Expert Systems with Applications 40 (2013) 2696–2703

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Evaluation of the OQuaRE framework for ontology quality

Astrid Duque-Ramos [a,*], Jesualdo Tomás Fernández-Breis [a], Miguela Iniesta [b], Michel Dumontier [c], Mikel Egaña Aranguren [d], Stefan Schulz [e], Nathalie Aussenac-Gilles [f], Robert Stevens [g]

[a] Facultad de Informática, Universidad de Murcia, Spain
[b] Facultad de Matemáticas, Universidad de Murcia, Spain
[c] Department of Biology, Carleton University, Canada
[d] Centre for Plant Biotechnology and Genomics, Technical University of Madrid (UPM), Spain
[e] Institute of Medical Informatics, Statistics and Documentation, Medical University Graz, Austria
[f] Institut de Recherche en Informatique de Toulouse, France
[g] Department of Computer Science, University of Manchester, UK

ARTICLE INFO

Keywords:
Ontology quality
Ontology evaluation
Ontology engineering
Software quality standard

ABSTRACT

The increasing importance of ontologies has resulted in the development of a large number of ontologies in both coordinated and non-coordinated efforts. The number and complexity of such ontologies make hard to ontology and tool developers to select which ontologies to use and reuse. So far, there are no mechanism for making such decisions in an informed manner. Consequently, methods for evaluating ontology quality are required. OQuaRE is a method for ontology quality evaluation which adapts the SQuaRE standard for software product quality to ontologies. OQuaRE has been applied to identify the strengths and weaknesses of different ontologies but, so far, this framework has not been evaluated itself. Therefore, in this paper we present the evaluation of OQuaRE, performed by an international panel of experts in ontology engineering. The results include the positive and negative aspects of the current version of OQuaRE, the completeness and utility of the quality metrics included in OQuaRE and the comparison between the results of the manual evaluations done by the experts and the ones obtained by a software implementation of OQuaRE.

1. Introduction

The increasing importance of ontologies has resulted in the development of a large number of ontologies in both coordinated and non-coordinated efforts. Ontologies for domains very different like life sciences, e-commerce, city logistics or geospatial data can be found in Ashburner et al. (2000), Beisswanger, Schulz, Stenzhorn, and Hahn (2008), Hepp (2008), Anand, Yang, van Duin, and Tavasszy (2012), or Tian and Huang (2012). The National Center for Biomedical Ontology (NCBO) BioPortal (http://bioportal.bioontology.org/) has more than two hundred biomedical ontologies and controlled vocabularies, the TONES repository (http://owl.cs.manchester.ac.uk/repository/) contains more than two hundred, and tools like Watson (http://kmi-web05.open.ac.uk/WatsonWUI/) or Swoogle (http://swoogle.umbc.edu/) give access to thousands.

Many ontology builders and application developers have usually several options for which ontology to use to support a particular intelligent software application or to reuse for building a new ontology. Existing ontologies could fit for such purposes with little (or even no) additional development if they have the appropriate content. We regard the ideal scenario as the one in which users are able to select ontology artefacts from a large repository, however, currently not the state of play. Nowadays, ontology builders lack support for making an informed decision, because standardized methods have not been developed for evaluating the quality of ontologies.

In this work, the quality of an ontology is its degree of conformance to functional and non-functional requirements and we assume that such conformance can be measurable. Current work in ontology evaluation can be classified according to the particular evaluation aim: ranking, correctness, or quality evaluation. The ranking approaches range from generic ontology rankings to the selection of the most appropriate ontology for a particular task (see, for instance Alani, Brewster, & Shadbolt, 2006; Lozano-Tello & Gomez-Perez, 2004; Tartir & Arpinar, 2007). The correctness category includes the approaches accounting for the formal correctness of the content represented in ontologies (see, for instance, Corcho, Gómez-Pérez, González-Cabero, & del Carmen Suárez-Figueroa, 2004; Guarino & Welty, 2004, chapter 8; Sleeman

* Corresponding author. Tel.: +34 868884613; fax: +34 868884151.
E-mail addresses: astrid.duque@um.es (A. Duque-Ramos), jfernand@um.es (J.T. Fernández-Breis), miniesta@um.es (M. Iniesta), michel_dumontier@carleton.ca (M. Dumontier), megana@fi.upm.es (M. Egaña Aranguren), stefan.schulz@medunigraz.at (S. Schulz), aussenac@irit.fr (N. Aussenac-Gilles), robert.stevens@manchester.ac.uk (R. Stevens).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.11.004

& Reul, 2006; Vrandecic, 2010). Finally, the quality category is related to the evaluation of ontology quality (see for instance Gangemi, Catenacci, Ciaramita, & Lehmann, 2006; Rogers, 2006; Sabou, Lopez, Motta, & Uren, 2006; Stvilia, 2007). While each has been addressed from different, heterogeneous ways, none has become standard. The need for standardized methods for evaluating the quality of an ontology remains unfulfilled.

In this work, we also approach ontology evaluation as a tool for helping developers to evaluate their ontologies in order to build trust for sharing and reusing ontologies, which is one of the main objectives of ontology evaluation according to Brank, Grobelnik, and Mladenic (2005). For this purpose, our approach is in line with the effort done by the approaches of the quality category, although we propose the adaptation of an already existing standard from the Software Engineering community. In Fernández-Breis, Egaña Aranguren, and Stevens (2009), we presented an approach based on ISO 9126, which is an international standard for software product evaluation, which was applied to two versions of the same biomedical ontology. This evaluation process had the problem of the excessive workload and dependence on human judgment, since the method did not provide computer-support to the evaluator. More recently, the ISO/IEC 25000:2005, which is a standard for Software product Quality Requirements and Evaluation known as SQuaRE ISO (2005), was used by us for evaluating the same two biomedical ontologies, but in an automatic manner Duque-Ramos, Lopez, Fernandez-Breis, and Stevens (2010). The results obtained in that work showed the benefits of using an automatic framework to support ontology evaluation processes, as presented in Duque-Ramos, Fernández-Breis, Stevens, and Aussenac-Gilles (2011).

