Ontology design for web sites recommendation in the health area

Regina Motz[1], Edelweis Rohrer[1]

Abstract. The query of web contents about the health area is an increasing common practice among users with different features. Some of them realize queries with a concrete aim, for example patients that suffer a disease or relatives of the patient, doctors, all of them demanding specific information about a disease. Others do it just by curiosity, for example to inform themselves about prevention of diseases. These users have different levels of academic training. Taken this fact into account, this paper presents the design of an ontology that aims to ease the task of determining the degree of adequacy that a given heath web site has for a given user, obtaining a recommendation of reading of the content for this user.

1 Introduction

This paper presents the design of an ontology that has as main aim to give a recommendation of health web sites for different user profiles. To achieve the aim posed, it is necessary to have a semantic structure whose main core associates quality levels to web contents, according to criteria established, and allows to issue a recommendation of reading to different user profiles. This leads to the definition of an ontology to abstract the most relevant concepts of the reality according to the four perspectives that web contents can be seen: healh domain, web sites, quality factors and user profiles. This ontology has a structure composed of a set of interrelated sub-ontologies that arises as a result of a process of conceptualization that follows guidelines presented in [1]. The start point of this process was the knowledge acquired from the site [WR1] and interviews with experts.Figure 1 shows the network of sub-ontologies that make up the ontology model. In this schema, sub-ontologies located at left have a greater degree of independence. Then, they are integrated so that each one reuses sub-ontologies located to its left.

[1] Instituto de Computación, Facultad de Ingeniería, Universidad de la República
[rmotz,erhrer]@fing.edu.uy

Health Domain and User Profile ontologies are totally independent. Health Domain ontology represents the application domain, it is conceptualized as an ontology of general character, so that it must be specialized for specific diseases, with the knowledge provided by experts. User Profile ontology represents the subject of the recommendation, i.e. users. It models different features of users that influence in the recommendation of web contents. Web Site ontology is created to model the object of the recommendation, i.e. documents queried by users. It reuses Health Domain ontology for a Specific Disease because that ontology represents the topic of documents published in web sites. The model of Quality ontology is defined to represent evaluation criteria according to different quality factors. It reuses Web Site ontology (and therefore Health Domain ontology) because the aim of Quality ontology is to classify documents of web sites according to established criteria. Quality ontology also must be extended for different diseases and quality factors, so that experts from medical and quality areas can define metrics of evaluation of contents for each quality factor.

Finally, Recommendation ontology is the structure which exploits ontologies defined according to the different perspectives, to represent the recommendation of a content for a given user. Those ontologies provide parameters that influence in the process of obtaining a recommendation level. The recommendation ontology also must be specialized depending on the disease, allowing experts of each one define metrics to be applied. It reuses User Profile ontology and Quality ontology for Specific Disease and Quality Factor (and therefore Site and Heath Domain ontologies).

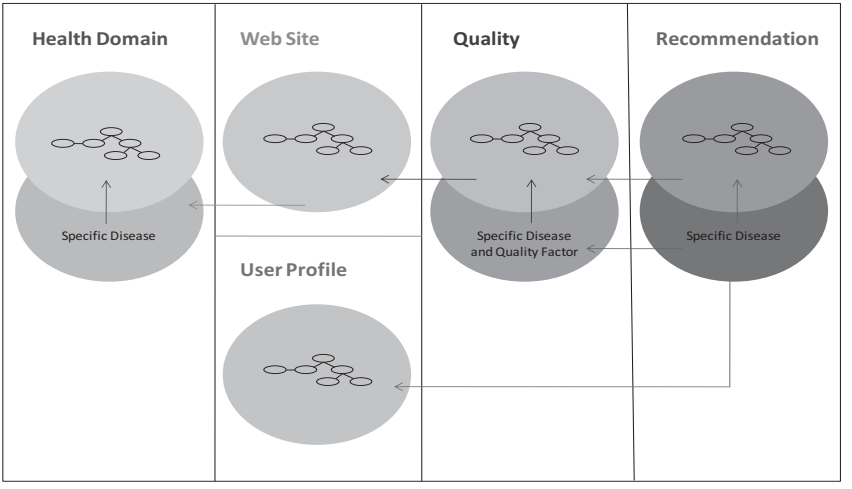

Figure 1. General schema of ontologies

The present design is being tested with a case of study of the health domain, as it is part of the project “Red Temática SALUS”. However, the design is suitable to be applied to different domains, such as e-learning or laws, so this is a more general proposal.

