



---

## 1<sup>st</sup> Year Project Report

---

Distribution: Public

---

# MedIEQ

Quality Labeling of Medical Web content using Multilingual  
Information Extraction

National Centre for Scientific Research "Demokritos"  
Teknillinen Korkeakoulu – Helsinki University of Technology  
Universidad Nacional de Educacion a Distancia  
Col.legi Oficial de Metges de Barcelona  
Zentralstelle der deutschen Ärzteschaft zur Qualitätssicherung in der Medizin  
Vysoka Skola Ekonomicka V Praze  
I-Sieve Technologies Ltd

2005107

February 2007

Project ref. no.	<i>2005107</i>
Project acronym	<i>MedIEQ</i>
Project full title	<i>Quality Labeling of Medical Web content using Multilingual Information Extraction</i>

Security (distribution level)	<i>Public</i>
Contractual date of delivery	<i>31 December 2006</i>
Actual date of delivery	<i>16 February 2007</i>
Deliverable number	<i>D3.1</i>
Deliverable name	<i>1<sup>st</sup> Year Project Report</i>
Type	<i>Report</i>
Status & version	<i>Final</i>
Number of pages	<i>14</i>
WP contributing to the deliverable	<i>WP1</i>
WP / Task responsible	<i>NCSR</i>
Other contributors	<i>TKK, UNED, WMA, AQUMED, UEP, I-sieve</i>
Author(s)	<i>V. Karkaletsis, A. Tsakonas, K. Stamatakis (NCSR)</i>
EC Project Officers	<i>Artur Furtado</i>
Keywords	<i>Project results, 1<sup>st</sup> year</i>
Abstract (for dissemination)	<i>This document describes the work carried out and the results obtained in the 1<sup>st</sup> year of the project, as well as the actions foreseen for the 2<sup>nd</sup> year of the project.</i>

## **Table of contents**

1. Introduction.....	4
2. Schema and tools for the labelling of health-related web content .....	6
3. Content analysis of health-related web resources .....	7
4. Creation and management of multilingual resources.....	8
5. Prototype labelling system.....	9
6. Dissemination of the results.....	10
7. Evaluation .....	12
8. Concluding Remarks.....	13

# 1. Introduction

The number of health information web sites is increasing day by day. It is known that the quality of these web sites, published by various providers, is very variable and difficult to assess. 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. The European Council supported an initiative within eEurope 2002 to develop a core set of “Quality Criteria for Health Related Websites”. 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:

- Filtering portals: the web resources are classified according to predetermined criteria and organized in groups in order to facilitate a quick access to quality reviewed information. Examples of this mechanism are the following: “Catalog and Index of French-speaking Medical Sites” (CISMEF), “Organising Medical Networked Information - The UK Gateway to reliable health information” (OMNI), “Agency for Quality in Medicine” (AQUMED).
- Third party accreditation: an organization evaluates actively the quality of the website according to a set of criteria. Compliance with those criteria is showed with a logo or trust mark on the homepage. HON Code of the Health on the Net Foundation, URAC Accreditation Program, Web Mèdica Acreditada are the most lively known quality seals.

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

In order for the health-related content labelling mechanisms to be more effective, 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.

The EC-funded (DG-SANCO) project MedIEQ is a technology project employing semantic web technologies for the description of web resources, content analysis technologies for collecting domain-specific web resources and extracting information from them. What we aim in this project is to examine their applicability in a specific area with special importance that of assisting the labelling of health related web resources.

MedIEQ will deliver tools that crawl the web to locate unlabelled health web resources in seven different European languages in order to examine their content and label it using a set of machine readable quality criteria. MedIEQ tools will monitor already labelled health web resources alerting labelling experts in case the resources’ content is updated against the quality criteria.

MedIEQ aims to advance current medical quality labelling technology capitalizing on the results of previous work on quality labelling and content analysis. The implementation of this objective will be based on the realisation of the following more specific objectives:

- Develop a schema for the quality labelling of health-related web content and provide the tools supporting the creation, maintenance and access of labelling data according to this schema;
- Specify a methodology for the content analysis of health-related web resources according to the MedIEQ schema and develop the tools that will implement it;
- Specify a methodology and develop the tools for the creation and maintenance of the multilingual resources that will support content analysis in health-related web resources;
- Develop a prototype labelling system and demonstrate it in 7 different languages and two labelling applications (third party accreditation, classification).

