A Corpus of OWL DL Ontologies

Nicolas Matentzoglu, Samantha Bail, and Bijan Parsia

School of Computer Science, University of Manchester, Manchester, UK
{matentzn,bails,bparsia}@cs.man.ac.uk

Abstract. Tool development for and empirical experimentation in OWL ontology engineering require a wide variety of suitable ontologies as input for testing and evaluation purposes. Empirical activities often resort to (somewhat arbitrarily) hand curated corpora available on the web, such as the NCBO BioPortal and the TONES Repository, or manually select a set of well-known ontologies. Results may be biased, even heavily, towards these datasets. Sampling from a large corpus of ontologies, on the other hand, may lead to more representative results. Current large scale repositories/web crawls are mostly uncurated, suffer from duplication and contain large numbers of ontology versions, variants, and facets, and therefore do not lend themselves to random sampling. In this paper, we describe the creation of a corpus of OWL DL ontologies using strategies such as web crawling, various forms of de-duplications and manual cleaning, which allows random sampling of ontologies for a variety of empirical applications.

Keywords: Ontology Engineering, empirical methods, corpus, OWL

1 Introduction

Since its introduction as an official W3C recommendation in 2004, the Web Ontology Language OWL [8] has become by far the most popular and prominent ontology language. An increasing amount of tool support for the ontology development cycle and a general shift from establishing the formal underpinnings of a now mature formalism to the development of applications calls for improved empirical evaluation methods. The empirical tradition in the wider area of ontology engineering is still rather young and underdeveloped, and faces a series of obstacles, such as a lack of both standard evaluation frameworks and standard corpora to produce comparable empirical results. Current empirical evaluations, such as OWL reasoner benchmarking, the utility of modularisation approaches, and studies on entailment related services such as explanations, often cherry-pick a few example ontologies or resort to hand crafted ontology repositories such as the NCBO BioPortal [5]. Alternatively, crawlers such as Swoogle [19] have collected huge amounts of semantic documents, contributing a lot to our understanding of the use of semantic web languages, and allowing us to catch a glimpse of the impact that OWL has on the web ontology landscape. While these datasets are certainly useful for many purposes, they do not necessarily

lend themselves to ontology research as they collect OWL *files* which may not individually correspond to distinct OWL *ontologies*.

In this paper we describe the challenges of gathering a large and *interesting*—in terms of ontology size and expressivity—corpus of OWL DL ontologies suitable for experimental purposes. We describe the gathering and cleaning process and characterise the corpus based on ontology metrics such as axiom and constructor usage, OWL profiles, and provenance data. Additionally, we review existing sources for experimental datasets and compare our corpus with such frequently used OWL corpora. While the gathering and curation procedures are still considered work in progress, the work presented in this paper brings us one step closer to sound empirical OWL research.

2 Preliminaries

2.1 The Web Ontology Language OWL

OWL is a successor of the web ontology language DAML+OIL [25], a description logic based ontology language with an RDF/XML syntax. The first version of OWL, which is based on the description logic $\mathcal{SHOIN}(D)$ and was described as a 'revision' of DAML+OIL, became an official W3C recommendation in February 2004.

OWL 2, the successor of OWL, comprises two species of different expressivities, namely OWL 2 DL and OWL 2 Full [8]. The underlying formalism of OWL 2 DL is the description logic $\mathcal{SROIQ}(D)$ [26]. While OWL 2 DL has the familiar description logic semantics, OWL 2 Full has an RDF-based semantics, which is a superset of the of the OWL 2 Direct Semantics.

OWL 2 profiles. There exist three named 'profiles' for OWL 2, syntactic subsets of OWL 2 DL that are tailored towards different applications, trading expressivity of the language for efficient reasoning. The *OWL 2 EL* profile is a tractable fragment of OWL 2 which is based on the description logic $\mathcal{EL}^{++}$ [13,11]. *OWL 2 QL* (Query Language), which is based on the *DL-Lite* family of description logics [12], has been defined for use in applications which focus on query answering over large amounts of instance data. Reasoning systems for ontologies in the *OWL 2 RL* (Rule Language) profile can be implemented using rule-based reasoning engines. All three profiles are essentially syntactic restrictions of OWL 2 DL and allow for polynomial time reasoning.

