Statistical Study About Existing OWL Ontologies From a Significant Sample as Previous Step for Their Alignment

Jorge Martinez-Gil, Enrique Alba, and José F. Aldana-Montes
Universidad de Málaga, Departmento de Lenguajes y Ciencias de la Computación
Boulevard Louis Pasteur 35, 29071 Málaga (Spain)
http://www.lcc.uma.es

***Abstract*—In this work, we present a proposal for characterizing the OWL ontologies available on the Web from a significant sample. We have conducted a study to review the specific characteristics of these ontologies paying attention to features which can be important from the point of view of the ontology alignment: language, sizes, number, and kind of entities that are represented in them. As a result, we offer some statistical data that can be helpful in order to understand the current situation of OWL ontologies in the Web and, therefore to guide the process of taking decisions when developing applications for aligning them.**

***Keywords*-ontologies; ontology alignment; semantic integration**

Measure what is measurable, and
make measurable what is not so.
– Galileo Galilei

I. Introduction

Ontologies have become one of the key enablers for the Semantic Web vision [1]. Ontologies try to represent knowledge (instead of data or information) in order that (Web) applications can perform more difficult tasks. Unfortunately, ontologies themselves are heterogeneous and distributed. Defined by different organizations or by different people in the same organization, ontologies can have vastly different characteristics [2]. So it is necessary to provide mechanisms in order to identify relations among them. This is the main task of the ontology alignment[1]. Ontology alignment has depthly studied and even, a lot of tools have been developed to deal with the problem [3]. But these tools are often developed without taking into account real knowledge from experts. In order to provide some hints to researchers about real problems we have conducted a study about ontologies available on the Web.

We introduce our work in more depth with a 5W approach which, in our humble opinion, summarizes our purpose.

What is this work about? We have conducted a study to review the specific characteristics of web ontologies paying attention to features which can be important from the point of view of the ontology alignment; such as their language, sizes or amount and kind of entities that are represented.

[1]In this work, we consider the expressions ontology alignment and ontology matching as synonyms

Why is this work useful? Considerable work has been made in the past on automating ontology alignment, either focusing on specific applications or aiming at providing a generic way for various applications. However, most of the state of the art automatic approaches are merely applicable for synthetic ontologies, and the effectiveness of these approaches decreases for real ontologies [4]. Now, we provide a statistical study about these real ontologies.

Where is this work applicable? Web ontologies are now in use in areas as diverse as Web Portals, Multimedia Collections, Design Documentation, Intelligent Agents, Web Services, and so on. Web ontologies are also the focus of much research into reasoning, language extensions, modeling techniques, and tool support that makes these various extensions and techniques accessible to users [5].

When can the results be applied? When developing knowledge management tools. Ontology alignment has been proposed as a way for finding solutions in scenarios where the semantic heterogeneity is a problem. So results for this study can be taken into account when developing solutions for information integration or distributed query processing.

Who can get benefited from it? Application developers. For example, only a few tools, called Partition Block Based[6], DSSIM[7], RIMOM[8], and PRIOR+[9], cares about the problem of deal with real large ontologies. From these tools, DSSIM manually partitions large ontologies into several smaller pieces, while RIMOM and PRIOR+ use simple string comparison techniques as alternatives, so are clearly solutions for improvement [6]. Partition Block Based matching is currently the only technique that is able to work with any kind of web ontologies. Rest of tools do not even take into account that ontologies can become larger.

The rest of this document is structured in the following way: Section 2 describes the related work. Section 3 presents briefly the preliminaries which are necessary to our approach. Section 4 contains the results of our statistical study. In Section 5, we make an interpretation of the results we have obtained. Finally, we summarize with the conclusions extracted from this study.

