Characterizing Knowledge on the Semantic Web with Watson

Mathieu d'Aquin, Claudio Baldassarre, Laurian Gridinoc,
Sofia Angeletou, Marta Sabou, and Enrico Motta*

Knowledge Media Institute (KMi), The Open University, United Kingdom
{m.daquin,c.baldassarre,l.gridinoc,s.angeletou,r.m.sabou,e.motta}@open.ac.uk

Abstract. Watson is a gateway to the Semantic Web: it collects, analyzes and gives access to ontologies and semantic data available online with the objective of supporting their dynamic exploitation by semantic applications. We report on the analysis of 25 500 ontologies and semantic documents collected by Watson, giving an account about the way semantic technologies are used to publish knowledge on the Web, about the characteristics of the published knowledge, and about the networked aspects of the Semantic Web. Our main conclusions are 1- that the Semantic Web is characterized by a large number of small, lightweight ontologies and a small number of large-scale, heavyweight ontologies, and 2- that important efforts still need to be spent on improving the published ontologies (coverage of different topic domains, connectedness of the semantic data, etc.) and the tools that produce and manipulate them.

1 Introduction

The vision of a Semantic Web, *"an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"* [3], is becoming more and more a reality. Technologies like RDF and OWL, allowing to represent ontologies and information in a formal, machine understandable way are now well established. More importantly, the amount of knowledge published on the Semantic Web – i.e, the number of ontologies and semantic documents available online – is rapidly increasing, reaching the critical mass required to enable the vision of a truly large scale, distributed and heterogeneous web of knowledge.

In a previous paper [4], we presented the design and architecture of Watson, a gateway to the Semantic Web. Watson is a tool and an infrastructure that automatically collects, analyses and indexes ontologies and semantic data available online in order to provide efficient access to this knowledge for Semantic Web users and applications. Besides enabling the exploitation of the Semantic Web, Watson can be seen as a research platform supporting the exploration of

* This work was funded by the Open Knowledge and NeOn projects sponsored under EC grant numbers IST-FF6-027253 and IST-FF6-027595

the Semantic Web to better understand its characteristics. This paper reports on the use of this infrastructure to provide quantitative indications about the way semantic technologies are used to publish knowledge on the Web, about the characteristics of the knowledge available online, and about the way ontologies and semantic documents are networked together.

A number of researchers have already produced analyses of the Semantic Web landscape. For example, [6] presents an analysis of 1 300 ontologies looking in particular at the way ontology language primitives are used, and at the distribution of ontologies into the three OWL species (confirming results already obtained in [2]). In [5], the authors of Swoogle present an analysis of the semantic documents collected by Swoogle. The forthcoming section shows complementary results to the ones presented in both these studies, based on a set of almost 25 500 semantic documents collected by Watson. In particular, in comparison with [5] that focuses on the Web aspects of the Semantic Web (number of files, provenance in terms of website and internet domain, RDF(S) primitive usage, etc.), we consider a more "Semantic Web" centric view, by providing an insight on characteristics like the expressiveness of the employed ontology languages, the structural and domain-related coverage characteristics of semantic documents, and their interconnections in a *knowledge network*.

2 Characterizing Knowledge on the Semantic Web with Watson

Below, we report on some of the results that have been obtained by collecting, validating and analyzing online ontologies and semantic documents. We focus on three main aspects in this study: the usage of semantic technologies to publish knowledge on the Web (Section 2.1), the characteristics of the knowledge published (Section 2.2) and the connectedness of semantic documents (Section 2.3).

Different sources are used by the Watson crawler to discover ontologies and semantic data (Google, Swoogle[1], *Ping the Semantic Web.com*[2], etc.) Once located and retrieved, these documents are filtered to keep only valid RDF based documents (by using Jena[3] as a parser). In addition, we have chosen to exclude RSS and FOAF files from the analysis. The main reason to exclude these documents is that RSS and FOAF together represent more than 5 times the number of other RDF documents in our collection. These two vocabularies being dedicated to specific applications, we believe that they would have introduced a bias in our characterization and therefore, that they should be studied separately. We consider here a set of almost 25 500 semantic documents collected by Watson.