2.2 Datasets used in practice

Single ontologies typically used for empirical work. There are a number of popular OWL ontologies commonly used for empirical research which we can roughly separate into 'in-use' ontologies, that is, ontologies built for use in a real-life application, and 'test' ontologies (for want of a better word) built for OWL research and tutorial purposes. Examples of such in-use ontologies include SNOMED CT, the GALEN Medical Knowledge Base, and the Gene Ontology

(GO). These ontologies are mostly large ontologies from the biomedical domain which are in active development and contain up to hundreds and thousands of classes and thousands of axioms, which makes them challenging candidates for OWL experiments. A special case of well-known ontologies is the National Cancer Institute thesaurus (NCIt) [6], a large medical ontology which has been in active development since 2003. The NCIt developers regularly release versions of the ontology, and an archive of over 100 versions is available online, which makes it a suitable collection for experiments studying the evolution of ontologies [22]. Frequently used test ontologies include the Pizza tutorial ontology, the Koala ontology, and the Lehigh University Benchmark ontology (LUBM). These ontologies are often built to exhibit specific features, such as coverage of a large number of OWL constructs as in the Pizza ontology, or the ability to scale the ontology to generate arbitrarily sized ABoxes, as is the case with the LUBM ontology. Many studies in the field of ontology debugging, for example, use small, hand-selected sets of test ontologies and in-use ontologies, such as Koala, MiniTambis, and Sweet-JPL, to evaluate the performance of their respective debugging techniques [28,32,37,20], or slightly more complex in-use ontologies [14,39,38], such as GALEN and GO. Large in-use ontologies, along with artificially scalable ontologies such as LUBM, are popular candidates for studies evaluating the performance of OWL reasoning algorithms, e.g. [36,15,31]. In cases where the authors justify their selection of ontologies, the use of 'established ontologies' seems a key argument [15]. For instance, Horrocks et al. [27] describe their benchmarking suite consisting of in-use ontologies including GO, GALEN, and NCIt as a set of 'standard test ontologies'. Similarly, Bock et al. argue [16] that their set of reasoner benchmarking ontologies was chosen because the ontologies are 'well established' and 'have been used in previous benchmarks', but go one step further by emphasising that the ontologies are representative (in terms of size and complexity) of the ontologies found in the Watson search engine, and therefore of the web ontology landscape in general.

Curated ontology repositories. There exists a number of well-known ontology repositories which are frequently used for empirical experimentation. In what follows, we will briefly describe some of the most prominent repositories and their applications in OWL research.

The *NCBO BioPortal* is an open repository of biomedical ontologies that provides access to ontologies from a variety of research groups [35]. As of April 2013, the repository contains 341 ontologies in various ontology formats. Due to its ontologies ranging widely in size and complexity, the BioPortal has become a popular corpus for testing OWL ontology applications in recent years, such as justification computation [24], reasoner benchmarking [29], and pattern analysis [33].

The *TONES* repository is a curated ontology repository which was developed as part of the TONES project as a means of gathering suitable ontologies for testing OWL applications. It contains 219 OWL and OBO ontologies and includes both well-known test ontologies and in-use ontologies which vary strongly

in size and complexity. While ontologies are occasionally added to the repository, it can be considered rather static in comparison with frequently updated repositories, such as BioPortal. The TONES ontologies are frequently used for empirical studies, either as a whole [29,40], by (semi-)randomly sampling from the set [30], or as a source of some of the individual ontologies mentioned above.

Similar to the TONES repository, the *Oxford ontology library* [9] is a collection of OWL ontologies gathered for the purpose of testing OWL tools. The library which was established in late 2012 currently contains 787 ontologies from 24 different sources, including an existing test corpus and several well-known in-use and test ontologies.

The *Protégé ontology library* [10] is a submission-based collection of ontologies linking to 95 OWL ontologies including some well-known test and in-use ontologies. While it is not used as frequently as the TONES repository (e.g. [41]), it fulfils a similar purpose of offering a selection of ontologies from a variety of domains.

Large-scale crawl based repositories. Crawl-based collections containing thousands and millions of files are popular sources of ontologies used in experiments.

