## *Management of Medical Website Quality Labels via Web Mining*

Vangelis Karkaletsis*
Institute of Informatics and Telecommunications, NCSR "Demokritos",
P.O. BOX 60228, Ag. Paraskevi 15310, Athens, Greece
Phone: + 30 21 0 6503197, fax: + 30 21 0 6532175, email: vangelis@iit.demokritos.gr

Konstantinos Stamatakis
Institute of Informatics and Telecommunications, NCSR "Demokritos",
P.O. BOX 60228, Ag. Paraskevi 15310, Athens, Greece
Phone: + 30 21 0 6503197, fax: + 30 21 0 6532175, email: kstam@iit.demokritos.gr

Pythagoras Karampiperis
Institute of Informatics and Telecommunications, NCSR "Demokritos",
P.O. BOX 60228, Ag. Paraskevi 15310, Athens, Greece
Phone: + 30 21 0 6503197, fax: + 30 21 0 6532175, email: pythk@iit.demokritos.gr

Martin Labský
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
Phone: +420 224095495, fax: +420 224095400, email: labsky@vse.cz

Marek Růžička
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
Phone: +420 224095495, fax: +420 224095400, email: ruzicka@vse.cz

Vojtěch Svátek
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
Phone: +420 224095495, fax: +420 224095400, email: svatek@vse.cz

Enrique Amigó Cabrera
ETSI Informática, UNED, C/ Juan del Rosal, 16, Ciudad Universitaria, 28040, Madrid, Spain
Phone: +34 913988651, fax: +34 913986535, email: enrique@lsi.uned.es

Matti Pöllä
Adaptive Informatics Research Centre, Helsinki University of Technology, Helsinki, Finland
Phone: +358 9 451 5115, fax: +358 9 451 3277, email: matti.polla@tkk.fi

Miquel Angel Mayer
Web Médica Acreditada, Medical Association of Barcelona (COMB), 08017 Barcelona, Spain
Phone: +34 9356788 35, fax: +34 935678836, email: mmayer.wma@comb.es

Dagmar Villarroel Gonzales
Agency for Quality in Medicine (AquMed), Berlin, Germany
Phone: +49 30 4005 2512, fax: +49 30 4005 2555, email: villarroelgonzales@azq.de

# *Management of Medical Website Quality Labels via Web Mining*


## Abstract

The WWW is an important channel of information exchange in many domains, including the medical one. The ever increasing amount of freely available healthcare-related information generates, on the one hand, excellent conditions for self-education of patients as well as physicians, but on the other hand entails substantial risks if such information is trusted irrespective of low competence or even bad intentions of its authors. This is why medical website *certification* (also called 'quality labeling') by renowned authorities is of high importance. In this respect, it recently became obvious that the labeling process could benefit from employment of web mining and information extraction techniques, in combination with flexible methods of web-based information management developed within the semantic web initiative. Achieving such synergy is the central issue in the *MedIEQ* project. The *AQUA* (Assisting QUality Assessment) system, developed within the MedIEQ project, aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts.


## 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, personal sites, health provider institutions, commercial sites, etc. (Mayer et.al., 2005). 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 affects their decisions about their health (Eysenbach, 2000; Diaz et.al., 2002). 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 (Soualmia et.al., 2003).

Although there are divergent opinions about the need for certification of health websites and adoption by Internet users (HON, 2005), different organizations around the world are working on establishing standards of quality in the certification of health-related web content (Winker et.al., 2000; Kohler et.al., 2002; Curro et.al., 2004; Mayer et.al., 2005). The European Council supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Websites" (EC, 2002). 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 websites. 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, certification systems, or any other initiative adopted by relevant parties, at European, national, regional or organizational level.

This stress on content quality evaluation contrasts with the fact that most of the current Web is still based on HTML, which only 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 websites' content. This "current web" must evolve in the next years, from a repository of human-understandable information, to a global knowledge repository, where information should be machine-readable and processable, enabling the use of advanced knowledge management technologies (Eysenbach, 2003). 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 on metadata (i.e. semantic annotations of the web content) (Berners-Lee et.al., 2001). These metadata can be expressed in different ways using the Resource Description Framework (RDF) language. 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 labeling 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 labeling process. Among the key ingredients for the latter are *web crawling* techniques that allow for retrieval of new unlabelled web resources, or *web spidering and extraction* techniques that facilitate the characterization of retrieved resources and the continuous monitoring of labeled resources alerting the labeling agency in case some changes occur against the labeling criteria.

The *AQUA* (Assisting QUality Assessment) system[1], developed within the MedIEQ project[2], aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts by making them computer-assisted. AQUA consists of five major components (each, in turn, incorporating several specialized tools): Web Content Collection (WCC), Information Extraction (IET), Multilingual Resources Management (MRM), Label Management environment (LAM), and Monitor Update Alert (MUA). While WCC, IET and MUA together constitute the web data analysis engine of AQUA, MRM provides them access to language-dependent medical knowledge contained in terminological resources, and LAM handles the generation, storage and retrieval of resulting labels. The user interface of AQUA allows for both entirely manual labeling and labeling based on the results of automatic analysis. In this chapter we will describe the challenges addressed and results achieved by applying the WCC and IET tools to raw web data, as well as the subsequent processes of quality label handling by LAM.

## Categories and Quality of Medical Web Content – a Survey

In order to investigate what types of Medical Web Content exist, at the beginning of the project we conducted a survey on a set of Greek health-related websites, classifying them into the following categories: "government organization", "healthcare service provider", "media and publishers", "patient organization / self support group", "pharmaceutical company / retailer", "private individual" and "scientific or professional organization". Apart from the categorization

[1] http://www.medieq.org/aqua/welcome.seam
[2] http://www.medieq.org

of these websites, we also collected additional information for them in order to construct a *medical web map*. The extra fields of information were the following: "last update", "language(s)", "title", "location", "description" and "keywords" of the website 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?".

**Table 1.** Categorization of Medical Web Content under review

| ***Categories*** | ***URLs*** | ***Percentage (%)*** |
|---|---|---|
| Government organizations | 15 | 2% |
| Healthcare service providers | 211 | 28% |
| Media and publishers | 64 | 9% |
| Patient organizations/ self support groups | 33 | 5% |
| Pharmaceutical companies/ retailers | 51 | 7% |
| Private individuals | 199 | 28% |
| Scientific or professional organizations | 110 | 15% |
| Universities/ research institutions | 40 | 6% |
| **Total** | **723** | **100%** |

We first collected a few thousands of URLs with the assistance of a *search engine wrapper*. The wrapper queried the Google search engine with several sets of health related keywords, in both Greek and English languages, and collected the resulting websites. From the English keywords' results we only kept those corresponding to websites 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 the remaining URLs, 723 websites were selected for having health-related content. These were then categorized according to the categories mentioned above. 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.