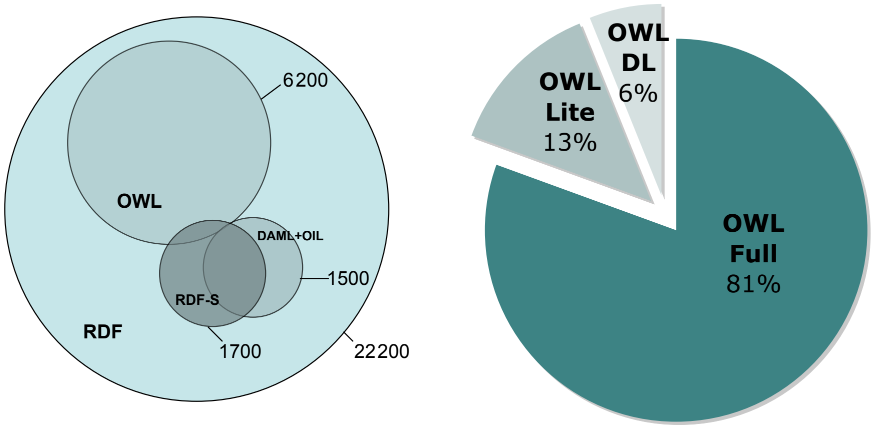
[1] http://swoogle.umbc.edu/

[2] http://pingthesemanticweb.com/

[3] http://jena.sourceforge.net/

2.1 Usage of Semantic Technologies

Semantic technologies such as OWL and RDF are now well established and commonly used by many developers. In this section, we look at the details of how the features provided by Semantic Web languages are exploited to describe ontologies and semantic data on the Web.



(a) **(b)**

Fig. 1. Usage of the ontology representation languages (a) and of the three OWL species (b).

Representation Languages. Watson implements a simple, but restrictive language detection mechanism. It is restrictive in the sense that it considers a document to employ a particular language only if this document actually *instantiates* an entity of the language vocabulary (any kind of description for RDF, a class for RDF-S, and a class or a property for OWL and DAML+OIL). Figure 1(a) provides a visualization of the results of this language detection mechanism applied on the entire set of semantic documents collected by Watson. A simple conclusion that can be drawn from this diagram is that, while the majority of these documents are exclusively considering factual data in RDF, amongst the ontology representation languages (RDF-S, OWL and DAML+OIL), OWL seems to have been adopted as standard. Another element that is worth to consider is the overlap between these languages. Indeed, our detection mechanism only considers a document to employ two different languages if it actually declares entities in both languages. For example, a document would be considered as being written in both RDF-S and OWL if it contains the definition of an *owl:Class* or an *owl:Property*, together with the definition of an *rdfs:Class*. According to this definition, the use of RDF-S properties like *rdfs:label* is not sufficient to consider the document as being written in RDF-S. Combining entities from two different meta-models, like for example OWL and RDF-S, can be problematic for the tools that manipulate the ontology (in particular, the inference mechanisms can become undecidable). These considerations have been taken into account in the design of OWL. As a consequence, unlike DAML+OIL documents, most of the OWL documents only employ OWL as an ontology language, leading to cleaner and more exploitable ontologies (see Figure 1(a)).

OWL is divided into three sub-languages, OWL Lite, OWL DL, and OWL Full, that represent different (increasing) levels of complexity. In this respect, the results obtained on the proportion of OWL documents of the three species are surprising (see Figure 1(b)): a large majority of the OWL ontologies are OWL Full. This confirms the results obtained by Wang et al. in [6] on a set of 1 300 ontologies. The explanation provided in [6] is that most ontologies fall into the OWL Full category because of simple syntactic mistakes. This intuition that documents are considered as OWL Full ontologies not because they use the expressive power of this sub-language is confirmed in the next paragraph, which looks at the expressiveness employed by ontologies.

Expressiveness. The Pellet reasoner[4] provides a mechanism to detect the level of expressiveness of the language employed in an ontology in terms of description logics (DLs). DLs are named according to the constructs they provide to describe entities, and so, to their expressive power. For example, the DL of OWL Lite is $\mathcal{ALCR}_+\mathcal{HIF}(D)$, meaning for example that it allows the description of inverse relations ($\mathcal{I}$) and of limited cardinality restrictions ($\mathcal{F}$).