Swoogle [19] is a crawl-based *semantic web search engine* that was established in 2004. The crawler searches for documents of specific filetypes (e.g. .rdf, .owl), verifies their status as a valid document of that type, and uses heuristics based on the references found in existing files to discover new documents. At the time of writing, Swoogle indexes nearly two million documents, and a search for ontologies (i.e. documents which contain at least one defined class or property) that match 'hasFiletype:owl' returns 88,712 results. While Swoogle is an obvious choice for gathering a large number of OWL ontologies for use in empirical studies [41,34,40], it does not have a public API and prevents result scraping, which makes it difficult to gain access to all search results. Furthermore, since the content is not filtered beyond removal of duplicate URLs, a random sample from Swoogle is most likely to return a set of small, inexpressive ontologies, or may be heavily biased towards ontologies from certain domains, as we will discuss in detail in Section 4.

Similar to Swoogle, *Watson* [17] is a search engine which indexes documents based on a web-crawler that targets semantic web documents. Watson uses filtering criteria in order to only include parseable documents and rank results according to their semantic *richness*, which is based on properties such as the expressivity of an ontology and the density of its class definitions. In addition to its web interface, Watson also provides APIs which allow users to retrieve lists of search results for a given keyword. At the time of its release, Watson was indexing around 25,500 documents; however, to the best of our knowledge, the service is no longer under active development.

The *Billion Triple Challenge* (BTC) dataset is an annually updated large dataset of RDF/XML documents used in the Semantic Web Challenge [2]. The 2011 set which contains 7.411 million RDF/XML documents crawled from the

web using various well-known Linked Data applications as seeds, such as DBPedia and Freebase. According to an analysis by Glimm et al. [21], the set contains just over 115,000 documents that contain a the rdfs:subClassOf predicate, which may be considered sufficient to class the document as an ontology. However, the authors identified that the corpus is biased towards several large clusters of documents from the same domain, which is indicated by the relatively small number of domains (109) that these potential ontologies originate from.

3 Gathering a corpus of OWL DL ontologies

While hand curated repositories often lack the potential for generalisability of claims, large-scale document collections suffer from a different problem: they typically contain many small and trivial OWL files as well as large numbers of duplicates, which means that a (naive) random sample is likely to introduce a heavy bias towards irrelevant cases for applications such as reasoner benchmarking and ontology profiling. If we want to make claims about *OWL ontologies on the web*, we need a way to obtain a set of *unique* non-trivial ontologies. In this section, we present our approach to addressing this issue by collecting a large amount of documents through web crawling and applying a series of filtering procedures. The focus of our work lies on the filtering steps applied to arrive at a set of (relatively) unique files with a high density of non-trivial OWL ontologies.

3.1 Data collection

The initial set of documents was collected using a standard web-crawler with a large seed list of URLs obtained from existing repositories and previous crawls. The sample obtained for this survey is preliminary in the sense that it is the result of only three weeks of downloading and crawling. We expect the results to improve gradually as the crawler collects more data, which also allows us to refine our heuristics for identifying OWL ontologies.

The seeds for the crawl were identified by a mixed strategy:

- Obtain seeds (336,414 direct links to potential ontology files) directly from a Google search, Swoogle, the OBO foundry, Dumontier Labs, and the Protégé Library, with the by far biggest amount of input coming from Swoogle. One thing to note is that we are not using the Swoogle cache, since we are interested in ontologies that are 'alive' and available on the web.
- Match URLs in a corpus obtained from a previous crawl in 2011 (43,006 URLs).
- Obtain data from APIs (currently only 413 seeds from BioPortal).

At the time of writing, the seeds for the crawler are spread across 3,413 domains (125 top level domains such as edu, com, org).

The crawler is based on crawler4j [3], a multi-threaded web crawler implemented in Java. We use a standard crawling strategy: broad and deep seeding, low crawling depth (3 levels) and looking for files with various extensions, for

example owl, owl2, rdf, rdfs, obo, owx, and variations thereof. Additionally, the crawler tests whether a link it followed might actually be an OWL file by using a set of syntactic heuristics (for example the OWL namespace declaration in all its syntactic variants), thus catching those OWL files that do not have a file extension (or a non-standard one). The crawler only identifies potential URLs, which are then passed on to a candidate downloader which is configured to download and re-download files in certain intervals. By preserving all the files obtained in this way, we hope to be able to analyse the evolution of the corpus at a later stage. In the short period of time that the crawler was running, 68,060 new candidate documents were discovered.

