



Evaluation of the first Prototype

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 Deliverable 16.1

February 2008

Project ref. no.	2005107
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 August 2007</i>
Actual date of delivery	<i>22 February 2008</i>
Deliverable number	<i>D-16.1</i>
Deliverable name	<i>Evaluation of the first Prototype</i>
Type	<i>Report</i>
Status & version	<i>Final</i>
Number of pages	<i>25</i>
WP contributing to the deliverable	<i>WP3</i>
WP / Task responsible	<i>AQUMED</i>
Other contributors	<i>WMA, NCSR, UNED, UEP, TKK)</i>
Author(s)	<i>Dagmar Villarroel (AQuMed) Miguel Angel Mayer (WMA) Silke Astroth (AQuMed) Angela Leis (WMA) Kostantinos Stamatakis (NCSR) Pythagoras Karampiperis (NCSR)</i>
EC Project Officer	<i>ArturFurtado</i>
Keywords	<i>Evaluation, quality labeling, recall, precision, usability</i>
Abstract (for dissemination)	<i>This document presents the first results of the evaluation of AQUA, which is the integrated system developed within the project MedIEQ. The purpose of AQUA is to support the work of the labeling expert. Since the primary goal of the evaluation of the 1st AQUA prototype 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, 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.</i>

Table of Contents

EXECUTIVE SUMMARY	4
1. INTRODUCTION	5
2. EVALUATION METHODOLOGY	7
2.1 Evaluation of the location and identification of unlabelled medical web sites	7
2.2 Evaluation of the accuracy of the automatic labeling.....	8
2.3 Usability evaluation.....	9
3. EVALUATION RESULTS	10
3.1 Evaluation of the location and identification of unlabelled medical web sites	10
3.2 Evaluation of the accuracy of the automatic labeling.....	12
3.3 Usability evaluation.....	14
4. EVALUATION RESULTS OF AQUA COMPONENTS	16
4.1 Web Content Collection (WCC) Toolkit	16
4.2 Information Extraction Toolkit (IET)	18
4.3 Multilingual Resources Management (MRM) Toolkit.....	19
5. CONCLUSIONS	23
APPENDIX A: LIST OF URLS USED FOR THE EVALUATION OF THE ACCURACY OF THE AUTOMATIC LABELING (ENGLISH WEB SITES)....	24
APPENDIX B: LIST OF URLS USED FOR THE EVALUATION OF THE ACCURACY OF THE AUTOMATIC LABELING (SPANISH WEB SITES)....	25

Executive Summary

The MedIEQ integrated system AQUA (Assisting QQuality Assessment) aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labelling experts by making them computer-assisted. AQUA incorporates several sub-systems and functionalities for the labelling expert. This document describes the evaluation results and analysis of the 1st AQUA prototype which integrates the first versions of the toolkits for content collection, information extraction and resources management, the tools for exploiting the RDF schema, and the interface for the labeling expert. This document also includes the results of the separate evaluations of the prototype components that have been performed in the corresponding WPs. The evaluation of MedIEQ at the level of its individual components (label management, web content collection, information extraction and multilingual resource management toolkits), was performed in the context of the corresponding work packages WP4, WP5, WP6 and WP7 respectively. In WP3, a systematic evaluation of the integrated system AQUA was conducted. The first prototype of AQUA supports the initial set of labeling criteria, in two languages (English and Spanish). The evaluation was conducted, for five of the labelling criteria contained in the first set, in English by AQUMED and in Spanish by WMA.

Since the primary goal of the evaluation of the 1st AQUA prototype 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, 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. As a result, the scope of the evaluation itself was refined (in comparison with the evaluation scope reported in D15 "Evaluation strategy") so as to provide more focused feedback towards this direction. The scope of this evaluation was the performance evaluation of AQUA on supporting the labelling process (i.e. identification of new unlabeled web resources and labelling of web resources), as well as, the usability evaluation of AQUA interface. The evaluation of the labelling performance on a real integration scenario in the usual work of a labelling organization was considered as the evaluation scope of the 2nd version of AQUA prototype.

1. Introduction

The AQUA system 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.

