An Empirical Study on How the Distribution of Ontologies Affects Reasoning on the Web

Hamid R. Bazoobandi, Jacopo Urbani, Frank van Harmelen, and Henri Bal

Dept. of Computer Science, Vrije Universiteit Amsterdam, NL
h.bazoubandi@vu.nl
{jacopo,frank.van.harmelen,bal}@cs.vu.nl

Abstract. The Web of Data is an inherently distributed environment where ontologies are located in (physically) remote locations and are subject to constant changes. Reasoning is affected by these changes, but the extent and significance of this dependency is not well-studied yet. To address this problem, this paper presents an empirical study on how the distribution of ontological data on the Web affects the outcome of reasoning. We study (1) to what degree datasets depend on external ontologies and (2) to what extent the inclusion of additional ontological information via IRI de-referencing and the *owl:imports* directive to the input datasets leads to new derivations.
We based our study on many RDF datasets and on a large collection of RDFa, and JSON-LD data embedded into HTML pages. We used both Jena and Pellet in order to evaluate the results under different semantics. Our results indicate that remote ontologies are often crucial to obtain non-trivial derivations. Unfortunately, in many cases IRIs were broken and the *owl:imports* is rarely used. Furthermore, in some cases the inclusion of remote knowledge either did not yield any additional derivation or led to errors. Despite these cases, in general, we found that inclusion of additional ontologies via IRIs de-referencing and *owl:imports* directive is very effective for producing new derivations. This indicates that the two W3C standards for fetching remote ontologies have found their way into practice.

Keywords: RDF, RDFa, JSON-LD, OWL, Reasoning, Web Of Data

1 Introduction

The Web contains large volumes of semantically annotated data encoded in RDF [19] or similar formats. Often, this data contains expressive ontologies that machines can leverage to perform reasoning and derive valuable implicit information. Since information re-usage is a corner stone of the Semantic Web [17], many datasets reuse ontologies that are already available rather than creating their own ones. These ontologies are distributed across the Web and the W3C standardized two mechanisms to retrieve them: IRIs de-referencing [6] and *owl:imports* [17].

The number and correctness of new derivations that reasoners produce depend on the availability and quality of these external ontologies. Therefore, it is crucial that reasoners can successfully retrieve them and that the union of external ontologies is still consistent. Unfortunately, the Web is an inherently distributed and uncoordinated environment where several factors may preclude the fetching and reusage of remote data. For example, remote ontologies might silently disappear or move to other locations, or independent authors may publish ontologies that contain syntactic and/or semantic mistakes [13]. All these possibilities can heavily affect the output of reasoning or even make reasoning impossible.

Although much effort has already been invested on studying the quality and accessibility of resources on the Web of Data (WoD) [5, 6, 4], to the best of our knowledge no work has ever studied how the distribution of ontological data on the web affects reasoning. The goal of this paper is to study this from a purely empirical perspective. To that end, we conduct a number of experiments and analyse the output of reasoning over a wide range of documents to offer a first preliminary answer to the following questions: *a)* how many derivations can reasoners derive from individual documents? *b)* To what extent do documents link to external ontologies and how accessible are such links? *c)* How many new derivations can reasoners derive after external ontologies are included and how can we characterize such derivations? *d)* To what extent does the inclusion of additional ontological data endanger reasoning? This paper presents a number of experiments to answer these questions.

As the input for our experiments, we took samples from *LODLaundromat (LODL)* [4], which is a large crawl of RDF documents from the WoD, and *Web Data Commons (WDC)* [25], which contains extracted RDFa, MicroData, and JSON-LD graphs embedded in the HTML pages. We conducted our experiments using Jena [22] and Pellet [27], two widely used reasoners, and performed two types of analyses: a quantitative analysis, which focuses on the number of derived triples; and a qualitative analysis, which looks into the relevance of derived triples with the input document.

We summarise below some key outcomes of our experiments. These will be discussed in the remainder of this paper with more details:

- In the majority of the cases, reasoning on a single document produces a small number of derivations that are mostly RDF or OWL axioms;
- Only a small number of IRIs were de-referencable. However, when IRIs could have been accessed, the inclusion of additional knowledge allowed reasoners to derive new triples. This finding highlights the importance of maintaining functioning links in the WoD;
- The directive *owl:imports* is used only in a very small number of documents (less than 0.2% documents of LODL and only on 121 graphs out of 500M in WDC). In the documents that use it, the (recursive) inclusion of the remote ontologies led to a significant increase of the number of derived triples. This demonstrate the potential of this mechanism;
- In a non-negligible number of cases, the inclusion of remote ontologies did not lead to the derivation of new triples. Also, we observed cases where the inclusion of external ontologies led to semantic conflicts (ABox or TBox conflicts) that failed Pellet. Additionally, in some cases Jena did not finish reasoning within 72 hours (despite the fact that on average the number of statements in input was fairly small).