The usage of a SQuaRE-based approach required the definition of components like quality model and quality metrics. A SQuaRE-based quality model is comprised of a series of quality characteristics like *functional adequacy* or *reliability*. Each characteristic has a series of subcharacteristics associated. For instance, *reusability* can be a subcharacteristic of *maintainability*. Furthermore, each quality subcharacteristic has a series of metrics associated. An example of quality metrics for ontologies could be the mean number of properties per class. Thus, in order to design OQuaRE following such principles, a few decisions had to be taken, like the sets of characteristics, subcharacteristics and metrics, and the association of metrics with subcharacteristics. Scores in SQuaRE are provided in the range 1 (worst) and 5 (best), and the communities that apply SQuaRE must define how the scores of the metrics are mapped onto such range and how such scores are combined to provide scores for the subcharacteristics and characteristics. In our previous works we mentioned that such decisions should be the result of some community agreement: consequently, the objective of this work will be to evaluate such decisions and provide new insights about how ontology quality evaluation should be performed. It is clear that if the community wants ontology quality evaluation to be a engineering activity, there must be a standardization on how such evaluation processes must be performed.

For such purpose, we will describe in this paper the evaluation experiment for assessing both the quality model and quality metrics used in OQuaRE. The experiment will be based on the informed judgment of experts in ontology construction. The experiment will be divided in two rounds: the first one will evaluate the quality model and the second one will evaluate the quality metrics. In both rounds the experts will be asked to apply the evaluation method in a different manner, in order for the authors to acquire a thorough understanding of the different OQuaRE components.

The OQuaRE framework and the evaluation method used in this work will be described in the following section. In Section 3, the results of the OQuaRE evaluation will be presented. Results will be discussed in Section 4, which will also include the conclusions of this work.

2. Methods

This section will describe the main instruments used in this evaluation exercise. We will address the main aspects of OQuaRE and how it has been applied so far (Section 2.1) and the methodological approach for carrying out the evaluation process (Section 2.2).

2.1. OQuaRE

OQuaRE is the Ontology quality evaluation (OQuaRE) framework based on the software product quality SQuaRE. OQuaRE aims at defining all the elements required for ontology evaluation: evaluation support, evaluation process and metrics. However, the current version of OQuaRE includes the quality model and the quality metrics. This means that parts like evaluation requirements and evaluation reports, to date, have not been addressed in depth.

The main objective of OQuaRE is to provide an objective, standardized framework for ontology quality evaluation, which could be applied in a number of situations in a similar way, so the strengths and weaknesses of ontologies can be identified. OQuaRE reuses and adapts the following SQuaRE characteristics to evaluate ontologies: reliability, operability, maintainability, compatibility, transferability and functional adequacy. Most quality subcharacteristics suggested by SQuaRE were also adapted to ontologies in OQuaRE. However, SQuaRE does not include the structural characteristic, which is important in ontologies, as it can be drawn from the number of state-of-the art approaches that use it. Therefore, we included it into OQuaRE. The complete description of the quality model can be found at http://miuras.inf.um.es/evaluation/oquare. Next, we describe the OQuaRE quality characteristics and subcharacteristics.

2.1.1. The quality characteristics

We describe as follows the quality characteristics of OQuaRE. The definition of some subcharacteristics can be found in Table 1.

- Structural: Formal and semantic properties that are important when evaluating ontologies. Some subcharacteristics are formalization, formal relations support, cohesion, tangledness, redundancy and consistency. These subcharacteristics account for the formal properties of the ontologies, the clarity of cognitive distinctions, the appropriate use of ontology modelling primitives and principles, etc.
- Functional adequacy: The capability of ontologies to provide concrete functions that have been identified in literature. An ontology is evaluated for this criterion according to the degree of accomplishment of functional requirements, that is, the appropriateness for its intended purpose according to state-of-the-art literature Stevens, Wroe, Gobel, and Lord (2008): reference ontology, controlled vocabulary, schema and value reconciliation, consistent search and query, knowledge acquisition, clustering and similarity, indexing and linking, results representation, classifying instances, text analysis, guidance and decision trees, knowledge reuse, inferencing, and precision.
- Reliability: Capability of ontologies to maintain their level of performance under stated conditions for a given period of time. Recoverability and availability are some of its subcharacteristics.
- Performance efficiency: Relationship between the level of performance and the amount of resources used, under stated conditions, taking into account elements such as the time response, or memory consumption. It has subcharacteristics like response time and resource utilization.
- Operability: Effort needed for use, and in the individual assessment of such use, by a stated or implied set of users, and it is measured through subcharacteristics such as learnability.

Table 1
Some subcharacteristics of OQuaRE.