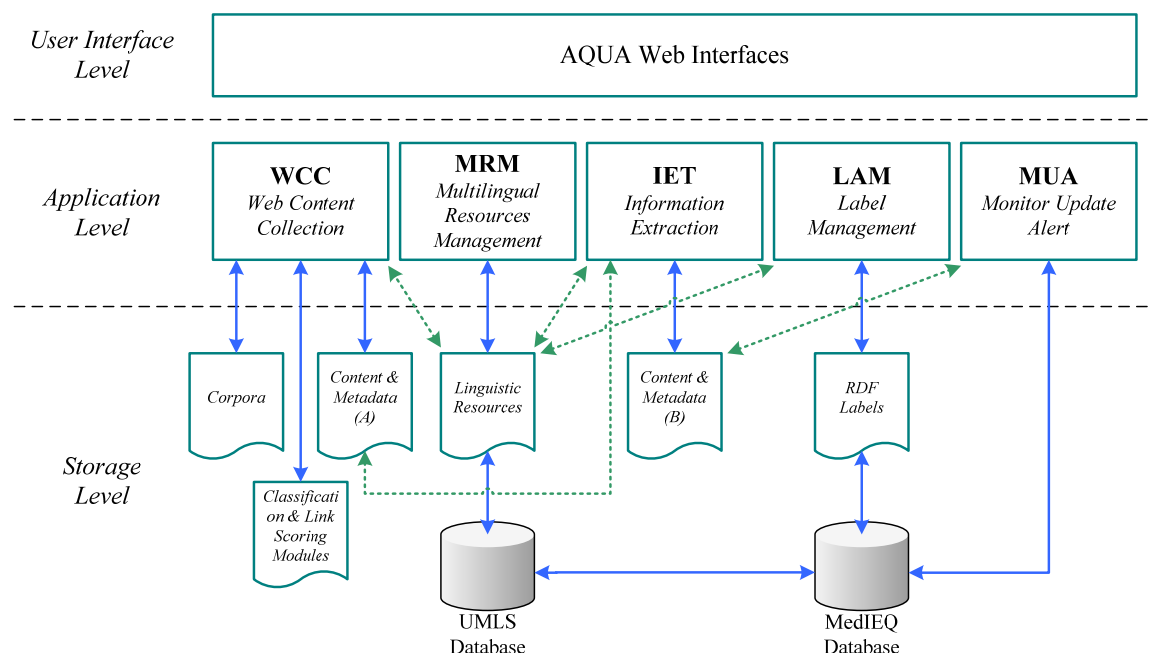


Figure 1. Architecture of the AQUA system.

AQUA incorporates several sub-systems (see application level in Figure 1) and functionalities for the labeling expert. Web Content Collection (WCC) identifies, classifies and collects online content relative to the criteria (proposed by the labeling agencies participating in the project). Information Extraction Toolkit (IET) analyses the web content collected by WCC and extracts attributes for MedIEQ compatible content labels. Label Management (LAM) generates, validates, modifies, compares content labels based on the schema proposed by MedIEQ. Multilingual Resources Management (MRM) toolkit gives access to health related multilingual resources; input from such resources is needed in specific parts of the WCC, IET and LAM toolkits.

This document describes the evaluation results and analysis of the 1st AQUA prototype which integrates the first versions of the toolkits for content collection, information extraction and resources management, the tools for exploiting the RDF schema, and the interface for the labeling expert. This document also includes the results of the separate evaluations of the prototype components that have been performed in the corresponding WPs.

The evaluation of MedIEQ at the level of its individual components (label management, web content collection, information extraction and multilingual resource management toolkits), was performed in the context of the corresponding work packages WP4, WP5, WP6 and WP7 respectively. In WP3, a systematic evaluation of the integrated system AQUA was conducted.

The first prototype of AQUA supports the initial set of labeling criteria (see D4.1), in two languages (English and Spanish). The evaluation was conducted, for five of the labelling criteria contained in the first set, in English by AQUMED and in Spanish by WMA.

Based on the evaluation strategy defined in Deliverable D15 and some operational modifications decided during the evaluation process, the evaluation took place from November 2007 until February 2008.

This document is structured as follows: In section 2, we present an overview of the evaluation methodology used, with emphasis on the refinements in comparison with the methodology reported in deliverable D15 "Evaluation strategy". Section 3 presents the detailed evaluation results and the feedback given by MedIEQ labelling organizations as an input to the continuous technical improvement action. Section 4 presents the evaluation results at the level of its individual components and Section 5 provides a summary of the conclusions extracted from the overall evaluation. Finally, in the corresponding appendixes the data (URLs) used for the evaluation of the first prototype of AQUA are presented.

2. Evaluation Methodology

Since the primary goal of the evaluation of the 1st AQUA prototype 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, 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. As a result, the scope of the evaluation itself was refined (in comparison with the evaluation scope reported in D15 "Evaluation strategy") so as to provide more focused feedback towards this direction. The scope of this evaluation was the performance evaluation of AQUA on supporting the labelling process (i.e. identification of new unlabeled web resources and labelling of web resources), as well as, the usability evaluation of AQUA interface.

According to the evaluation methodology presented in detail in deliverable D15, the evaluation strategy consists of two main phases:

- Evaluation I, responsible for the evaluation of the different tasks available in AQUA. This phase comprises four parts, in accordance with the four main functional aspects of AQUA:
 - Location and identification of unlabelled medical web sites
 - Automatic labeling
 - Automated monitoring and alerting system
 - Usability aspects
- Evaluation II, responsible for the evaluation of the performance of AQUA when it is integrated in the day-to-day work of the participating agencies.

For the evaluation of the 1st version of AQUA, Evaluation phase I was implemented, since the evaluation of the labeling performance on a real integration scenario in the usual work of a labeling organization (Evaluation II) was considered as the evaluation scope of the 2nd version of AQUA prototype.

Moreover, since according to the technical implementation plan, the automated monitoring and alerting system was not implemented in the 1st version of AQUA, the corresponding evaluation task, that is, the evaluation of the accuracy of the automated monitoring and alerting system, has been also planned as a part of the evaluation of the 2nd version of AQUA prototype.

In order for the reader to be easier to follow the evaluation results presented in this document, this section presents an overview of the exact methodological steps used for the evaluation of the 1st version of AQUA.

2.1 Evaluation of the location and identification of unlabelled medical web sites

With respect to deliverable D15, no operational modification was introduced for the evaluation of the location and identification of unlabelled medical web sites.

Since this task applies only within the activities of AQUMED the corresponding evaluation was conducted only by AQUMED in English, as follows:

- Step 1: The evaluators were asked to perform two searches about (a) ischemic heart disease and (b) breast cancer using the crawler available in AQUA (the proposed keywords and directories are listed in D15).
- Step 2: The achieved results were checked in order to find out if they are relevant for the query. The criteria to identify a relevant web site are described in D15.
- Step 3: Precision was calculated using the following formula:
$$\text{Number of correct values} / \text{number of total values}$$

2.2 Evaluation of the accuracy of the automatic labeling

With respect to deliverable D15, for the evaluation of the accuracy of the automatic labeling the following operational modifications were introduced:

1. A list of 25 web sites for each language was used for the evaluation of the 1st version of AQUA. For the evaluation of the 2nd version of AQUA a different list consisting of 75 web sites will be used.
2. The accuracy of the automatic labelling in the 1st version of AQUA was measured for a given set of 5 criteria. The evaluation of the accuracy of the rest criteria will be the subject of the evaluation of the 2nd version of AQUA which will support the full set of labelling criteria.