The present report describes the work carried out and the results obtained in the 1<sup>st</sup> year of the project, as well as the actions foreseen for the 2<sup>nd</sup> year of the project. Sections 2-7 present the results (technical or not) directly related to the above objectives:

- Schema and tools for the labelling of health-related web content
- Content analysis of health-related web resources
- Creation and management of multilingual resources
- Prototype labelling system
- Dissemination of the results
- Evaluation

Section 8 concludes presenting some remarks based on the results achieved so far.

## 2. Schema and tools for the labelling of health-related web content

Building upon the experience from the recently finished EC-funded Quatro<sup>1</sup> project, in which two MedIEQ partners were involved (NCSR, WMA) and the workings of the W3C Web Content Labelling Incubator Group<sup>2</sup> (NCSR is actively involved in this initiative), we decided to use the RDF language and more specifically the RDF-CL<sup>3</sup> model to create the vocabulary for the medical labels.

Our aim is not to suggest the MedIEQ vocabulary as “the only one to use” but to show instead the value of machine readable labels, reusing existing RDF vocabularies and creating new criteria under a MedIEQ vocabulary only when our needs are not covered. So, the emphasis is on the technology and not on the labelling criteria included in the vocabulary.

The 1<sup>st</sup> version of the MedIEQ vocabulary was developed containing the first set of labelling criteria (see Deliverable D4.1). The selection of these criteria is based on the comparison and analysis of the criteria currently used by the participating labelling agencies WMA and AQuMed, the eEurope quality criteria of health web content guidelines and a label agency of international reference as Health on the Net Foundation (HON). The terms selected capture important aspects of health related content and form the case study for MedIEQ technology partners (for more details, see Deliverable D4.1). Partners have started discussing the final version of the vocabulary which extends the initial set of criteria (to be announced via the project’s web site by the end of March 2007).

The label management toolkit (LAM) is under development and a first version has already been developed. LAM functionalities will make it easy for the users to create, update and delete labels, to monitor changes in labels and properly import and export labels to the system. All these will be available via a user friendly web interface integrated in the MedIEQ system AQUA look and feel and be accessible from it.

---

<sup>1</sup> <http://www.quatro-project.org/>

<sup>2</sup> <http://www.w3.org/2005/Incubator/wcl>

<sup>3</sup> <http://www.w3.org/2004/12/q/doc/content-labels-schema.htm>

### 3. Content analysis of health-related web resources

The Web content collection and extraction methodology involves the following main steps:

1. Crawling: searching the web to identify on-line resources with health related content. Such a search is performed by the Crawler tool exploiting existing search engines and web directories.
2. Spidering: Health-related web resources either known or identified by the Crawler, are explored. All internal links in these resources are scored and the most promising followed, every visited resource's content is classified according to the labelling criteria issued and fit web resources are locally stored.
3. Extraction: Retrieved web resources are processed by the information extraction tools in order to locate and extract information relevant to the labelling criteria. For instance, this may result in the extraction of contact details from relevant web pages. The extracted items are then used as input to the label database (see deliverable D8).

The 1<sup>st</sup> version of the Web Content Collection (WCC) toolkit has been prepared. It includes the first versions of the following components (see Appendix II):

- Crawler, for locating unlabeled web resources
- Spider, for navigating in a web site (labelled or not) in order to locate interesting web resources according to the labelling criteria
- Corpus formation tool (CFT), for collecting corpus necessary for the training of the content classifiers employed by the crawler and the spider
- Trained module generator (TMG), for training the classifiers using the collected corpus
- Content classification component (CCC), for running the trained content classifiers

For details on content collection methodology see deliverable D6.