The rest of the article explains in detail each sub-ontology showed in Figure 1, in Sections 2, 3, 4, 5 and 6. Section 7 presents some conclusions and future works.

2 Health Domain Ontology

The structure of this sub-ontology can be seen in Figure 2. Concepts identified in this model arise from the analysis of the site [WR1]. Despite the ontology is the representation of information about a specific disease (Alzheimer), querying sites about other diseases and visualizing the concepts identified, leads to infer that the resulting model can be considered as a general ontology for any disease. If the domain is analyzed more deeply, new concepts will arise from the study of the terminology used in other diseases, which can be introduced.

As showed in Figure 2, for each connection which links an instance of the Disease concept to instances of other concepts, an inverse relationship is defined. That decision is taken because for each instance of any concept related to the Disease concept, such as Diagnostic or Treatement, it is necesary to obtain the corresponding instance of the Disease concept, e.g. it is necessary to infer the disease that corresponds to the treatement. In this way the relationship "controls" is added, with domain Treatment and range Disease. This allows to deduce, for any instance of any concept, the disease that corresponds to it.

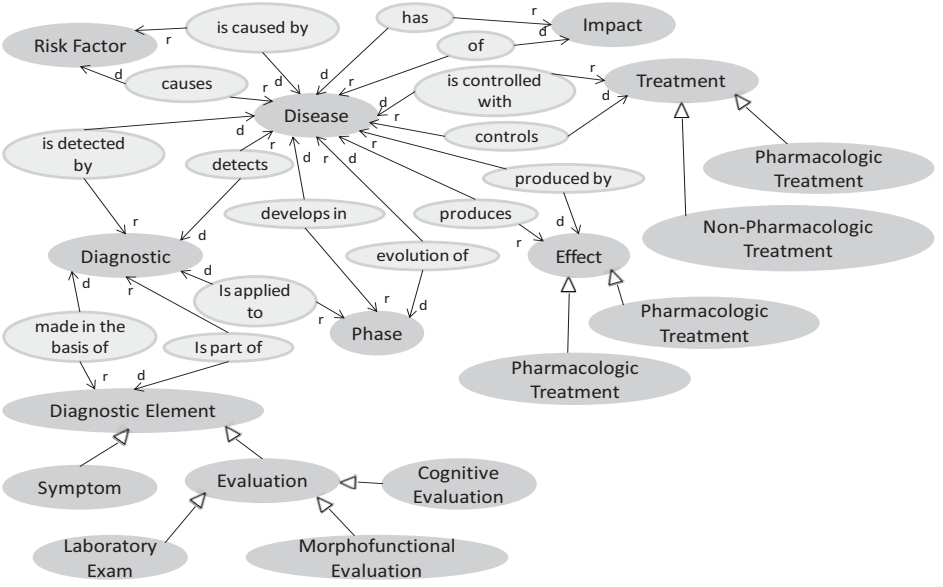

Figure 2. Health Domain Ontology.