For the evaluation of the accuracy of the automatic labeling, a number of web sites was labeled manually. The results achieved with these two methods were compared. The evaluation was performed in both English and Spanish as follows:

- Step 1: 25 web sites in each language were chosen according to the criteria described in D15
- Step 2: Each web site was manually reviewed in order to determine the values for the criteria under evaluation. The criteria used for the evaluation were: "1.2. Resource title", "3.3. Target / intended audience", "5.1. Virtual consultation available", "6.4. Advertisement present" and "7. Other seal" (for the criteria definitions see D4.2).
- Step 3: Review tasks were created following the instructions in the User Guide.
- Step 4: After finishing a review task AQUA sent out an e-mail to inform the evaluator that the results of the review task are available.
- Step 5: AQUA proposes some positive examples (links to pages) for the possible values of each criterion. The evaluator has to check the links until s/he finds a positive example. Then s/he decides which one(s) is/are the appropriate value(s). We used the following procedure for each criterion:
 - *Resource title*: beginning with the first link in the list the evaluator checks the proposed titles and the corresponding links until an adequate one is found
 - *Target / intended audience*: for each proposed value the evaluator checks the proposed links beginning with the first one until a positive example is found
 - *Virtual consultation available*: same procedure as above.

- *Advertisement present*: the evaluator checks the first link for “true”. Only if s/he does not find a positive example, s/he checks the proposed links for “false”.
 - *Other seal*: same procedure as Target / intended audience.
- Step 6: The manually created labels were updated with the new values found during the semi-automatic labeling. As a result, AQUA creates an additional version of the label.
 - Step 7: Both versions are compared using the automatic comparison function which is also available in AQUA. For the gold standard, the manually created label is used. If the values from the semi-automatic label and the manual label are concordant, it is considered as “correct”, if not, it is considered “wrong”
 - Step 8: We calculated for each criterion the precision using following formula:

Number of correct values / number of total values

2.3 Usability evaluation

With respect to deliverable D15, no operational modification was introduced for the usability evaluation. The evaluation was performed in both English and Spanish as follows:

- Step 1: Scenarios were presented to the 4 evaluators, 2 for Spanish and 2 for English interfaces. Each scenario represents a different task available in AQUA. The scenarios were described in the User Guide.
- Step 2: The evaluators were asked to perform the 4 scenarios only with support of the User Guide. To perform each task, a few web sites were given as examples. The evaluators were free to use them or to choose other ones for the experiment.
- Step 3: After that they filled a questionnaire and gave their feedback about the usability of the user interface for each task.

3. Evaluation Results

This section presents the detailed evaluation results, as well as, the feedback given by MedIEQ labelling organizations as an input to the continuous technical improvement action.

3.1 Evaluation of the location and identification of unlabelled medical web sites

Based on the evaluation methodology, for each search task (namely, for “ischemic hearth disease” and “breast cancer”) three different cases were tested (namely, (a) Using only keywords, (b) Using only web directories and (c) Using key words and web directories). The results of this evaluation are presented in Table 1. With respect to cases (b) and (c), due to the large number of results returned (shown in parentheses, in italics, in the corresponding tasks 2, 3, 5, 6), the labeling expert of AQUMED examined only a sub-set of them (the first 126 in task 2, the first 110 in task 3, the first 179 in task 5, the first 101 in task 6) to give an indication of the crawler performance.

Table 1: Evaluation Results of the AQUA Crawler

TASK	TOPIC	Which Keywords I have used?	Which Web dirs I have used?	Nr. of pos. web resources	Precision
1	Ischaemic/ischemic heart disease	myocardial infarction, coronary bypass OR revascularisation	-	32/40	0.8
2	Ischaemic/ischemic heart disease	-	<ul style="list-style-type: none"> • http://directory.google.com/Top/Health/Conditions_and_Diseases/Cardiovascular_Disorders/ • http://uk.dir.yahoo.com/Health/Diseases_and_Conditions/Angina_Pectoris/ 	33/126 (651)	0.26
3	Ischaemic/ischemic heart disease	myocardial infarction, coronary bypass OR revascularisation	<ul style="list-style-type: none"> • http://directory.google.com/Top/Health/Conditions_and_Diseases/Cardiovascular_Disorders/ • http://uk.dir.yahoo.com/Health/Diseases_and_Conditions/Angina_Pectoris/ 	40/110 (691)	0.36
4	Breast cancer	Breast cancer OR cancer of breast, Mastectomy AND breast cancer, Breast conserving surgery	-	90/108	0.83
5	Breast cancer	---	<ul style="list-style-type: none"> • http://directory.google.com/Top/Health/Conditions_and_Diseases/Cancer/Breast/ • http://uk.dir.yahoo.com/Health/Diseases_and_Conditions/Breast_Cancer/Mastectomy/ 	71/179 (351)	0.40
6	Breast cancer	Breast cancer OR cancer of breast, Mastectomy AND breast cancer Breast conserving surgery	<ul style="list-style-type: none"> • http://uk.dir.yahoo.com/Health/Diseases_and_conditions/Breast_Cancer • http://directory/google.com/Top/Health/Conditions_and_Diseases/Cancer/Breast 	64/101 (459)	0.63

It must be noted that when using directories, the user is able to define whether to constraint the number of the crawler results or not, by using the proper syntax; which is to put in front of a directory's url, when configuring the Crawler, either an [S] or an [L]. The first sets the Crawler to continue retrieving urls found deeper in the directory structure, while [L] sets the Crawler to retrieve urls only from one page. In the results presented above (tasks 2, 3, 5 and 6 in table 1), the [S] indicator was employed. In that way, the Crawler explored, in each task, all subdirectories of the initial directories, which increased the number of irrelevant urls.

