

HEALTH-RELATED WEB CONTENT: QUALITY LABELLING MECHANISMS AND THE MEDIEQ APPROACH

Vangelis Karkaletsis, Kostas Stamatakis, Vangelis Metsis,
Vassiliki Redoumi, Dimitris Tsarouhas

National Centre for Scientific Research (NCSR) "Demokritos",
Institute of Informatics & Telecommunications,
15310 Aghia Paraskevi Attikis, Athens, Greece,
Tel: +30 210 6503197, Fax: +30 210 6532175,
e-mail: {vangelis, kstam, [vmetsis](mailto:vmetsis@iit.demokritos.gr)}@iit.demokritos.gr

ABSTRACT

As the number of medical web sites in various languages increases, it is more than necessary to implement control measures that give the consumers adequate guarantee that the health web sites they are visiting, meet a minimum level of quality standards and that the professionals offering the information on the web site are responsible for its contents. The paper describes the existing labelling mechanisms, presents the main objectives of the EC-funded project MedIEQ, and the tools that will be implemented, and discusses the results from an initial survey on the Greek medical web using some of the project tools.

Keywords: Medical web content, quality labelling, semantic web technologies, web content analysis, Greek medical web content.

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, published by various authorities, is very variable and difficult to assess [1]. At the same time, the necessity to implement control measures that give the consumers adequate guarantee that the health web sites they are visiting meet a minimum level of quality standards and that the professionals offering the information on the web site are responsible for its contents, is increasing. Different organizations around the world are currently working on establishing quality labelling criteria for the accreditation of health-related web content [1-8]. The European Council supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Websites" [9]. However, self-adherence to such criteria is nothing more than a claim with little enforceability. It is necessary to establish rating mechanisms which exploit such labelling criteria.

There are two major mechanisms in medical quality labelling. The first one is based on third party accreditation: a web site is assessed by a labelling agency and, if certain criteria are met, a label is assigned and added to the web site. The second mechanism is based on classification and filtering: medical web sites are reviewed by experts and characterized against certain criteria; some of them are filtered depending on their characterization; the rest are organized into web directories to facilitate access by health information consumers.

Both mechanisms, as currently applied, present drawbacks. As for the first mechanism, the added label is not machine-processable (such that a web browser or a search engine could locate, parse, "understand" and display its characteristics in a human readable way). Furthermore, in both mechanisms, it is difficult for labelling authorities to monitor already labelled sites as well as to find new unlabelled ones.

Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labelling, the recently started EC-funded project MedIEQ¹ aims to pave the way towards the automation of quality labelling process in medical web sites. MedIEQ will deliver tools that crawl the Web to locate unlabelled medical web sites in seven different European languages in order to examine their content using a set of machine readable quality criteria. MedIEQ tools will monitor already labelled medical sites alerting labelling experts in case the sites' content is updated against the quality criteria.

The project started by building the current web medical map of the participating countries. The initial survey, carried out on the Greek medical web, using some of the project tools, as well as the results of this survey are discussed in this paper.

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 Web Médica Acreditada (WMA)² for Spain and Latin America [1]. The WMA criteria include all the quality criteria specified in the context of the eEurope 2002 initiative [2]. The first level of these criteria includes Identification, Content, Confidentiality, Advertising and Sponsorship, Virtual Consultation, Non compliance³.

However, the specification of labelling criteria is not enough on its own. As noted in the introduction, self-adherence to such criteria is nothing more than a claim with little enforceability. It is necessary to establish rating mechanisms which exploit such labelling criteria.

There are two major mechanisms in medical quality labelling. The first one is based on third party accreditation: a web site is assessed by a labelling agency and, if certain criteria are met, a label is assigned and added to the web site. This is the model used, among others, by WMA. The second mechanism is based on classification and filtering: medical web sites are reviewed by experts and characterized against certain criteria; some of them are filtered depending on their characterization; the rest are organized into web directories to facilitate access by health information consumers. This is the approach of the Agency for Quality in Medicine (AQuMed)⁴. Both WMA and AQuMed are participating in the MedIEQ project.

Both mechanisms, as currently applied, present drawbacks. As for the first mechanism, the added label is not machine-processable (such that a web browser or a search engine could locate, parse, "understand" and display its characteristics in a human readable way)..Technology for creating machine processable labels requires the establishment of common labelling vocabularies and machine processable schemas as well as the use of semantic web technologies for enabling the label's parsing by web browsers or search engines. The efficient presentation of the label's content to the user via web browsers and search engines will promote the use of labels to the general public.