II. Related Work

To the best of our knowledge, only a few statistical studies about ontologies have been performed in the past. However, none of them have been conducted from the point of the view of the ontology alignment when collecting features from the ontologies. This is a brief summary of them:

- Wang et al. [10] described an algorithm to extract features from real world ontologies in order to obtain a benchmark useful for developers who wish to build software for this kind of ontologies.
- Tempich and Volz [11] used a set of ontologies for collecting information about entities in order to build reasoners. By examining their own data, they proposed to cluster ontologies into five categories.
- Magkannaraki et al. [12] collected information in order to detect problems (missing typing, namespace problems, wrong vocabularies, and so on) from ontologies.
- Bechhifer and Volz [13] conducted a new study by using 277 OWL ontologies in order to obtain the expressivity of them. They showed that many of these OWL Full[2] ontologies (a little restrictive kind of OWL ontology) are OWL Full because of missing type triples, and can be easily patched syntactically.
- Wang et al. [14] extended the work in [13]. They collected a much larger size of samples and applied similar analysis to attempt to patch these OWL Full files. In addition, they shown how many OWL Full files can be coerced into much more restrictive types.
- Finally, Warren [15] paid attention to ontologies in the public domain as their continuing availability in order to monitor the ongoing projects for developing ontologies.

The novelty of our work in relation to these studies is that we have conducted a study to find the characteristics of existing public web ontologies paying attention to features such as their language, sizes or amount and kind of entities that are represented. In our opinion, these characteristics are useful from the point of view of the developer of ontology alignment tools who frequently has to take decisions related to the ontologies these tools have to deal with.

III. Preliminaries

OWL Web Ontology Language [16] is the most common language for representing web ontologies. OWL has been designed to be used by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content. Components from OWL ontologies are defined now. It is neccesary to bear in mind these concepts because they are going to be the object of our study.

[2]OWL Full is a kind of OWL ontology designed to be compatible with RDF Schema

Definition 1 (Class). *A class is a kind of ontology entity that defines a group of individuals that belong to this class because they share some properties.*

Example 1. `Jorge` , `Enrique` and `José` are members of the class `Person`. Classes can be organized in a specialization hierarchy using `subClassOf`.

In general, there is a most general class named Thing that is the class of all individuals and is a superclass of all classes. There is also a most specific class named Nothing that is the class that has no instances and a subclass of all classes [16].

Definition 2 (Property). *A Property is a kind of ontology entity that states relationships between individuals or between individuals and data values.*

There are two kinds of properties: a) Object Property and b) Datatype Property. The first kind can be used to relate an instance of a class to another instance of other class. The second can be used to relate an instance of a class to an instance of a datatype.

Example 2. For example, the property `wasBorn` is an Object Property. Because it can be used to link an instance of a class for representing `People` to other instance of a class representing `Places` . For example (: denotes instance), `:Marta wasBorn :Madrid.`

On the other hand, the property `hasAge` is a Data Property because it can be used to relate an instance of the class representing `People` to an instance of the datatype Integer. For example (: denotes instance), `:Marta hasAge 29.`

Definition 3 (Individual). *An Individual is a kind of ontology entity that is an instance of one o more classes, and properties may be used to relate one individual to another.*

Example 3. An individual named `Jorge` may be described as an instance of the class `Person` and the property `hasNationality` may be used to relate the individual `Jorge` to the individual `Spanish` .

A. Data collection

We have used the international version of the Google[3] search engine to collect our OWL ontologies. We have taken the 300 first ontologies that are indexed for the query *filetype:owl*. Unlike other works, where *toy ontologies*[4] are discarded, we have not discarded any kind of ontologies.

[3]http://www.google.com

[4]Several authors use the term *toy ontology* for naming those kind of ontologies that are not useful, i.e. examples, tests and, so on

Language	#Ontologies	Percentage
English	250	0.833
Neutral	12	0.040
Spanish	10	0.033
German	10	0.033
Portuguese	5	0.016
Internationalized	5	0.016
French	3	0.010
Italian	2	0.007
Dutch	2	0.007
Polish	1	0.003