3.2 Evaluation of the accuracy of the automatic labeling

3.2.1 Evaluation for English web sites

The list of the web sites used for this evaluation is presented in Appendix A. From this initial list consisting of 25 web sites, a web site was excluded from the evaluation, since it was not available at the time of the evaluation execution. The results of this evaluation are presented in Table 2.

Table 2: Evaluation Results of the Automatic Labeling (English Web Sites)

Criterion	Possible values	Nr. values	Nr. correct values	Precision
1.2. Resource title		24	13	0.54
3.3. Target / intended audience	Adult Children Professional	31	26	0.84
5.1. Virtual consultation available	True False	24	22	0.92
6.4. Advertisement present	True False	24	23	0.96
7. Other seal	N/A HON	24	24	1

The main remarks received by the evaluators during the execution of this evaluation, were the following:

- Regarding criterion “1.2 Resource title”, in 6 resources the system proposed only one value with only one relevant link. The proposed link corresponded always to the homepage. In that cases the evaluator accepted the unique proposed value and did not check any sub site.
- Regarding criterion “3.3 Target/intended audience” AQUA proposed correct examples for all the values, although the sub page that stated explicitly the target audience was usually not found.
- Regarding criterion “5.1 Virtual consultation” and “6.4 Advertisement present”, for some resources a number of positive suggestions were proposed by the system, that were not correct. In these cases the evaluator selected the negative value “False”.

3.2.2 Evaluation for Spanish web sites

The list of the web sites used for this evaluation is presented in Appendix B. 24 Spanish websites were selected. Four websites, from the initial list, didn't work and they were replaced with other ones (see Appendix 2). The results of this evaluation are presented in Table 3.

Table 3: Evaluation Results of the Automatic Labeling (Spanish Web Sites)

Criterion	Possible values	Nr. values	Nr. correct values	Precision
1.2. Resource title		24	20	0.84
3.3. Target / intended audience	Adult Children Professional	24	12	0.50
5.1. Virtual consultation available	Pos Neg	24	16	0.67
6.4. Advertisement present	Pos Neg	24	14	0.58
7. Other seal	N/A HON	24	24	1

The main remarks received by the evaluators during the execution of this evaluation, were the following:

- Regarding criterion “3.3 Target/intended audience” AQUA proposed correct examples for all the values, although the sub page that stated explicitly the target audience was usually not found.
- Regarding criterion “5.1 Virtual consultation” and “6.4 Advertisement present”, for some resources a number of positive suggestions were proposed by the system, that were not correct. In these cases the evaluator selected the negative value “False”.

3.3 Usability evaluation

For the execution of this evaluation two groups of evaluators were used. The first group consisted of evaluators (Silke Astroth and Elena Mohl) with no familiarization with AQUA. The second group consisted of evaluators (Dagmar Villarroel, Miguel Angel Mayer and Angela Leis) with experience in the use of AQUA.

The results of this evaluation phase were in practice the feedback given by MedIEQ labelling organizations as an input to the continuous technical improvement action. As a consequence, in this section we present the response received from all the evaluators as well as the progress of the relevant technical actions. The results of this evaluation are presented in Table 4.

Table 4: AQUA Usability Evaluation Results

Evaluator Comment	Priority	Applied in AQUA v1.1	Applied in AQUA v2.0	Implementation Status
Scenario 1: Manual generation of content labels				
The option to remove more than one web resource at a time would be useful	low		✓	pending
The menu item "Add a web resource" should be more visible	high		✓	pending
It would be useful to choose, how the web resources are listed, e.g., by clicking on "Title" they would be listed alphabetically, by clicking on "creation date" from the newest to the oldest, etc.	high	✓		implemented
If we have many resources, it would be useful to show them in more than one page, e.g. only 20 resources at each page.	high		✓	pending
The space for the list of "My web resources" is narrow. A possible solution would be to put the menu on the left site at the top as "folders"	high		✓	pending
The values of target audience criterion could be slightly modify as follow: "professionals", "patient/adult" and "patient/child"	high	✓		implemented
In "My Labeled Web resources" when the name of the website is longer, the name is cut.	low	✓		implemented
For all criteria, as well as, for the Host restrictions and explanation should be added	high	✓		implemented
In the manual generation of the label the user interface has errors when viewed by Internet Explorer	high	✓		implemented
Scenario 2: Computer-assisted labelling or asking AQUA to review unlabelled resources				
The number of proposed values is too high to review. We suggest developing a filter so as to present only the best examples per value (e.g. max. 5 links per proposed value).	high	✓		implemented
The information, when a task was created is missing (maybe as another column)	low		✓	pending
The members of a same organization cannot see the Review Tasks of the other members. It is desired that all the reviewers have access to all created Review Tasks.	high		✓	pending
An indication (user notification) that only the selected resources will be spidered should be helpful	low		✓	pending
The option to add comments to the labels would be useful	high		✓	pending

Evaluator Comment	Priority	Applied in AQUA v1.1	Applied in AQUA v2.0	Implementation Status
Scenario 2: Computer-assisted labelling or asking AQUA to review unlabelled resources (cont.)				
When comparing labels, it would be useful to see on the top when they have been created	low		✓	pending
Scenario 4: Search for new, unlabeled web resources				
The information, when a task was created is missing (maybe as another column)	low		✓	pending
The menu item "go to search properties" should be more visible	high		✓	pending
General Comments				
The response time of the system is slow	high	✓		implemented
When a criterion represents a date (e.g. last update) a date picker component should be available	high	✓		implemented
When the system needs time for processing, a please wait message should be presented to the user.	high	✓		implemented

It must be noted that for evaluation scenario 3 (Editing a label created with assistance of AQUA) no comments were received from the evaluators.

