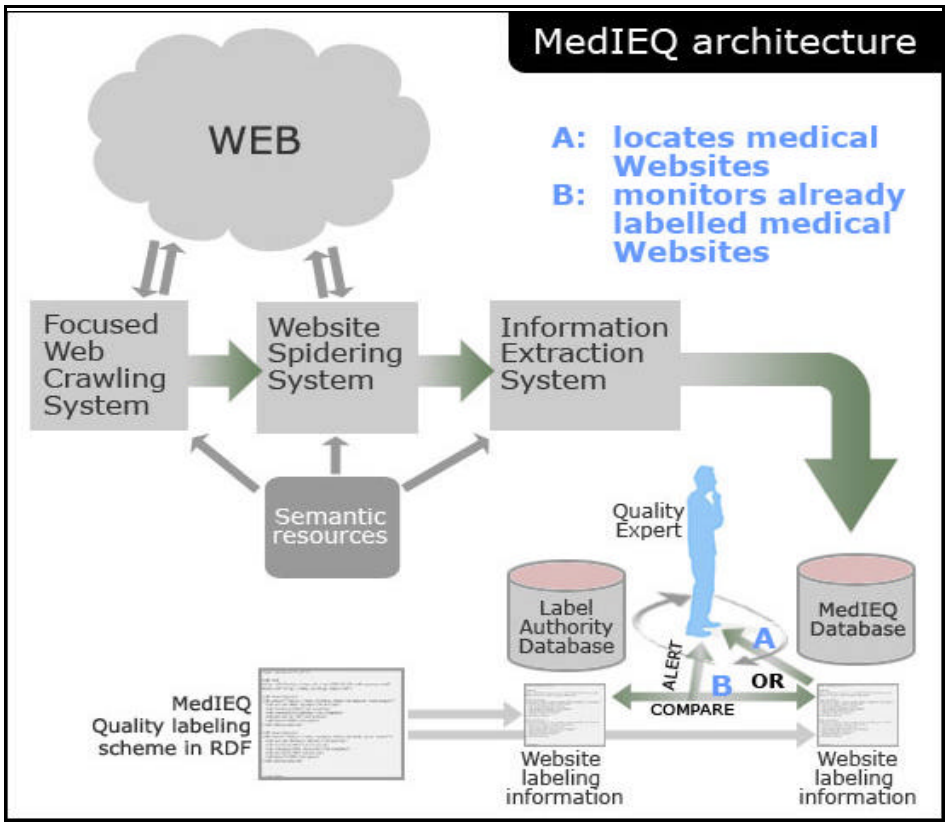

Figure 1. MedIEQ architecture using semantic web resources and label authorities.

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. [20]

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.[20]

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.[20]

3. Conclusions

Since the number of medical websites as well as the patient interest for this information grow it is necessary to find some mechanisms to guarantee and control the quality of them.

The main problem that these mechanisms face is the need for a continuous review and control of the accredited or classified web sites that means a huge amount of human effort. WMA, as third-party accreditation system, for instance, periodically reviews manually the accredited web sites to renew the quality label. On the other hand, in AQuMED, as filtering and rating system, website directories are periodically updated due to the addition of new sites and changes in the characterization of the already visited ones.

Up to now there is not working a standard RDF schema for medical web sites. MedIEQ will put forward a specific medical metadata vocabulary, making use of the experience in previous projects in this area, the EC Quality Criteria for Health Related Websites, [13] the W3C standards as the RDF Content labels schema [29] that is developed in QUATRO, and other standardized vocabularies as the Dublin Core Metadata Initiative [30] and FOAF project.[31] On the other hand, previous initiatives didn't use spidering tools technologies that enable the

automation of the rating process, such as information extraction techniques that allow the continuous monitoring of labeled web sites alerting the labeling agencies (LA) in case some changes occur against the labeling criteria, alerting experts the sites content is updated against the quality criteria, thus facilitating the work of medical quality labeling agencies. [28]

The resulting technology presented by MedIEQ 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

MedIEQ is a project funded by the European Union under the Programme of community action in the field of Public Health (2003-2008) and it is made up of: National Center for Scientific Research "Demokritos", Greece (the coordinator), Institute of Informatics and Telecommunications, Software and Knowledge engineering Laboratory (I-sieve Technologies Ltd.), Greece; Universidad Nacional de Education a Distancia, (UNED), Spain; Web Médica Acreditada (WMA) of the Medical Association of Barcelona (COMB), Spain; Agency for Quality in Medicine (AQuMED), Germany; the University of Economics in Prague (UEP), Cezch Republic; Helsinki University of Technology (HUT), Finland; Geneva University Hospitals - Service of Medical Informatics (HUG), Switzerland.

References

[1] Mayer MA, Leis A, Sarrias R, Ruíz P. Web Mèdica Acreditada Guidelines: realiability and quality of health information on Spanish-Language websites. In: R. Engelbrecht et al (Eds).Connecting Medical Informatics and Bioinformatics. Proceedings of the 19th International Congress of the European Federation for Medical Informatics (CD-ROM). Geneve, Switzerland. .Vol I, No. 1, 2005. p.1287-92.