Table I
SUMMARY OF THE MOST USED LANGUAGES FOR DEVELOPING ONTOLOGIES

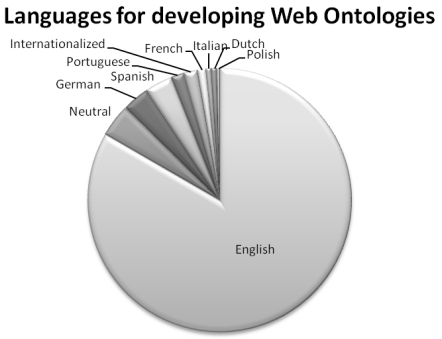

Figure 1. Representation of the most used languages for developing ontologies

Moreover, we have included ontologies that are bad-formed. In case we have to deal with a bad-formed ontology, its contribution to data collected will be ignored. Protegé[5] has been used to count the entities contained on the ontologies. The collection task was was done until march 2009.

IV. STATISTICAL STUDY

In this section, we perform a statistical study to understand several characteristics from OWL ontologies. These are the aspects to research and their justification:

- *Language chosen for developing the OWL ontologies.* This aspect is important because it can help designers to take decisions related to the inclusion of background knowledge support.
- *Size of the files where OWL ontologies are contained.* This aspect is important when designing input components for ontology alignment tools.
- *Amount and nature of the entities represented on the OWL ontologies.* Understanding this fact can help designers when taking decisions about the inclusion of ontology matching algorithms.
- *Classification of the ontologies according to the statistical data obtained.* We think that it is a very important too, because it can help us to decide when a ontology is small, when is medium size, and when is large from a strictly statistical point of view.

In Table 1 we can see the absolute number and the percentage of ontologies available on the Web for a specific language.

Figure 1 is the graphical representation for Table 1. English is the most used language used for developing existing ontologies, followed by German and Spanish.

Size of the files where ontologies can be contained could seem irrelevant: there are comments, overhead, and so on. But in practice, programmers have to build applications that accept as input this kind of files. So, although this characteristic has not a strong importance from a theoretical point of view, it is useful in the practice. Table 2 shows

[5]http://protege.stanford.edu/

Method	Result (Kb)
Average Mean	204.26
Standard Deviation	976.18
Mode	5.00
Median	43.00
Variance	952934.27

Table II
STATISTICAL SUMMARY OBTAINED FROM THE SIZES OF THE FILES WHERE ONTOLOGIES ARE CONTAINED

a statistical summary obtained from the sizes of the files where ontologies are contained.

The average size for the file where an ontology is contained is 204.26 Kb. The standard deviation and variance are so high, so the dispersion is high. The most repeated size in the sample is a file of 5 Kb. An the median (central value) is much lower that the average mean.

Figure 2 shows an histogram for representing the size for the `owl` files that contains the web ontologies. Ontologies has been grouped in 250 Kb multiples. The last bar represents the amount of ontologies larger than 1000 Kb that we have found.

Figure 3 represents the size distribution for the files. The logarithmic function seems to be the most appropriate to do that. The equation that tries to represent the trend of the empirical data can be seen in the graphic. The quality of this function when representing the sizes of the ontologies is 93.94 percent.

Figure 4 represents the distribution of the total existing entities. We have obtained that the 48% of entities are classes, 43% are individuals, 6% are object properties, and only 3% are datatype properties.

Table 3 summarizes the information related to entities that are represented into the ontologies. We can notice that the dispersion of data is very high. Moreover, the big difference between the average mean and the median tell us that there is a larger number of small ontologies than large ontologies.

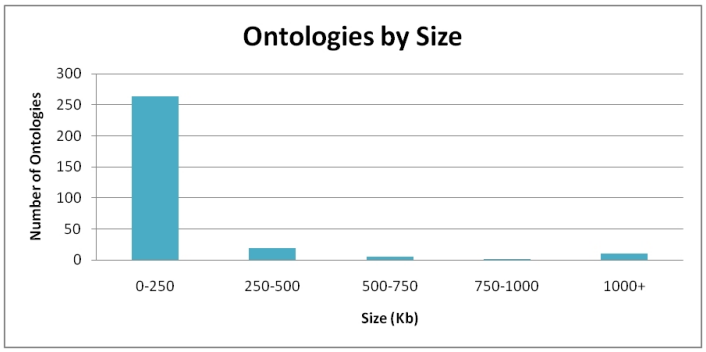

Figure 2. Histogram for representing the size for the files that contains the web ontologies