As mencioned, Health Domain ontology for a Specific Disease is a specialization of Health Domain ontology that gathers particular information about a disease, provided by the

expert. In this extension subclasses are defined, whose instances correspond to this disease. For example, the Alzheimer Diagnostic subclass is defined as the set of all individuals of the Diagnostic class that have a relationship "detects" with the "Alzheimer" instance of the Disease class. The definition of the Alzheimer Diagnostic subclass in OWL language is showed below:

```
<owl:Class rdf:ID="AlzheimerDiagnostic">
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:parseType="Collection">
                <owl:Restriction>
                    <owl:onProperty rdf:resource="&p1;detects"/>
                    <owl:hasValue rdf:resource="#Alzheimer"/>
                </owl:Restriction>
                <owl:Class rdf:about="&p1;Diagnostic"/>
            </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
</owl:Class>
```

This means that the axioms establishing the necessary and sufficient conditions to be satisfied by instances of subclasses for a specific disease, use inverse relationships introduced in the Health Domain ontology.

For each disease it is possible to specify a different specialization. Each specialization can extend Health Domain ontology adding more complex structures, with new specific concepts and relationships, according to the required expressivity.

To ensure that Health Domain ontology include all the medical terminology and no concept is omitted, it would be convenient to study the possibility of integrating these first concepts identified with some medical taxonomic structure or general ontology. Such structure must provide all the medical terminology and if it is available, the semantic structure that organizes the terminology. GALEN, UMLS and ON9 are some of biomedical information databases that are available. Of these, UMLS has been analyzed more deeply, because it integrates different terminological databases in a common semantic structure named Semantic Network. Therefore, it is the candidate ontology to be reused. Moreover, the integration mechanism to be used must be defined. It could be *merging*, *mapping*, or some intermediate solution, which can be consulted in [2] and [3]. If merging is applied, a possible solution could be to extract the subset of the Semantic Network of UMLS which has to do with the Health Domain and combining it with the structure illustrated in Figure 2, to obtain a new ontology. Mapping, on the contrary, is an integration process that should establish correspondences between the structure of Figure 2 and suitable concepts of the Semantic network of UMLS, to obtain the required information of it (concepts and terminology), preserving original ontologies.

3 Web Site Ontology

Web Site ontology represents the content of web sites, i.e. the set of documents published, which must be classified according to quality criteria, associating quality levels that allows them to be recommended to different users. Figure 3 shows the design of this sub-ontology. It has a central concept Document with different attributes and relationships that link it with other concepts such as Author, Source and the union of concepts of Health Domain ontology representing the topic of the document. With regard to the source of documents, if it is a site, the model expresses if it has quality certification or not.

Two subclases of the Document concept are defined: Doc. Page and Doc. Type Not Page. It is necessary to isolate documents which are web pages, because they have certain features that distinguish them from the rest, like the fact that they are part of the Page concept. Then, the Doc. Page concept must be a subclass of the concept Page, which includes all instances that represent web pages (they can be documents or not). But the implementation with Protégé 3.4, does not allow the representation of the multiple inheritance relationship from the Doc. Page concept to Document and Page concepts. Hence, a rule must be written to implement the relationship Doc. Page "is a" Page.

```
DocPage(?x)  → Page(?x)
```

Moreover, a rule must be written to control that the instance of the Doc. Page Concept that has as source a web site (an instance of the Site concept) is an instance of the Page concept and it is linked to the Site Concept through the relationship "is part of".

```
DocPage(?x)∧Site(?y)∧hasSource(?x, ?y) → Page(?x)∧isPartOf(?x, ?y)
```