The majority of websites belong to the *healthcare service provider* category (211 URLs) and to the *private individual* category (199 URLs). This fact reveals that in Greek medical web, the private sector is dominant (which seems reasonable), while the websites coming from the public sector like government organizations and universities/research institutions are a 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, at the time of the survey, only three websites had a *quality seal*, namely, HON Code (HON, 2001) and all of them belong to the *scientific or professional organization* 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 labeling aspect. Thus, the quality of the content of Greek medical websites appears to be doubtful. To support this, note that the HTML tags for "description" and "keywords" (which the crawler reads automatically), were found as either empty or containing misleading information in most Greek medical pages, while, for example, a quick look into a portion of the German medical web showed the opposite. Concluding, only few Greek medical websites conform to the biggest part of the selected criteria as to be considered of good quality.

We also conducted analogous but less elaborate studies for other 'less-spoken' languages that are involved in the MedIEQ project but not covered by the partner labeling agencies, namely Czech and Finnish. The first observations of the Czech and Finnish medical web maps seem to confirm the hypotheses formed based on the analysis of Greek websites detailed above.

Thus, the establishment of mechanisms/infrastructures for the quality certification of health related websites is quite critical. Its positive role would amount to forcing health content providers to the following directions:

- For *existing* online medical content: conform to generally accepted quality criteria defined by experts. For online medical content *planned* 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* websites, already trusted by health information consumers, would clearly boost the opinion that the web is not an advertising-oriented 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 attracting a larger amount of users.

The MedIEQ project aims to directly contribute to this direction.

## State of the Art in Health Web Quality Labelling

Two major approaches currently exist concerning the labeling of health information in the internet: a) *filtering portals* (organizing resources in health topics and providing opinions from specialists on their content) and b) *third-party certification* (issuing certification trustmarks or seals once the content conforms to certain principles). In general, and in both approaches, the labeling process comprises three tasks that are carried out entirely or partially by most labeling agencies:

- *Identification* of new web resources: this could happen either by active web searching or on the request of the information provider, i.e. the website responsible actively asks for the review in order to get a certification seal.
- *Labeling* of the web resources: this could be done with the purpose of awarding a certification seal or in order to classify and index the web resources in a filtering portal.
- *Re-reviewing* or *monitoring* the labeled web resources: this step is necessary to identify changes or updates in the resources as well as broken links, and to verify if a resource still deserves to be awarded the certification seal.

This is the general case; eventually, any particular agency can integrate additional steps which may be necessary in its work. The two labeling agencies participating in MedIEQ, Agency for Quality in Medicine – AQuMed (http://www.aezq.de) and Web Mèdica Acreditada - WMA (http://wma.comb.es), represent the two approaches mentioned above: AQuMed maintains a filtering portal while WMA acts as a third-party certification agency.

The indexing and labeling process in AQuMed consists of five steps:

1. *Inclusion of a new resource.* There are two ways through which a new resource can be identified for indexing in AQuMed database. The first one is through internet search and the second one is through a direct request from the information provider. The websites are selected according to general criteria: content, form and presentation should be serious, authorship, sponsorship and creation/update date should be clear, and only websites without commercial interest should be indexed.
2. *Website classification.* Previously unlabelled websites are classified into four groups: treatment information, background information, medical associations/scientific organizations and self-help/counseling organizations. Only the sites with treatment information proceed to the next step.
3. *Evaluation.* Sites with treatment information are evaluated using the DISCERN (DISCERN) and Check-In (Check-In) instruments. DISCERN is a well-known user guidance instrument, and Check-In was developed by AQuMed in collaboration with the "Patient Forum" of the German Medical Association. Check-In is based on DISCERN and the AGREE (AGREE, 2004) instrument for critical evaluation of medical guidelines.
4. *Confirmation.* The database administrator has to confirm the result of the evaluation. It can be modified, erased, or simply confirmed.
5. *Feedback to the information provider.* AQuMed sends an e-mail with the result of the evaluation in the case of sites with treatment information and with the information about the admission into the AQuMed database in the case of other categories.

AQuMed's database is periodically populated through new internet searches and is regularly examined for broken links. The evaluated web resources are also periodically re-reviewed in order to identify changes against the criteria or other updates.

Similarly, the complete certification process in WMA consists of the following four steps:

1. The person in charge of a website sends a (voluntary) request to WMA in order to initiate the process. Using the online application form, the interested party provides certain information to WMA and has the chance to auto-check the WMA criteria based on the Code of Conduct and the Ethical Code;
2. The WMA Standing Committee assesses the website based on the WMA criteria (medical authorship, updating, web accessibility, rules in virtual consultation, etc.), and issues recommendations;
3. WMA sends a report to the person in charge who implements the recommendations;
4. When the recommendations have been implemented, it is possible to obtain the seal of approval. In such a case, WMA sends an HTML seal code to be posted on the

accredited website. In addition, WMA includes the site's name and URL to the index of accredited websites and an RDF file is generated.

## Experimental Collection of Labeling Criteria

In the MedIEQ project we decided to develop a representative collection of *labeling criteria*, which would reflect the needs of the *labeling agencies* involved in the project consortium and at the same time provide an adequate proof of concept for our general methodology for computer-assisted labeling. It is important to stress that the methodology and software tools are to a large degree independent of the concrete criteria and thus could be easily adapted to different criteria used by various agencies. Such adaptation is also eased by the fact that the criteria specification was also influenced by the analysis of criteria used by other organizations such as HON, and thus has significant overlap with them.

The set of labeling criteria used in MedIEQ (36 in total, organized in 10 different categories) is shown in Table 1. For each of these criteria, the AQUA system aims to identify and extract relevant information to be proposed to the expert (i.e. automatically provide information otherwise searched for manually within the site). The expert can accept or modify AQUA's suggestions and generate a quality label on the fly.