The 1<sup>st</sup> version of the information extraction toolkit (IET) is under development. IET will provide a uniform interface to multiple IE engines which will be used within MedIEQ. The IET will also provide functionalities to define, run and monitor the progress of extraction tasks, to export and visualize extracted information and to manage extraction data models. So far, experiments have been performed with a single prototype IE engine code-named Ex, based on extraction ontologies, which have been integrated into the IET. For details on the information extraction methodology, see deliverable D8.

## 4. Creation and management of multilingual resources

The Multilingual Resources Management (MRM) Toolkit will be a set of tools/components for the management of linguistic resources in different languages. The toolkit will allow to import linguistic and semantic resources in pre-defined supported formats into a repository and to access them in order to manipulate their contents for the project needs (for instance, in order to configure the content collection and extraction components using the appropriate keywords).

The key resource used by the MRM Toolkit is the Unified Medical Language System (UMLS) package, specifically one of its components: the UMLS Metathesaurus. Using UMLS is an effective way of obtaining and updating the resources, since it is distributed as a unique package. In addition, it already contains the standard version of MeSH controlled vocabulary in most of the languages involved in the MedIEQ project. In order to access, manipulate and integrate all the resources available in UMLS within the MRM Toolkit, we will also use the UMLSKS services. Thus, the MRM Toolkit will work as an extra-layer on top of the UMLS datasets and tools in order to allow communication between the resources repositories and the other components.

The MRM toolkit is currently under development. The 1<sup>st</sup> version is due to month 15 of the project (end of March 2007). An early version is to be delivered before month 15 in order to examine its integration within the MedIEQ system (AQUA). Among the actions foreseen in the next months, we are also evaluating the possibility of integrating along with the MeSH hierarchy, the SNOMED-CT medical terminology. We are also testing the Medical Text Indexer (MTI) as an alternative method of discovering MeSH headings for citation titles and abstracts and suggesting indexing terms.

The architecture and the main functionalities of MRM are presented in deliverable D10.

## 5. Prototype labelling system

The prototype MedIEQ labelling assisting system (also called AQUA, from Assisting Quality Assessment) consists of 5 subsystems or toolkits:

1. label management toolkit (LAM) manages (generates, validates, modifies, compares) machine readable labels based on the RDF-CL model;
2. web content collection toolkit (WCC) identifies, classifies and collects on-line content relative to a number of machine readable quality criteria (according to the proposed vocabulary in the MedIEQ schema);
3. information extraction toolkit (IET) analyses the web content collected and extracts attributes for MedIEQ compatible content labels;
4. multilingual resources management toolkit (MRM) gives access to health-related multilingual resources; input from such resources is needed in specific parts of both the WCC and IET toolkits;
5. monitor-update-alert toolkit (MUA) handles a few auxiliary but important jobs, like the configuration of monitoring tasks, the MedIEQ database's entries updates, the alerts to labelling experts when important differences occur during monitoring existing quality labels.

All data necessary to the different subsystems as well as to the overall AQUA system are stored in:

- the MedIEQ repository,
- the MedIEQ database,
- the UMLS database

An initial version of the AQUA interface for the labelling expert has been released. More details are given in the deliverable D12.

The 1<sup>st</sup> version is due to month 18 of the project (end of June 2007). This will integrate the first versions of all subsystems and will cover two languages (English, Spanish). Evaluation of the integrated system will start right after according to the evaluation strategy presented in Deliverable D15.

## 6. Dissemination of the results

The dissemination and exploitation strategy will be continuously reviewed and refined until the end of the project (see Deliverable D17).

The following are the most important dissemination activities during the 1<sup>st</sup> year of the project.

The project web site was set up at <http://www.medieq.org>. The public area of the project site contains information on the project objectives and partners as well as a set of the public documents produced so far (project leaflets, press releases in 8 languages, papers accepted for publication, presentations in conferences). A link is also provided to the recently ended project QUATRO (DG INFSO, Safer Internet programme) and the QUATRO tools that will be exploited in MedIEQ for demonstrating project results to the visitors of labelled web sites.

