Quality Labelling of Medical Web Content: the MedIEQ proposal

Vangelis Karkaletsis[1], *Miquel Angel Mayer*[2]

[1]National Centre for Scientific Research (NCSR) "Demokritos", Institute of Informatics & Telecommunications, Athens, Greece, vangelis@iit.demokritos.gr

[2]Web Médica Acreditada (WMA), Medical Association of Barcelona (COMB), Spain, mmayer.wma@comb.es

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 presents the current mechanisms for labelling medical web content and introduces the work done in the EC-funded projects Quatro and MedIEQ. Quatro has defined a vocabulary for quality labels and a schema to deliver them in a machine-processable format. MedIEQ aims to support the work of medical quality labelling organisations providing tools to monitor already labelled medical sites as well as to locate unlabelled sites and examine them against a set of pre-defined labelling criteria.

Keywords

Medical web content, quality labelling, semantic web technologies, web content analysis.

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. [1]. On the other hand, patients continue to find new ways of reaching health information and more than four out of ten health information seekers say the material they find affect their decisions about their health [2, 3]. However, it is difficult for health information consumers, such as the patients and the general public, to assess by themselves the quality of the information because they are not always familiar with the medical domains and vocabularies [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 supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Websites" [13]. The specific aim was to specify a commonly agreed set of simple quality criteria on which Member States, as well as public and private bodies, may build upon for developing mechanisms to help improving the quality of the content provided by health related web sites. These criteria should be applied in addition to relevant Community law. As a result, a core set of quality criteria was established. These 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.

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 such as the ones specified within the eEurope 2002 initiative, 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, which specifies how to layout the content of a web page addressing human readers. HTML as such cannot be exploited efficiently by information retrieval techniques in order to provide visitors with additional information on the web sites' content. This "current web" must evolve in the next years, from a repository of human-understandable information, to a global knowledge repository, where the information should be machine-readable and processable, enabling the use of advanced knowledge management technologies [15]. This change is based on the exploitation of Semantic Web technologies. The Semantic Web is "an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation" based in metadata (i.e. semantic annotations of the web content) [16]. These metadata can be expressed in different ways as the Resource Description Framework (RDF) language[1]. RDF is the key technology behind the Semantic Web, providing a means of expressing data on the web in a structured way that can be processed by machines.

In order for the medical quality labelling mechanisms to be successful, they must be equipped with semantic web technologies that enable the creation of machine-processable labels as well as the automation of the labelling 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 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 rating mechanisms. Section 3 presents the on-going work in the EC-funded project Quatro for the definition of a common vocabulary of quality labelling criteria and the development of a machine processable labelling schema. Section 4 presents the main objectives of the recently started EC-funded project MedIEQ which, building on Quatro results and continuing the work of previous relevant projects, aims to pave the way towards the automation of quality labelling process in medical web sites exploiting multilingual information retrieval and extraction techniques. 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[2]. This initiative includes all the labelling criteria specified in the content of the eEurope 20002 initiative. The first level of these criteria is presented in Table 1.

The "Identification" criterion concerns the provision of information such as the site ownership, contact information, professionals involved in case the site offers consultation services, category of the site such as associations for patients, medical centres, training, virtual consultation, scientific societies, information for consumers, etc. The "Content" criterion

[1] http://www.w3.org/RDF/

[2] http://wma.comb.es

enforces the provision of information on the updates made to the site, the authors of the medical resources, references to bibliography, etc. In addition, it examines the structure of web site in terms of its accessibility, the clear identification of internal and external links, as well as the accuracy of the scientific content provided and the target audience. "Confidentiality" examines the site's policy on the use of consumer data (i.e. if such a policy exists, whether the site informs the visitor on the specific policy it adopts, whether the site states that it respects the confidentiality laws). "Advertising and Sponsorship" examines whether the site contains advertising, if this is clearly distinguished from the scientific content, whether the site is sponsored and in such a case if it provides information on the sponsor policy. The "Virtual consultation" criterion examines the provision of services for virtual consultation of health users and/or professionals, for chat and news, and whether the site warns its visitors on the limits and use of these services. Finally, the "Non-compliance" criterion examines the potential misuse of the label (if, for instance, the label has expired).

Table 1 WMA labelling criteria

Identification
Content
Confidentiality
Advertising and Sponsorship
Virtual Consultation
Non compliance

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. 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, among others, by 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 the Agency for Quality in Medicine (AQuMed)[3]. Both WMA and AQuMed are participating in the MedIEQ project.

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 web site in order to initiate 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) to express acceptance of these recommendations;
2. The WMA Standing Committee assesses the web site based on the WMA criteria;
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 web site, as well as adding its name and URL to the index of accredited websites.