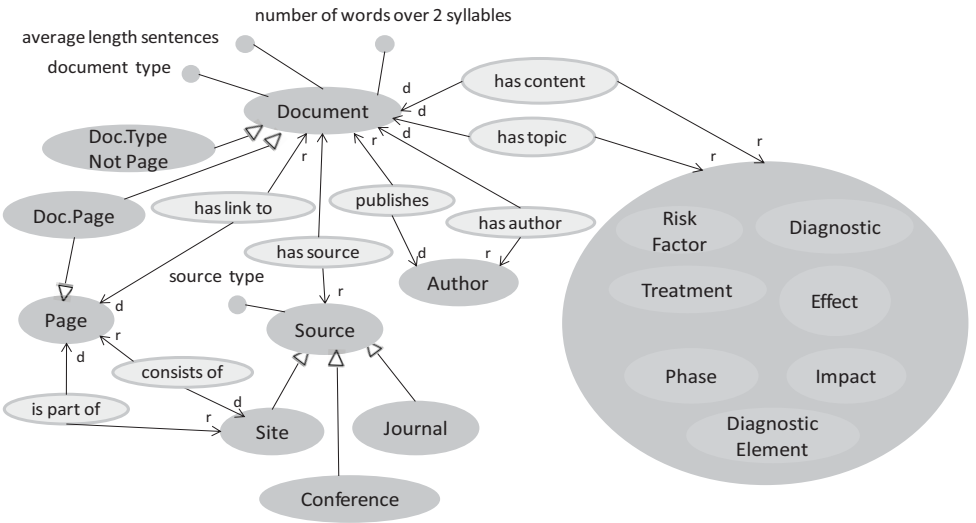

Figure 3. Web Site Ontology

4 User Profile Ontology

In the model of this sub-ontology, which can be seen in Figure 4, three different aspects of users are expressed: academic level, the role played (doctor, researcher, patient, carer, relative, etc.) and the age range. These user features are going to influence the comprehension and interests of them with regard to medical documents.

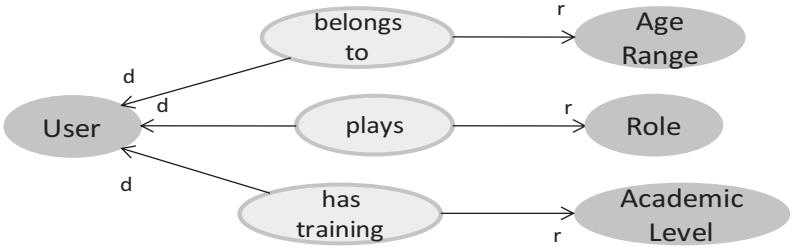

Figure 4. User Profile Ontology

5 Quality Ontology

The quality of a web site has been measured through different factors. Some of the traditional ones are: navigation, user interface aspects, legibility (size of letter, colors, images), performance aspects (time it takes to access to the site content), etc. The approach of this work is analyzing the quality that arises of the information value that the site provides, its adecuacy for the use the reader wishes to give to it and in which degree it satisfies his expectations. From this point of view, the quality of the site depends on its context of use and its final consumer.

Following this approach, the knowledge acquired with experts allowed to choose some quality factors as most likely to be measured. These factors are described briefly below.

Believability

From [4] two definitions are extracted.

“Believability: the extent to which data is regarded as true and credible”.

“Reputation: the extent to which data is highly regarded in terms of its source or content”.

The former is a general definition that expresses the meaning of data ‘s believability, while the latter talks about data properties (source, content) to be considered to evaluate if a document is believable or not.

About this factor, it is important to take into account the existence of sites with certified quality labels, such as HON [WR2], WIS [WR3] and WMA [WR4], which means

that documents linked by these sites will be evaluated with a higher level of quality than the contents of sites that have no certification.

Timeliness

In [4] the following definition can be found:

"Timeliness: the extent to which data is sufficiently up-to-date for the task at hand"

Regarding this factor, what really matters is measuring the freshness of data published, rather than de publication date.

Readability

[5] is a research of different readability metrics that have been created for different domains and user profiles. It sets the following definition:

"Readability is what makes some texts easier to read than others".

The same work mentions the definition of G. Harry McLaughlin (1969), creator of the SMOG readability formula:

"The degree to which a given class of people find certain reading matter compelling and comprehensible."

There are a lot of readability formulas created for different authors, like FOG and SMOG grade levels, that reached good results when they were tested [5].

$$\text{FOG grade level} = 0.4\ (\text{average sentence length} + \text{hard words}) \quad \textbf{(1)}$$

$$\text{SMOG grade level} = 3 + \sqrt{\text{polysyllable count}} \quad \textbf{(2)}$$

Completeness

From [4] the following definition is taken:

"Completeness: the extent to which data is not missing and is of sufficient breadth and depth for the task at hand".