**Table 2.** The set of criteria examined in MedIEQ.

| *ID* | *Criterion Name* | *Description* |
|---|---|---|
| ***1. Resource Defining Information*** | | |
| 1.1 | Resource URI | Includes information identifying/describing the resource. Concerning the resource URI: a) whether the resource's URI is valid or not and b) in case it redirects to external domains, are these domains between those specified when the resource was added? The rest is information like the resource's last update, its title and the language(s) in which content is provided. |
| 1.2 | Resource title | |
| 1.3 | Resource last update | |
| 1.4 | Resource language(s) | |
| ***2. Ownership / Creatorship*** | | |
| 2.1 | Organization name(s) (owner) | The user should know who is behind the resource in order to judge by himself the credibility of the provided information. Therefore, information like the name(s) of the organization(s) providing the information and the type of this(these) organization(s) should be available. At the same time, the name(s), title(s) (e.g. MD, PhD, Dr, etc.) and contact details of website responsible(s), to contact in case of questions on health related issues, as well as the name(s) and contact details of the webmaster(s) should be available. |
| 2.2 | Organization type(s) (owner) | |
| 2.3 | Responsible name(s) | |
| 2.4 | Responsible title(s) | |
| 2.5 | Responsible(s) contact details | |
| 2.6 | Webmaster name(s) | |
| 2.7 | Webmaster(s) contact details | |
| ***3. Purpose / mission*** | | |
| 3.1 | Purpose / mission of the resource provided | It has to be clear for the user which is the goal and motivation of the provided information and for what kind of users it was created e.g. adults, children, people with diabetes, etc. |
| 3.2 | Purpose / mission of the owner(s) provided | |

| | | |
|---|---|---|
| 3.3 | Target / intended audience(s) | |
| 3.4 | Statement declaring limitation of the provided information | |
| ***4. Topics / Keywords*** | | |
| 4.1 | Topics / Keywords (UMLS) | Mapping of the resource's content to concepts from the UMLS Metathesaurus. |
| ***5. Virtual consultation*** | | |
| 5.1 | VC service available | A virtual consultation (VC) service is an online service allowing the user to ask questions and/or send/upload information on health related issues asking for advice. The name(s) and details of the person(s) responsible(s) for this service should also be clearly mentioned. Moreover, a declaration that VC is only a supporting means that cannot replace a personal consultation with a physician should be provided. |
| 5.2 | VC responsible name(s) | |
| 5.3 | VC responsible(s) contact details | |
| 5.4 | Statement declaring limitation of the VC service | |
| ***6. Funding / Advertising*** | | |
| 6.1 | Statement declaring sources of funding (sponsors, advertisers, etc.) | Health web resources should disclose possible conflicts of interest. For this reason it is important to know how and by whom a web resource is funded. If there are any sponsors, it has to be clear who they are. Furthermore, it should be stated that sponsors do not have any influence on the content. Additionally, it has to be known whether the web resource hosts or not advertising material in whatever format. In case that happens, such material should be clearly distinguished from informative material. Furthermore, information on resource's policy with regard to advertising must be easily accessible and clear. |
| 6.2 | Name(s) of funding (sponsoring) organization(s) | |
| 6.3 | Statement declaring limitation of influence of sponsors on content | |
| 6.4 | Advertising present | |
| 6.5 | Are advertisements clearly separated from editorial content? | |
| 6.6 | Policy with regard to advertisement | |
| ***7. Other Seal or Recommendation*** | | |
| 7.1 | Other seal(s) present | Are there other seals identified in the resource? Indicates that the resource already conforms to other, known quality criteria. Identifiers for other seals: a) Real seals: WMA, HONcode, pWMC, URAC, eHealth TRUST-E, AFGIS, b) Filtering health portals (a resource is recommended by): AQUMED, Intute, WHO ("Vaccine Safety Net") |
| 7.2 | Which other seal(s)? | |
| ***8. Information Supporting Scientific Content*** | | |
| 8.1 | References, bibliography (with links to literature) | Regarding the provided specialized health information (scientific parts of the resource) it is relevant to know if it is based on scientific books, medical journal articles, etc. For this, scientific articles or documents should include a references or bibliography section. Additionally, it is important to know if such information is up-to-date (publication and last modification dates are required) and who is the author of such content (author(s) name(s) and contact details are required for pages/documents providing scientific information). |
| 8.2 | Publication / creation date | |
| 8.3 | Last revision / modification date | |
| 8.4 | Author name(s) | |
| 8.5 | Author(s) contact details | |
| 8.6 | Editorial policy | |
| ***9. Confidentiality / privacy policy*** | | |
| 9.1 | Explanation on how personal data | Internet users are much concerned about protection of their |

| | | |
|---|---|---|
| | (visitor coordinates, e-mail messages, etc.) is handled | privacy and personal data. For this reason the resource should provide a confidentiality/privacy policy ensuring that personal data (visitor coordinates, e-mail messages, etc.) is safely handled, describing how these data are handled. |
| ***10. Accessibility*** | | |
| 10.1 | Accessibility level | The resource is examined upon various accessibility criteria and information on its accessibility level (whether the resource is of level A, AA or AAA) is deduced. |

# The AQUA System Overview

## *Development Objectives*

Taking into account WMA and AQuMed approaches, the AQUA tool (Stamatakis et. al., 2007) was designed to support the main tasks in their labeling processes, more specifically:

1. Identification of unlabelled resources having health-related content;
2. Visit and review of the identified resources;
3. Generation of content labels for the reviewed resources;
4. Monitoring the labeled resources.

Compared to other approaches that partially address the assessment process (Griffiths et. al., 2005; Wang & Liu, 2006), the AQUA system is an integrated solution. AQUA aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts by making them computer-assisted. The steps towards this objective are the following:

Step 1: Creating machine readable labels by:

- Adopting the use of the RDF model (W3C, 2004) for producing machine-readable content labels; at the current stage, the RDF-CL model (W3C, 2005) is used. In the final version of AQUA, another model called POWDER, introduced by the recently initiated W3C Protocol for Web Description Resources (POWDER) working group (W3C, 2007), will be supported.
- Creating a vocabulary of criteria, consolidating on existing ones from various Labeling Agencies; this vocabulary is used in the machine readable RDF labels.
- Developing a label management environment allowing experts to generate, update and compare content labels.

Step 2: Automating parts of the labeling process by:

- Helping in the identification of unlabelled resources.
- Extracting from these resources information relative to specific criteria.
- Generating content labels from the extracted information.
- Facilitating the monitoring of already labeled resources.

Step 3: Putting everything together; AQUA is implemented as a large-scale, enterprise-level, web application having the following three tiers:

- The user tier, including the user interfaces for the labeling expert and the system administrator
- The application tier where all applications run
- The storage tier consisting of the MedIEQ file repository and the MedIEQ database.