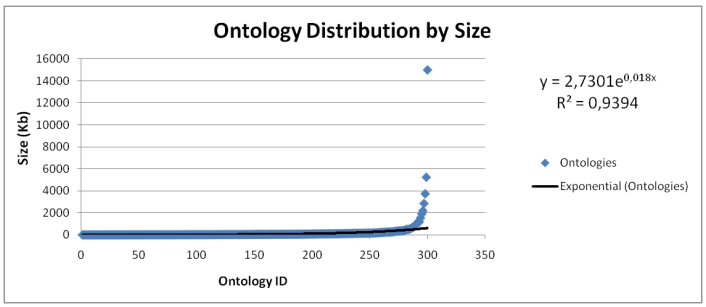

Figure 3. Distribution of the sizes of the ontologies

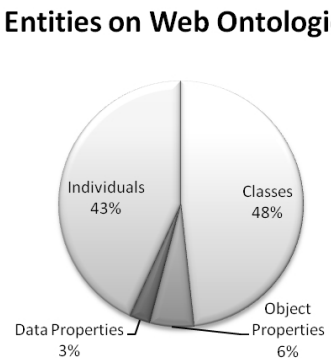

Figure 4. Percentage of entities represented in the ontologies from the collected sample

We think that entities from large ontologies make a key contribution to increase the value for the average mean. Maximum and minimum values are the largest and the smallest number of entities respectively.

In Table 4, we have partitioned the sample in five equivalence classes. These equivalence classes are non-exclusive, thus, a given ontology can belong to one if we attend at its classes and also to another if we attend at its individuals. We have named to these classes in the following way: a) Very Small Ontologies, b) Small Ontologies, c) Medium Ontologies, d) Large Ontologies, and e) Very Large Ontologies.

V. Discussion

This section is about an interpretation of the results we have obtained. The section is divided in subsections corresponding to the four most important aspects of the study: a) Languages for developing the ontologies, b) distribution of the sizes of the ontologies, c) entities contained in the ontologies, and d) classification into categories.

A. *About the languages of the ontologies*

Most of the ontologies from our sample (83.3%) are in English. This overwhelming majority of this language for developing ontologies gives us an evidence that most of the knowledge contained on the Web is in English. It is neccesary also to mention the effort for developing neutral ontologies when possible (for describing very precise domains where entities can be represented using codes, for example). German and Spanish languages are important too, but they are far from the first. Internationalized ontologies, thus, the kind of ontologies where entities are in several languages, represents a marginal amount of the existing ontologies currently available. But, what does all mean for a developer? Well, ontology matching developers who only include support for English dictionaries in their tools will cover the most of the real cases. This percentage could be higher as they include support for the rest of languages.

	#Classes	#Object Properties	#Data Properties	#Individuals
Average Mean	384.73	46.16	21.82	343.62
Standard Deviation	1856.75	86.93	46.93	1801.95
Mode	2.00	0.00	0.00	0.00
Median	52.50	17.50	7.00	11.00
Variance	3447517.65	7557.73	2202.52	3247017.29
Maximum	23141.00	950.00	557.00	17943.00
Minimum	0.00	0.00	0.00	0.00

Table III
STATISTICAL DATA RELATED TO ENTITIES THAT ARE REPRESENTED FROM THE ONTOLOGIES COLLECTED

	#Classes	#Object Properties	#Data Properties	#Individuals
Very Small Ontologies	0-12	0-3	0	0
Small Ontologies	13-29	4-11	1-4	1-4
Medium Ontologies	30-75	12-30	5-11	5-26
Large Ontologies	76-160	31-60	12-34	27-172
Very Large Ontologies	171-23141	61-950	35-557	173-17943

Table IV
PARTITION OF THE SAMPLE ACCORDING TO EQUIVALENCE CLASSES

B. About the sizes of the ontologies

Sizes of the ontologies follow a long tail distribution (also known as Zipf distribution or Pareto distribution). That it is to say, the size of ontologies is very small for a big proportion of the population of the distribution and this size is increased gradually for the rest of ontologies. The main characteristic of this kind of distribution is known as the 80/20 rule. Thus, the 80 percent of the population is small, and the other 20 percent is distributed along a long tail of sizes that are increased gradually.