Total		**OWL**		**OWL Full**	
DL	*Nb Documents*	*DL*	*Nb Documents*	*DL*	*Nb Documents*
$\mathcal{AL}$(D)	21375 (84%)	$\mathcal{AL}$(D)	3644 (59%)	$\mathcal{AL}$(D)	3365 (78%)
$\mathcal{AL}$	2455 (10%)	$\mathcal{AL}$	1406 (23%)	$\mathcal{AL}$	281 (6.5%)
$\mathcal{ALH}$(D)	293 (1%)	$\mathcal{ALCF}$(D)	105 (1.5%)	$\mathcal{ALCF}$(D)	68 (1.5%)
$\mathcal{ALCF}$(D)	105 (<1%)	$\mathcal{ALC}$	94 (1.5%)	$\mathcal{ALH}$(D)	44 (1%)
$\mathcal{ALH}$	102 (<1%)	$\mathcal{ALH}$(D)	54 (<1%)	$\mathcal{ALCOF}$(D)	28 (<1%)
$\mathcal{ALC}$	101 (<1%)	$\mathcal{ALCOF}$(D)	43 (<1%)	$\mathcal{ALC}$	27 (<1%)

Table 1. Most common classes of expressiveness employed by semantic documents, on the entire set of semantic documents collected by Watson, on the sub-set of OWL ontologies and on the sub-set of OWL Full ontologies.

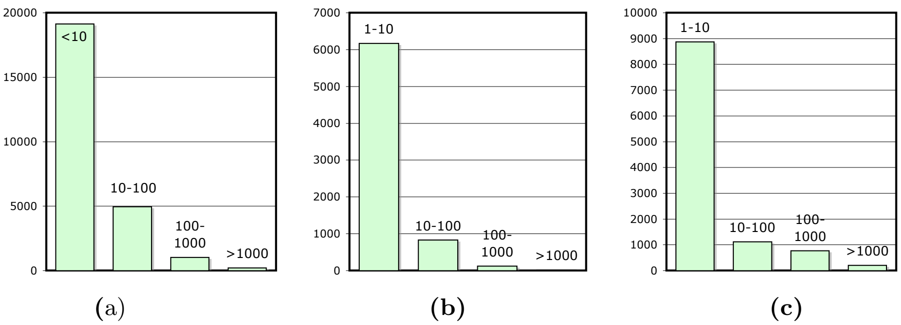
Using this mechanism allows us to assess the complexity of semantic documents, i.e., how they employ the expressive power provided by ontology representation languages. Indeed, the analysis presented in Table 1 shows that the advanced features provided by the ontology representation languages are rarely used. $\mathcal{AL}$ is the smallest DL language that can be detected by Pellet. Only adding the use of datatypes (D) and of hierarchies of properties ($\mathcal{H}$) to $\mathcal{AL}$ is sufficient to cover 95% of the semantic documents. It is worth mentioning that these two elements are both features of RDF-S.

[4] `http://www.mindswap.org/2003/pellet/`

Looking at the results for OWL and OWL Full ontologies (second and third parts of Table 1), it appears that the division of OWL in Lite, DL and Full, which is based on the complexity and on the implementation cost, is not reflected in practice. Indeed, the fact that most OWL Full ontologies employ only very simple features confirms the intuition expressed in the previous paragraph: while these ontologies would get the disadvantages of using OWL Full, they do not actually exploit its expressiveness. Moreover, while one of the most popular feature of OWL, the possibility to build enumerated classes ($\mathcal{O}$), is only permitted in OWL DL, transitive and functional properties ($\mathcal{R}+$), which are features of OWL Lite, are rarely used.[5]

2.2 Structural and Topic Coverage Characteristics of Knowledge on the Semantic Web

One important aspect to consider for the exploitation of the Semantic Web concerns the characteristics of the semantic documents in terms of structure and topic coverage. In this section, we report on the analysis of these aspects from the data provided by the Watson repository with the objective of helping users and developers in knowing what they can expect from the current state of the Semantic Web.



(a) (b) (c)

Fig. 2. Number of semantic documents (y axis) in 4 categories of size, in terms of the total number of entities (a), classes (b), and individuals(c).