## *System Architecture*

AQUA addresses a complex task. However, various design and implementation decisions helped MedIEQ partners keep AQUA extensible and easy to maintain. The main characteristics of its implementation include:

a) open architecture,

b) accepted standards adopted in its design and deployment,

c) character of large-scale, enterprise-level web application, and

d) internationalization support.

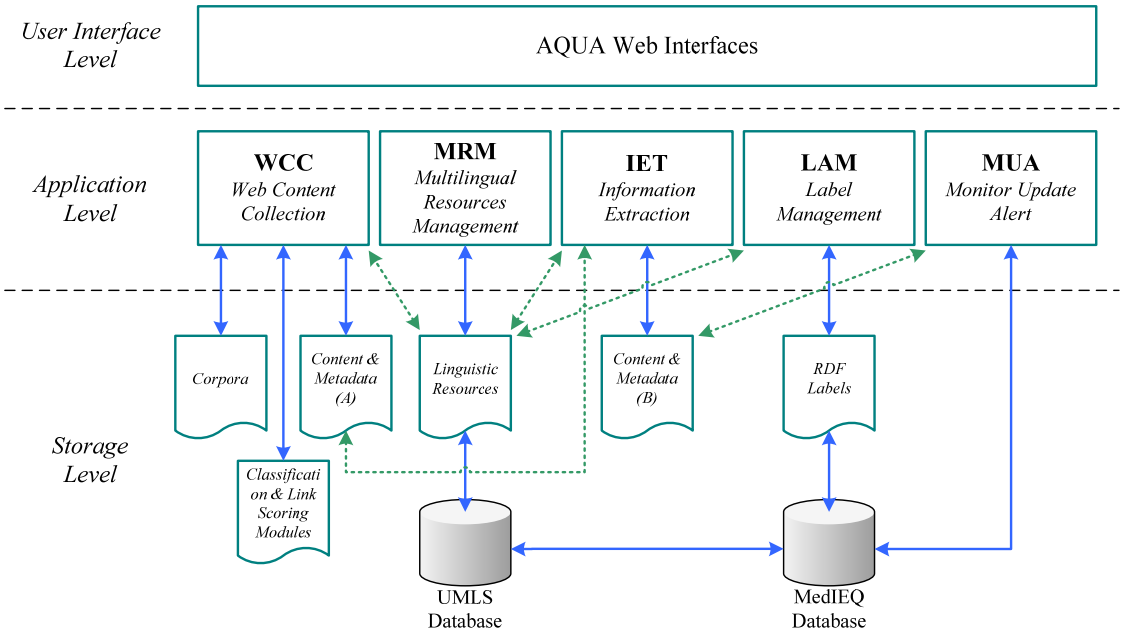



**Figure 1.** Architecture of the AQUA system.

AQUA incorporates several subsystems (see the application level in Figure 1) and functionalities for the labeling expert. The *Web Content Collection* (WCC) component identifies, classifies and collects online content relative to the criteria proposed by the labeling agencies participating in the project. The *Information Extraction Toolkit* (IET) analyses the web content collected by WCC and extracts attributes for MedIEQ-compatible content labels. The *Label*

*Management* (LAM) component generates, validates, modifies and compares the content labels based on the schema proposed by MedIEQ. The *Multilingual Resources Management* (MRM) gives access to health-related multilingual resources; input from such resources is needed in specific parts of the WCC, IET and LAM toolkits. Finally, *Monitor-Update-Alert* (MUA) handles auxiliary but important jobs like the configuration of monitoring tasks, the MedIEQ database updates, or the alerts to labeling experts when important differences occur during the monitoring of existing content labels.

While the first prototype, made operational in autumn 2007, only addresses the certification of new resources and covers two languages (English and Spanish), the full version of the system will also enable monitoring of already labeled resources and will cover 7 languages in total.

Figure 1 shows all the possible data flows in AQUA (dashed arrows): a) From WCC to IET: pages collected by WCC, once undergone a first-level extraction by WCC (extraction of metadata 1), are then forwarded to IET for further processing (extraction of metadata 2); b) From IET to MUA: MUA takes all metadata collected by both WCC and IET and updates the MedIEQ database; c) From MRM to WCC, IET, LAM: custom vocabularies generated by the MedIEQ users through MRM interface, can be accessed from other toolkits (WCC, IET, LAM), where the user may need them.

The following two sections are devoted to a more detailed description of AQUA, namely of its (manual) label management components and of its automated labeling support components.

# AQUA LAM Component: Creating Machine-Readable Labels

## *Representation Formalism for Machine-Readable Labels*

To make content labels machine-readable the use of the RDF model is adopted. At the current stage, the RDF-CL model is used. The RDF-CL model was issued by the EC-funded project Quality Assistance and Content Description (QUATRO) (www.quatro-project.org); it is currently being refined by the W3C Protocol for Web Description Resources (POWDER) working group (W3C, 2007). POWDER is expected to be completed before the end of the MedIEQ project and the plan is to use it in the final version of AQUA.

## *User Interaction with the Label Management Environment*

The *label management* interface and associated tools, together called LAM, allows experts to generate, update and compare content labels. From within the LAM user interface the user is able to a) generate new RDF labels from information automatically extracted by other AQUA tools, b) manually fill the relevant fields and generate new RDF labels, c) edit and update existing RDF labels, and d) compare RDF labels among themselves.

The user interface to generate/edit a label is a *web form* (see Figure 2) with input boxes, single and multiple select boxes, links and buttons. It is split into two distinct areas. The first part lets the user *constrain* the application of the label to certain hosts by explicitly declaring the host URIs or by adding regular expressions that properly identify them. Multiple hosts can be defined. Regular expressions for more fine-grained addressing can be defined as well. These definitions can be combined via the union and intersection operators and thus create rules that link different parts of a web resource with different labels.