Developers of ontology alignment tools can use this characteristic for taking decisions about the percentage of real ontologies that is covered by their tools. That it is to say, developing a tool for dealing with the 80 percent of the ontologies is easy but, dealing with the rest of the population of ontologies becomes more difficult in a gradual way.

C. About the entities represented in the ontologies

According to our study, we have a *web of classes*. Classes are designed to contain individuals but, nowadays, we have more classes (or groups of individuals) than individuals. One possible explanation could be that ontologies are frequently used as models for interoperability purposes, instead of annotating resources. In order to the Semantic Web may become real, ontologies should begin to be used more intensively for annotating resources. Related to the small number of properties, maybe ontologies are not enough expressive and are unfortunately still often reduced to some kind of ligthweight models like taxonomies.

What is the lesson that a developer can learn from this? Well, it seems a good idea to design algorithms which uses individuals for comparing the classes to which they belong. However, these tools are not going to find many individuals yet.

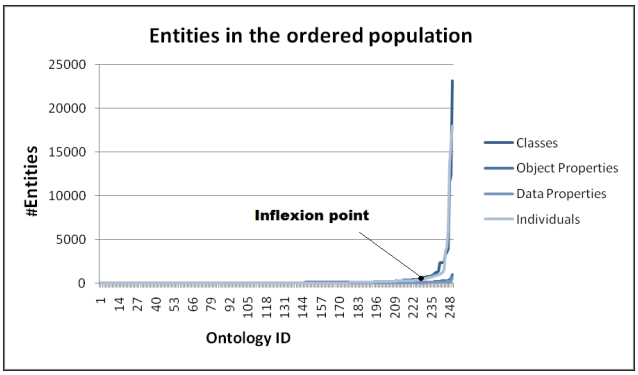

Figure 5. Inflexion point tells us where the linear trend for entities is broken, and therefore where we can begin to call Very Large to ontologies

D. About the classification into categories

If we attend to the results, we can realize of an annoying fact. Could be an ontology considered very large with 171 classes? Well the answer is not clear. Firstly, from a strictly statistical point of view, an ontology with 171 classes has a larger number of classes than the 80 percent of existing ontologies. But it is neccesary that this ontology may have at least 61 object properties, 24 data properties and 173 individuals to be considered as a complete very large ontology. However, experience tells us that it still seems to be a medium size ontology.

Maybe we should use the average size of an OWL ontology. We have that according to the average mean, a medium ontology has 384.73 classes. So we could consider an ontology with a larger number of classes as a large ontology, at least, larger that the mean. The problem consists in that the number of classes still seems to be insufficient to be considered as a big one.

We think that the solution to the problem can be found by inspection of the Figure 5. We can see that entities follow a linear trend in most part of the figure, but this trend is

broken in a point (called inflexion point) where the number of classes and individuals begin to grow in an exponential way. We think that it is reasonable to consider this inflexion point, where an explosion of classes and individuals can be appreciated, as the limit for separating very large ontologies from the rest. This inflexion point tell us that the limit could be near to 1500 classes or 1500 individuals.