4. Evaluation results of AQUA components

This section presents the evaluation results at the level of AQUA's individual components. The evaluation results are grouped by AQUA's main toolkits, which correspond to the different technical Work Packages.

4.1 Web Content Collection (WCC) Toolkit

4.1.1 Evaluating the Focused Crawler

The Crawler's role in AQUA is to assist the labelling expert in the identification of new unlabelled web resources having health-related content. Crawler's overall performance in 6 search tasks is analysed in section 3 of this document. Here, we summarize our evaluation results on Crawler's enhancing mechanism: Crawler's content classification component (results presented also in D7.1).

We trained two content classifiers, one for content in English and another for content in Greek. For this, we used two corpora, an English one, consisting of 1976 pages (944 pos & 1032 neg samples) and a Greek one, consisting of 1707 pages (869 pos & 838 neg), all manually annotated.

Three different classifiers provided by the Weka¹ classification platform have been tested. These are SMO, Naïve Bayes and Flexible Bayes (Naïve Bayes with kernel estimation) and during the tokenization process all the HTML tags of the HTML documents were removed. All 1-grams, 2-grams and 3-grams were produced and the best of them according to information gain were selected² (more details on the evaluation methodology can be found in D7.1).

Table 5: Classification performance results for content classification in English.

	1-grams						1/2/3-grams					
	Tags removed			Tags not-removed			Tags removed			Tags not-removed		
	Prec.	Rec.	F-m.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.
NB	0.75	0.63	0.68	0.78	0.47	0.59	0.89	0.28	0.42	0.84	0.35	0.50
FB	0.73	0.55	0.62	0.77	0.42	0.54	0.79	0.38	0.52	0.78	0.33	0.48
SMO	0.75	0.61	0.67	0.75	0.49	0.59	0.78	0.55	0.64	0.74	0.46	0.57

Table 6: Classification performance results for crawling in Greek.

	1-grams						1/2/3-grams					
	Tags removed			Tags not-removed			Tags removed			Tags not-removed		
	Prec.	Rec.	F-m.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.
NB	0.65	0.94	0.77	0.66	0.93	0.77	0.64	0.94	0.76	0.62	0.95	0.75
FB	0.64	0.94	0.76	0.63	0.96	0.77	0.64	0.95	0.77	0.60	0.98	0.74
SMO	0.70	0.88	0.78	0.78	0.87	0.81	0.69	0.91	0.79	0.72	0.91	0.80

In the above results (tables 5 and 6) we notice the low performance of the content classifiers in both languages. This is probably justified by the fact that is difficult, even for humans, in various cases to assess whether a website has health-related content or not. This is supported by the fact that in English, where we had a "tighter" policy in

¹ <http://www.cs.waikato.ac.nz/ml/weka/>

² The final list of the selected n-grams may contain 1-grams, 2-grams and 3-grams.

manual annotation, considering what is health-related and what is not, the performance of the classifier is even lower.

The distinction between health and non-health web sites seems difficult when using only the content of the home page. Possible future approach is to use also the content of a few more pages deeper in the website to improve classification performance.

4.1.2 Evaluating Spider's performance

As the Spider is launched directly from AQUA, it is possible that several users run several sessions of the Spider at the same time (each session containing about 4-8 parallel spidering threads). For best performance, it was important to improve, wrt Spider's 1st version, Spider's resource management and effectiveness in downloading. In cases of very large web sites (i.e. when having archives of articles they may have more than 2000 unique web pages), spidering should remain efficient in terms of both hardware resources management and downloading speed. For evaluation purposes, we set up a limit of 100 unique web pages per website.

Comparison of results from both versions is presented in table 7. Note that the 1st version of spider had to stop in both attempts due to insufficient memory in specified time, whereas the final version looks to be stable on "infinite" number of web sites and it stops after retrieving all their contents.

Table 7: AQUA Spider Evaluation Results

Version	Web sites	Web pages	Time	Resource limit
1st	150	10000	1,5 hours	500 MB RAM for JVM
1st	180	15000	2,5 hours	800 MB RAM for JVM
Final	1600 (2,6 % reach limit)	45000	8 hours	800 MB RAM for JVM

It should be noted that only in 2,6% of the 1600 web sites spidered by the final version, the limit of 100 unique pages was reached. The memory requirements mentioned above may seem to be very high. However, we shouldn't forget that content classification is performed while spidering, which necessitates to constantly hold in memory the Content Classification Modules (CCMs). This takes almost half of Spider's memory usage.

Similarly, aiming to improve Spider's effectiveness, link-scoring has been examined in the 1st version of the spidering tool. The idea was to classify links before visiting the corresponding pages (similar to classifying web page content), in order to decide whether it's worth or not to finally visit those pages. Due to the fact that the evaluation results were not promising, the overall idea of link-scoring was abandoned.

4.1.3 Evaluating the Trained Module Generation tool & Content Classification Component (TMG/CCC)

TMG/CCC is a component for the generation of CCMs (content classification modules). It takes as input collections of content (web pages in ascii format) and can use different ML techniques to produce classification models.