STRUCTURAL
Formalization: An efficient ontology has to be built on top of a formal model to support reasoning.
Formal relations support: Most ontologies only have formal support for taxonomy. The usage of additional formal theories would be a positive indicator.
Cohesion: An ontology has a high cohesion if the classes are strongly related.
Tangledness: This measures the distribution of multiple parent categories, so that it is related to the existence of multiple inheritance, which is usually a sign of suboptimal design.
FUNCTIONAL ADEQUACY
Schema and value reconciliation: An ontology can provide a common data model that can be applied to particular views for their reconciliation and integration. Ontologies facilitate the achievement of semantic interoperability if they are able to provide the semantic context for data and information.
Consistent search and query: The formal model of the ontology allows for better querying and searching methods. The ontology structure can guide search processes if they provide a semantic context to evaluate the data wanted by the users. This semantic context is not just provided by the concepts, but also by all the machine computable properties and axioms.
Knowledge reuse: The degree to which the knowledge of an ontology can be used to build other ontologies.
Knowledge acquisition: Ontologies can be seen as templates for generating the forms by which instances are acquired.
MAINTAINABILITY
Modularity: The degree to which the ontology is composed of discrete components such that a change to one component has a minimal impact on other components.
Reusability: The degree to which an asset (part of) the ontology can be used in more than one ontology, or in building other assets.
Analysability: The degree to which the ontology can be diagnosed for deficiencies or causes of failures (inconsistences), or for the parts to be modified to be identified.

- Compatibility: The ability of two or more ontologies to exchange information and/or to perform their required functions while sharing the same hardware or software environment. Replaceability and interoperability are examples of subcharacteristics according to SQuaRE.
- Maintainability: The capability of ontologies to be modified for changes in environments, in requirements or in functional specifications. Some subcharacteristics are modularity, reusability, analysability, changeability, modification stability and testability.
- Transferability: degree to which the ontology can be transferred from one environment (e.g., operating system) to another. Portability and adaptability are example of its subcharacteristics.
- Quality in use: Degree to which an ontology used by specific users meets their needs to achieve specific goals. It has subcharacteristics associated with usability and flexibility in use.

2.1.2. The quality metrics

OQuare does not attempt to develop new metrics for ontology quality evaluation but to reuse and adapt successful metrics from both ontology and software engineering communities. In this sense, some well-known metrics from Objected Oriented Programming have been included in its adapted form to ontologies (see, for instance Chidamber & Kemerer, 1994; Li & Henry, 1993). In addition to this, metrics for the structural properties like the ones presented in Yao, Orme, and Etzkorn (2005) and Tartir and Arpinar (2007) have been reused from the ontology community. Table 2 contains the definition of these metrics for OWL ontologies.

2.1.3. Experiences with OQuaRE

OQuaRE has been applied in two case studies in the last years, and the main results of such can be found at http://miuras.inf.um.es/evaluation/oquare. It was used for the evaluation of two versions of the Cell Type Ontology (see Bard, Rhee, & Ashburner, 2005). A series of ontologies of units of measurement were evaluated. In both cases, the ontologies were firstly manually evaluated, and then automatically evaluated with a home-made software tool developed to run the evaluation experiments. Such tool takes an OWL ontology as input and generates a numeric report as output. This report includes the scores for all the characteristics, subcharacteristics and metrics involved in the evaluation process. The analysis of such results was done in two ways. First, the quality of the ontologies was evaluated and a series of strengths and weaknesses were identified. Second, we identified that doing a manual evaluation was perceived by the participants as very hard and that the results without human intervention did not seem too different. However, there might be several reasons for this, since an evaluation of the appropriateness of the OQuaRE configuration that leads to such results had not been externally evaluated. By OQuaRe configuration we mean the quality characteristics, quality subcharacteristics, quality metrics, which subcharacteristics contribute and how much to a characteristic, which metrics contribute and how much to a subcharacteristic, and how score are scaled in the [1,5] range. Those are the main issues under investigation in this paper.

2.2. The evaluation method

In this section we describe the experimental approach used for evaluating OQuaRE, which follows the Goal/Question/Metric method van Solingen and Egon (1999).

2.2.1. Objectives

The research objectives of this experiment are to evaluate the appropriateness, completeness and usability of the quality model and quality metrics of OQuaRE. We also investigate the feasibility of automatic evaluation processes, through the comparison of manually and automatically calculated results. To achieve such objectives the following goals were defined:

- Appropriateness of the quality subcharacteristics.
- Difficulty in applying the quality subcharacteristics.
- Relevance of metrics for measuring a particular subcharacteristic.
- Usefulness of metrics for measuring a particular subcharacteristic.
- Similarity between the results of the fully manual, metrics-supported manual, and fully automatic evaluation methods.

2.2.2. Participants

The set of participants were selected by applying the following criteria: all the participants should be experts in ontology construction, should be active and interested in ontology quality, should be from different research institutions and from different countries, should represent different "ontology modelling schools", and should not have played an active role in the design of OQuaRE. An additional constraint was imposed by the case study: given that the ontology under evaluation in the experiment was biomedical, background and expertise in biomedical ontologies was required. Three experts were selected for this experiment.

2.2.3. Experimental material

In this section we describe the ontology, documents and tools used in the experiment:

- Ontology: The existence of an ontology of units was considered important by the W3C Semantic Web Best Practices and Development working group (http://www.w3.org/2003/12/swa/swbpd-charter). Since then, a series of ontologies of units of measurement have been developed. Ontologies of units are relevant not only for biomedical domains, but also in any other science and engineering domain (see Rijgersberg, Wigham, & Top, 2011). In Duque-Ramos et al. (2011), a series of ontologies of units were evaluated, and we have selected one of them for this experiment, the Measurement Units Ontology SWEET 2.0 Scientific Units Ontology (SCIUNITS), which can be found at http://sweet.jpl.nasa.gov/2.0/sciUnits.owl. This ontology has 326 classes, 54 object properties, 22 data properties, 103 individuals, 103 class assertion axioms, 134 object property assertion axioms and 76 data property assertion axioms. This ontology was selected because its size and design was found appropriate for this experiment, given our objectives and the constraints associated with the effort.
- Documents: The OQuaRE website was used by the experts as the reference for retrieving all the information about the framework, the quality model, the quality metrics, etc. That website also hosted a page describing the experiment and the forms to be filled by the experts.
- Tools: A home-made software tool that implements the OQuaRE metrics was used. This tool takes an OWL ontology as input and generates a numeric report as output. This report includes the scores for all the characteristics, subcharacteristics and metrics involved in the evaluation process. This tool was not used by the participants for the evaluation although they were provided with the corresponding scores.