In general, our findings are encouraging because they indicate that remote knowledge (fetched either with IRI de-referencing or via *owl:imports*) does lead to new valuable derivations. However, we have also witnessed several problems which show that further research is still very much needed.

This paper is structured as follows: Sec. 2 reports on the experimental settings; Sec. 3 presents the results of the experiments where reasoning was applied without fetching remote ontologies, Sec. 4 presents the results after we de-referenced IRIs and Sec. 5 after we imported ontologies via *owl:imports*. Finally, Sec. 6 reports on related work while Sec. 7 concludes the paper. An extended version of this paper is available as technical report at `http://hbi250.ops.few.vu.nl/iswc2017/survey/iswc2017_tr.pdf`

2 Experimental setup

In coming sections, we run a number of experiments to analyze how the inclusion of external ontologies from the web affects the outcome of reasoning. However, Prior to that, in this section, we first provide some information about the setup of our experiments and define some terms that we are going to use in the rest of this paper.

Inputs On the Web, semantically-annotated data can be encoded either as RDF knowledge graphs (which are serialized in a number of files) or be embedded in HTML pages. Therefore, we considered two large collections of both types: *LODLaundromat* (LODL) [4] and the *Web Data Commons* (WDC) dataset [25]. LODL contains a collection of RDF files that were either crawled from online archives or submitted to the system. At the time we conducted this study, the collection consisted of 500K RDF files from more than 600 domain names. The WDC dataset contains RDFa, Microdata, Microformat, and JSON-LD data extracted from HTML pages. We use the 2015 crawl which provides about 541M named graphs from more than 2.7M domain names. We chose these two datasets because they are, as far as we know, the largest available collections of semantically-annotated data available on the Web. We must stress, however, that neither of these two collections offers any guarantee of representativeness. As far as we know, no crawled collection from the Web can make such claim. They simply represent the best approximation that we have available.

In this paper, we refer to sets of triples which are locally available as *documents*. For LODL, a document corresponds to a RDF file. For WDC, a document corresponds to the set of triples in a named-graph (the named graph is the URI of the webpage from which the triples were extracted). We refer to the number of triples contained in a document as its *size*. We used two reasoners, *Pellet* [27] (version 2.3.1) and the *OWLMiniReasoner* of Jena [22] (version 3.1.0), to evaluate reasoning under different computational logic. We use these two reasoners (instead of, for instance, more scalable solutions like RDFox [24], VLog [28] or WebPIE [29]) because they are well-tested implementations and work under different semantics. The *OWLMiniReasoner* reasoner in Jena works under the RDF semantics and supports an incomplete fragment of OWL Full that omits the forward entailments of *minCardinality*/*someValuesFrom* restrictions (detailed list of the supported constructs is available online[1]). In contrast, Pellet supports a sound but incomplete OWL DL reasoning (i.e., SROIQ(D)) [27] and we use it to

[1] `https://jena.apache.org/documentation/inference/#owl`

perform ABox DL reasoning[2]. Once again, we refer to the online documentation for a detailed list of the supported constructs. Each reasoner is launched with the default settings. The only modification is that we disabled the automatic *owl:imports* inclusion for both reasoners in all experiments.

We refer to the terms and axioms defined in the RDF [19], RDFS [7], OWL [17], and XSD [12] specifications as *standard terms* and *standard axioms* respectively. A standard predicate is a standard term that appears as predicate in a triple. We assume that standard terms and axioms are locally available (but not part of the input document) because in practice the reasoners have stored a local copy.

Reasoning In this paper, reasoning is used to derive new conclusions. The reasoning procedure is simple and equivalent for both reasoners: First, we load a set of triples G into the reasoner. We refer to the set G as the *input* of the reasoning process. In some experiments, G equals to a document while in others it will include also some remotely fetched triples. Then, we query the reasoner with the SPARQL query `SELECT ?s ?p ?o { ?s ?p ?o }`, which is meant to retrieve all the triples the reasoner can derive. Each answer returned by the reasoner is translated into a RDF triple $\langle ?s\ ?p\ ?o\rangle$. Let G' be the set of all returned triples. We call every triple $t \in G' \setminus G$ a *derived* triple and refer to the set $G' \setminus G$ as the *set of derived triples* or *derived triples*, or in short *derivations*. Clearly, this set will be different depending on the used reasoner. We would like to stress that the purpose of our experiments is *not* to compare the output of two reasoners but to analyse their output w.r.t. the inclusion/exclusion of remote ontological information.