The final version of TMG/CCC uses only Machine Learning (ML) techniques. The idea of combining ML with heuristics has been abandoned, as evaluation showed that what was gained in performance was insignificant. At the same time, tool's configuration for heuristics was generating disproportional overhead (manual

selection of relevant keywords was necessary in every different classification task and for every language).

Evaluation of the CCMs has been performed in both languages supported so far: English and Spanish (while evaluation only in English had been performed for the 1st version of TMG/CCC). The Support Vector Machine (SMO) classifier algorithm (implemented by the Weka package) has been adopted as it showed the best performance in average in both languages. Evaluation results in English were analytically presented in D7.1. In table 9, we present the final results in both languages (for details in the evaluation methodology, see D7.1 & D7.2).

Table 9: Evaluation results for English & Spanish (note that in Spanish, the “Children” target is not considered since there are not enough health pages for training).

Category	English			Spanish		
	Precision	Recall	Fm	Precision	Recall	Fm
CI	0,84	0,96	0,9	0,8	0,65	0,72
AD	0,87	0,8	0,83	0,77	0,72	0,75
VC	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	0,8	0,78	0,79	-	-	-
Professional	0,77	0,81	0,79	0,62	0,63	0,62

Several learning schemes, decision trees, naive Bayes and support vector machine were tested. Since the documents to be classified are relatively heterogeneous, a great amount of features (words) appear. Consequently, the SMO learner based on support vector machine generated the best results. The feature extraction process was complemented with a feature reduction based on information gain (1000 features).

As we expected, the most difficult accreditation criterion for classification purposes is the target audience, being the also the most subjective categorization criterion. The results over Spanish corpora are much lower. It can be due to several variables: the corpus annotation, the characteristics of Spanish web pages, the stemming, stop word removing, etc. The children target audience for Spanish pages is avoided because there not were enough samples for training.

4.2 Information Extraction Toolkit (IET)

In this section we present extraction results for contact information in the English language and in two evaluation modes: loose and strict (for details see section 3.3 of D9.1). The data set consists of 109 HTML documents which were all manually classified as contact pages. The data set contains 146 HTML files as some documents include frames or iframes. In total, the documents contain 6930 annotated named entities of the following types:

- title (medical title, e.g. MD),
- name (person name, e.g. J. Smith, John Smith, Smith),
- street (street name incl. number, e.g. 22 Oak Lane),
- city (city name, e.g. New York City, NYC),
- region (any geographical unit other than city or country, e.g. counties or boroughs),

- zip (zip code, e.g. 162 00, 94105-2099),
- country (e.g. Denmark),
- phone (phone numbers including extensions, e.g. +420 463 928 281),
- email (e.g. smith@uep.cz, smith at uep dot cz),
- organization (full organization name, e.g. Medical Center at the York Hospital),
- department (name of a unit inside organization, e.g. Department of Preventive Medicine)

Table 10: Initial evaluation results for contact information extraction in English.

	Counts		strict mode			loose mode		
	gold	auto	prec	recall	F	prec	recall	F
Title	743	859	0.71	0.82	0.76	0.78	0.86	0.82
Name	2197	1711	0.66	0.51	0.58	0.74	0.56	0.64
Street	120	100	0.62	0.52	0.56	0.85	0.67	0.75
City	346	534	0.47	0.73	0.57	0.48	0.76	0.59
Region	213	36	0.83	0.14	0.24	0.83	0.14	0.24
Zip	117	154	0.59	0.78	0.67	0.67	0.85	0.75
Country	478	726	0.58	0.89	0.70	0.59	0.89	0.71
Phone	899	785	0.97	0.84	0.90	0.99	0.87	0.93
Email	517	512	1.00	0.99	1.00	1.00	0.99	1.00
Organization	1115	717	0.57	0.37	0.44	0.81	0.51	0.63
Department	185	112	0.51	0.31	0.38	0.85	0.45	0.59
Overall	6930	6085	0.70	0.62	0.66	0.78	0.68	0.72

From the above results (table 10) we can conclude to the following points:

- Co-reference resolution (to be implemented in the final version of IET) is expected to improve results for name and organization. Currently, it is often the case that e.g. "American Association of Pediatrics" is recognized correctly but "AAP" is missed. Typically, abbreviations like "AAP" would be mentioned more times in the document, leading to very low precision on that document. E.g. 9 occurrences of "AAP" and a single "American Association of Pediatrics" lead to 10% precision. For person names, a similar example could be "Dr. Mark Bacon" vs. "Mark" or "Bacon".
- Looking at the data, automatic induction of formatting wrappers should significantly improve results esp. recall for named entities like name. First version of this is implemented but needs tuning.

4.3 Multilingual Resources Management (MRM) Toolkit

From the MRM toolkit we evaluate one component, the MRM Browser. That's because all the other components of the MRM toolkit are actually auxiliary to the Browser. Given that MRM is not included in the evaluation of the 1st AQUA prototype, a usability evaluation of the MRM interfaces is planned to be included in the usability evaluation of the final release of AQUA. Therefore, we here summarize our remarks after comparing the MRM Browser to two other existing browsing approaches:

- the NLM (the US National Library of Medicine) MeSH Browser ³ and
- the SNOMED Browser ⁴ by BT.

Two experiments have been implemented for this: (a) Term retrieval and (b) Browsing hierarchies.

³ <http://www.nlm.nih.gov/mesh/MBrowser.html>

⁴ <http://www.jdet.com/>

4.3.1 Term retrieval

The first tested functionality is the term retrieval process. Given a query containing a set of words, the MeSH Browser retrieves all concepts containing at least one of the query words, and ranks the retrieved terms alphabetically. On the other hand, in the most flexible search strategy of the SNOMED Browser, terms that contain words that sound like the words entered are returned. The MRM Browser instead returns all terms containing at least one query word ranking the retrieved results according to (a) the number of covered terms and (b) the term length.