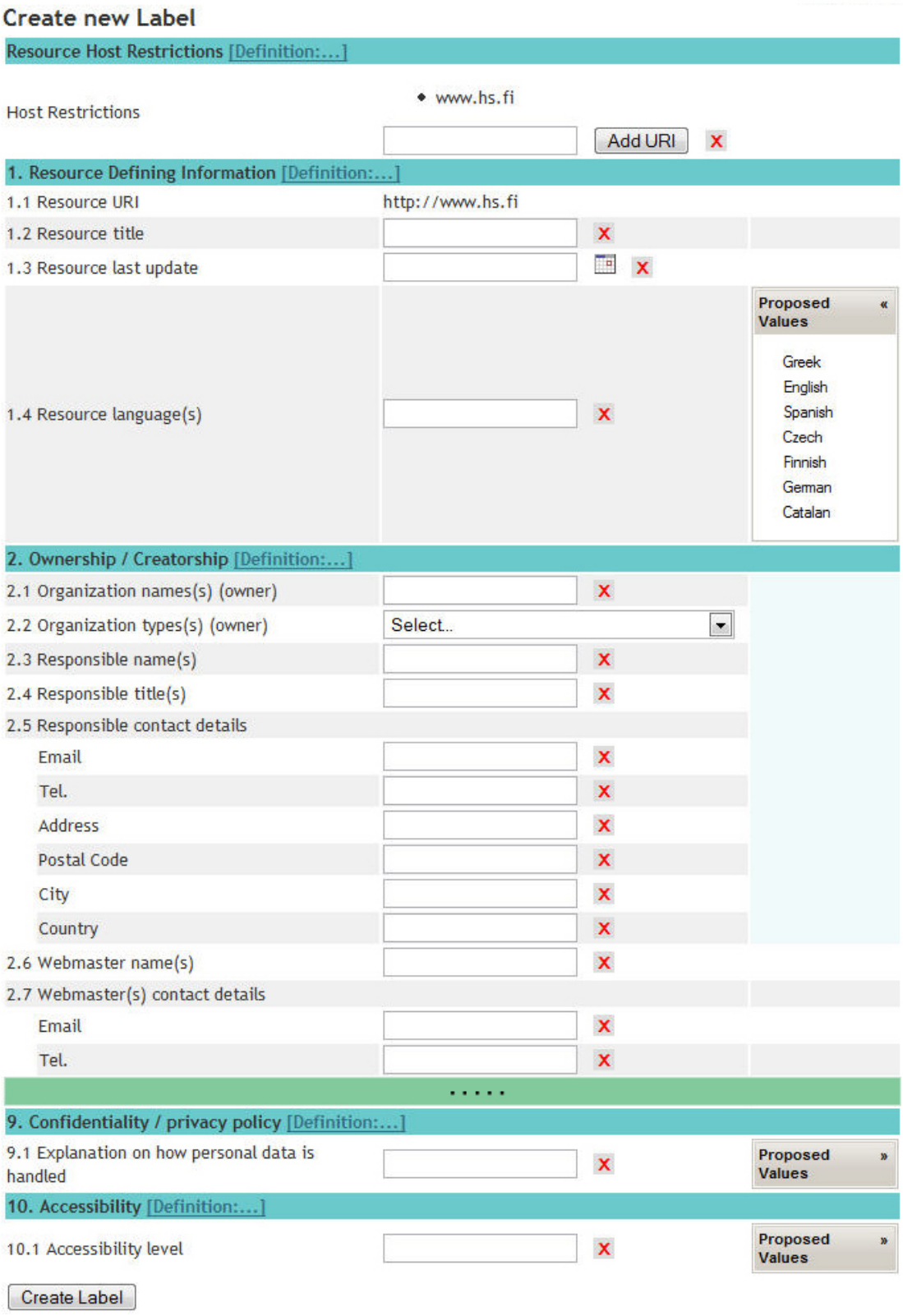



**Figure 2.** The AQUA label management environment (LAM) interface

The second part is where the *label properties* are assigned *values*. The label properties are the actual descriptors of a web resource, mapping the labeling criteria. A set of label descriptors can be linked with a set of host restrictions defined in the first part. Related properties are grouped to make the user filling them easier.

Once the user has filled the label metadata, restrictions and properties, s/he can *save* the label. There is a notification field that alerts the user if the label already exists in the system, and its changes are tracked by the AQUA version control system. In this case the user can save the label as a *revision* of an existing label. If the label is new, the user just selects to save the label. In both cases the user has the option to download an RDF/XML serialized form of the label. This serialized label can be assigned to the web resource by the site webmaster.

# AQUA WCC+IET: Automating Parts of the Labeling Process

## *Locating Unlabeled Web Resources*

The AQUA *crawling* mechanism is part of the *web content collection* environment (WCC) (Stamatakis et. al., 2007). Its AQUA interface is shown in Figure 3. The Crawler searches the Web for health-related content that does not have a content label yet (at least not a label found in MedIEQ records). It is a meta-search engine that exploits results returned from known search engines and directory listings from known Web directories. All collected URLs from all sources are merged and filtered, and a pre-final URLs list is returned. The merging / filtering process: a) removes possible duplicates, b) ignores sub-paths of URLs already in list, and c) removes URLs already having a content label (the Crawler consults the MedIEQ database for this).


**Figure 3.** Configuring the MedIEQ Crawler from the AQUA interface

The crawling process becomes even more focused with the aid of a *content classifier*, inductively trained to distinguish health content from non-health content. This classification component visits every URL from the merged / filtered pre-final URL list and checks its contents, thus filtering out some more entries.

The current version of the AQUA Crawler queries Google and Yahoo! *search engines* (with terms proposed by the user) and explores Web directories (again proposed by the user). By merely using general-purpose search engines, the Crawler inevitably inherits their shortcomings. Therefore, aiming to further enhance our Crawler, we also include two search mechanisms specialized to the *health domain*: one provided by HON (www.hon.ch) and another by Intute's Health and Life Sciences branch (www.intute.ac.uk). The Crawler interface is shown in Figure 3.