Categorization of derivations In order to perform a more fine-grained analysis of the derived triples, we categorize them based on the complexity of reasoning process that produces them into the following four disjoint categories:

- **Type1** derivations are derivations that contain *only standard terms*. Typically, triples in this category are the tautologies extracted from these languages (e.g.,⟨*rdf:subject rdf:type rdf:Property*⟩).
- **Type2** derivations contain exactly one non-standard term that appears in one or more triples in the input set (e.g., ⟨*:resource rdf:type rdf:Resource*⟩).
- **Type3** derivations contain two non-standard terms that appear in the same input triple (e.g., if the input contains the triple ⟨*:ClassA owl:equivalentClass :ClassB*⟩ then a *Type3* derivation could be ⟨*:ClassB rdfs:subClassOf :ClassA*⟩).
- **Type4** derivations contain two or more non-standard terms that never appeared in the same input triple (e.g., if the input contains the triples ⟨*:resource rdf:type :ClassA*⟩ and ⟨ *:ClassA rdfs:subClassOf :ClassB*⟩ then a *Type4* triple could be ⟨ *:resource rdf:type :ClassB*⟩).

The reason behind such classification is that *Type1* derivations should be easy to return. *Type2* and *Type3* derivations are less easy because they require one pass on the data (*Type3* have the additional complexity that the reasoner might need to change the ordering of the terms). The derivation of *Type4* triples usually requires a join between multiple triples, and thus their derivation is computationally more demanding.

[2] TBox and ABox are terms from Description Logics. TBox triples encode 'schema' information which is crucial for reasoning while ABox triples encode assertional information.

Most non-trivial implicit knowledge that reasoners derive are usually of *Type2*, *Type3* or *Type4*.

Failures In some experiments, the reasoners were unable to complete the reasoning process. Causes for failure varied between a limited scalability of the algorithms/implementation, syntactic errors [5], and ontological inconsistencies [26]. Please note that the notion *ontological inconsistency* usually includes *unsatisfiability*, *incoherence*, or *inconsistency*. However, because reasoners do not crash as a result of *unsatisfiability* and *incoherence*, in this paper we ignore them, and whenever we use the term *ontological inconsistency* or in short *inconsistency*, we refer to conflicting assertions (ABox) or axioms (TBox) that make reasoning impossible and cause reasoner to abort the process.

A complete analysis of the failures is beyond the scope of this paper. Here, we say that a reasoner *failed* (or that a *failure* occurred) when the reasoner did not terminate successfully the reasoning process. We classify failures either as *exceptions*, which occurred when the reasoner had prematurely terminated (e.g., because of an inconsistency or a syntactic error in the input), or as *timeouts* in case the reasoner did not conclude the inference within 72 hours.

Computing infrastructure Many experiments required several hours to finish. To speed up the execution, we launched several of them in parallel using the DAS4 cluster[3]. Each machine in the cluster has 24G of memory and two quad-core 2.4 GHz CPUs.

Data and source code All data, source code to run the experiments, and all derived triples are available at `http://hbi250.ops.few.vu.nl/iswc2017/survey/`. We believe that publishing all results of our experiments is useful also because such results can be used as inputs for other studies.

3 Local Reasoning

First, we intend to evaluate how many new triples the reasoners can derive from local data. But what can be considered as "local" in the Web of Data? One possibility is to consider all the RDF datasets that are stored on the same website as local. Unfortunately, there are several repositories that contain datasets from several other locations. Another possibility is to assume that local data is stored in files that share the same prefix (e.g. *dbpedia-01.gz*, *dbpedia-02.gz*), but this is a rather weak heuristic which does not always hold in practice. For the LODL dataset, we eventually concluded that the best solution was to consider as "local" only the triples that are contained in a single document (i.e., a RDF file for LODL and a named graph for WDC), because documents are the minimal storage units that are always entirely available on the same physical location. Thus, we will compute how many new triples reasoners can derive from single documents.

