A Snapshot of the OWL Web

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 and detailed characterisations of real ontologies. 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 selected sets of well-known ontologies. Findings of surveys and results of benchmarking activities 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 and web crawls are mostly uncurated and suffer from duplication, small and (for many purposes) uninteresting ontology files, and contain large numbers of ontology versions, variants, and facets, and therefore do not lend themselves to random sampling. In this paper, we survey ontologies as they exist on the web and 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 standardisation by the W3C in 2004, the Web Ontology Language OWL[1] has become a widely used language for representing ontological knowledge. OWL ontologies are used across a wide spectrum of domains, ranging from chemistry to bio-health informatics and medical data. There exists an increasing amount of tool support for OWL ontologies, such as OWL reasoners, ontology editors, ontology browsers and visualisation tools, as well as numerous approaches to tasks such as ontology mapping, debugging, and modularisation. Testing and evaluation of proposed techniques and tools form an important part of the development process, and while there are some tools that are specifically tailored towards certain ontologies (such as, for example, the Snorocket reasoner[2] which is aimed at classifying the SNOMED CT ontology [15]), most tools are aimed at general OWL ontologies. One of the core decisions required for a sound empirical

[1] http://www.w3.org/TR/owl2-overview/
[2] http://protegewiki.stanford.edu/wiki/Snorocket

methodology is the selection of a suitable dataset and some clarity about that choice and its implications. In particular, choice of data set can threaten both the *internal* validity (i.e. whether a found correlation indicates a causal relation) and *external* validity (i.e. the extent to which the result can be generalised).

Current empirical evaluations, such as OWL reasoner benchmarking and studies on the effectiveness of various debugging techniques, often cherry-pick a few example ontologies or sample from ontology repositories such as the NCBO BioPortal.[3] Alternatively, crawlers such as Swoogle [6] 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 crawl-based 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 characterise the landscape of OWL ontologies found on the web, with a focus on using *collections* of ontologies for OWL tool development and evaluation purposes. We describe the challenges of gathering a large and meaningful corpus of OWL DL ontologies that is suitable for such experimental tasks, which is based on an automated web crawl combined with several filtering steps to identify OWL ontologies based on heuristics, and to take into account duplicates, versions, and facets of ontologies. We discuss the characteristics of the corpus, such as axiom and constructor usage, OWL profiles, and provenance data, and compare it to several other collections of OWL ontologies that are frequently used (or designed for) testing purposes. The purpose of this paper is twofold: first, we provide insights into the landscape of OWL ontologies found on the web nearly a decade after OWL became an official standard. Second, we highlight the issues faced when selecting suitable test corpora in the OWL tool development process and aim to support tool developers in making informed decisions when choosing test collections.

2 Preliminaries and background

In this section, we will give a very brief introduction to the web ontology language OWL 2 and the OWL 2 profiles. We then discuss the use of OWL ontology collections in empirical evaluations.

2.1 The Web Ontology Language OWL

OWL 2 [3], the latest revision of the Web Ontology Language OWL, comprises two species of different expressivities, namely OWL 2 DL and OWL 2 Full. The underlying formalism of OWL 2 DL is the description logic $\mathcal{SROIQ}(D)$ [10]. While OWL 2 DL has the familiar description logic semantics (*Direct Semantics*), OWL 2 Full[4] has an RDF-based semantics, which is a superset of the OWL

[3] `http://bioportal.bioontology.org/`

[4] `http://www.w3.org/TR/owl2-rdf-based-semantics/`

2 Direct Semantics; OWL reasoners, however, are restricted to ontologies in (a subset of) OWL DL.

There exist three named 'profiles' for OWL 2, which are 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 $\mathcal{EL}^{++}$ [2]. *OWL 2 QL* (Query Language) which is based on the *DL-Lite* family of description logics [1], 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.

2.2 Datasets used in practice

A wide range of empirical ontology research requires access to a somehow 'interesting' set of ontologies as input to experiments. Empirical studies involving OWL tools and techniques frequently make use of existing ontologies and ontology repositories for test and evaluation purposes. In order to put our work into context with empirical OWL research, we will give an overview of some of the curated OWL ontology repositories and large-scale collections that are commonly used for empirical evaluations.

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 which invites submissions from OWL researchers. As of April 2013, the repository contains 341 ontologies in various ontology formats including the full set of OBO Foundry[5] ontologies. Due to its ontologies ranging widely in size and complexity, BioPortal has become a popular corpus for testing OWL ontology applications in recent years, such as justification computation [9], reasoner benchmarking [11], and pattern analysis [13].