We have implemented two experiments. In the first one we look for concept identifiers (CUI) in SNOMED Browser using as query the corresponding concept strings in MeSH. We have seen that most of strings associated to the same CUI are different in both resources. The retrieved results when using our Browser (over SNOMED resource) are compared against those when using the SNOMED Browser. The second experiment is analogous. We look for concepts in MeSH Browser using the corresponding concept strings in SNOMED, comparing the retrieved results with those when using the MRM Browser (over MeSH resource). This is a sensible way to emulate user queries that do not match exactly with the strings in the browsed resource, considering the language variability.

Figures 2 and 3 present the average rank position of retrieved concepts. The results suggest that the MRM term retrieval process allows finding concepts with less effort. For instance, in many cases, when using the MeSH Browser, a required concept is found even after the first 100 returned. However, in most cases, when using the MRM Browser, the concept is within the 15 ranked first (see Figure 3).

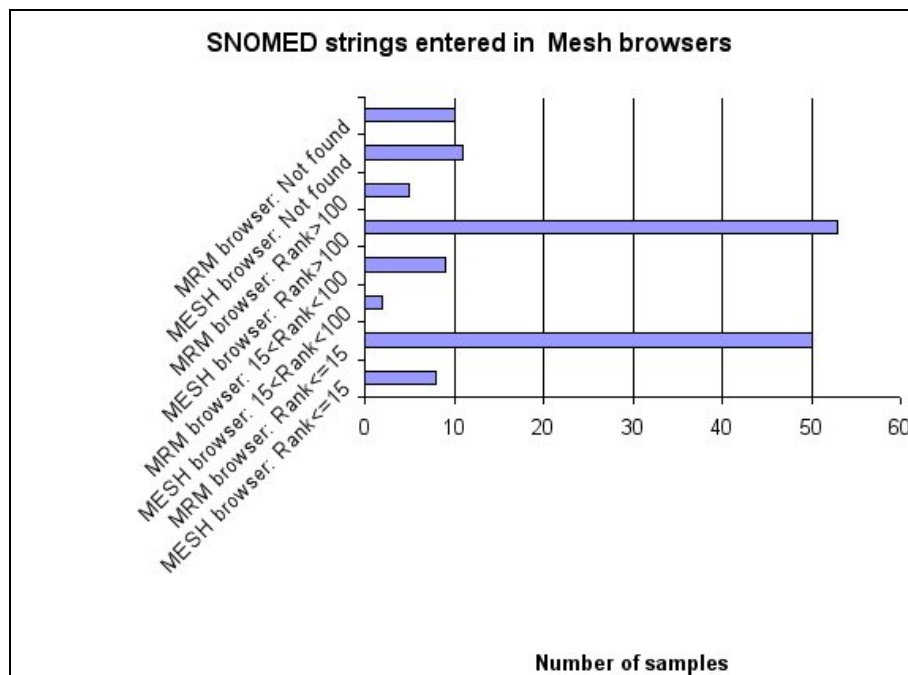


Figure 2. Evaluation results for browsing MeSH resource

Figure 3 presents the results of browsing SNOMED using concepts from MeSH. The SNOMED Browser returns “not found” if some query term cannot be aligned. In order to consider these cases, we remove systematically the last word of the query until some concepts are returned. Even so, “not found” concepts are more frequent when using the SNOMED Browser than using the MRM browser over the SNOMED resource. In addition, finding the term before the 50th concept is more frequent when using the MRM Browser.

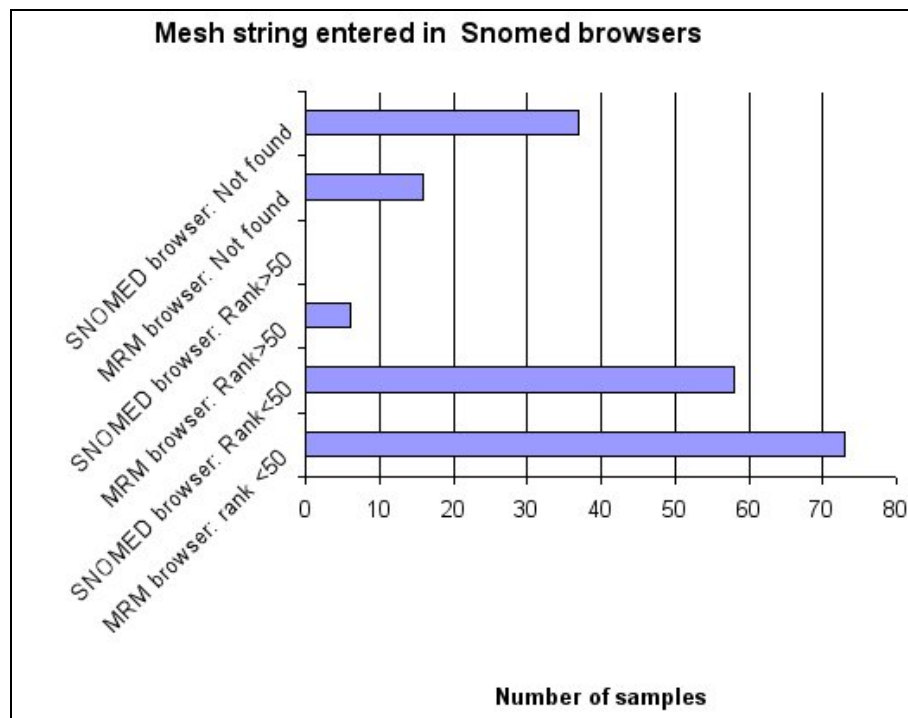


Figure 3. Evaluation results for browsing SNOMED resource

4.3.2 Browsing hierarchies

The second functionality analysed is the appropriateness of the browsing process across concept relations. The MRM toolkit offers the following advantages over other existing browsers:

- A unified interface for multiple medical resources in several languages.
- Tackles different hierarchical contexts of concepts separately, while other browsers as the MeSH Browser, show all contexts simultaneously, entering redundant concepts in the user view (see Figure 4). In addition, the MRM Browser allows jumping from one concept to others without changing the hierarchical concept.