At first, this factor was conceptualized with a high level of detail. The experience of the experts leads to redefine this concept, discarding the idea of establishing metrics with complex conditions, because they would evaluate the correctness rather than the completeness of the contents.

Once defined the quality factors, it is handled the possibility of establishing an order in evaluating them, so that if the result of measuring a given factor produces a quality level lower than a predetermined threshold, in some way the document is discarded and the evaluation of other factors does not continue. For example, if a document fails to reach a certain level of believability or confidence in the source, the evaluation process must not measure the quality factor completeness.

From the present analysis of quality factors, the main conclusion that arises is that the real possibility of measuring factors of quality and the depth or maximum level of detail that can be reached must be defined by the expert.

The modeling of all aspects related to quality assessment, results in a general ontology as a core model, the Quality ontology. This ontology aims to provide a basis for adaptive systems, whatever quality factor and domain the application treats with. This ontology does not define specific metrics to be applied, as this task must be the responsibility of the expert in the domain and quality factors. This leads to the specialization of the initial ontology, according to the disease and the quality factor to be assessed. Each extension should be carried out by the designer of ontologies and the expert, working together. Then, one or more ontologies extending the core are obtained, in which process the expert provides metrics to be applied, generally expressed through rules.

Figure 5 shows the Quality ontology. This model contains a relevant concept, Quality Factor, around which two main aspects are represented: the dependency among quality factors and metrics to measure them.

The dependency among quality factors, as mentioned, defines an order in the evaluation of the factors. This is expressed through the relationship “depends on” with the Quality Factor concept as domain and range. To infer when a quality factor must be evaluated, according to results of evaluations of previous factors, it is added the “acceptable” attribute to the Quality Factor Level concept. This attribute indicate if a quality level of a factor enables the evaluation of quality factors which depend on it (acceptable) or does not allow to continue with such evaluation (not acceptable).

With regard to metrics, subclasses are defined to classify metrics that measure the same quality factor and it is also expressed that metrics can depend on the topic of the document. The Quality Factor level is also specialized for each factor and it is associated to the Document concept.

During the conceptualization process, the idea of obtaining a general quality level for web contents is analyzed, by combining quality levels of all factors which have been evaluated. This idea is finally discarded, because the possible instances of the general quality level concept are multiple combinations of the instances of the Quality Factor Level concept, for each quality factor considered. For example, it is difficult to determine which would be the meaning of the general Quality level concept if it expresses the combination of a low level of timeliness, a high level of believability and a medium level of readability. It seems more useful to evaluate individual quality factors separately for a given document, considering the purpose of issuing a recommendation for a certain user profile.