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, varying 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 [11,16], by (semi-)randomly sampling from the set [12], or as a source of individual ontologies.

Similar to the TONES repository, the ***Oxford ontology library***[6] is a collection of OWL ontologies gathered for the purpose of testing OWL tools. The

[5] `http://www.obofoundry.org/`

[6] `http://www.cs.ox.ac.uk/isg/ontologies/`

library, which was established in late 2012, currently contains 793 ontologies from 24 different sources, including an existing test corpus and several well-known in-use and test ontologies.

The ***Protégé ontology library***[7] 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. [17]), 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. While the two largest collections, Swoogle and Watson, seem to be no longer under active development, the Billion Triple Challenge dataset is still updated annually.

Swoogle [6] 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. In April 2013, Swoogle indexed 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' returned 88,712 results. While Swoogle is an obvious choice for gathering a large number of OWL ontologies for use in empirical studies (e.g. [17,14,16]), it does not have a public API and prevents result scraping in order to reduce server load, 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 the Section 3.2.

Similar to Swoogle, ***Watson*** [4] 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 valid (that is, parseable) documents and ranks 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.[8] 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. [7], the set contains just over 115,000 documents that contain a the rdfs:subClassOf predicate, which

[7] `http://protegewiki.stanford.edu/wiki/Protege\_Ontology\_Library`
[8] `http://challenge.semanticweb.org/`

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* ontologies (at least to some degree). For our corpus, we consider ontologies that are in OWL DL (and not mere RDFS), contain some logical content and are parseable by the OWL API. 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 with a a high density of (relatively) unique OWL DL ontologies. Table 1 shows an overview of the individual steps in the data curation procedure and the numbers of files filtered out in each step.

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 as follows:

- 336,414 URLs of potential ontology files obtained directly from a Google search, Swoogle, OBO foundry, Dumontier Labs,[9] and the Protégé Library.
- 43,006 URLs obtained from an experimental crawl in 2011.
- 413 ontologies downloaded from the BioPortal REST API.

The crawler is based on crawler4j,[10] a multi-threaded web crawler implemented in Java. We use a standard crawling strategy (broad and deep seeding, low crawling depth, i.e. 3 levels), searching for files 'typical' extensions, e.g. owl, rdf, obo, owl.xml, and variations of the type owl.txt, owl.zip, etc. Additionally, the crawler tests whether a link it followed might actually be an OWL file by using

[9] `http://dumontierlab.com/`
[10] `http://code.google.com/p/crawler4j/`

a set of syntactic heuristics (e.g. 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 that attempts to download files in certain intervals. In the short period of time that the crawler was running, 68,060 new candidate documents where discovered. A large number of candidates in the seeds were not retrievable, due to, amongst others, unreachable domains or possibly restrictions for crawler access.

3.2 Data curation

Identifying valid OWL files Many surveys of documents on the web acknowledge the necessity of preprocessing crawl results in order to remove irrelevant and duplicate files. Our pipeline for identifying candidate OWL files from the files gathered by the crawler is as follows:

1. Attempting to load and parse files with the OWL API can be computationally expensive, especially for non-OWL files for which the API tries out every possible parser before failing, and for large OWL files. Thus, we applied syntactic heuristics to filter out documents
   - that were clearly not OWL (less than six lines of text, first fifty lines contain the `<html>` tag),
   - or did not contain any OWL declaration (in any syntax) or OBO format version in the first sixty lines.

   This step reduced the initial dataset from 268,944 files to 231,839. A random (statistically significant) sample of 1,037 files that we attempted to load with the OWL API revealed that approximately 11% of the thus removed files were falsely identified as not being OWL.
2. The next step was the removal of byte-identical files. We used Apache Commons IO[11] to determine file stream identity. 43,515 files were grouped into clusters of byte-identical files.
3. Next, all remaining unique files were loaded and saved with the OWL API [8]. Relatively few files (4,590) where not loadable due to parser errors, while 31 did not terminate loading in practical time. After this step, the corpus contained 213,462 valid OWL files.
4. We then removed further duplicates by excluding 6,142 files that have a byte-identical OWL/XML serialisation. The result of the curation pipeline to this point is a set of 207,230 unique (in terms of byte-identical duplicates) and valid OWL files.

Note that we consider the loss of ontologies which cannot be parsed by the OWL API to be negligible, since this API is the most comprehensive of its kind, covering most types of OWL syntaxes and all OWL 2 constructs.