3.2 Data curation

Identifying valid OWL DL files. Many surveys on documents on the web acknowledge the necessity of doing some preprocessing to mitigate the large bias introduced by imperfect seeding strategies or single ontologies distributed across large numbers of multiple files. Our pipeline for identifying candidate OWL files from the crawl is as follows:

1. Loading and parsing files can be computationally expensive, especially when dealing with large file sizes. We applied syntactic heuristics to filter out documents which were clearly not OWL by looking for strings that regularly occur in different serialisations of the language. That way, we reduced our initial dataset from 268,944 files to 231,839. A random statistically significant sample of 1,037 from the files that were identified as non-OWL revealed that this filtering step yielded around 11% false negatives.
2. The next step was the removal of byte identical files. We used Apache Commons IO [1] to determine file stream identity. 43,515 files were grouped into clusters of byte identical files, and removed from the corpus.
3. The next process was to load and save all unique files with the OWL API [23], a widely used library for manipulating and serialising ontologies. Relatively few files (4,590) were not loadable at all (throwing exceptions), while 31 did not terminate loading. After this step, the corpus contained 213,462 valid OWL files.
4. We then narrowed down the corpus further by excluding files that have a byte identical OWL/XML serialisation.
5. The result of the curation pipeline to this point is a set of 207,230 unique and valid OWL files.

File based manual cleaning. One of the main difficulties of gathering a corpus of *ontologies* rather than a corpus of arbitrary OWL files is the problem of identifying what exactly constitutes a single ontology. This results from the non-standard ways of publishing ontologies: (1) There may exist several different versions of an ontology. These can be either subsequent versions which have been released in sequence (e.g. version 1.0, 1.1, ...), or slightly modified *variants*, such as 'light' or 'full'. (2) Single ontologies may be distributed over multiple files (e.g. DBPedia, Semantic Media Wikis) or published in modules contained in

individual files (faceted publishing). In order to identify versions and distributed ontologies, we applied a manual cleaning strategy: a random sample of 100 ontologies was drawn from the corpus by one of the authors, and grouped by filesize and file name patterns in order to identify clusters of files. If an identified cluster contained only trivial files (such as pages of a Semantic Media Wiki or proofs from Inference Web), all files belonging to the cluster (based on the domain and filename pattern) were removed from the corpus. This process was repeated until the sample appeared heterogeneous enough. In this process, the sample was reduced to just above 19,000 files.

Domain based manual cleaning. The file based manual cleaning worked well for weeding out the most prominent clusters of trivial files. Looking at the distribution of domains in our sample, we decided to go one step further and inspect clusters that were not captured easily by this method. We grouped the files by the domain source and inspected the biggest ones manually to remove files that have no or only trivial usage of the OWL language (e.g. owl:sameAs). Some large contributor domains were eliminated almost entirely (productontology.com), others required more careful attention (sweet.jpl.nasa.gov for example provides subsequent versions of each ontology, we decided to keep the latest ones). While this seems a lot of manual work, without this step the corpus would have been heavily biased towards the fairly inexpressive representatives of the big clusters. The biggest clusters were contributed by Semantic Media Wiki pages (146,866 files), Inference Web [4] metadata (19,042), and the New York Times 'subject headings' dataset [7] (10,438).

3.3 Corpus description

Having applied the filtering steps described above, the filtered corpus of OWL ontologies obtained from the crawl contained 9,871 files of which 9,827 files could be loaded.[1] Out of these, 208 were empty (either no axioms, or no entities in the signature, including annotation properties) and another 3,207 fell under RDF(S), and a further 1,865 were not in the OWL 2 DL profile, i.e. were OWL Full. Note that one of the causes for an ontology being in OWL Full is missing entity declarations which we consider to be a minor error; thus, before checking the profile of the ontologies, we 'repaired' this error by inserting missing entity declarations. The final set of valid and non-empty OWL DL ontologies consists of 4,547 documents. In this section, we will describe some of the properties of this corpus, including provenance information, axiom and constructor usage, and OWL profiles. The full set of metadata as well as the corpus are available online.[2]

Provenance. We count 728 distinct domains providing ontology files for the final set, spread across 52 top level domains. The distribution of top level

[1] Since the ontologies were *not* merged with their imports closure at the time of downloading, some ontologies failed loading due to missing imports only months later.

[2] `http://owl.cs.manchester.ac.uk/owlcorpus`

Table 1: Mean, median, min, max, and total numbers of axioms and entities for the root ontologies ($\mathcal{O}_r$) and their imports closure ($\mathcal{O}$).