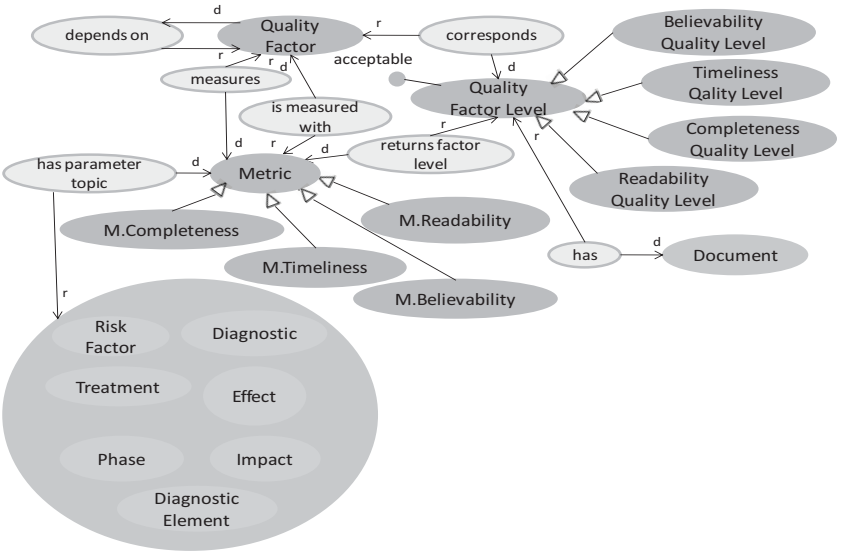

Figure 5. Quality Ontology

As expressed previously, the Quality ontology must be specialized according to the disease and the quality factor to be evaluated. The model of Figure 6 extends the Quality ontology for Alzheimer disease and believability factor. It adds concepts and relationships to the general Quality ontology, to represent parameters of the specific metric. Concepts added are source and author parameters to believability metric.

To represent different metrics of quality factors, the use of SWRL rules is proposed, according to the following arguments:

- It allows the model to express in a clear way the function that metrics play in the process of assigning a quality level for each document. Through rules it can be expressed that from parameters given by certain document properties (expression in the left of the symbol →) the quality level is infered (in the right of →). Below an example of a SWRL rule for the evaluation of the believability quality factor is showed.

```
Document(?d)∧Site(?s)∧QualityLabel(?l)∧hasCertification(?s,?l)∧
hasSource(?d,?s) → hasQualityFactorLevel(?d,Confiable)
```

- As the metric definition is responsibility of experts, the syntax of writing that contains the antecedent and the consequent seems to be an intuitive and ordered mechanism for users who are not specialist in ontology design. They have not to explore the model to decide where to define axioms, etc.

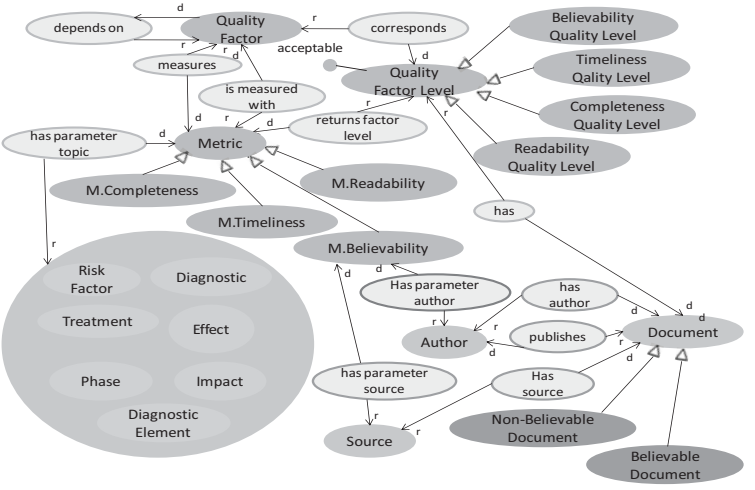

Figure 6. Quality Ontology for Alzheimer and Believability Factor

6 Recommendation Ontology

Figure 7 shows the structure of this ontology, which is general, it defines a basis to model recommendation ontologies focused to a specific disease.

One of the most relevant concepts in this ontology is Reading of Content, which establishes a relationship between two main concepts: Document and User. Reading of Content expresses the action of reading a document executed by a user, so that User is the subject and Document is the object of the recommendation.

Other central concept is Recommendation Metric, that represents if it is recommended or not, and in which degree, that a user read a document. Input parameters (i.e. aspects which influence in the recommendation) are: features of users (Profile User ontology is reused), quality levels of the document for each quality factor (Quality ontology is reused) and the topic of the document. The result of applying this metric allows the model to categorize instances of the Reading of Content concept, associating them a recommendation level. The method of calculation of the level is not expressed in this ontology, these formulas are defined in the specialization of the ontology, for a specific disease. This task is responsibility of the domain expert, too.