2.2.4. Data collection

As mentioned, this experiment was designed in two rounds, each of them having different objectives. Both rounds served us for collecting the data needed for analysing our evaluation goals. In the first round, the experts were asked:

- Read the instructions and get used to the OQuaRE quality model, that is, characteristics and subcharacteristics.
- Apply OQuaRE manually to the target ontology, that is, assigning values in the range [1,5] to each subcharacteristic.
- Answer a questionnaire which contained questions about the difficulty and appropriateness of each subcharacteristic.

In the second round, the experts were asked for:

- Read the instructions and get used to the OQuaRE quality metrics. This step includes to download from the website the scores for each metric calculated by our tool.
- Apply OQuaRE to the target ontology, that is, assigning values in the range [1,5] to each subcharacteristic. For that purpose, they could use the metrics score as a reference.
- Answer a questionnaire which contained questions about the usefulness and relevance of each metric.

2.2.5. Data analysis

The techniques used in the analysis of the data collected depend on the nature of data and on the particular goal. We have compared the results of the manual, computer-supported and automatic evaluation methods using descriptive statistics to obtain a basic description of the samples. ANOVA procedures have been used to test whether there are significant differences between sample means.

With respect to the data collected in round 1, we have applied agglomerative hierarchical clustering to the mean of the answers provided by the experts about the difficulty of getting a score for the quality subcharacteristics in the manual approach and the appropriateness of the quality subcharacteristics for grouping the subcharacteristics into classes. Hence, similar ones in terms of difficulty and appropriateness are in the same class. The procedure has been similar for round 2, looking for clusters of metrics based on their usefulness and relevance. Finally, k-means algorithms have been used to describe particular behaviours of the subcharacteristics in the three evaluation methods.

3. Results

In this section, we present the main results of this experiment. The complete results and the materials used in this experiment can be found at the following website: http://miuras.inf.um.es/evaluation/oquare/SCIUNITS/Experiment.html.

3.1. Hypotheses, variables and experimental design

Hypotheses and variables are established according to the research goals. The variables *Difficulty* and *Appropriateness* are obtained from the sample means of the values given by the experts about difficulty and appropriateness of the subcharacteristics in the questionnaire of round 1.

Our main statistical or null hypothesis says that there is no statistical difference on the mean score by using either the manual, manual assisted or the automatic method. The variables to contrast this hypothesis are obtained from the mean of the assigned values by the experts in each quality subcharacteristic in manual and assisted manual evaluation methods and the scores obtained by the automatic evaluation method, which are called M1, M2 and M3, respectively.

We have chosen an experimental design which tries to identify sources of variability in the experimental units (subcharacteristics), reduce the effect of such sources and improve the precision of the answers to questions of interest. This experimental design involves two factors, the method and the evaluated subcharacteristic. The method of evaluation will be the main factor whose measures (scores) are matched by subcharacteristic producing homogeneous blocks. In other words, the design is a multisample generalization of matched-pairs design. The subcharacteristics with no assigned value in M3 will be discarded in those procedures that need complete blocks.

3.2. Evaluating the answers of the expert

Before obtaining the variables of our study, we analysed the answers obtained from the methods M1 and M2. One way ANOVA applied to the corresponding data allows concluding that there are some pairs of sample means statistically different (P-value < 0.0001). The only two significant differences can be observed between Expert 1 with respect to Expert 2 and Expert 3 in M2. On the other hand, the experts do not modify significantly their mean judgment from M1 to M2.

Fig. 1 shows a basic description of the answers of experts in both methods by means of a boxplot, where the symbol inside the box represents the mean sample, and 95 percent confidence intervals for the means.

Table 2
Definition of the metrics used in OQuaRE.

Lack of Cohesion in Methods (LCOMOnto): the length of the path from the leaf class to Thing, divided by the total number of paths in the ontology
Weighted Method Count (WMCOnto): Mean number of Datatype Properties, Object Properties and subclasses per class
Depth of Inheritance Tree (DITOnto): Length of the largest path from Thing to a leaf class of the ontology
Number of Ancestor Classes (NACOnto): Mean number of superclasses per leaf class
Number of Children (NOCOnto): Mean number of the direct superclasses per class minus the subclasses of Thing
Response for a class (RFCOnto): Number of Datatype Properties and Object Properties that can be directly accessed from the class
Number of properties (NOMOnto): Mean number of Datatype Properties and Object Properties per class
Properties Richness (PROnto): Ratio of the number of Datatype Properties and Object Properties defined in the ontology divided by the number of subclasses, Datatype Properties and Object Properties
Attributes Richness (AROnto): Number of restrictions of the ontology divided by the number of classes
Relationships per class (INROnto): Mean number of subclasses per class
Inheritance relationships richness (CROnto): Mean number of individuals per class
Annotation Richness (ANOnto): Mean number of annotation properties per class