However, establishing machine-processable labels is not enough. In both mechanisms, labelling authorities must be equipped with technologies that support the monitoring of already labelled sites as well as the detection of unlabelled ones. This requires the use in practice of web content analysis technologies, such as crawling for detecting medical web sites, spidering for locating inside those sites web pages relevant to the labelling criteria examined, and information extraction for acquiring data from the located web pages that correspond to the labelling criteria, and which will be either compared to existing labelling data or will be stored in order to be validated and enriched by the labelling experts.

The MedIEQ project

¹ <http://www.medieq.org/>

² WMA (<http://wma.comb.es/>) is the medical labelling initiative of the Medical Association of Barcelona (<http://www.comb.cat/>)

³ For details on WMA criteria, visit <http://wma.comb.es/eng/codi.htm>

⁴ <http://www.aeqz.de/>

MedIEQ (Quality labeling of Medical Web Content Using Multilingual Information Extraction) continues the work of previous projects in the area of medical quality labelling (MedCERTAIN⁵, MedCIRCLE⁶ and WRAPIN. 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).

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 will involve components for the following tasks:

- Crawling: crawl the Web to locate interesting web sites [10].
- 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 [11].
- 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 [11].
- Data storage: The extracted information is stored in a database according to the specification of the medical quality labelling schema.

The processes of continuous review and control of labelled medical web sites and locating new unlabelled medical web sites are absolutely essential to assure the quality of health knowledge disseminated through the Web. MedIEQ aims at the development of a labelling platform to assist the work of labelling experts, increasing in turn the number of labelled medical sites and improving their monitoring.

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.

In the case of AQUAMED, 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 classification into web directories.

A survey of the Greek medical web

We conducted an initial survey to categorize Greek health-related web sites to a number of categories agreed by the project partners. These categories are: “government organization”, “healthcare service provider”, “media and publishers”, “patient organization / self support group”, “pharmaceutical company / retailer”, “private individual”, “scientific or professional organization”.

⁵ <http://www.medcertain.org/>

⁶ <http://www.medcircle.org/>

Apart from categorization, we collected additional information on every web site, in order to construct a Greek medical web map. It's worth mentioning that the extra fields of information we suggested correspond to a subset of existing quality criteria from both quality agencies which participate in MedIEQ. These fields are then proposed to the consortium and all project partners agreed upon their use (other countries medical maps will finally have similar structure). The adopted fields are: "last update", "language(s)", "title", "location", "description" and "keywords" of the web site but also "trust marks: are they present or not", "trustworthiness (a first estimation on the quality of the medical content: is it reliable?)", "advertisements: are they present or not?".

Below, we present, some difficulties we had during categorization and we comment some of the characteristics of the profile of the Greek medical web as this emerges from our survey. We finally estimate the necessity of the adoption of accreditation methods for online medical content through quality labelling mechanisms in the Greek reality.

We first collected a few thousands of URLs with the assistance of a search engine wrapper. The wrapper queried the Google search engine⁷ with several sets of health related keywords, in both Greek and English languages, and collected the resulting web sites. From the English keywords' results we kept only those corresponding to web sites originated from Greece. On the resulting Greek URLs' list, an automated filtering procedure was applied, where duplicates, overlapping and other irrelevant URLs were removed. 1603 URLs remained. Checking manually, one-by-one, all the 1603 URLs, we finally kept only 723 web sites having health-related content. This was our corpus.

We then categorized them according to the categories mentioned above. We additionally collected information (corresponding to quality criteria) for every web site of the map. The crawling software (developed for the purposes of the project), based on machine learning and heuristic methods, extracted the machine detectable information, which is "last update", "language(s)", "title", "location", "description" and "keywords".

Apparently, the 723 sites examined do not cover the totality of the Greek medical web content. However, they comprise a fair sample of that, which allowed us to make some useful observations with regard to this content. In future rounds we intend to crawl the Greek web more extensively in order to create the full map of the Greek medical web sites.

Categorization of Greek web medical content

During our effort to correlate the URLs found with the above mentioned categories, we found out that there are sites that can be placed under several categories. For instance, a site could be categorized as "private individual" and "healthcare service provider" at the same time. The same occurs with the categories "media and publishers" and "private individual". These because most of the monitored web sites do not provide a clear authorship; it is not obvious whether the web sites belong to a private individual or to a healthcare service provider. However, assigning a web site into more than one category, finally gives more information for the site's content.

Based on the agreed categories our results are distributed as follows:

Categories	URLs	Percentage %
Government organizations	15	2%
Healthcare service provider	211	28%
Media and publishers	64	9%
Patient organizations/self support groups	33	5%
Pharmaceutical company/retailer	51	7%
Private individuals	199	28%
Scientific or professional organizations	110	15%
Universities/research institutions	40	6%
Total	723	100%

⁷ <http://www.google.com/>

The majority of web sites belong to healthcare service provider category (211 URLs) and to the private individual category (199 URLs). This fact reveals that in Greek medical web private sector is dominant (which seems reasonable), while the web sites coming from the public sector like government organizations and universities/research institutions are the minority (54 URLs). Furthermore, it is remarkable that a great portion (110 URLs) of the Greek medical web belongs to scientific/professional organizations.