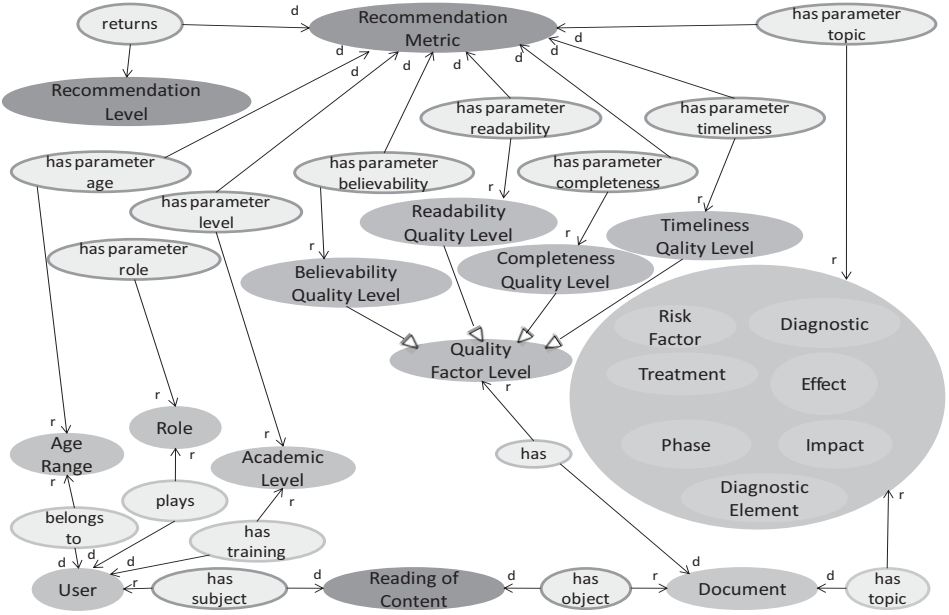

Figure 7. Recommendation Ontology

A Recommendation ontology for a Specific Disease, which extends the general ontology, defines the calculation of the recommendation level through a metric expressed by using SWRL rules.

7 Conclusions and future works

The main aim of this design is to obtain a flexible model that does not depend on particular mechanisms of evaluation of web sites contents, such as a specific quality metric. Whenever it is required to evaluate documents applying a different metric for a quality factor or about another disease, new extensions of Quality and Recommendation ontologies must be implemented, keeping up the core of the model intact, except for general modifications.

Starting from the design presented, good practices of Ontology Engineering lead to evaluate the model in an interaction between ontology engineers and domain experts. From this evaluation it is expected to obtain a feedback to reach a final refinement of structures which compose the model.

In the future, the integration of the Health Domain Ontology with UMLS is an aim, to use the medical terminology and be able to obtain a more wide view of the domain. Another future goal is to carry out an evaluation of tools that facilitate the application of a distributed and collaborative methodology of ontology design. It should be very useful to have a tool to

compare models of specialized ontologies and then to decide the introduction of changes in the general ontology. This practice would facilitate the evolution process of general ontologies avoiding they become out-to-date.

References

[1] Edelweis Rohrer. Study of Methodologies of Ontology Design and Development and application to a case of study of web sites evaluation in the health. Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. To be published.

[2] Jos de Bruijn, Marc Ehrig, Cristina Feier, Francisco Martín-Recuerda, François Scharffe, Moritz Weiten. Ontology mediation, merging and aligning. Semantic Web Technologies. Published Online**:** 3 Jul 2006. Editor(s): John Davies, Rudi Studer, Paul Warren. Print ISBN: 9780470025963 Online ISBN: 9780470030332 DOI: 10.1002/047003033X Copyright © 2006 John Wiley & Sons, Ltd.

[3] Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho. Ontological Engineering. Springer Verlag, 2002/2003.

[4] L. Pipino, Y. Lee, R. Wang. Data quality assessment. Communications of the ACM 4 (2002) 211-218.

[5] William H. DuBay. The Principles of Readability. Costa Mesa, CA: Impact Information, 2004.

Web References

[WR1] http://www.alzheimermed.com.br/

[WR2] http://www.hon.ch/

[WR3] http://www.portalesmedicos.com/web_interes_sanitario/index.htm

[WR4] http://wma.comb.es/