Cluster detection One of the main difficulties of gathering a corpus of *ontologies* rather than a corpus of arbitrary OWL files is the problem of identifying

[11] `http://commons.apache.org/io/`

Table 1: Summary of the curation pipeline.

Document state	Removed	Size after
Retrieved		268,933
Passed heuristic	37,094	231,839
Passed OWL API, de-duplicated (byte identity)	18,377	213,462
De-duplicated (byte identity after common serialisation)	6,142	207,320
Systematically manually filtered	197,449	9,871
Non-OWL 2 DL and empty ontologies filtered	5,324	4,547

what exactly constitutes a single ontology. This results from the different non-standard ways of publishing ontologies:

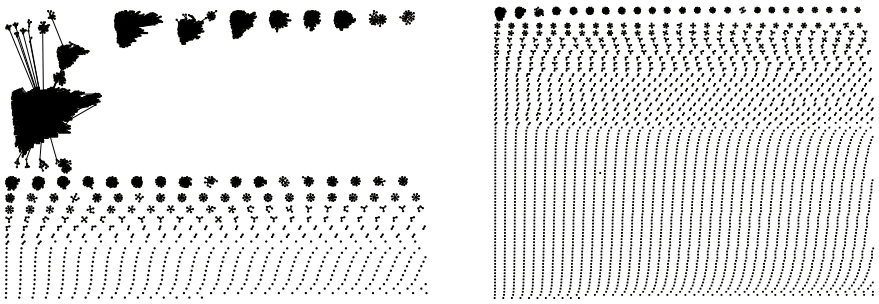
1. There may exist several different *versions* of an ontology. These can be either subsequent versions which have been released in sequence (e.g. version 1.0, 1.1, ...), or slightly modified *variants*, such as 'light' or 'full'.
2. Single ontologies may be distributed over multiple files (e.g. DBPedia, Semantic Media Wikis) or published in modules contained in individual files (*faceted* publishing). The individual files are often very small and describe only trivial fragments of larger OWL ontologies.

In order to identify clusters of versions, variants, and distributed ontologies, we applied two filtering steps based on similar file sizes and file names, and based on the source of the OWL file.

File name and file size patterns First, a random sample of 100 ontologies was repeatedly drawn from the corpus, and grouped by file size and file name patterns in order to identify clusters of files. If an identified cluster contained large groups of very similar 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 file name pattern) were removed from the corpus. This process was repeated until a random sample of 100 ontologies appeared heterogeneous enough, i.e. did not contain large numbers of files with obviously similar file names and sizes. In this process, the sample was reduced from 207,230 to just above 19,000 files, which is a reduction by more than 90%.

Domain names The file based cluster detection worked well for weeding out the most prominent clusters of distributed ontologies. We then grouped the remaining files by the domain source and inspected the biggest clusters of domains manually to remove files that have no usage (including mere usage of 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, of which we decided to keep the latest ones).

The largest clusters identified in the cluster detection stage were data generated by various Semantic Media Wikis (146,866), files containing formulas, rules, and related metadata from the Inference Web (19,042). Other notable clusters were generated by the New York Times subject headings SKOS vocab-



(a) Before manual cluster removal. (b) After manual cluster removal.

Fig. 1: Similarity graphs before (sample) and after manual cluster removal.

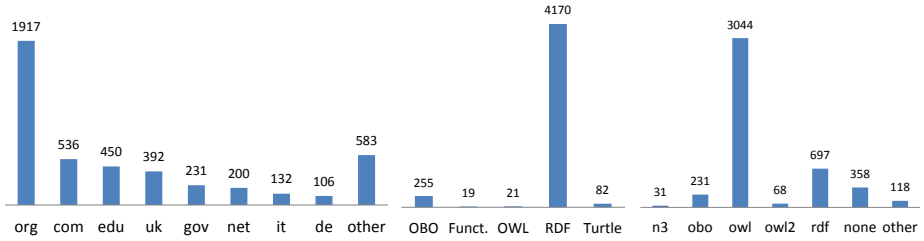
ulary (10,438), the UniProt Protein Knowledge Base (5,580 RDF files), as well as files describing instance data of the well-known Friend of a Friend (FOAF) vocabulary (2,312). In total, the clustering process removed over 50% of the ontologies in the crawl set, reducing the corpus to 9,871 files which we presumed to be largely cluster-free.