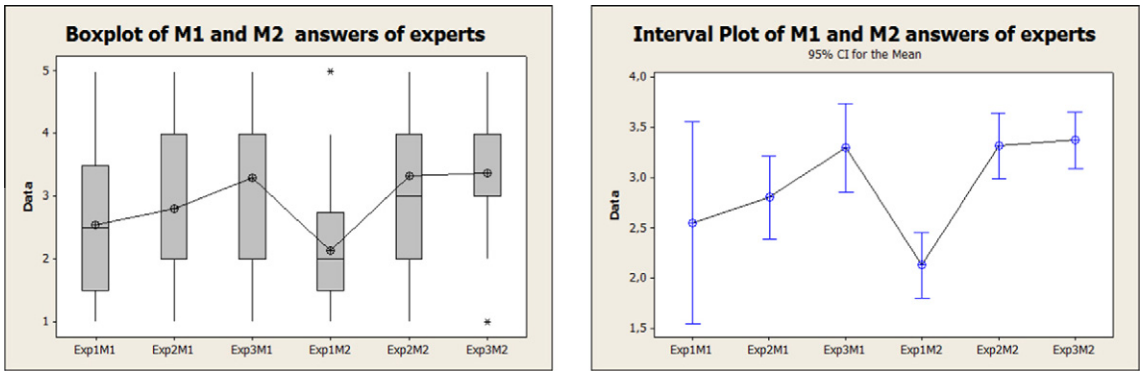

Fig. 1. Boxplots representing a basic description of the answers of experts in both methods and 95 percent confidence intervals for the means.

3.3. Comparison of the evaluation methods

3.3.1. Univariate description

We described the experimental data through summary statistics and graphics. Table 3 shows the basic statistics for the scores given to the case study ontology by using the three methods.

Those statistics are represented in the multiple boxplot in Fig. 2, where the line in the middle of each box corresponds to the median and the sides correspond to the first and third quartiles, respectively. The mean of M3 variable is greater than mean of M2 and M1 but we investigated whether some of these differences are significant. We analyse it in the next section.

3.3.2. Two way ANOVA

We have used the two way analysis of variance procedure with a single observation per cell, that is, by means of an additive model, to analyse whether the observed means scores of the three evaluation methods are statistically different, isolating and removing from the error term the variation attributable to the blocks (subcharacteristics). For this purpose, only complete blocks have been used, thus discarding subcharacteristics with missing values in M3.

The difference between the means of M1 and M3 methods is significant at 0.05 significance level but smaller than 0.04 (P-value = 0.044 in two-way additive ANOVA). Through this test we also obtain that the block factor does not explain a significant proportion of the total variance (P-value = 0.129 associated with this source of variation). According to this, if we only take into consideration the main factor there is not enough evidence that the means are different. Hence, the means score of the three methods are not statistically different at 0.05 of significance level (P-value = 0.055 in one-way ANOVA layout).

Given that the normality and homoscedasticity assumptions of ANOVA model cannot be accepted with this data set, non-parametric procedures were applied to obtain that the observed medians are not statistically different (P-value = 0.119 in Friedman test blocked by subcharacteristic).

The previous results show the lack of evidence to reject the null hypothesis, thus we can conclude that neither the mean nor the median depend on the evaluation method.

Table 3
Basic statistics for the scores of the three methods (1 = lowest; 5 = highest).

Method	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
M1	2.76	1.19	1.00	2.00	2.50	3.50	5.00
M2	3.08	0.68	2.00	2.66	3.00	3.66	5.00
M3	3.33	1.16	1.00	2.66	3.00	4.50	5.00

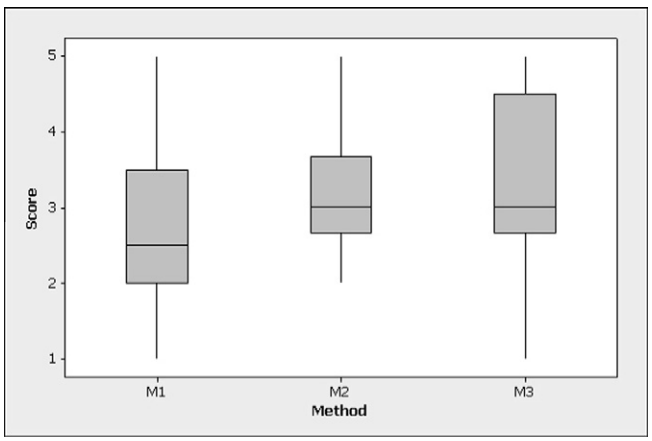

Fig. 2. Boxplot representing basic statistics of scores methods.

3.4. Trivariate description and classification

We calculated the matrix of sample correlations of M1, M2 and M3 variables, which are shown in Table 4 together with their associated *P*-values. The lack of correlation between M1 and M3 and between M2 and M3 and the significant correlation between M1 and M2 are useful because they mean that we could represent the experimental units (subcharacteristics) in two dimensions, discarding only a small proportion of the variation in the data and giving a proper interpretation of the new variables in terms of M1, M2 and M3.

Finally, we attempted to classify the subcharacteristics in three homogeneous groups according to their scores in M1, M2 and M3 by means of k-means clustering. The final partition of the iterative process produces three groups whose centroids appear in Table 5. The subcharacteristics belonging to Cluster 1 tend to have a profile with higher score than the mean in the three methods. The subcharacteristics belonging to Cluster 2 have a high score in M3 and have low score in M1 and M2. The subcharacteristics belonging to Cluster 3 tend to have a profile with lower score than the mean in the three methods.

3.5. Round 1: the quality model

We want to classify the subcharacteristics into clusters in such a way the profile of subcharacteristics in the Difficulty and Appropriateness variables in the same cluster are very similar, whereas the profile of subcharacteristics in different clusters are very different.