In the case of AQuMed, the medical web sites are selected according to some criteria of the Health on the Net (HON) Code [5] and are classified in four categories: treatment information, background information, self-help and counselling organisations and finally medical organizations. After that, the treatment information is evaluated according to the DISCERN guidelines[4] and with the CHECK-IN instrument[5]. Patients have access to this information through the website http://www.patienten-information.de/.

[3] http://www.aezq.de

[4] http://www.discern.org.uk/

Both rating mechanisms, as they are currently applied, present two major drawbacks for the work of the labelling experts involved. The label they add is not actually machine-processable such that a web browser (or a search engine) could locate it inside the browsed (or the retrieved) web page and parse it in order to "understand" its content and present it to the page's visitor. 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. Imagine, for instance, using one of the known search engines to retrieve information on a specific treatment, and receive search results annotated with content labels without having to visit the site itself. The users will, most probably, visit the labelled sites and ignore the unlabelled ones. This will press medical content providers to add machine-readable labels, issued by established labelling authorities, in their sites, promoting the labels use and improving the quality of medical information and services provided through the WWW.

However, establishing machine-processable labels is not enough. 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 following sections present two on-going EC-funded projects. The first one, named Quatro, concerns establishing machine-processable labels for various domains (medical domain is one of the cases examined). The second one, MedIEQ, is a recently started project that builds on Quatro work aiming at the development of technologies to support the work of labelling authorities in monitoring labelled sites and locate unlabelled ones.

3. Quatro vocabulary and labelling schema

Quatro (Quality Assurance and Content Description)[6] is an on-going EC-funded project which aims to provide a common vocabulary and machine readable schema for quality labelling of web content, 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.

Quatro's work on providing a platform for machine-understandable quality labels, also called trustmarks, is part of the Semantic Web activity. Quatro adds to the picture in two ways: by providing a way in which any number of web resources can easily share the same description; by providing a common vocabulary that can be used by labelling authorities. By basing the labels on RDF, Quatro is effectively promoting the addition of data on the web that a wide variety of other applications can use to build trust in a given resource.

At the time of writing this paper, the details of the Quatro vocabulary have been finalized and the complete vocabulary is available on the Quatro site and elsewhere, both as a plain text document and an RDF schema. It will be available for free usage by any Labelling Authorities (LAs) as they see fit. The project's vocabulary 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.

[5] http://www.patienteninformation.de/content/informationsqualitaet/download/check_in.pdf
[6] http://www.quatro-project.org

- 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.

LAs 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.

Work is now underway to develop applications to make use of the machine-readable labels:

- An application for checking the validity of machine-readable labels found in web resources. A label's validity is checked against the corresponding information found in the LA's database. Furthermore, Quatro also enables, for some cases, the checking of label's validity against the content of the web resource. The application is implemented as a proxy server, named QUAPRO.
- A browser extension, named ViQ, which enables the visual interpretation of label found in the web resource requested by the user, according to QUAPRO results. A user is therefore able to see that a site has a label and be notified on the label's validity and content.
- A wrapper for search engines' results, named LADI, which indicates the presence of label(s) on the web sites listed. This will be available for inspection by clicking an icon adjacent to the relevant result. As in the case of ViQ, label validation and user notification will be performed by QUAPRO.

One of the case studies in Quatro concerns the labelling of medical web sites through the involvement of the WMA labelling authority, which is also participating in the MedIEQ project.

4. MedIEQ

MedIEQ (Quality labeling of Medical Web Content Using Multilingual Information Extraction)[7] continues the work of previous projects in the area of medical quality labelling (MedCERTAIN[8], MedCIRCLE[9] and WRAPIN[10]). 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) established a third-party rating system to select high quality information medical web sites 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 [17]. The main objective of WRAPIN (Worldwide online Reliable Advice to Patient and Individuals) was to make available a tool to determine information quality by automatically checking a document against matching sources from databases of known quality.

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;

[7] http://www.medieq.org
[8] http://www.medcertain.org/
[9] http://www.medcircle.org/
[10] http://www.wrapin.org/

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 (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 [18].