In order to illustrate the effects of the manual cluster removal, Figure 1 shows two graphs describing the (pairwise) similarity between the ontologies in the corpus before the clustering (on a random sample of 4,547 out of 207,230 ontologies) in Figure 1a and after the clustering (the final 4,547 ontologies, as described in the next section) in Figure 1b. Our notion of similarity is described in section 4.5. We can see that the degree of similarity within the corpus before the filtering is significantly higher than after the cluster removal, with 2,815 connected components before the cluster removal, compared to 521 in the final corpus. Also, the amount of ontologies with no or only few similarity relations is considerably lower after the cluster removal (higher degree of uniqueness).

OWL DL filtering Having applied the filtering steps described above, the remaining corpus of OWL ontologies obtained from the crawl contained 9,871 files of which 9,827 files could be loaded.[12] Out of these, 208 were empty (either no axioms, or no entities in the signature, including annotation properties) and 3,207 fell under RDF(S). A further 1,865 ontologies were not in the OWL 2 DL profile for reasons other than missing declarations. We consider missing declarations to be minor violations and thus decided to simply inject them to ensure a more meaningful profile membership (an ontology with a missing class declaration should still be in DL if it was in DL without it).

Apart from missing class- (77.4%), annotation- (67.8%), object property- (34.4%) and data property declarations (15.1%), the main reason for the remaining 1,865 ontologies not falling into OWL DL was the use of reserved vocabulary, most prominently for class IRIs, which occurred in 62.5% of the ontologies and

[12] Since the ontologies were *not* merged with their imports closure at the time of downloading, some ontologies failed loading due to missing imports during the analysis.



(a) Top level domains. (b) Syntaxes. (c) File extensions.

Fig. 2: Provenance data of the ontologies in the crawl corpus.

for object properties in 16.5%. Further, a number of ontologies (up to 5%) suffered from various other invalid IRI problems, such as using an IRI for both a datatype and a class (5.2%) and using non-absolute IRIs (2.8%). The remaining issues were caused by the use of non-simple properties in cardinality restrictions. All these violations—which affected almost one fifth of the 9,827 valid OWL files we gathered—cause common OWL DL reasoners to either reject or only process parts of the ontologies.

3.3 Provenance data

Domain sources In the final corpus of 4,547 valid and non-empty OWL files, we count 728 distinct domains (an average of 6.5 ontologies per domain), spread across 52 top level domains. The distribution of top level domains is very similar to the one determined by a Swoogle study characterising the semantic web in 2006 [5]. As Figure 2a shows, '.org' contributes almost half of the documents (42%), followed by '.com' (12%) and '.edu' (10%).

File extensions and syntax Figures 2b and 2c show an overview of the OWL syntaxes and file extensions used for the published OWL files. We can see that the vast majority of ontologies were originally serialised in RDF/XML (4,170), while only a fraction (less than 1%) were published as OWL/XML files. The most frequent file extension used was .owl (67% of the files), followed by .rdf (15%) and .obo (5%). Interestingly, it appears that only a single file in the corpus had the extension .owx, the recommended extension for OWL/XML serialisations.[13]

4 Comparison of OWL collections

In order to put our crawl-based OWL corpus in context with existing collections of OWL ontologies, we compare its basic ontology metrics against four commonly used datasets. The datasets were selected based on their popularity and intended use as test corpora, as discussed in Section 2.2; thus, some of the less prevalent

[13] `http://www.w3.org/TR/owl2-xml-serialization/`

Table 2: Entity usage (average, median, maximum) in the five collections.

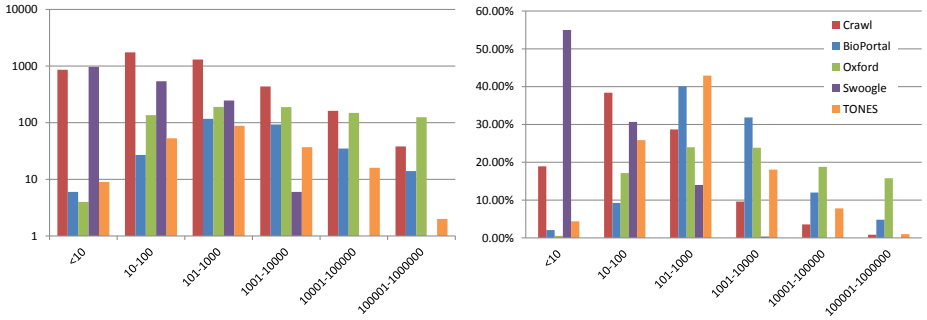
		Crawl	BioPortal	Oxford	Swoogle	TONES
Classes	avg	1,320	11,534	5,652	16	763
	med	27	470	209	9	138
	max	518,196	847,760	244,232	5,104	524,039
Object properties	avg	43	37	43	11	34
	med	8	7	10	15	8
	max	4,951	1,390	964	251	922
Data properties	avg	14	9	5	16	13
	med	1	0	0	18	0
	max	2,501	488	1,371	133	708
Individuals	avg	484	1,075	3,810	29	163
	med	1	0	0	15	0
	max	604,209	232,646	466,937	855	178,308
Logical axioms	avg	3,789	28,050	49,990	60	1,332
	med	69	958	729	8	256
	max	740,559	1,163,895	2,492,725	5,098	1,100,724