Cluster Analysis operates on the matrix of Euclidean distances between all pairs of characteristics and proceeds sequentially yielding a nested arrangement of characteristics in groups, in a hierarchical agglomerative process. We have also chosen Ward's linkage method to determine the distance between two clusters. We obtained a partition in three clusters or groups of subcharacteristics with a maximum of distance 3.46 between them. Table 6 shows the centroid of the clusters and the grand centroid. The subcharacteristics belonging to Cluster 1 have a profile similar to the grand centroid and can be described as appropriate and difficult. The subcharacteristics of Cluster 2 can be described as difficult and not appropriate. Finally, the subcharacteristics of Cluster 3 can be described as appropriate and not difficult.

Table 7 shows the subcharacteristics bellowing to each group by applying of hierarchical cluster to *Difficulty* and *Appropriateness* variables.

Table 4
Sample correlation of M1, M2 and M3.

	M1	M2
M2	0.649	
P-value	0.000	
M3	−0.095	0.001
P-value	0.624	0.994

Table 5
The correlations between the original and new variables.

Variable	Cluster centroid			Grand centroid
	Cluster 1	Cluster 2	Cluster 3	
M1	3.66	1.00	2.12	2.68
M2	3.48	2.00	2.88	3.07
M3	3.97	5.00	2.61	3.33

Table 6
Cluster Analysis using *Difficulty* and *Appropriateness* variables.

Variable	Cluster centroid			Grand centroid
	Cluster 1	Cluster 2	Cluster 3	
Difficulty	3.34	4.75	1.95	2.90
Appropriateness	3.84	2.16	4.28	3.87

Table 7
The subcharacteristics included in each cluster.

Cluster 1	Cluster 2	Cluster 3
Formalization	Cohesion	Tangledness
Formal relations support	Recoverability	Consistency
Redundancy	Learneability	Controlled vocabulary
Reference ontology	Helpfulness	Text analysis
Schema and value reconciliation		Guidance and decision trees
Consistent search and query		Knowledge reuse
Knowledge acquisition		Inference
Clustering and similarity		Availability
Indexing and linking		Modularity
Results representation		Analysability
Reusability		Modification stability
Changeability		Testability
Structural accuracy		Cycles
Domain coverage		Classifying instances
Precision		Ease of use
Error detection		Interoperability
Response time		Effectiveness
Resource utilization		Efficiency
Portability		Context extensibility
Replaceability		
Satisfaction		
Context conformity		
Adaptability		

Table 8
Cluster Analysis using the variables *Usefulness* and *Relevance*.

Variable	Cluster centroid			Grand centroid
	Cluster 1	Cluster 2	Cluster 3	
Relevance	2.83	1.74	3.72	2.47
Usefulness	2.44	1.32	3.51	2.09

Table 9
The associations included in Cluster 1, which can be considered with intermediate relevance and usefulness.

Subcharacteristic	Metric Cluster 1
Consistent search and query	ANOnto, INROnto, Formalization
Knowledge acquisition; clustering and similarity	RROnto
Formal relations support, indexing and linking	
Results representation	CROnto
Knowledge reuse	ANOnto, Formalization, LCOMOnto
Learneability	WMCOnto, LCOMOnto, NOMOnto, CBOnto, NOCOnto
Modularity	WMCOnto
Reusability	DITOnto, NOCOnto
Analysability	CBOnto
Changeability	WMCOnto, DITOnto, LCOMOnto, RFCOnto, NOMOnto
Modification stability	WMCOnto, CBOnto, LCOMOnto, RFCOnto, NOCOnto
Replaceability	NOCOnto, NOMOnto
Adaptability	WMCOnto, RFCOnto, CBOOnto

Table 10
The associations included in Cluster 2, which can be considered with low relevance and usefulness.

Subcharacteristic	Metric Cluster 2
Cohesion; availability	LCOMOnto
Redundancy	ANOnto
Schema and value reconciliation; consistent search and query	RROnto, AROnto
Knowledge acquisition	ANOnto, NOMOnto
Clustering and similarity; indexing and linking; results representation; guidance and decision trees; knowledge reuse	AROnto
Text analysis	Formalization
Recoverability	WMCOnto, DITOnto, NOMOnto, LCOMOnto
Learneability	RFCOnto
Reusability	WMCOnto, RFCOnto, NOMOnto, CBOnto
Analysability	WMCOnto, DITOnto, LCOMOnto, RFCOnto, NOMOnto
Testability	WMCOnto, DITOnto, LCOMOnto, RFCOnto, NOMOnto, CBOnto
Replaceability	WMCOnto, DITOnto
Adaptability	DITOnto

Table 11
The associations included in Cluster 1, which can be considered relevant and useful.

Subcharacteristic	Metric Cluster 3
Tangledness	TMOnto
Controlled vocabulary	ANOnto
Schema and value reconciliation	Formalization, Consistency
Indexing and linking; guidance and decision trees	INROnto
Knowledge reuse	INROnto, NOMOnto, Consistency
Modularity	CBOnto
Changeability	CBOnto, NOCOnto

3.6. Round 2 results: the quality metrics

We analysed the answers provided by the experts about the relevance and usefulness of metrics in the questionnaire of round 2. Given that one metric may be associated with multiple subcharacteristics, the experts might have considered one metric useful or relevant for some subcharacteristics but not for some other ones. For each pair (subcharacteristic, metric), the experts were asked to give a value between 1 (lowest) and 5 (highest).