Size. As already mentioned, Watson has collected almost 25 500 distinct semantic documents (by distinct we mean that if the same file appears several times, it is counted only once, see Section 2.3). Within these documents, about 1.1 million distinct entities (i.e. classes, properties, and individuals having different URIs) have been extracted.

[5] Considering only features not handled by RDF-S (i.e. excluding $\mathcal{ALH}(\mathcal{D})$), $\mathcal{O}$ is the third most used feature of OWL with 236 ontologies, after $\mathcal{C}$ (748) and $\mathcal{F}$ (598), while $\mathcal{R}+$ is last with only 31 ontologies.

An interesting information that can be extracted from this analysis is that ontologies on the Semantic Web are generally of very small size. Indeed, the average number of entities in semantic documents is around 43, that is far closer to the minimum size of semantic documents (1 entity) than to the bigger one (more than 28 000 entities). Looking more in detail, it can be seen that the Semantic Web is in fact characterized by a large number of very small documents, and a small number of very large ones (see Figure 2(a)). It is worth mentioning that, as shown in Figures 2(b) and 2(c), this observation is valid for both ontological knowledge and factual data.

Measures	**Value**
Total number of classes	161 264
Total number of properties	76 350
Total number of individuals	984 526
Total number of domain relations	32 572
Total number of sub-class relations	106 729
Total number of instance relations	1 114 795
average **P-density** (number of properties per class)	**0.20**
average **H-density** (number of super-classes per class)	**0.66**
average **I-density** (number of instances per class)	**6.9**

Table 2. Measures of density over the Watson repository.

Density. One way to estimate the richness of the representation in semantic documents is to rely on the notion of density. Extending the definition provided by [1], we consider the density of a semantic entity to be related to its interconnection with other entities. Accordingly, different notions of density are considered: the number of properties for each class (P-density), the number of super-classes for each class (H-density), and the number of instances for each class (I-density). In the case of P-Density, a class is considered to possess a property if it is declared as the domain of this property. It is worth mentioning that none of these measures takes inheritance into consideration: only directly stated relations are counted. Computing these measures on the whole Watson repository (see Table 2) allows us to conclude that, on average, ontology classes are described in a lightweight way (this correlates with the results obtained in the previous section concerning the expressiveness of the employed language). More precisely, the P-density and H-density measures tend to be low on average, in particular if compared to their maximum (17 and 47 respectively). Moreover, it is often the case that ontologies would contain a few “central”, richly described classes. This characteristic cannot be captured by simply looking at the average density of the collected entities. Therefore, we looked at the maximum density within one ontology (i.e. the density of the densest class in the ontology). The *average maximum P-density in ontologies that contain domain relations* is still low (1.1), meaning that, in most cases, classes may at most possess only 1 property, if any. Similar results are obtained for H-density (1.2 average maximum H-density in ontologies having sub-class relations).

Another straightforward conclusion here is that the amount of instance data is much bigger than the amount of ontological knowledge in the collected semantic documents. It is expected that the Semantic Web as a whole would be built on a similar ratio of classes, properties and individuals, requiring ontology based tools to handle large repositories of instances.

Topic Coverage. Understanding the topic coverage of the Semantic Web, i.e. how ontologies and semantic documents relate to generic topic domains like health or business, is of particular importance for the development of semantic applications. Indeed, even if it has already been demonstrated that the Semantic Web is rapidly growing [5], we cannot assume that this increase of the amount of online knowledge has been achieved in the same way for every application domain.

The Watson analysis task includes a mechanism that categorizes ontologies into the 16 top groups of DMOZ[6]. Each category is described by a set of weighted terms, corresponding to the name of its sub-categories in DMOZ. The weight $w(t) = \frac{1}{l(t)} \times \frac{1}{f(t)}$ of a term t is calculated using the level $l(t)$ of the corresponding sub-category in DMOZ and the number of times $f(t)$ the term is used as a sub-category name. In this way, a term would be considered as a good descriptor for the category (has a high weight) if it is high in the corresponding sub-hierarchy and if is is rarely used to describe other categories. The *level of coverage* of a given ontology to a given category then corresponds to the sum of the weight of the terms that match (using a simple lexical comparison) entities in the ontology.