The following talks, papers presentations/publications have been done:

- V. Karkaletsis, invited speaker on the Public Health Programme at the eHealth 2006 High Level Conference, Session “Global trends and perspectives”, Malaga, 12 May, 2006 (<http://www.ehealthconference2006.org/>)
- V. Karkaletsis, invited talk of on “Quality Labeling of Web Content” at the 3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI 2006), 9 June, Athens, Greece (<http://www.icsd.aegean.gr/aiai2006/>)
- V. Karkaletsis, presentation of the MedIEQ project (“Quality Labelling of Health related Web Content: the MedIEQ project”) at the Joint Meeting of the DG SANCO Health Systems Working Party and DG INFSO eHealth Working Group, 22 June, Luxembourg
- Mayer MA., invited talk at the Workshop “La acreditación de calidad de los sitios web de salud” at the Autonoma University of Barcelona. InCom-UAB. Barcelona, Spain, 9 October 2006.
- J. Kosek, P. Nalevka, “Relaxed—on the Way Towards True Validation of Compound Documents”, Proceedings of 15th International World Wide Web Conference (WWW’2006), 23-26 May 2006, Edinburgh, Scotland.
- M.A. Mayer, coordinated the submission of a paper on the MedIEQ project to ICMCC Event 2006 (<http://www.icmcc2006.org>), Hague, Netherlands, 7-9 June 2006.
- V. Karkaletsis, K. Stamatakis, V. Metsis, V. Redoumi, D. Tsarouhas, “Health-related Web Content: quality labelling mechanisms and the MedIEQ approach”, Proceedings of the 4th International Conference on Information Communication Technologies in Health (ICICTH-2006), July 13-15, 2006, Samos Island, Greece (<http://www.ineag.gr/ICICTH/index.html>)
- J. Kosek, M. Labsky, J. Nemrava, M. Ruzicka, V. Svatek, “The MedIEQ project: evaluation of medical web resources with the help of information extraction” (In Czech), Datakon, the Annual Database Conference, October 2006, Brno, Czech Republic.
- Mayer MA, Leis A, Ruiz P, Karkaletsis V, Stamatakis K. MedIEQ: metadatos y sistemas de extracción semántica de información sanitaria en Internet y su aplicación en estrategias de calidad. Herramienta para la mejora de la calidad asistencial. Proceedings of the XI Congreso Nacional de Informática Médica. Informed 2006, Murcia: 67-72.
- Mayer MA, Leis A, Karkaletsis V, Vilarroel D. El proyecto europeo MedIEQ (Quality Labelling of Medical Web content using Multilingual Information Extraction): la Web Semántica al servicio de los usuarios de salud. Proceedings of the

VIII Jornadas de Gestión de la Información. Asociación Española de Documentación Científica (SEDIC), Madrid: 43-51.

NCSR was actively involved as a sponsoring organisation in the W3C Incubator Group on Content Labels (WCL) (<http://www.w3.org/2005/Incubator/wcl/>). WCL members are currently working for the formation of a W3C working group, named Protocol for Web Description Resources (POWDER) Working Group, that will exploit WCL results towards a relevant W3C recommendation (see [http://www.w3.org/2006/12/powder\\_charter.html](http://www.w3.org/2006/12/powder_charter.html)). The aim is to follow, at a later stage of MedIEQ, the data model for content labels to be proposed by the group.

At the beginning of the 2<sup>nd</sup> year of the project, the partners have already scheduled certain dissemination activities, for which they are currently working on, and they are exchanging views for more activities in the coming months. The scheduled activities involve the submission of papers to conferences/workshops, lectures on MedIEQ, as well as the organisation of an initial demo of AQUA during the “eHealth week 2007: From Strategies to Applications”, April 16-20, 2007, in Berlin. For more details, see Deliverable D17.

## 7. Evaluation

The evaluation of MedIEQ at the level of its individual components (label management, web content collection, information extraction and multilingual resource management toolkits), will be performed in the context of the corresponding work packages WP4, WP5, WP6 and WP7 respectively.

A systematic evaluation of the integrated system AQUA is foreseen within WP3.

The objectives of the evaluation strategy are:

- evaluate the performance of each of the three main tasks of the labelling process, i.e. identification of new unlabeled web resources, labelling of web resources and monitoring of labelled resources;
- evaluate the usability of AQUA interface;
- evaluate the labelling performance within a scenario for AQUA integration in the usual work of a labelling agency.