sets (e.g. the Protégé library) were excluded. The statistics are given here to allow a comparison between the collections, but no statement is made about which dataset is 'better', as this obviously depends heavily on the purpose. Importantly, the collections in this section are largely left untouched and are *not* curated in the way the Web Crawl was: they may even contain OWL Full and RDFS. The only criterion for inclusion apart from availability was parseability by the OWL API.

The BioPortal and TONES snapshots are from November 2012 and include those OWL and OBO files that could be downloaded and loaded by the OWL API. Files that could be retrieved, but not parsed, usually suffered from unresolvable 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, of which 1,757 could be loaded. The last collection is a snapshot of the Oxford ontology library from April 2013. The final sets were: Crawl (4,547), BioPortal (292), Oxford (793), Swoogle sample (1,757) and TONES (205). For the reasons discussed in section 3.2, missing entity declarations were injected prior to metrics gathering in all cases.

4.1 Entity usage

Classes, properties, individuals Table 2 shows a detailed overview of the average, median, and maximum values of the relevant logical entities occurring in the five collections (minimum numbers were 0 in all collections, thus they are not listed in the table). Swoogle clearly stands out as a collection with comparatively small numbers of entities per ontology. On average, both the BioPortal and Oxford collections contain very large numbers of logical axioms and classes, with the Oxford collection also containing several ontologies that are particularly heavy on individuals. In comparison, the crawl corpus contains ontologies with on average significantly fewer classes than the curated repositories.



(a) Absolute distribution.

(b) Relative distributions.

Fig. 3: Distribution of ontology sizes, binned by number of logical axioms.

Logical axioms In addition to the logical axiom counts given in Table 2, Figures 3a and 3b show a comparison of the ontology sizes in the five collections, sorted into six size bins ranging from less than 10 to over 100,000 logical axioms. We can see that the majority of ontologies in the crawl-based collections (Crawl and Swoogle) are in the lower two bins of fairly small ontologies (less than 100 axioms), whereas the other three collections roughly follow a normal distribution (given this particular binning). On closer inspection we find that the Swoogle snapshot still contains a large number of trivial files from the Semantic Media Wiki, which significantly adds to the number of small ontologies. In the case of the Oxford library and TONES this is likely to be due to the editors explicitly selecting a range of 'interesting' (i.e. medium to large) ontologies.

4.2 Constructors and axiom types

Constructors Figure 4 shows a comparison of the constructor usage in the five collections (as returned by the OWL API). In the crawl corpus, we can see that beyond the basic constructors in $\mathcal{AL}$ (intersection, universal restrictions, existential restrictions of the type $\exists r.\top$, and atomic negation) which are used by the majority (88%) of ontologies in the crawl, property-based constructors, such as inverse properties $\mathcal{I}$ (35% of ontologies) and property hierarchies $\mathcal{H}$ (30%), are the most prevalent across the crawl corpus. Perhaps surprisingly, full existential restriction (of the type $\exists r.C$ for a possibly complex expression C) are only used in 16% of the ontologies. Furthermore, only a very small number of ontologies make use of qualified number restrictions $\mathcal{Q}$ (5%) and complex property hierarchies $\mathcal{R}$ (4%), which might be explained by the fact that they were only introduced with OWL 2.