Data Collection In our context, performing reasoning on every document is neither feasible nor desirable. The infeasibility is due to the large number of documents in LODL and WDC. We estimated that even if we could use all machines of our cluster it would take months to finish the computation. The undesirability comes from the fact that more than two thirds of the documents in LODL are fetched from two sources

[3] `http://www.cs.vu.nl/das4`

			#Documents per domain				#Triples per Documents			
	#Domains	#Documents	Max	Average	Median	Min	Max	Average	Median	Min
$\mathtt{LODL}_1$	510	673	7	1.3	1	1	5.2M	80.9K	70	1
$\mathtt{WDC}_1$	67K	74K	8	1.35	1	1	10.6K	20.5	6	1

Table 1. Statistics about the samples of $\mathtt{LODL}_1$ and $\mathtt{WDC}_1$ used for local inference.

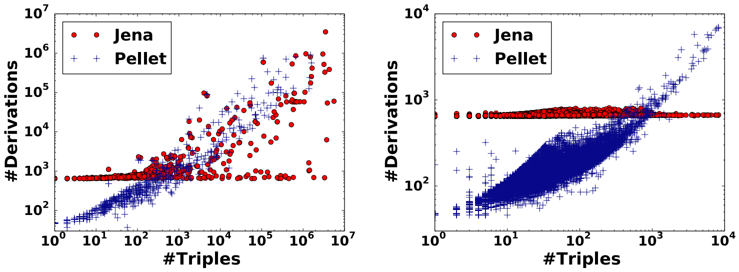


Fig. 1. Number of derived triples w.r.t document size on $\mathtt{LODL}_1$ (left) and $\mathtt{WDC}_1$ (right).

– *sonicbanana.cs.wright.edu* and *worldbank.270a.info* – while in the WDC dataset there is a significant difference between the number of documents from popular domain names such as *wordpress.com* and the ones from less popular sources.

With such large skew in terms of provenance, aggregations over the entire datasets will be strongly biased towards a few sources. While a simple random sampling strategy would be enough to reduce the input to a manageable size, it would be ineffective in removing the bias. To avoid this second problem, we first perform a random sampling over domain names with the sample size determined by the Cochram's formula [8] with a confidence level of 95%, and less than 0.5% margin of error. Then, from every selected domain name, we randomly picked as many documents as the logarithmic transformation of number of documents from that domain. This is a well-known methodology for sampling from skewed sources [32]. We call $\mathtt{LODL}_1$ the sample extracted from LODL, and $\mathtt{WDC}_1$ the sample extracted from WDC. Statistics about them are reported in Tab. 1.

One surprising number in Tab. 1 is the relatively small average number of triples in documents in $\mathtt{LODL}_1$. In fact, 673 documents are indeed only a small fraction of all documents in the LODL collection. Such aggressive reduction is due to the relatively low number of domains in the collection and the extreme skew of the distribution of files among them. With such input, we are forced to select only a few documents per domain, otherwise we would be unable to construct a sample without skew. We believe this is the fairest methodology in order to present results which are most representative (i.e., cover the largest number of sources). If the reader is interested in biased results, we report in the TR the results obtained with a larger randomly selected sample.

Reasoning Results We launched Pellet and Jena over both samples, and report in Fig. 1 the number of derivations in relation with the size of the documents. We can draw a few interesting considerations from these results: First, the number of Pellets' derivations is proportional to the size of the input documents. This occurs both with $\mathtt{LODL}_1$ and

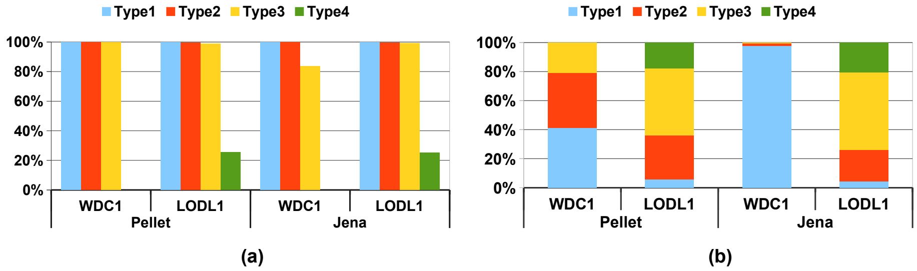


Fig. 2. (a) Percentage of documents that yielded derivations of each type. (b) Ratio of each derivation type w.r.t total number of derivations.