We also noticed that only three web sites have a quality seal (namely HONCode [2]) and all of them belong to the scientific or professional organizations category. We could argue that the non conformance to trust mark quality criteria characterizes the Greek medical web as a whole which demonstrates that Greek online medical content providers are not familiar with the quality labelling aspect. Thus, the quality of the content of Greek medical web sites appears to be doubtful. To support this, note that the html tags for “description” and “keywords” (which the crawler reads automatically), found either empty or contain misleading information in most Greek medical pages, while, for example, a quick look in a portion of the German medical web showed the opposite.

It was found that 146 out of the 723 web sites do contain advertisements relevant or not to medical content. There has been a lot of discussion about advertising and sponsorship in web sites. Our estimation is that the presence of advertisements and sponsorship in a web site don't necessarily affect its quality, given that advertisements are clearly distinguished from the scientific content and information on sponsorship policy is provided.

Most of the Greek healthcare service providers' websites as well as those of individual doctors do not provide health related information or online consultation services or public fora where discussion on medical issues could be held. They serve self-advertisement purposes, as they mainly provide descriptions on services and contact details.

It must be noted that we also found a number of pharmaceutical companies' web sites which mainly advertise and sell nutrition supplements without being accredited from the Greek National Organization for Medicines⁸. Finally, there were some online fora from unknown provider, where conversations on medical subjects, mainly between patients, were held; such sites could not be placed under any defined category. The above types of sites will be used for project-internal purposes only, and they won't be included in the medical map that we will publicise. This is due to the fact that we do not want to include possibly harmful web sites in a public map. It must also be noted that with respect to sites of complementary/alternative medicine, we decided to include them in the map only when they are clearly accredited by medical doctors.

Concluding, only few Greek medical web sites conform to the biggest part of the selected criteria so as to be considered of good quality. We could say that accreditation for medical web content, in our country, constitutes a high priority need. Possible establishment of an independent mechanism for the quality accreditation of Greek health web sites, could force health content providers to the following directions:

- For already existing online medical content: conform to generally accepted quality criteria defined by specialists. For online medical content scheduled to be published: designed to adapt to specific standards (presence of detailed information on the content provider, authorship information, last update, contact data, etc.).
- High quality web sites, trusted by health information consumers, would clearly boost the opinion that the web is not an advertising or dangerous space, but a powerful source of information and must be considered as such. In the same direction, the national medical sector could be motivated to develop web resources of quality, extending the usefulness of the medium and eventually attract a larger amount of users.

Concluding Remarks

Since the number of medical websites as well as the patients' interest for such information is growing, it seems necessary to establish mechanisms to control their quality.

⁸ <http://www.eof.gr/>

The resulting technology of 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.

Regarding the Greek medical web, our survey showed that most sites do not meet a standard quality level. At the same time, the majority of them use the web space for self-advertising purposes and not as a medium of communication between professionals and patients. The concept of a web site providing useful and responsible medical information is rare in the Greek Health web.

We believe that an accreditation procedure could constitute the main motive for Greek medical content providers and the medical society to change their current perspective against the need for and the value of the web.

References

- [1] Mayer MA, Leis A, Sarrias R, Ruíz P. Web Mèdica Acreditada Guidelines: reliability and quality of health information on Spanish-Language websites. In: Engelbrecht R et al. (ed.). Connecting Medical Informatics and Bioinformatics. Proc of MIE2005 (2005), 1287-92.
- [2] Health on the Net Foundation (HONCode). Home page. Available from: <http://www.hon.ch> .
- [3] 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 283 (12) (2000), 1600-1606.
- [4] Hi-Ethics, Inc. Health Internet Ethics: Ethical Principles for offering Internet Health services to consumers. Available from: <http://www.hiethics.com/Principles/index.asp> .
- [5] 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 10(6) (2002), 515.
- [6] URAC. Health Web Site Accreditation. Home page. Available from: <http://webapps.urac.org/websiteaccreditation/default.htm> .
- [7] CISMef. <http://www.chu-rouen.fr/cismef/>
- [8] Curro V, Buonomo 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 29(2) (2004), 95-107.
- [9] 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.
- [10]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", WDA 2003, Edinburgh, UK (2003), 75-78.
- [11]V. Karkaletsis, C.D. Spyropoulos, C. Grover, M.T. Paziienza, 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), Valencia, Spain (2004), 725 - 729.