	Mean		Median		Min		Max		Sum	
	$\mathcal{O}_r$	$\mathcal{O}$	$\mathcal{O}_r$	$\mathcal{O}$	$\mathcal{O}_r$	$\mathcal{O}$	$\mathcal{O}_r$	$\mathcal{O}$	$\mathcal{O}_r$	$\mathcal{O}$
Anno. prop.	9		4		0		135		40,299	
Axioms	12,261		194		2		3,337,397		55,750,395	
Logical axioms	3,789		69		1		740,559		17,229,616	
ABox axioms	1,552	1,621	0	1	0	0	739,274	739,274	7,055,335	7,370,895
TBox axioms	2,224	2,581	39	79	0	0	652,361	65,2361	10,110,790	11,735,859
RBox axioms	15	26	0	0	0	0	7,136	7,169	67,352	116,556
Signature	1,653	1,874	65	86	1	1	604,939	604,939	7,517,520	8,522,566
Classes	1,122	1,320	18	27	0	0	518,196	518,196	5,099,366	6,001,695
Object prop.	26	44	5	8	0	0	4,900	4,951	115,911	197,784
Data prop.	11	14	0	1	0	0	2,453	2,501	50,590	63,665
Individuals	484	484	1	1	0	0	604,209	604,209	2,199,854	2,201,124
Datatypes	3	3	2	2	0	0	51	51	11,500	115,48

domains is very similar to the one that was determined by a study characterising the semantic web in 2006 [18]. '.org' contributes almost half of the relevant documents (1,917), followed by '.com' (536) and '.edu' (450).

Regarding the formats the ontologies were published in, we inspected the file extensions and the syntax used in the corpus: in terms of syntax used, the majority of ontologies were originally serialised in RDF/XML (4,170), 255 ontologies were distributed as OBO 1.2 flat files, 82 in Turtle syntax, 21 in OWL/XML, and 19 in OWL Functional Syntax. The file extensions used were .owl (3,119), .rdf (706), .obo (230), and other (492).

Entities and axioms. Table 1 shows an overview of the entity and axiom statistics in the corpus for the root ontologies ($\mathcal{O}_r$) and their imports closure ($\mathcal{O}$). The most notable observation to be made is the relatively low median for all relevant entity statistics. This can be explained by the fact that there are still a good amount of small and potentially redundant files in corpus.

The ratio of TBox and ABox axioms in the ontologies (including imports closure) is shown in Table 3. Almost half of the ontologies in the corpus are 'schema only' and contain only TBox axioms, while over 11% are 'data only' and contain only ABox axioms. On average, the TBox makes up 72.2% of the axioms in an ontology.

Profiles and constructors used. Figure 1.b shows the distribution of constructors used in the ontologies in the corpus (as returned by the OWL API). We can see that beyond the basic constructors available in the description logic $\mathcal{AL}$, property-based constructors such as inverse properties and property hierarchies are the most prevalent across the corpus. On the other hand, only a very small number of ontologies make use of qualified number restrictions and complex property hierarchies, which might be explained by the fact that they were only introduced with OWL 2.

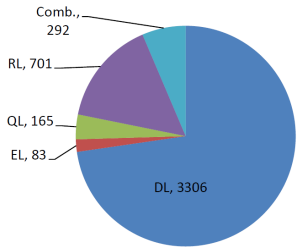

Fig. 1.a: Distribution of profiles. 'Comb': any combination of RL, QL, and EL.

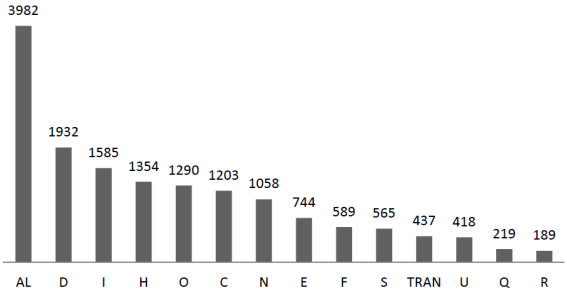

Fig. 1.b: Distribution of constructors ordered by number of ontologies that use the constructor.

Table 2: Distribution of profiles across size bins. 'Profiled': total number of ontologies in some profile.