Regarding the other collections, the Swoogle snapshot only makes use of very few constructors, leaving out most of the more expressive ones. Looking at the axiom type usage in Table 3, this may be explained by the fact that the Swoogle snapshot contains mainly assertion axioms which, in this case, only contain atomic entities and no complex constructors. On the other end of the

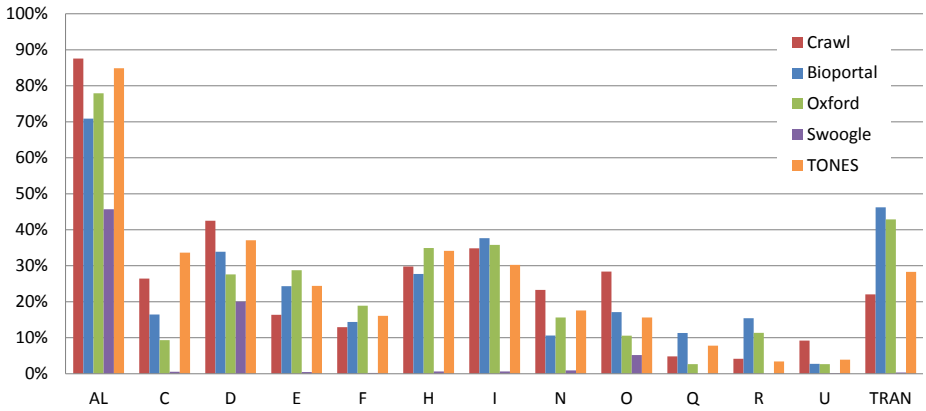


Fig. 4: Frequency of OWL constructor usage in the five collections. Bar height indicates the proportion of ontologies in a collection that use the constructor.

spectrum, the remaining collections contain similarly large numbers of ontologies using property-related constructors such as transitive properties (TRAN), inverse properties $\mathcal{I}$, and property hierarchies $\mathcal{H}$. While there is no general trend towards a 'most complex' collection, we can find high numbers of ontologies with transitive properties in the BioPortal and Oxford collections, whereas the crawl corpus contains a comparatively large number of ontologies with nominals $\mathcal{O}$ (28%) and unqualified number restrictions $\mathcal{N}$, and all three curated collections (BioPortal, Oxford, and TONES) contain (proportionally) more full existential restrictions $\mathcal{E}$ than the crawl. As with the crawl corpus, the least used constructors in all collections are qualified number restrictions $\mathcal{Q}$ and complex property hierarchies $\mathcal{R}$, along with the union ('or') operator, which occurs in less than 10% of ontologies in all collections.

Axiom types Table 3 shows an overview of the most frequent axiom types (in terms of total usage in all collections, not taking into account entity declarations).[14] We can see that by far the most frequently used axiom types in the crawl corpus are AnnotationAssertion and SubClassOf axioms. Domain and range axioms on object properties also occur in nearly half of the ontologies in the corpus; interestingly, their frequency is roughly pairwise identical across all collections, which may indicate that ontology developers generally add domain and range axioms together when introducing object properties. As we have already seen in the discussion on constructors, object property related axiom types such as subproperties, transitive and inverse properties occur frequently in between one fifth and nearly half of the ontologies in the different collections (with the exception of Swoogle). Class related axioms, such as DisjointClasses and EquivalentClasses, can be found equally often in the four collections. This shows

[14] Note that annotations were removed during the BioPortal download and serialisation process; thus, the corpus does not contain any AnnotationAssertion axioms.

Table 3: Axiom type usage as proportion of ontologies that use an axiom type.

	Crawl	BioPortal	Oxford	Swoogle	TONES
SubClassOf	77.0%	96.9%	79.4%	5.6%	92.2%
AnnotationAssertion	78.1%	-	88.9%	36.4%	68.3%
ClassAssertion	44.8%	30.0%	69.7%	35.5%	26.3%
ObjectPropertyRange	47.2%	36.5%	45.8%	1.0%	44.4%
ObjectPropertyDomain	45.6%	38.2%	44.1%	0.9%	43.4%
EquivalentClasses	36.8%	37.9%	44.6%	1.1%	45.4%
SubObjectPropertyOf	30.1%	40.6%	44.4%	0.6%	34.6%
TransitiveObjectProperty	22.0%	46.1%	42.9%	0.3%	28.3%
DisjointClasses	31.0%	42.3%	24.8%	0.3%	41.0%
InverseObjectProperties	31.1%	33.4%	32.8%	0.6%	27.3%
DataPropertyRange	31.5%	29.7%	15.1%	0.6%	27.3%
FunctionalObjectProperty	18.4%	29.4%	24.3%	0.2%	26.8%
DataPropertyDomain	29.6%	27.3%	13.7%	0.5%	22.9%
ObjectPropertyAssertion	17.2%	13.3%	18.5%	26.2%	12.2%
FunctionalDataProperty	15.2%	21.2%	5.4%	0.2%	20.0%
DataPropertyAssertion	13.0%	9.6%	10.8%	19.0%	6.8%

that, while the clear majority of axioms are fairly 'trivial' SubClassOf and ClassAssertion axioms, more complex axiom types occur frequently in these OWL ontologies.