[2] Eysenbach G. Consumer health informatics. BMJ 2000; 320 (4); 1713-16.

[3] Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients´use of the Internet for medical Information. J Gen Intern Med 2002; 17(3):180-5.

[4] Soualmia LF, Darmoni SJ, Douyère M, Thirion B. Modelisation of Consumer Health Information in a Quality-Controled gateway. In: The New Navigators: from Professionals to Patients. Baud R. et al. (ed.) Proc of MIE2003 701-706.

[5] Analysis of 9th HON Survey of Health and Medical Internet Users Winter 2004-2005. Available from: http://www.hon.ch/Survey/Survey2005/res.html.

[6] Risk A, Dzenovagis J. Review of Internet health information quality initiatives. J Med Internet Res 2001; 3(4):e28.

[7] Health on the Net Foundation (HONCode). Home page. Available from: http://www.hon.ch.

[8] Winker MA, Flanagan A, Chi-Lum B, . Guidelines for Medical and Health Information Sites on the Internet: principles governing AMA web sites. American Medical Association. JAMA 2000; 283 (12): 1600-1606.

[9] Hi-Ethics, Inc. Health Internet Ethics: Ethical Principles for offering Internet Health services to consumers. Available from: http://www.hiethics.com/Principles/index.asp.

[10] Kohler C, Darmoni SD, Mayer MA, Roth-Berghofer T, Fiene M, Eysenbach G. MedCIRCLE - The Collaboration for Internet Rating, Certification, Labelling, and Evaluation of Health Information. Technology and Health Care, Special Issue: Quality e-Health. Technol Health Care 2002; 10(6): 515.

[11] URAC. Health Web Site Accreditation. Home page. Available from: http://webapps.urac.org/websiteaccreditation/default.htm.

[12] Curro V, Buonuomo PS, Onesimo R, de RP, Vituzzi A, di Tanna GL, D'Atri A. A quality evaluation methodology of health web-pages for non-professionals. Med Inform Internet Med 2004;29(2):95-107

[13] European Commission. eEurope 2002: Quality Criteria for Health related Websites. Available from: http://europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf.

[14] Wilson P. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the Internet. BMJ 2002; 321: 598-602.

[15] Eysenbach G. The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?. J Healthc Techn Manag 2003; 5: 194-212.

[16] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American, May 2001.

[17] Palmer Sean B. The Semantic Web: an Introduction. Available from: http://infomesh.net/2001/swintro/.

[18] World Wide Web Consortium (W3C). Available from: http://www.w3.org.

[19] Manola F, Miller E. RDF Primer. W3C Recommendation 10 February 2004. Available from: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.
[20] Mayer MA, Karkaletsis V, Archer P, Ruiz P, Stamatakis K, Leis A. Quality labelling of medical web content. Health Informatics Journal 2006;12:81-87.
[21] MedIEQ (Quality labeling of Medical Web Content Using Multilingual Information Extraction). Internet homepage. Accessible in: http//www.medieq.org.
[22] Eysenbach G, Köhler C, Yihune G, Lampe K, Cross P, Brickley D. A framework for improving the quality of health information on the world-wide-web and bettering public e-health: The MedCERTAIN approach. In: Haux R, Patel V, Hasmann A (eds.) 2001; Medinfo01, Proceedings of the Tenth World Congress on Medical Informatics: 1450-1454.
[23] Gaudinat A. & Boyer C. WRAPIN (Worldwide online Reliable Advice to Patients and Individuals). In MEDNET 2003, The 8th Annual World Congress on the Internet and Medicine, Geneva.
[24] QUATRO. Archer P. and Quatro Project Members. Quatro - a metadata platform for trustmarks. Proceedings del International Conference on Dublin Core and Metadata Applications, Madrid 2005.
[25] Eysenbach G, Köhler C, Yihune G, Lampe K, Cross P, Brickley D. A metadata vocabulary for self- and third-party labelling of health websites: Health Information Disclosure, Description and Evaluation Language (HIDDEL). Proc AMIA Annu Fall Symp 2001; 169-173.
[26] Archer P and Quatro members. Quatro – a metadata platform for trustmarks. Proc Int Conf on Dublin Core and Metadata Applications. Madrid 2005. p. 211-214.
[27] DISCERN online. Quality criteria for consumer health information. Available from: http://www.discern.org.uk/.
[28] Stamatakis K, Karkaletsis V, Paliouras G, Horlock J, Grover C, Curan JR, Dingare S. Domain-specific web site identification: the CROSSMARC focused web crawler. In Proceedings of the Second International workshop on Web Document Analysis (WDA 2003), August 2003, Edinburgh 75-78.
[29] RDF Content Labels:Schema Description. Available from: http://www.w3.org/2004/12/q/doc/content-labels-schema.htm.
[30] Dublin Core Metadata Initiative (DCMI). Internet homepage. Available from: http://dubincore.org/.
[31] The Friend of a Friend (FOAF). Internet homepage. Available from: http://www.foaf-project.org/.