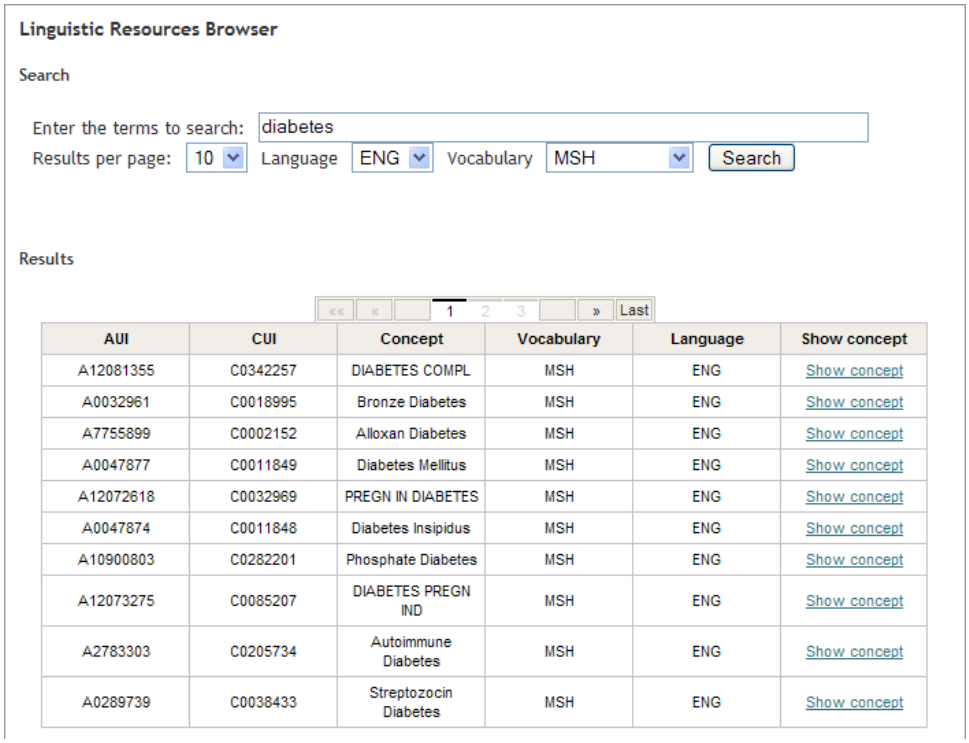
## *Browsing Medical Knowledge Sources*

One of the main requirements when working with medical web resources, is to identify and classify them based on standardized medical terms. Such terms (knowledge sources) have been globally defined by the Unified Medical Language System (UMLS) (www.nlm.nih.gov/research/umls/). UMLS provides a wide set of linguistic health resources, well maintained and up-to-date, containing health concepts and relations between concepts and between resources.

AQUA incorporates a module called Multilingual Resources Management Toolkit (MRM) that aims to support labeling experts in:

- easily accessing and browsing selected "knowledge sources" form the variety that UMLS provides.
- creating new, custom resources, to better support the labeling process

MRM is an environment from which linguistic resources, either UMLS-supported or not (custom or user generated) in different languages can be managed. MRM provides a user-friendly environment for accessing and managing both UMLS "knowledge sources" and custom resources (see Figure 4).




| AUI | CUI | Concept | Vocabulary | Language | Show concept |
|---|---|---|---|---|---|
| A12081355 | C0342257 | DIABETES COMPL | MSH | ENG | Show concept |
| A0032961 | C0018995 | Bronze Diabetes | MSH | ENG | Show concept |
| A7755899 | C0002152 | Alloxan Diabetes | MSH | ENG | Show concept |
| A0047877 | C0011849 | Diabetes Mellitus | MSH | ENG | Show concept |
| A12072618 | C0032969 | PREGN IN DIABETES | MSH | ENG | Show concept |
| A0047874 | C0011848 | Diabetes Insipidus | MSH | ENG | Show concept |
| A10900803 | C0282201 | Phosphate Diabetes | MSH | ENG | Show concept |
| A12073275 | C0085207 | DIABETES PREGN IND | MSH | ENG | Show concept |
| A2783303 | C0205734 | Autoimmune Diabetes | MSH | ENG | Show concept |
| A0289739 | C0038433 | Streptozocin Diabetes | MSH | ENG | Show concept |

**Figure 4.** Browsing Medical Knowledge Sources with AQUA

## *Spidering the Website*

While the Crawler proceeds from the initial user's content collection requirement to the identification of a relevant website as a whole, the *Spider*, in turn, examines individual pages of the site. The sites whose URLs are obtained from the Crawler are processed by the Spider one-by-one in several independent threads. Unreachable sites/pages are revisited in next run.

Since not all the pages of a web site are interesting for the labeling process, the Spider utilizes a content classification component that consists of a number of *classification modules* (statistical and heuristic ones). These modules decide which pages contain interesting information. Each of them relies on a different classification method according to the classification problem on which it is applied. Pages identified as belonging to classes relevant to the labeling criteria are stored locally in order to be exploited by the Information Extraction subsystem.

One of the main classification modules of the Spider is the "UMLS/MeSH categoriser", called POKA. POKA (http://www.seco.tkk.fi/tools/poka/) is a tool for automatic extraction of ontological resources (RDF, OWL, SKOS) from text documents. In the MedIEQ framework, POKA is used to find relations between medical web content and medical vocabularies such as MeSH to facilitate categorization of web resources. The POKA system is used as a component of the web spidering tool where the spider traverses health web sites by gathering internal links and visiting the corresponding web pages one by one. POKA is then harnessed to find medical terminology inside these pages by matching content with the MeSH vocabulary.

### *Extracting Information Relative to Criteria*

MedIEQ continues and builds upon the work of previous projects in the area of *information extraction* (IE) (Karkaletsis et.al. 2004; Rainbow, 2005; Labsky & Svatek, 2006). The AQUA IE toolkit (IET) employs a set of components responsible for the extraction of elementary information items found in each document and for the integration of these items into a set of semantically meaningful objects called *instances*. An instance (of certain general class) can be for example the set of contact information about a health provider or the set of bibliographic information about a scholarly resource referred to on the website.

The core IE engine currently used within IET is the *Ex* system (Labsky et al., 2007), which relies on combination of so-called *extraction ontologies* with exploiting local *HTML formatting regularities* and the option of embedding *trainable classifiers* to perform selected extraction subtasks. IET is built as a generic information extraction toolkit that supports changes and additions to the utilized labeling schemes. In this way, IET can also be used for IE using third-party labeling schemes and within other domains.

### *Monitoring of Already Described Resources*

Another part of AQUA, called MUA (from Monitor-Update-Alert), handles problems such as the *configuration* of *monitoring tasks*, the necessary MedIEQ *repository updates* and the *alerts* to labeling experts when important differences (relative to the quality criteria) occur during the monitoring of previously labeled sites. MUA thus extends the functionality of the content collection and extraction toolkits by shifting from a one-shot scenario to that of continuous monitoring.

MUA is currently in its design phase. Fully functional implementation is envisaged in the late phase of the MedIEQ project (mid-2008).

## Preliminary evaluation of AQUA

### *Locating Unlabeled Web Resources*

In this section, we summarize evaluation results on Crawler's content classification component. For this evaluation, we used an English corpus, consisting of 1976 pages (944 positive & 1032 negative samples), all manually annotated. Three different classifiers have been tested (SMO, Naïve Bayes and Flexible Bayes). All 1-grams, 2-grams and 3-grams were produced and the best of them according to information gain were selected (see Table 3). Best performance was achieved with 1-grams and HTML tags removed.