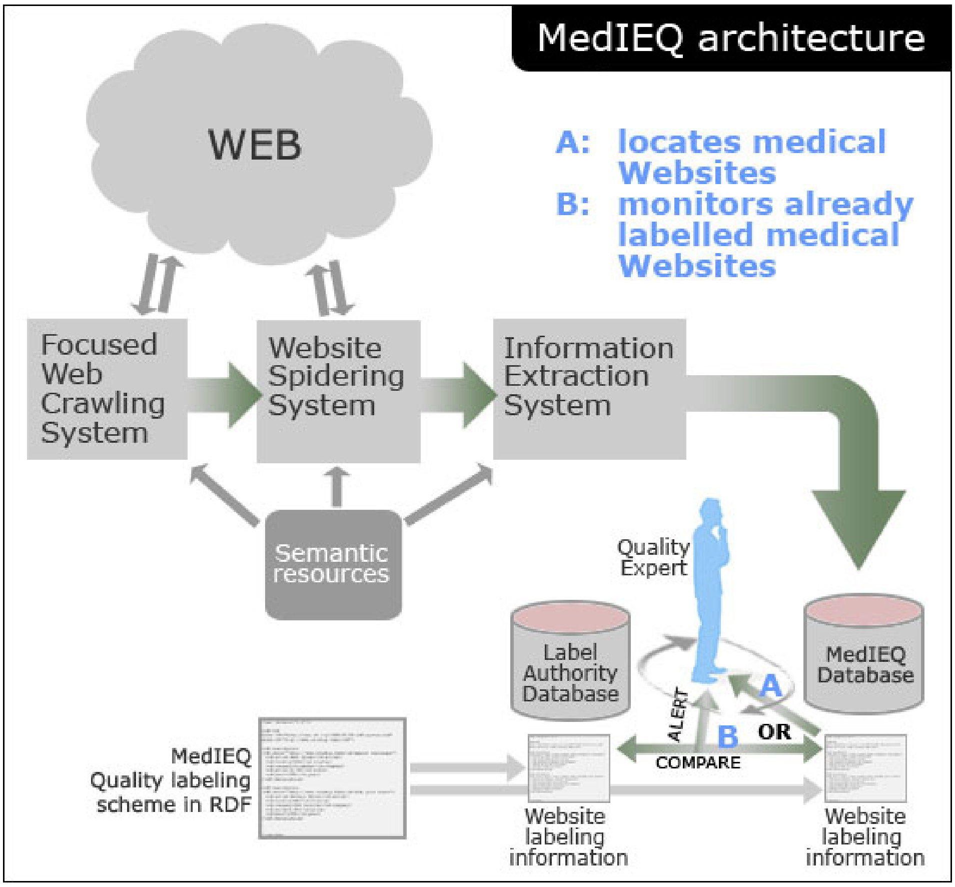
MedIEQ architecture
WEB
A: locates medical Websites
B: monitors already labelled medical Websites
Focused Web Crawling System
Website Spidering System
Information Extraction System
Semantic resources
Quality Expert
Label Authority Database
MedIEQ Database
ALERT
A
B
OR
COMPARE
MedIEQ Quality labeling scheme in RDF
Website labeling information
Website labeling information

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

- 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 [18].

- 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 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. 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.

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 separately 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.

5. Concluding Remarks

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

The main problems that these mechanisms face are related to the lack of machine-processable labels that can be identified and parsed by search engines or web browsers, as well as the need for continuous review and control of already characterised (accredited or classified) web sites and the location of ones that have not been characterised yet, tasks that currently require 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, web site directories are periodically updated due to the addition of new sites and changes in the characterization of the already visited ones.

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.

Acknowledgements

This research was partially funded by the EC through the DG SANCO “Public Health” project MedIEQ (Quality labeling of Medical Web Content Using Multilingual Information Extraction). MedIEQ involves the following partners: National Centre for Scientific Research "Demokritos", Greece (coordinator); Universidad Nacional de Education a Distancia, Spain; Web Medica Acreditada of the Medical Association of Barcelona, Spain; Agency for Quality in Medicine, Germany; University of Economics in Prague, Czech Republic; Helsinki University of Technology, Finland; I-sieve Technologies Ltd., Greece; Geneva University Hospitals - Service of Medical Informatics, 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: Engelbrecht R et al. (ed.). Connecting Medical Informatics and Bioinformatics. Proc of MIE2005 (2005), 1287-92.

[2] Eysenbach G. Consumer health informatics. BMJ 320 (4) (2000), 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 17(3) (2002), 180-5.

[4] Soualmia LF, Darmoni SJ, Douyère M, Thirion B. Modelisation of Consumer Health Information in a Quality-Controled gateway. In: Baud R et al. (ed.). The New Navigators: from Professionals to Patients. Proc of MIE2003 (2003), 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] Health on the Net Foundation (HONCode). Home page. Available from: http://www.hon.ch .

[7] 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.

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

[9] 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.

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

[11] CISMeF. http://www.chu-rouen.fr/cismef/

[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 29(2) (2004), 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 321 (2002), 598-602.

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

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

[17] 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.

[18] 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.