Ontology size	EL	QL	RL	Profiled
<10	9.9%	12.0%	52.8%	57.7%
10-100	4.2%	6.6%	19.5%	23.5%
101-1000	4.6%	6.7%	9.1%	15.3%
1001-10000	5.9%	12.1%	8.5%	21.3%
10001-100000	4.9%	15.4%	7.4%	22.2%
100001-1000000	5.3%	2.6%	7.9%	15.8%

Table 3: Proportion of TBox axioms in the corpus.

TBox size	% of corpus
100%	47.7%
75 – 99%	15.3%
56 – 74%	9.5%
46 – 55%	3.2%
26 – 45%	4.8%
1 – 25%	8.2%
0%	11.3%

Figure 1.a shows the overall distribution of the OWL 2 profiles across the corpus; a more detailed view of the profile distribution across bins of different ontology sizes is shown in Table 2. We can see that around three quarters of the DL ontologies in the corpus are not in any of the three OWL 2 profiles, while the majority of profiled ontologies are in the RL profile. The distribution of profiles does not differ significantly across the different ontology sizes, although overall the number of ontologies in *some* profile is higher for very small (less than 10 axioms) ontologies. This is not surprising, as the use of fewer axioms also limits the potential use of non-profile constructors.

3.4 Corpus comparison

In order to put our collection in context with existing corpora of OWL ontologies, we have compared the basic metrics against commonly used sets. For space reasons, datasets included in this study were mostly selected based on their popularity as test corpora; thus, some of the more recent collections (e.g. the Oxford ontology library) and less prevalent sets (e.g. the Protégé library) were excluded for the time being. The BioPortal and TONES snapshots used here are from November 2012. The BioPortal dataset does not reflect the entire repository

but only those OWL and OBO files that could be downloaded through the REST API. Out of the 219 ontologies currently in TONES, only 205 were loadable (mainly due to missing imports). The third dataset is a sample from a Swoogle snapshot from May 2012 containing OWL and SKOS ontologies. We drew a statistically significant random sample (99% confidence, confidence interval 3) of 1,839 files from the Swoogle snapshot, out of which 1,757 could be loaded by the OWL API. The last dataset under consideration here is the version archive of the NCI thesaurus, consisting of 106 subsequent versions.

An overview of the metrics of the different datasets (ontologies including imports closure) is shown in Table 4. Overall, the entity numbers of the crawl corpus lie between the Swoogle sample and the manually curated collections. However, we also see that mean numbers of entities (logical axioms, signature size) of the two crawl-based collections (Swoogle and our crawl corpus) are significantly smaller than those of the manually curated repositories, BioPortal and TONES. With respect to the ontology profiles, both BioPortal and TONES contain fairly high numbers of ontologies in the EL profile (over 50% and 28%, respectively); this may be due to the presence of large biomedical ontologies which are restricted to the EL profile in order to guarantee tractable reasoning. On the other hand, both repositories also contain a large number of OWL 2 Full ontologies, which generally cannot be handled by current OWL reasoners.

Finally, Figures 2.a and 2.b show how the five datasets compare in terms of ontology sizes. Both crawl-based collections contain a larger proportion of small ontologies (with over 50% of Swoogle's ontologies containing less than 10 axioms), while the manually curated repositories TONES and BioPortal contain a high proportion of large ontologies.

Table 4: Comparison of ontology metrics of the five OWL datasets.

	Crawl	BioPortal	TONES	Swoogle	NCIt
Number of documents	4,547	293	205	1,757	106
Axioms and Entities					
Logical axioms (mean)	3,789	28,050	10,109	60	119,277
Logical axioms (max)	740,559	1,163,895	1,100,724	5,098	195,967
Signature (mean)	1,874	12,657	6,871	82	77,595
Signature (max)	604,939	847,761	524,117	5,118	123,301
Classes (mean)	1,320	11,534	5,861	16	66,449
Classes (max)	518,196	847,760	524,039	5,104	95,701
Profile density					
EL	5.6%	50.7%	28.3%	17.6%	0.0%
QL	8.5%	33.9%	18.1%	57.6%	4.7%
RL	21.2%	24.7%	12.7%	90.3%	0.0%
OWL 2 DL ontologies	100.0%	82.9%	77.6%	96.7%	100.0%

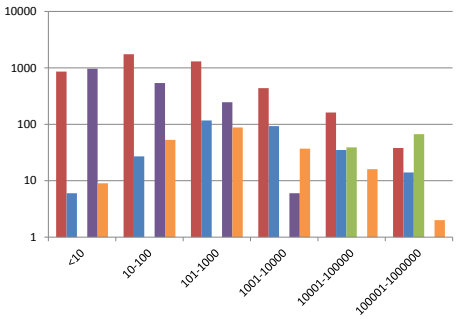

Fig. 2.a: Absolute numbers of ontologies in size bins across the five collections.