We applied k-means clustering to the mean of those answers for classifying those pairs (subcharacteristic, metric) in homogeneous groups according to their profiles in *Relevance* and *Usefulness*. The centroids of the three clusters obtained appear in Table 8. The pairs (subcharacteristic, metric) belonging to Cluster 1 tend to have profile with scores slightly higher than the mean in *Relevance* and *Usefulness* (see Table 9). The pairs belonging to Cluster 2 have profile with scores lower than the mean in *Relevance* and *Usefulness* (see Table 10). Finally, the pairs belonging to Cluster 3 have profile with scores greater than the mean in *Relevance* and *Usefulness* (see Table 11).

4. Discussion and conclusions

In this paper we have presented and evaluated the OQuaRe framework for ontology quality evaluation. The results have been presented in the previous section and the major findings and conclusions associated with such results are presented and discussed next.

4.1. Evaluation methods and evaluations of the experts

The ontology of the case study has been evaluated using three different methods. First, the experts were asked to evaluate manually the ontology, with no support except their own expertise and resources. Second, the experts were asked to repeat the process with the support of a series of pre-calculated metrics. Finally, the ontology was given a score by using only those metrics. The results show that we cannot say that there are significant differences between the scores of the three methods, and that the scores with the support of the metrics are higher. The scores of the second evaluation can be higher than in the first one due to several reasons. The experts might have become more familiar with the subcharacteristics due to the previous evaluation or due to the fact that the availability of the metrics may have contributed to a more precise understanding of the subcharacteristics and has provided additional information to the experts. The interesting result is that the same behaviour was found in the results of all the experts and that, even the mean of M2 is higher, there is no evidence of statistical difference, and there is no statistical difference between the scores in M1 and M2 of the experts when measured individually.

4.2. The quality model

According to the results, three clusters of quality subcharacteristics have been identified by grouping them in terms of difficulty and appropriateness. The largest cluster is Cluster 1, which includes the subcharacteristics that are moderately appropriate and moderately difficult, with means between 3 and 4, whereas the smallest is Cluster 2, which includes the difficult and not very appropriate ones. Consequently, it can be said that the set of subcharacteristics included in the quality model are of interest for the participating experts.

The experts found that the definition of some subcharacteristics should be improved since their understandability was part of the difficulty they found in their application. The lack of knowing the intended contexts of use of the ontology was key in the difficulty found by the experts in scoring some subcharacteristics. In the current model, the functional adequacy characteristics include subcharacteristics which might account for that, but this does not mean that they are the intended uses for the ontology under evaluation. SQuaRE includes a quality requirements module which has not been defined so far in OQuare and which would fulfill such requirement. Ontology developers would then determine which properties of the ontology should be evaluated, the functional requirements and the competency questions. The experts also made suggestions about the subcharacteristics. For instance, a division of structural accuracy into a series of subcharacteristics is proposed for being a broad notion. This deserves further study and analysis since some suggestions might be considered metrics instead of subcharacteristics.

4.3. The quality metrics

The objective of the second round was to evaluate the OQuaRE quality metrics and their usage in the framework. The associations between metrics and subcharacteristics have also produced three clusters. Cluster 1 includes the associations with intermediate relevance and usefulness, Cluster 2 contains the low relevance and usefulness metrics, and Cluster 3 contains the relevant and useful associations. It can be said that Clusters 1 and 3 contain positively considered associations and more than 50% of the associations belong to such clusters. Most metrics have been considered useful and relevant for at least one subcharacteristic, but three of them have been considered not useful for all its associations. Such results suggest that a redesign of the set of metrics is needed.

The experts consider that some metrics makes sense only in a given context of use, and that making such distinction is important because, in the absence of additional information, it might be not possible to assess some task-based metrics. This affects specially to the functional category subcharacteristics and their metrics associated, which explains the bad results of such metrics associations. Thus, this result seems to be in line with the discussion of round 1.

The experts found difficult some metrics because of their definition in an OWL-independent way, whereas the ontologies under evaluation are OWL. This is due to the fact that OQuaRE has been designed to be applicable to ontologies in different formats and with different level of formalization. In this way, the experts miss some metrics relevant for OWL ontologies. They suggested new metrics, like relations in use or new associations between metrics and subcharacteristics, like ANOnto for Learnability. In some cases, they recommended having a limited number of metrics per subcharacteristic. This is consistent with the result obtained about the low relevance and usefulness of many pairs (metric, subcharacteristic).

4.4. Evolution of OQuaRE

According to the results, OQuaRE should be improved in several ways before being a candidate for supporting ontology evaluation processes. On the one hand, OQuaRE should be extended with the quality requirements module, which would allow for determining potential contexts of use that would be useful for the human evaluators. Given its relation with SQuaRE, OQuaRE provides a family of evaluation methods. New methods can be defined by just adapting the number of subcharacteristics associated with each characteristic, the number of metrics associated with each subcharacteristics, and the corresponding weights. This would then satisfy the expert suggestions of evaluation profiles, since different combinations could be associated with the different contexts of use of the ontologies. The set of subcharacteristics should also be refined, which can also be done given the philosophy of SQuaRE-based approaches.

It should be noted that our goal differ from other approaches like Oh and Yeom (2012), which focus on the evaluation of ontology modularization, although we plan to see how their results could be used to improve our framework. We are currently extending the home-made tool that has been used in this experiment which would allow the definition of different evaluation methods based on the OQuaRE framework, including profiles for OWL ontologies. This seems a sensible way of proceeding according to the experts suggestions, and this would require to select not only generic metrics that could be applied to ontologies but also metrics specifically designed for OWL ontologies, like the ones provided by http://owl.cs.manchester.ac.uk/metrics/ or the Ontology Pitfall Scanner Poveda, Suárez-Figueroa, and Gomez-Perez (2010). In addition to this, the work developed by the Ontology Usage Framework http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2011_ApplicationFramework_Synthesis is certainly of interest for OQuaRE since they attempt to define a set of ontology usage characteristics and ontology value metrics that could be mapped onto ours.