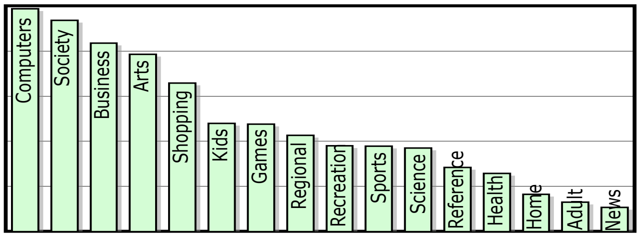


Fig. 3. Relative coverage of the 16 topics corresponding to the top categories of the DMOZ topic hierarchy.

This simple mechanism allows us to compute a rough overview of the relative coverage of these 16 high level topics on the Semantic Web. Among the semantic documents collected by Watson, almost 7 000 have been associated to one or several topics (have a non null level of coverage on some topics). Figure 3 describes the relative coverage of the 16 considered topics. In this figure, the y axis corresponds to the sum of the levels of coverage of all ontologies for the considered topic. The actual numbers here are not particularly significant, as we

[6] `http://dmoz.org/`

are more interested in the differences in the level of coverage for different topics. As expected, it can be seen that, while some topics are relatively well covered (e.g. computers, society, business), others are almost absent from the collected semantic documents (home, adult, news). Also, when comparing these results to the distribution of web documents within the DMOZ hierarchy, it is interesting to find that, according to this categorization, the coverage of these topics on the "classical Web" is also rather unbalanced (with categories varying from 31 294 to 1 107 135 documents), but that the order of the topics according to coverage is very different (computers for example is the 6^{th} category in coverage).

Finally, by looking at the level of coverage of each ontology, the *power law* distribution that has been found for other characteristics (size, expressiveness) also applies here: a few semantic documents have a high level of coverage, often with respect to several topics, whereas the large majority have a very low level of coverage, with respect to one or two topics only.

2.3 The Knowledge Network

While the Web can be seen as a network of documents connected by hyperlinks, the Semantic Web is a network of ontologies and semantic data. This aspect also needs to be analyzed, looking at the semantic relations linking semantic documents together.

Connectedness. Semantic documents and ontologies are connected through references to their respective namespaces. While the average number of references to external namespaces in the documents collected by WATSON seems surprisingly high (6.5), it is interesting to see that the most referenced namespaces are very often hosted under the same few domains (`w3.org`, `stanford.edu`, `ontoworld.org`, etc.)[7] This seems to indicate that a small number of large, dense "nodes" tend to provide the major part of the knowledge that is reused.

Another element of importance when considering the inter-connection between online semantic data is whether the URIs used to describe entities are *dereferenceable*, i.e., wether the namespaces to which they belong correspond to an actual location (a reachable URL) from which descriptions of the entities can be retrieved. Several applications, like Tabulator[8] or the *Semantic Web Client Library*[9] are indeed based on this assumption: that the Semantic Web can be *traversed* through dereferenceable URIs. However, among the semantic documents that explicitly declare their namespace, only about 30% correspond to actual locations of semantic documents, which means that these applications can only access a restricted part of the Semantic Web.

Redundancy. As in any large-scale distributed environment, redundancy is inevitable on the Semantic Web and actually contributes to its robustness: it

[7] It is important to remark here that the references to the namespaces of the representation languages, such as RDF and OWL, were not counted.

[8] `http://www.w3.org/2005/ajar/tab`

[9] `http://sites.wiwiss.fu-berlin.de/suhl/bizer/ng4j/semwebclient/`

is useful for an application to know that the semantic resources it uses can be retrieved from alternative locations in case the one it relies on becomes unreachable. As already mentioned, the 25 500 documents collected by WATSON are distinct, meaning that if the same file is discovered several times, it is only stored and analyzed once, even if WATSON would keep track of all its locations. On average, every semantic document collected by WATSON can be found in 1.27 locations, meaning that around 32 350 URLs actually address semantic data or ontologies. Ingnoring this simple phenomenon, like it is the case for example with the analysis described in [5], would have introduced an important bias in our analysis.