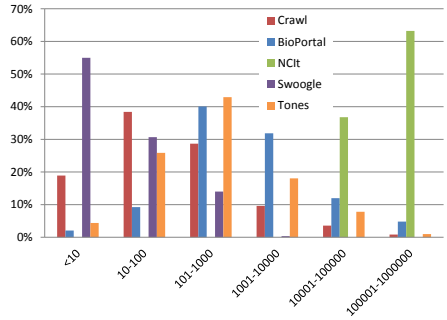

Fig. 2.b: Proportions of ontologies in size bins across the five collections.

4 Conclusions and future work

In this paper, we have presented an approach to creating a corpus of OWL DL ontologies for testing and evaluation purposes. While the actual corpus is still work in progress, we have already assembled an interesting—in terms of size, logical axiom counts, OWL constructs, and OWL 2 profiles—set of OWL DL ontologies. Apart from laying the foundations for an ontology repository, the main focus of this paper was to present an approach to gathering a reasonably interesting set of ontologies in a principled way.

While we believe that we have laid the foundations for a corpus of interesting ontologies, we acknowledge some of the limitations our current collection strategy: (i) resource limitations (essentially memory allocated to the Java Virtual Machine) might have caused some very big ontologies to have been dismissed; (ii) the standard problem of reaching the Hidden or Deep Web; (iii) our manual curation steps are not easily repeatable; and (iv) problems with imports resolution. The main limitations of our approach stem from general problems with web crawling, since it is unlikely that we will manage to index *all* OWL ontologies that are reachable on the web. However, we expect that a stronger focus on meta crawling (i.e. crawling search engines) and more extensive (manual) repository reviewing will gradually expand our seed. Replacing the manual filtering steps with automated ones is a major concern for the future.

In terms of future directions, we focus on two main aspects. Firstly, we want to find out more about the ontologies in our and other corpora. We are currently investigating ways to use ontology features to describe corpora in order to determine how representative a corpus is for the general population of ontologies of a certain type. The second goal is to establish a repository of OWL ontologies that allows researchers to retrieve specific samples of ontologies for various empirical tasks. One common problem for ontology researchers is the retrieval of a set of ontologies of a particular characteristic, for example *'a set of 100 large ontologies in* $\mathcal{EL}^{++}$*'*. We plan to provide the infrastructure that makes it possible to retrieve datasets that can also be made permanently accessible to other researchers, thus aiding the reproducibility of empirical experimentation.