**Table 3.** Classification performance results for content classification

| | 1-grams (Tags removed) | | |
|---|---|---|---|
| | **Prec.** | **Rec.** | **Fm.** |
| NB | 0.75 | 0.63 | **0.68** |
| FB | 0.73 | 0.55 | **0.62** |
| SMO | 0.75 | 0.61 | **0.67** |

The relatively low performance of the content classifiers is justified by the fact that is difficult, even for humans, in various cases to assess whether a website has health-related content or not.

## *Spidering the Website*

The classification mechanism our Spider exploits has been examined using statistical classification techniques, for the criteria listed in Table 4. In addition, for the last criterion, a method based on heuristic detection was examined.

**Table 4.** The MedIEQ criteria upon which our classification components were evaluated

| ***Criterion*** | ***MedIEQ approach*** |
|---|---|
| The target audience of a website | Classification among three possible target groups: adults, children and professionals |
| Contact information of the responsible of a website must be present and clearly stated | Detection of candidate pages during the spidering process and forwarding for information extraction |
| Presence of virtual consultation services | Detection of parts of a website that offer such services during the spidering process |
| Presence of advertisements in a website | Detection of parts of a website that contain advertisements during the spidering process |

Several learning schemes, decision trees, naive Bayes and supported vector machines (SMO) were tested. The performance of the SMO classifier, which provides the best results, is presented in Table 5. As expected, the most difficult criterion for classification purposes is the target audience, as being a highly subjective one.

**Table 5.** SMO performance

| | **English** | | | **Spanish** | | |
|---|---|---|---|---|---|---|
| **Category** | **Precision** | **Recall** | **Fm** | **Precision** | **Recall** | **Fm** |
| Contact Info | 0.84 | 0.96 | 0.90 | 0.80 | 0.65 | 0.72 |
| Advertisements | 0.87 | 0.80 | 0.83 | 0.77 | 0.72 | 0.75 |
| Virtual Consultation | 0.87 | 0.87 | 0.87 | 0.75 | 0.58 | 0.65 |
| Adults | 0.78 | 0.75 | 0.77 | 0.65 | 0.64 | 0.65 |
| Children | 0.80 | 0.78 | 0.79 | - | - | - |
| Professional | 0.77 | 0.81 | 0.79 | 0.62 | 0.63 | 0.62 |

## *Extracting Information Relative to Criteria*

Table 6 shows preliminary results for extraction of *contact information*. Data sets were collected through website crawling and spidering, contact pages were identified and manually annotated for English (109 HTML pages), Spanish (200) and Czech (108). The collections contained roughly 6900, 4400 and 20000 named entities, respectively. The *contact extraction* ontologies (one per language with shared common parts) were developed based on seeing 30 randomly chosen documents from each dataset and evaluated using the remaining documents. Extraction ontologies utilize nested regular patterns at word, character and HTML tag level. They also refer to gazetteers such as lists of city names, common first names and surnames. Each ontology contained about 100 textual patterns for the context and content of attributes and also for the single extracted 'contact' class, attribute length and data type constraints and several axioms. For the results below we did not exploit trainable classifiers; their meaningful combination with the manually authored extraction knowledge is still work-in-progress, and when applied standalone, their results were so far slightly inferior to those achieved via extraction ontologies. We attribute this observation to small amount and large heterogeneity of training data.

The effort spent on developing and tuning the ontologies was about 2-3 person-weeks for the initial, English ontology, and 1-2 person weeks for its customization to Spanish and Czech. In the strict mode of evaluation, only exact matches are considered to be successfully extracted. In the loose mode, partial credit is also given to incomplete or overflown matches; e.g. extracting 'John Newman' where 'John Newman Jr.' was supposed to be extracted will count as a 66% match (based on overlapping word counts). Table 6 shows results in 'strict/loose' order. Some of the performance numbers below may be impacted by a relatively low inter-annotator agreement (English and Spanish datasets are still being cleaned to remove inconsistencies).

**Table 6.** Results of IET for contact information[3]

| | English | | | Spanish | | | Czech | | |
|---|---|---|---|---|---|---|---|---|---|
| **Attribute** | **Precision** | **Recall** | **Fm** | **Precision** | **Recall** | **Fm** | **Precision** | **Recall** | **Fm** |
| Degree/Title | 0.91 | 0.95 | 0.93 | - | - | - | 0.98 | 0.97 | 0.98 |
| Name | 0.89 | 0.91 | 0.91 | 0.85 | 0.89 | 0.87 | 0.97 | 0.99 | 0.98 |
| Street | 0.73 | 0.78 | 0.75 | 0.77 | 0.77 | 0.78 | 0.94 | 0.96 | 0.95 |
| City | 0.98 | 0.96 | 0.97 | 0.98 | 0.98 | 0.97 | 0.89 | 0.87 | 0.88 |
| Zip | 0.88 | 0.92 | 0.89 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Country | 0.98 | 1.00 | 0.99 | 0.98 | 1.00 | 0.96 | 0.97 | 0.91 | 0.95 |
| Phone | 0.98 | 0.97 | 0.97 | 0.93 | 0.94 | 0.93 | 0.99 | 1.00 | 0.99 |
| Email | 1.00 | 1.00 | 1.00 | 0.98 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |
| Company | 0.70 | 0.73 | 0.70 | - | - | - | - | - | - |
| Department | 0.60 | 0.69 | 0.64 | - | - | - | - | - | - |
| **Overall** | 0.90 | 0.91 | 0.92 | 0.91 | 0.94 | 0.92 | 0.96 | 0.97 | 0.98 |

[3] At the time of writing, degrees were not annotated as part of the Spanish collection and results for company and department names for Spanish and Czech were still work in progress.

### *AQUA Usability Evaluation*

The 1st AQUA prototype was also evaluated by the labelling organizations participating in the MEDIEQ project (namely, WMA and AQUMED). The primary goal of this evaluation was to conclude with a functional prototype that has the potential to be fully integrated within the day-to-day activities of a labelling organization. To this end, a parallel technical improvement action took place, refining given functionalities. The main objective of the extra technical improvement action was to enhance the overall system workflow, so as to better match the day-to-day practice. The specifications for these technical refinements were given by an iterative feedback process with the MedIEQ labeling organizations, during the evaluation. It must be noted that the current interim version of AQUA was well received by both labelling organizations participating in the Usability Evaluation testing, and that they expressed their confidence that AQUA will be fully integrated within their day-to-day labelling activities.