At a more fine-grained level, descriptions of entities can also be distributed and do not necessarily exist in a single file. Pieces of information about the same entity, identified by its URI, can be physically declared at different locations. Indeed, among the entities collected by WATSON, about 12% (approximately 150 000) are described in more than one place.

URI duplication. In theory, if two documents are identified by the same URI, they are supposed to contribute to the same ontology, i.e. the entities declared in these documents are intended to belong to the same conceptual model. This criterion is consistent with the distributed nature of the Semantic Web in the sense that ontologies can be physically distributed among several files, on different servers. However, even if this situation appears rarely (only 60 URIs of documents are "non unique"), in most cases, semantic documents that are identified by the same URI are not intended to be considered together. We can distinguish different situations leading to this problem:

Default URI of the ontology editor. `http://a.com/ontology` is the URI of 20 documents that do not seem to have any relation with each other, and that are certainly not meant to be considered together in the same ontology. The reason for this URI to be so popular is that it is the default namespace attributed to ontologies edited using (some of the versions) of the OWL Plugin of the Protégé editor[10]. Systematically asking the ontology developer to give an identifier to the edited ontology, like it is done for example in the SWOOP editor[11], could avoid this problem.

Mistaken use of well known namespaces. The second most commonly shared URI in the WATSON repository is `http://www.w3.org/2002/07/owl`, which is the URI of the OWL schema. The namespaces of RDF, RDF Schema, and of other well known vocabularies are also often duplicated. Using these namespaces as URIs for ontologies is (in most cases) a mistake that could be avoided by checking, prior to giving an identifier to an ontology, if this identifier has already been used in another ontology.

Different versions of the same ontology. A third common reason for which different semantic documents share the same URI is in situations where an ontology evolves to a new version, keeping the same URI (e.g., `http://lsdis.cs.uga.edu/proj/semdis/testbed/`). As it is the same ontology, it

[10] `http://protege.stanford.edu/`

[11] `http://www.mindswap.org/2004/SWOOP/`

seems natural to keep the same URI, but in practice, this can cause problems in these cases where different versions co-exist and are used at the same time. This leads to a need for recommendations of good practices on the identification of ontologies, that would take into account the evolution of the ontologies, while keeping different versions clearly separated.

3 Conclusion

The main motivation behind Watson is that the Semantic Web requires efficient infrastructures and access mechanisms to support the development of a new kind of applications, able to exploit dynamically the knowledge available online. We believe that a better understanding of the current practices concerning the fundamental characteristics of the Semantic Web is required. In this paper, we have reported on the analysis of the 25 500 distinct semantic documents collected by Watson, giving an account about the way semantic technologies are used to publish knowledge on the Web, about the characteristics of the published knowledge, and about some of the networked aspects of the Semantic Web. Our main conclusions are 1- that the Semantic Web is characterized by a large number of small, lightweight ontologies and a small number of large-scale, large-coverage and heavyweight ontologies, and 2- that important efforts still need to be spent on improving published ontologies (coverage of different domains, connectedness of the semantic data, etc.) and the tools that produce and manipulate them.

Many other aspects and elements could have been analyzed, and the research work presented here can be seen as a first step towards a more complete characterization of the Semantic Web. In particular, we only considered the characterization of the *current* state of the Semantic Web, analyzing a *snapshot* of the online semantic documents that represent the Watson repository. In the future, we plan to also consider the dynamics of the Semantic Web, looking at how the considered characteristics evolve over time.

References

1. H. Alani, C. Brewster, and N. Shadbolt. Ranking Ontologies with AKTiveRank. In *Proc. of the International Semantic Web Conference, ISWC*, 2006.
2. S. Bechhofer and R. Volz. Patching Syntax in OWL ontologies. In *Proc. of International Semantic Web Conference, ISWC*, 2004.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.
4. M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. Watson: A Gateway for the Semantic Web. In *Proc. of European Semantic Web Conference, ESWC, Poster Session*, 2007.
5. L. Ding and T. Finin. Characterizing the Semantic Web on the Web. In *Proc. of International Semantic Web Conference, ISWC*, 2006.
6. T. D. Wang, B. Parsia, and J. Hendler. A Survey of the Web Ontology Landscape. In *Proc. of the International Semantic Web Conference, ISWC*, 2006.