References

1. Apache Commons - Commons IO. http://commons.apache.org/io/. Accessed: 2013-04-17.
2. Billion Triple Challenge - dataset. http://challenge.semanticweb.org/. Accessed: 2013-04-16.
3. crawler4j - open source web crawler for Java. http://code.google.com/p/crawler4j/. Accessed: 2013-04-17.
4. Inference Web - a knowledge provenance infrastructure. http://inference-web.org/wiki/Main_Page. Accessed: 2013-04-16.
5. NCBO BioPortal - home. http://bioportal.bioontology.org/. Accessed: 2013-04-17.
6. NCIt - National Cancer Institute thesaurus. http://ncit.nci.nih.gov/. Accessed: 2013-04-16.
7. New York Times - linked open data. http://data.nytimes.com/. Accessed: 2013-04-16.
8. OWL 2 Recommendation. http://www.w3.org/TR/owl2-overview/. Accessed: 2013-04-17.
9. Oxford Ontology Repository. http://www.cs.ox.ac.uk/isg/ontologies/. Accessed: 2013-04-16.
10. Protege Ontology Library. http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library. Accessed: 2013-04-16.
11. OWL 2 profiles. http://www.w3.org/TR/owl2-profiles/, 2009. Accessed: 2013-04-16.
12. A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *J. of Artificial Intelligence Research*, 36:1–69, 2009.
13. F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope. In *Proc. of IJCAI-05*, pages 364–369, 2005.
14. F. Baader, R. Peñaloza, and B. Suntisrivaraporn. Pinpointing in the description logic $\mathcal{EL}^+$. In *Proc. of KI'07*, pages 52–67, 2007.
15. M. Babik and L. Hluchy. A testing framework for OWL-DL reasoning. In *Proc. of SKG-08*, 2008.
16. J. Bock, P. Haase, Q. Ji, and R. Volz. Benchmarking OWL reasoners. In *Proc. of AREA-08*, 2008.
17. M. d'Aquin, C. Baldassarre, L. Gridinoc, M. Sabou, S. Angeletou, and E. Motta. Watson: supporting next generation semantic web applications. In *Proc. of WWW-07*, 2007.
18. L. Ding and T. Finin. Characterizing the semantic web on the web. In *Proc. of ISWC-06*, 2006.
19. L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proc. of CIKM-04*, 2004.
20. J. Du and G. Qi. Decomposition-based optimization for debugging of inconsistent OWL DL ontologies. In *Proc. of KSEM-10*, pages 88–100, 2010.
21. B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres. OWL: yet to arrive on the web of data? In *Proc. of LDOW 2012*, 2012.
22. R. S. Gonçalves, B. Parsia, and U. Sattler. Concept-based semantic difference in expressive description logics. In *Proc. of ISWC-12*, 2012.
23. M. Horridge and S. Bechhofer. The OWL API: a Java API for OWL ontologies. *Semantic Web J.*, 2(1):11–21, Jan. 2011.

24. M. Horridge, B. Parsia, and U. Sattler. Extracting justifications from BioPortal ontologies. In *Proc. of ISWC-12*, 2012.
25. I. Horrocks. DAML+OIL: a description logic for the semantic web. *IEEE Data Engineering Bulletin*, 25(1):4–9, 2002.
26. I. Horrocks, O. Kutz, and U. Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of KR-06*, 2006.
27. I. Horrocks, B. Motik, and Z. Wang. The HermiT OWL reasoner. In *Proc. ORE 2012*, 2012.
28. A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler. Debugging unsatisfiable classes in OWL ontologies. *J. of Web Semantics*, 3(4):268–293, 2005.
29. Y.-B. Kang, Y.-F. Li, and S. Krishnaswamy. Predicting reasoning performance using ontology metrics. In *Proc. of ISWC-12*, pages 198–214, 2012.
30. C. M. Keet. Detecting and revising flaws in OWL object property expressions. In *Proc. of EKAW 2012*, 2012.
31. D. C. Martinez, A. Krisnadhi, F. Maier, K. Sengupta, and P. Hitzler. Reconciling OWL and rules. Technical report, Kno.e.sis Center, Wright State University, 2011.
32. T. Meyer, K. Lee, R. Booth, and J. Z. Pan. Finding maximally satisfiable terminologies for the description logic ALC. In *Proc. of AAAI-06*, pages 269–274, 2006.
33. E. Mikroyannidi, N. A. A. Manaf, L. Iannone, and R. Stevens. Analysing syntactic regularities in ontologies. In *Proc. of OWLED-12*, 2012.
34. T. A. T. Nguyen, R. Power, P. Piwek, and S. Williams. Measuring the understandability of deduction rules for OWL. In *Proc. of WoDOOM-12*, 2012.
35. N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, May 2009.
36. J. Z. Pan and E. Thomas. Approximating OWL-DL ontologies. In *Proc. of AAAI-07*, pages 1434–1439, 2007.
37. G. Qi and A. Hunter. Measuring incoherence in description logic-based ontologies. In *Proc. of ISWC/ASWC-07*, pages 381–394, 2007.
38. K. Shchekotykhin, G. Friedrich, and D. Jannach. On computing minimal conflicts for ontology debugging. In *Proc. of MBS-08*, 2008.
39. B. Suntisrivaraporn, G. Qi, Q. Ji, and P. Haase. A modularization-based approach to finding all justifications for OWL DL entailments. In *Proc. of ASWC-08*, pages 1–15, 2008.
40. A. Third. Hidden semantics: what can we learn from the names in an ontology? In *Proc. of INLG*, pages 67–75, 2012.
41. T. D. Wang, B. Parsia, and J. Hendler. A survey of the web ontology landscape. In *Proc. of ISWC-06*, pages 682–694, 2006.