Acknowledgements

This work has been possible thanks to the funding of the Spanish Ministry of Science and Innovation through Grant TIN2010-21388-C02-02 and co-funded by the FEDER Programme. Mikel Egaña Aranguren is funded by the Marie Curie Cofund programme (FP7).

References

Alani, H., Brewster, C., & Shadbolt, N. (2006). Ranking ontologies with aktiverank. *The Semantic Web-ISWC 2006*, 1–15.

Anand, N., Yang, M., van Duin, J., & Tavasszy, L. (2012). Genclon: an ontology for city logistics. *Expert Systems with Applications, 39*, 11944–11960.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics, 25*, 25.

Bard, J., Rhee, S., & Ashburner, M. (2005). An ontology for cell types. *Genome Biology, 6*, R21.

Beisswanger, E., Schulz, S., Stenzhorn, H., & Hahn, U. (2008). Biotop: an upper domain ontology for the life sciencesa description of its current structure, contents and interfaces to obo ontologies. *Applied Ontology, 3*, 205–212.

Brank, J., Grobelnik, M., & Mladenic, D. (2005). A survey of ontology evaluation techniques. In Proceedings of the conference on data mining and data warehouses (SiKDD 2005).

Chidamber, S., & Kemerer, C. (1994). A metric suite for object oriented design. *IEEE Transactions on Software Engineering*, 467–493.

Corcho, O., Gómez-Pérez, A., González-Cabero, R., & del Carmen Suárez-Figueroa, M. (2004). ODEVAL: a tool for evaluating RDF(S), DAML+OIL and OWL concept taxonomies. In *AIAI* (pp. 369–382).

Duque-Ramos, A., Fernández-Breis, J. T., Stevens, R., & Aussenac-Gilles, N. (2011). Oquare: a SQuaRE-based approach for evaluating the quality of ontologies. *Journal of Research and Practice in Information Technology, 43*, 159–173.

Duque-Ramos, A., Lopez, U., Fernandez-Breis, J., & Stevens, R. (2010). Towards an square-based quality evaluation framework for ontologies. In *Workshop on ontology quality, EKAW 2010*.

Fernández-Breis, J., Egaña Aranguren, M., & Stevens, R. (2009). A quality evaluation framework for bio-ontologies. In *ICBO 2009: proceedings of the 2009 international conference on biomedical ontology* (pp. 136–139). University at Buffalo, NY: Nature Precedings.

Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. (2006). Modelling ontology evaluation and validation. In *ESWC* (pp. 140–154).

Guarino, N., & Welty, C. A. (2004). An overview of ontoclean. In *Handbook on ontologies* (pp. 151–172). Springer.

Hepp, M. (2008). Goodrelations: an ontology for describing products and services offers on the web. *Knowledge Engineering: Practice and Patterns*, 329–346.

ISO (2005). ISO 2005, software engineering – software product quality requirements and evaluation (SQuaRE) – guide to SQuaRE (ISO/IEC 25000). Geneva, Switzerland: International Organization for Standardization.

Li, W., & Henry, S. (1993). Object-oriented metrics that predict maintainability. *Journal of Systems and Software, 23*, 11–122.

Lozano-Tello, A., & Gomez-Perez, A. (2004). Ontometric: a method to choose the appropriate ontology. Ontological analysis, evaluation, and engineering of business systems analysis methods. *Journal of Database Management, 15*, 1–18 [Special issue].

Oh, S., & Yeom, H. Y. (2012). A comprehensive framework for the evaluation of ontology modularization. *Expert Systems with Applications, 39*, 8547–8556.

Poveda, M., Suárez-Figueroa, M., & Gomez-Perez, A. (2010). Common pitfalls in ontology development. *Current Topics in Artificial Intelligence*, 91–100.

Rijgersberg, H., Wigham, M., & Top, J. (2011). How semantics can improve engineering processes: a case of units of measure and quantities. *Advanced Engineering Informatics, 25*, 276–287.

Rogers, J. E. (2006). Quality assurance of medical ontologies. *Methods of Information in Medicine, 45*, 267–274.

Sabou, M., Lopez, V., Motta, E., & Uren, V. (2006). Ontology selection: ontology evaluation on the real semantic web. *International EON Workshop, Evaluation of Ontologies for the web, co-located with WWW20006.*

Sleeman, D., & Reul, Q. (2006). Cleanonto: evaluating taxonomic relationships in ontologies. In *Proceedings of 4th international EON workshop on evaluation of ontologies for the web*.

Stevens, R., Wroe, C., Gobel, C., & Lord, P. (2008). Applications of ontologies in bioinformatics. In S. Staab & R. Studer (Eds.), *Handbook on ontologies in information systems* (pp. 635–658). Springer.

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday*, 12.

Tartir, S., & Arpinar, I. B. (2007). Ontology evaluation and ranking using ontoqa. In *ICSC '07: proceedings of the international conference on semantic computing* (pp. 185–192). Washington, DC, USA: IEEE Computer Society.

Tian, Y., & Huang, M. (2012). Enhance discovery and retrieval of geospatial data using soa and semantic web technologies. *Expert Systems with Applications, 39*, 12522–12535.

van Solingen, R., & Egon, B. (1999). *The goal/question/metric method*. McGraw-Hill Education.

Vrandecic, D. (2010). Ontology evaluation. Ph.D. thesis, Institute of Applied Informatics and Formal Description Methods AIFB.

Yao, H., Orme, A., & Etzkorn, L. (2005). Cohesion metrics for ontology design and application. *Journal of Computer science, 1*, 107–113.