4.3 Datatypes

Regarding the usage of datatypes, we found that a very small number of built-in datatypes occur frequently in the five collections, whereas the remaining types are only used rarely. The most frequently used datatypes are rdf:plainLiteral (between 25.9% in BioPortal[15] and 82.1% of the ontologies in the Oxford corpus) and xsd:string datatypes (between 26.8% in the Swoogle snapshot and 59.7% in our crawl corpus). In the Swoogle snapshot, the general datatype usage is lower than in the other collections, with a maximum of only 36.6% of ontologies using rdf:plainLiteral. Interestingly, however, the ontologies in the Swoogle snapshot make more frequent use of xsd:integer (25.3%), xsd:dateTime (25.1% of ontologies), and xsd:decimal (24.2%) than the other collections, which all range between only 1.5% and 10.7% for these types. Finally, across the collections, most other built-in datatypes occur in a small number of ontologies, with the exception of rdfs:literal, which can be found in 18% of the ontologies in the crawl corpus, and xsd:anyURI, which is used in over a third (37.8%) of the ontologies in the Oxford collection.

4.4 OWL profiles

As mentioned in Section 2, the OWL 2 profiles are relevant for OWL reasoners, which are only compatible with OWL DL ontologies, or may be tailored towards

[15] Due to the removal of annotations in the BioPortal download it is likely that the figures for the BioPortal collection are lower than they would be with annotations.

Table 4: OWL 2 profiles in the collections.

	Crawl	BioPortal	Oxford	Swoogle	TONES
Full	0.0%	17.1%	15.3%	3.2%	22.4%
DL	100%	82.9%	84.7%	96.8%	77.6%
EL only	1.8%	16.4%	3.0%	0.0%	9.8%
EL total	4.0%	29.4%	3.2%	0.6%	19.0%
QL only	3.6%	0.7%	0.9%	0.1%	0.0%
QL total	4.8%	33.2%	9.2%	57.5%	18.0%
RL only	15.4%	1.0%	18.0%	33.3%	2.0%
RL total	19.6%	22.9%	27.1%	90.3%	11.7%
DL only	72.7%	29.8%	53.6%	5.8%	46.8%

Table 5: Overlap comparison between the collections.

Corpus 1	Corpus 2	Sim.	Con.
Crawl	BioPortal	10.9%	17.6%
Crawl	TONES	11.8%	19.1%
TONES	BioPortal	11.9%	17.9%
Swoogle	Oxford	12.5%	17.3%
Swoogle	BioPortal	14.2%	15.9%
Swoogle	TONES	15.0%	17.1%
Crawl	Swoogle	15.1%	36.3%
Oxford	BioPortal	16.7%	23.2%
Crawl	Oxford	16.8%	24.2%
TONES	Oxford	19.1%	27.0%

a specific subset of OWL 2 DL. Table 4 shows an overview of the profiles for the different ontologies. Note that the profiles are not exclusive, that is, an ontology in one profile may also be in the other profiles; thus, we distinguish between ontologies which are in one profile only, and the total proportion of ontologies in a profile (including other profiles). 'DL only' denotes the proportion of OWL DL ontologies that do not fall into any of the three sub-profiles.

Across the collections, the level of OWL DL ontologies is fairly high (minimum 77.6% in TONES), whereas the occurrence of ontologies in the OWL profiles varies strongly. We can see immediately that the majority of ontologies in the crawl corpus does not fall into any of the sub-profiles EL, QL, or RL, whereas the Swoogle ontologies are largely in a combination of the RL and QL profiles (due to them being fairly inexpressive), with only a fraction (5.8%) being more expressive. A comparatively large number of ontologies (16.4%) in BioPortal fall into the OWL 2 EL (only) profile, which is likely caused by the presence of many large bio-medical ontologies in the corpus that are explicitly designed to be in EL. On the other hand, there are almost no QL or RL only ontologies in BioPortal.

4.5 Overlap analysis

In order to determine the to which extent the different collections overlap (i.e. shared ontologies), we performed a pairwise comparison of the ontologies in each of the five collections based on two measurements: a) two ontologies are *similar* if the overlap (the intersection of the signatures divided by the union of the signatures) is at least 90%. b) There exists a *containment* relation between two ontologies $\mathcal{O}_1$, $\mathcal{O}_2$, if $sig(\mathcal{O}_1) \subseteq sig(\mathcal{O}_2)$ or $sig(\mathcal{O}_2) \subseteq (\mathcal{O}_1)$. As shown in Table 5, the pairwise similarity overlap (Sim.) between the collections ranges between 10.9% (crawl vs. BioPortal) and 19.1% (TONES vs. Oxford repository). The containment overlap (Con.) between the collections is significantly higher, ranging between 15.9% (Swoogle vs. BioPortal) and 36.3% for the containment relations between the crawl corpus and the Swoogle sample, which is likely to be caused by the heavy use of Swoogle results as seeds for the web crawler.