<p>Respiratory Tract Diseases [C08]</p> <p>Lung Diseases [C08.381]</p> <p>Lung Diseases, Obstructive [C08.381.495]</p> <ul style="list-style-type: none">▶ Asthma [C08.381.495.108]Bronchitis [C08.381.495.146] +Pulmonary Disease, Chronic Obstructive [C08.381.495.389] +
<p>Respiratory Tract Diseases [C08]</p> <p>Respiratory Hypersensitivity [C08.674]</p> <ul style="list-style-type: none">Alveolitis, Extrinsic Allergic [C08.674.055] +Aspergillosis, Allergic Bronchopulmonary [C08.674.060]▶ Asthma [C08.674.095]Asthma, Exercise-Induced [C08.674.095.110]Status Asthmaticus [C08.674.095.880]Rhinitis, Allergic, Perennial [C08.674.810]Rhinitis, Allergic, Seasonal [C08.674.815]
<p>Immune System Diseases [C20]</p> <p>Hypersensitivity [C20.543]</p> <p>Hypersensitivity, Immediate [C20.543.480]</p> <p>Respiratory Hypersensitivity [C20.543.480.680]</p> <ul style="list-style-type: none">Alveolitis, Extrinsic Allergic [C20.543.480.680.075] +Aspergillosis, Allergic Bronchopulmonary [C20.543.480.680.085]▶ Asthma [C20.543.480.680.095]Asthma, Exercise-Induced [C20.543.480.680.095.110]Status Asthmaticus [C20.543.480.680.095.880]Rhinitis, Allergic, Perennial [C20.543.480.680.791]Rhinitis, Allergic, Seasonal [C20.543.480.680.795]

Figure 4. Redundant information in MeSH Browser.

5. Conclusions

The results of the evaluation of both the location and identification of unlabelled medical web sites, as well as, the accuracy of the automatic labelling were satisfactory. By using only the proposed by AQUA links, it was possible to identify the right value in more than 80% of the different labelling cases.

In general the evaluation results were promising for the further development of AQUA as an assisting system for the labelling experts. However, for the evaluation of the final version of AQUA towards its integration within the day-to-day activities of a labelling organization, the evaluation of the full set of labelling criteria, as well as, the application of AQUA in a real day-to-day practice scenario is required. This will be the main objective of the evaluation of the final version of AQUA. To this end, and based on the experience gained from the evaluation of the 1st AQUA version, an update of the evaluation plan will be needed, so as to focus on the practical aspects of the day-to-day activities of a labelling organization rather than the functional specifications of AQUA.

Appendix A: List of URLs used for the evaluation of the accuracy of the automatic labeling (English Web Sites)

1. <http://www.patient.co.uk/>
2. <http://www.vasculitisfoundation.org/>
3. <http://www.clevelandclinic.org>
4. <http://www.uhrad.com/>
5. <http://www.britishlivertrust.org.uk/>
6. <http://www.allayurveda.com/>
7. <http://www.curezone.com/>
8. <http://healthlink.mcw.edu/>
9. <http://www.wrongdiagnosis.com/>
10. <http://www.drgreene.org/>
11. <http://pathweb.uchc.edu>
12. <http://www.aacap.org/>
13. <http://www.aafp.org/>
14. <http://www.helpguide.org/>
15. <http://www.rcpsych.ac.uk/>
16. <http://www.xps.org/>
17. <http://www.cdc.gov/>
18. <http://www.eatright.org/>
19. <http://gamma.wustl.edu/>
20. <http://www.niams.nih.gov/>
21. <http://www.guideline.gov/>
22. <http://www.who.int/>
23. <http://www.urologychannel.com/>
24. <http://www.ncemi.org/>
25. <http://www.nlm.nih.gov/>

Appendix B: List of URLs used for the evaluation of the accuracy of the automatic labeling (Spanish Web Sites)

1. <http://www.svneumo.org/>
2. <http://www.buromedicos.com/>
3. <http://www.geodental.net/>
4. <http://www.webpediatria.com/>
5. <http://www.termasworld.com/index.php?lang=es>
6. <http://www.galenicom.com/>
7. <http://www.pardal.net/>
8. <http://centromedicolapaloma.com/>
9. <http://www.mgz.cl/>
10. <http://www.solohijos.com/>
11. <http://www.sap.org.ar/>
12. <http://www.estheticmon.com/>
13. <http://www.cicatrizando.es>
14. <http://www.urovirtual.net>
15. <http://www.saenzdecabazon.com/esp/fresp.html>
16. <http://www.diabetesymas.com/>
17. <http://www.paidopsiquiatria.com/>
18. <http://www.grupoaulamedica.com/web/index2.cfm>
19. <http://www.directoriomedico.com.ve/>
20. <http://www.inicia.es/de/MedicoRural/>
21. <http://www.vhebron.es/vhesp.htm>
22. <http://www.clinicaclaros.com/>
23. <http://www.geocities.com/bebesano/index.html>
24. <http://www.odontodos.net/>