VI. Conclusions

In this work, we have surveyed a significant sample of OWL ontologies available on the Web. The end goal of this work is to provide some information about characteristics that can be interesting from the point of view of the ontology alignment. As conclusion of this work, we can remark several interesting points:

1) Most of the ontologies from our sample (83.3%) are in English. It exists a big difference in relation to the second most used language: neutral (4%), thus, ontologies which only contain technical words that are not attribuible to any language. German and Spanish languages are the third most used languages when developing OWL ontologies, but their use is marginal in comparison with English.
2) Size for existing OWL ontologies tends to follow a long tail distribution. According to the heuristic formulated by Pareto for this kind of distributions, this means that the 80 percent of the population is small and, the other 20 percent is distributed along a tail of sizes that are increased slowly and gradually.
3) We have studied the nature and distribution of entities represented on the ontologies and we have found that classes are the most represented entity. Therefore, we have more groups of individuals than individuals themselves on the Web. This is an evidence that ontologies are not being used intensively for annotating resources or, at least, that are not being populated.
4) Finally, we have been able to establish a five-class classification of ontologies according to the kind and number of entities that they contain. We have ordered and partitioned the set of ontologies and we have obtained five non-exclusive equivalence classes and the conditions that are necessary to test in order to determine if a given ontology belongs to them. We have discussed about the existence of a inflexion point where linear trend for the growth of entities is broken. We have proposed to use this inflexion point in order to differentiate Very Large Ontologies from the rest.

As future work, we propose to use the results of this study to develop applications that can address the problem of aligning real ontologies. We think that the statistical data that we have provided can guide to developers when taking design decisions for their ontology alignment tools.

Acknowledgments

This work have been funded by the Spanish Ministry of Sciences and Innovation (MICINN) and FEDER under contracts TIN2008-04844 and TIN2008-06491-C04-01 and CICE, Junta Andalucia, under contracts P07-TIC-02978 and P07-TIC-03044.

References

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

[2] J. Li, J. Tang, Y. Li and, Q. Luo. RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Trans. Knowl. Data Eng. 21(8): 1218-1232 (2009)

[3] J. Euzenat, P. Shvaiko. Ontology Matching. Springer-Verlag, 2007

[4] P. Shvaiko, J. Euzenat, F. Giunchiglia and, B. He. Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007) Busan, Korea, November, 2007

[5] A. Kalyanpur, B. Parsia and, J. Hendler. A Tool for Working with Web Ontologies. Int. J. Semantic Web Inf. Syst. 1(1): 36-49 (2005)

[6] W. Hu, Y. Qu and, G. Cheng: Matching large ontologies. A divide-and-conquer approach. Data Knowl. Eng. 67(1): 140-160 (2008)

[7] M. Nagy, M. Vargas-Vera and, E. Motta. DSSim managing uncertainty on the semantic web, in: Proceedings of ISWC + ASWC Workshop on Ontology Matching, 2007, pp. 160-169.

[8] M. Mao and Y. Peng. The Prior+: results for OAEI campaign 2007, in: Proceedings of ISWC + ASWC Workshop on Ontology Matching, 2007, pp. 219-226.

[9] J. Tang, J. Li, B. Liang, X. Huang, Y. Li and, K. Wang. Using Bayesian decision for ontology mapping, Journal of Web Semantics 4 (4) (2006) 243-262

[10] T. Wang, B. Parsia and, J. Hendler. A Survey of the Web Ontology Landscape. International Semantic Web Conference 2006: 682-694

[11] S. Bechhofer and R. Volz. Patching Syntax in OWL Ontologies. International Semantic Web Conference 2004: 668-682

[12] A. Magkanaraki, S. Alexaki, V. Christophides, and D. Plexousakis. Benchmarking RDF Schemas for the Semantic Web. International Semantic Web Conference 2002: 132-146

[13] C. Tempich and R. Volz. Towards a benchmark for Semantic Web reasoners - an analysis of the DAML ontology library. EON 2003

[14] S. Wang, Y. Guo, A. Qasem, and J. Heflin. Rapid Benchmarking for Semantic Web Knowledge Base Systems. International Semantic Web Conference 2005: 758-772

[15] R. Warren, Ontologies: Where are we at?, Knowledge-Based Bioinformatics Workshop, September 2005.

[16] OWL, Ontology Web Language, http://www.w3.org/TR/owl-features/, 2008.