5 Conclusions and future work

In this paper, we have presented an overview of the OWL ontology landscape with a focus on the application of different collections for empirical evaluations. We presented an approach to creating a large yet interesting corpus of OWL DL ontologies suitable for testing and evaluation purposes, characterised the corpus, and compared it to other existing collections of OWL ontologies, such as the NCBO BioPortal and a random sample from the Swoogle search engine. We have seen that drawing a random sample from an unfiltered crawl-based collection may be representative for the general population of OWL files 'found on the web', however, it does not yield relevant data to be used for measuring, for example, reasoner performance on 'actual' ontologies. The direct comparison of these ontology metrics allows OWL tool developers to make an informed decision when selecting a suitable collection of OWL ontologies for testing purposes, while it also shows that a careful filtering procedure of a crawl-based corpus brings the resulting set closer to curated repositories in terms of ontology size and expressivity.

While we believe that we have laid the foundations for a large, crawl-based repository of ontologies for empirical evaluations, we acknowledge some of the limitations our current collection strategy suffers from:

1. Resource limitations (essentially memory allocated to the Java Virtual Machine) might have caused a few very big ontologies to have slipped through in the initial steps of the curation procedure.
2. Web crawlers may not reach the Hidden or Deep Web.
3. The manual curation steps are not easily repeatable.
4. Problems with unavailable ontology imports.

The main limitations of our approach stem from general problems with web crawling, since it is unlikely that we will be able 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. With the insights we have gained into general cluster characteristics, we aim to replace the manual filtering procedures by automated ones. The problem of unavailable ontology imports can be easily solved by downloading the imports closure of an ontology in the crawling and ontology validation process.

In addition to improving the crawling and validation strategies and the analysis of the actual content of the ontologies, we focus on establishing 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 OWL 2 EL ontologies with more than 100 axioms'*. We plan to provide an 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. A prototype of this repository can be found at `http://owl.cs.manchester.ac.uk/owlcorpus`.

Nicolas Matentzoglu is supported by a CDT grant by the UK EPSRC.

References

1. A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *J. of Artificial Intelligence Research*, 36:1–69, 2009.
2. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of IJCAI-05*, pages 364–369, 2005.
3. B. Cuenca Grau, I. Horrocks, B. Motik, B. Parsia, P. F. Patel-Schneider, and U. Sattler. OWL 2: The next step for OWL. *J. of Web Semantics*, 6:309—322, 2008.
4. 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.
5. L. Ding and T. Finin. Characterizing the semantic web on the web. In *Proc. of ISWC-06*. Springer Verlag, 2006.
6. 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.
7. B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres. OWL: yet to arrive on the web of data? In *Proc. of LDOW 2012*, 2012.
8. M. Horridge and S. Bechhofer. The OWL API: a Java API for OWL ontologies. *Semantic Web J.*, 2(1):11–21, Jan. 2011.
9. M. Horridge, B. Parsia, and U. Sattler. Extracting justifications from BioPortal ontologies. In *Proc. of ISWC-12*, 2012.
10. I. Horrocks, O. Kutz, and U. Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of KR-06*, 2006.
11. Y.-B. Kang, Y.-F. Li, and S. Krishnaswamy. Predicting reasoning performance using ontology metrics. In *Proc. of ISWC-12*, pages 198–214, 2012.
12. C. M. Keet. Detecting and revising flaws in OWL object property expressions. In *Proc. of EKAW 2012*, 2012.
13. E. Mikroyannidi, N. A. A. Manaf, L. Iannone, and R. Stevens. Analysing syntactic regularities in ontologies. In *Proc. of OWLED-12*, 2012.
14. 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.
15. K. A. Spackman. SNOMED RT and SNOMED CT. Promise of an int. clinical ontology. *M.D. Computing*, 17(6):29, 2000.
16. A. Third. Hidden semantics: what can we learn from the names in an ontology? In *Proc. of INLG*, pages 67–75, 2012.
17. T. D. Wang, B. Parsia, and J. Hendler. A survey of the web ontology landscape. In *Proc. of ISWC-06*, pages 682–694, 2006.