## Concluding Remarks

Other attempts to automatically assess health information in the internet exist but address the assessment process only partially. The Automated Quality Assessment procedure (AQA) (Griffiths et. al., 2005) ranks depression websites merely according to their evidence-based quality. The Automatic Indicator Detection Tool (AIDT), presented in a recent study (Wang & Liu, 2006), is suggested as a complementary instrument for the assessment of health information quality. AIDT is evaluated upon the automatic detection of pre-defined indicators that correspond to a number of technical quality criteria. However, AIDT focuses on a narrow scope of extraction techniques only, and does not address the assessment process as a whole. In contrast, the AQUA approach seems to be unique in covering the whole workflow of labeling agencies and employing a comprehensive and flexible collection of automated tools.

Assessing the quality of health-related information published on the internet is a task with great importance for the quality of the healthcare itself, due to a large proportion of patients as well as medical practitioners nowadays using the internet as a high-coverage information resource. It is at the same time a complex task as it has to examine the conjunction of a number of different aspects. Various initiatives around the world have attempted to codify these aspects into criteria, principles, codes of conduct, etc. Health specialists review online health resources and label them, either by issuing certification trustmarks or by including them in a thematic health portal. However this work can be proven quite tedious even for experienced users. Additionally, as it currently relies on manual effort, the labeling process is very time-consuming. Instruments to assist certain parts of the work exist; they however focus on specific problems only and none of them addresses the assessment process as a whole. In this context, efforts such as the MedIEQ project bring will wide reusability to content labels in the health domain by giving them machine-readable semantics and by providing services, such as those of the AQUA system, for creating and exploiting these machine-readable labels.

From the knowledge technology research viewpoint, the added value of MedIEQ is in employing existing technologies in a novel application: the automation of the labeling process in health-related web content. These technologies are *semantic web* technologies for describing web resources and *web search* (crawling and spidering) *and mining* (classification and information extraction) technologies for collecting domain-specific web content and extracting information from it. Experimental results for the mining components, investigating the performance of

different inductive-learning-based as well as knowledge-engineering-based methods, are promising.

## References


AGREE, Appraisal of Guidelines Research and Evaluation (AGREE), 2004. Available Online at: http://www.agreecollaboration.org/instrument/

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

Check-In, Available online at: http://www.patienten-information.de/content/informationsqualitaet/informationsqualitaet/images/check_in.pdf

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.

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.

DISCERN, DISCERN: Quality criteria for consumer health information. Available Online at: http://www.discern.org.uk/.

EC, European Commission. eEurope 2002: Quality Criteria for Health related Websites, 2002. Available Online at: http://europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf.

Eysenbach G. Consumer health informatics. BMJ 320 (4) (2000), 1713-16.

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

Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. J Med Internet Res. 2005 Dec 30;7(5):e59.

HON, Health on the Net Foundation, HONCode, 2001. Available Online at: http://www.hon.ch

HON, Health on the Net Foundation, Analysis of 9th HON Survey of Health and Medical Internet Users Winter 2004-2005, 2005. Available Online at: http://www.hon.ch/Survey/Survey2005/res.html

Karkaletsis V, Spyropoulos CD, Grover C, Pazienza MT, Coch J, Souflis D. A Platform for Crosslingual, Domain and User Adaptive Web Information Extraction. In Proceedings of the European Conference in Artificial Intelligence (ECAI); 2004; Valencia, Spain; p. 725-9.

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.

Labsky M, Svatek V. Information Extraction with Presentation Ontologies. In: ESWC'06 Workhshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro; 2006 June.

Labsky M., Svatek V., Nekvasil M., Rak D.: The Ex Project: Web Information Extraction using Extraction Ontologies. In: Proc. PriCKL'07, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery. Warsaw, Poland, October 2007.

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.

Rainbow, University of Economics Prague, Knowledge Engineering Group, Reusable Architecture for Intelligent Brokering Of Web information access (Rainbow), 2005. Available Online at: http://rainbow.vse.cz/descr.html

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.

Stamatakis K, Chandrinos K, Karkaletsis V, Mayer M.A, Gonzales D.V, Labsky D.V, Amigó E, Pöllä M. AQUA, a system assisting labelling experts assess health web resources. In Proceedings of the 12th International Symposium for Health Information Management Research (iSHIMR 2007), Sheffield, UK, 18-20 July, (2007), 75-84.

Stamatakis K, Metsis V, Karkaletsis V, Ruzicka M, Svátek V, Amigó E, Pöllä M. Content collection for the labeling of health-related web content. In Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 07), LNAI 4594, Amsterdam, 7-11 July, (2007), 341-345.

W3C, Resource Description Framework (RDF), 2004. Available Online at: http:// www.w3.org/TR/rdf-schema/

W3C, RDF-Content Labels (RDF-CL), 2005. Available Online at: http://www.w3.org/ 2004/12/q/doc/content-labels-schema.htm

W3C, Protocol for Web Description Resources (POWDER), 2007. Available Online at: http://www.w3.org/2007/powder/

Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the Web. Int J. Med Inform. 2006 May 31.

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

# Appendix: Key Terms and Their Definitions

| *Terminology* | *Definition* |
|---|---|
| Crawling | A web crawler is a program or automated script which browses the World Wide Web in a methodical, automated manner. This process is called web crawling. Web crawlers are mainly used to create a copy of all the visited pages for later processing. |
| Information Extraction | Automatic assignment of meaning to elementary textual entities and possibly more complex structured objects. |
| Metadata | Data that describes information about either online or offline data. Information that characterizes the who, what, where, and how related to data collection. Often, the information refers to special tagged fields in a document that provide information about the document to search engines and other computer applications. Web pages often include metadata in the form of meta tags. Description and keywords meta tags are commonly used to describe the Web page's content. Most search engines use this data when adding pages to their search index. |
| Resource Description Framework (RDF) | Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model, but which has come to be used as a general method of modeling information through a variety of syntax formats. The RDF metadata model is based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. |
| Semantic Web | The Semantic Web is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. It derives from W3C director Tim Berners-Lee's vision of the Web as a universal medium for data, information, and knowledge exchange. At its core, the semantic web comprises a set of design principles, collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed as prospective future possibilities that have yet to be implemented or realized. Other elements of the semantic web are expressed in formal specifications. |
| Spidering | A web spider is a complementary mechanism/tool to a web crawler. Web crawlers are mainly used to create a copy of all the visited pages for later processing, whereas, web spiders are used to gather specific types of information from Web pages. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. |
| Web Mining | Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the application that uses data mining to analyse and discover interesting patterns of user's usage data on the web. Web content mining is the process to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site. |