The first prototype of AQUA will support the first set of labelling criteria defined in Deliverable D4.1, in two languages (English and Spanish).

The details of evaluation strategy are described in Deliverable D15. This strategy involves two parts:

- Evaluation I: evaluates each of the three tasks named above and the usability of AQUA interface;
- Evaluation II: evaluates MedIEQ tools in the context of a scenario, where AQUA is partially integrated in the day-to-day work of the participating agencies.

## 8. Concluding Remarks

Concluding the 1<sup>st</sup> year project report, the main results during this year are the following:

- The 1<sup>st</sup> version of the MedIEQ labelling schema. This includes a vocabulary of 11 labelling criteria which are used in the MedIEQ machine readable labels implementing the adopted RDF-CL model.
- The 1<sup>st</sup> version of the Label Management toolkit (LAM). This provides tools for label generation, uploading, validation, editing, versioning and comparison.
- The 1<sup>st</sup> version of the Web Content Collection (WCC) toolkit. WCC provides tools for web crawling and spidering of web sites.
- An early version of the Information Extraction toolkit (IET) is being developed. IET provides tools for extracting from the textual content of web resources important information, according to the criteria specified in the MedIEQ schema.
- An early version of the Multilingual Resources Management toolkit (MRM) is being developed. MRM provides tools for accessing, editing, and exploiting UMLS resources in the project languages.
- An early version of the integrated MedIEQ system (AQUA) has been released. This is being constantly improved based on the partners' comments.

During the 2<sup>nd</sup> year, two milestones are the following:

- Delivery of the final version of the vocabulary extending the initial set of labelling criteria which are currently under discussion among the partners (end of March 2007). It must be stressed, as it was already done previously, that our aim is not to suggest this vocabulary as “the only one to use” but to show instead the value of machine-readable labels built on the basis of a common model. MedIEQ is a technology project aiming to promote the use of machine readable labels, where different vocabularies can be used, as well as the use of technology to support the work of labelling experts in labelling new content and monitoring the already labelled one.
- Delivery of the 1st version of the integrated system AQUA (end of June 2007). This is to be evaluated by the labelling authorities participating in the project, WMA and AQuMed, according to an already defined evaluation strategy. This strategy involves also the evaluation of AQUA tools in the daily practice of the labelling authorities.

In order to achieve these objectives and increase the project's impact, we consider very important the cooperation with other labelling authorities that do not participate in the project. This is one of the reasons that led us to form the project's Advisory Committee (AC) which involves experts in the field, among which representatives of three labeling authorities: Health on the Net (HON), Internet Content Rating Association (ICRA) and Catalog and Index of French-speaking Medical Sites (CISMEF). In addition, the AC provides us another instrument to increase the impact of project results.

We would also like to stress the fact that MedIEQ is progressing in parallel with other relevant projects and initiatives, with which it has strong links, and which can help significantly to promote its work. MedIEQ machine readable labels are based so far on the RDF-CL model proposed by the recently ended EC-funded project QUATRO, in which two MedIEQ partners were involved (NCSR, WMA). This model was studied in the context of the W3C Web Content Labelling Incubator Group (WCL), in which the MedIEQ coordinator was actively involved as sponsoring organisation. WCL will most probably lead to a W3C

working group (POWDER - Protocol for Web Description Resources) aiming to produce a content labelling model as a W3C recommendation. MedIEQ will use this model, at its final stage, for creating machine readable labels.

Furthermore, MedIEQ will exploit QUATRO tools to actually read the labels once created. NCSR which was responsible for the QUATRO proxy (QUAPRO) continues to maintain it, and WMA is in the process of converting several of its labels in order to be readable by the QUATRO tools.

In general, there is a growing interest for semantic web applications and MedIEQ is such an application in a domain presenting significant interest, both from a technological and a societal point of view, due to the large number of health-related web resources and the impact of their content to the society. MedIEQ consortium aims to develop technology that will support the work of labelling authorities, increasing the number of labelled health-related web sites and improving their monitoring.