$\texttt{WDC}_1$. The number of the derivations produced by Jena was instead more constant. It only starts to grow proportionally with the largest $\texttt{LODL}_1$ documents.

Fig. 2a shows the percentage of documents that yielded triples in each of the four categories outlined in Sec. 2. We see that all documents led to *Type1* and *Type2* derivations, regardless of the reasoner used. We inspected samples of the triples in each category and found that most *Type1* triples are RDFS and OWL axioms, while most of the *Type2* derivations are triples that describe resources or predicates, e.g., both reasoners always derive that predicates are instances of *rdf:Property*. In general, almost all documents have also led to *Type3* derivations. The only exception was $\texttt{WDC}_1$ in combination with Jena since in this case almost 20% of the documents did not return any *Type3* derivation. We manually inspected a sample of *Type3* derivations and found that they resemble to *Type2* information in the sense that they also describe predicates and resources. For example, the statement “a property is a *rdfs:subPropertyOf* itself” is a *Type3* statement that both reasoners have frequently derived.

We observed that Jena did not derive any *Type3* triples if the document contained only standard predicates. Instead, Pellet frequently derived *Type3* triples that state that classes are equivalent/subclass of themselves. In contrast to Pellet, we noticed that Jena always returned about 600 *Type1* triples regardless of the actual input. This explains why the number of derivations tends to be constant for Jena in Fig. 1: It mainly consists of 600 *Type1* statements plus some *Type2* or *Type3* statements that describe resources and predicates. To explain this more clearly, we show in Fig 2b the ratio of each derivation type against total number of derivations in the samples. We see from the figure that Jena on $\texttt{WDC}_1$ only produces *Type1* derivations, thus the total number of derivations tends to remain constant for each document. The situation is different for $\texttt{LODL}_1$ where the sizes of documents vary considerably. There, a smaller number of all derivations is of *Type1* which indicates that Jena derives significantly more derivations of other types. Interestingly, we notice from Fig. 2a that all $\texttt{WDC}_1$ documents derive *Type2* triples and about 80% of them derive *Type3* derivations with Jena. However, in Fig. 2b we see that these triples are fewer than *Type1* triples. This means that each documents led to only few *Type2* and *Type3* statements while the largest number of derivations is of *Type1*.

Finally, we observed that neither reasoner was able to derive *Type4* triples from $\texttt{WDC}_1$, while for $\texttt{LODL}_1$ only 24% of the documents yielded such derivation. This sug-

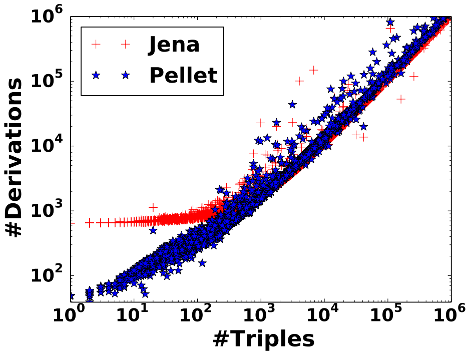


Fig. 3. Number of derivations vs document size in the *skewed* LODL sample

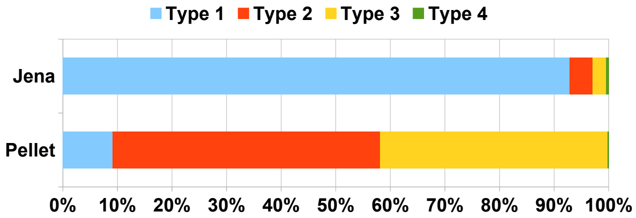


Fig. 4. Ratio of each derivation type w.r.t the total number of derivations in the *skewed* LODL sample

gests that in general most of the derivations that we can obtain from single documents are sort of "descriptions" of the terms in the dataset (e.g., a predicate is an instance of *rdfs:Property*, a class is an instance of *rdfs:Class*, etc.).

In order to provide more insight into how a skewed sample may affect our conclusions, here we compare the results of our experiments over non-skewed $\texttt{LODL}_1$ sample, with that of the skewed sample of LODL. Fig. 3 shows the number of derivations relative to the number of triples in the input document for the skewed LODL sample. By comparing this figure to the Fig. 1 for $\texttt{LODL}_1$ we observe that (especially for Jena results) the two graphs are different. Furthermore, we compare the ratio of derivation type w.r.t the total number of derivations in the skewed sample in Fig. 4. By comparing this figure to Fig. 2b for $\texttt{LODL}_1$ documents, we observe a significantly different view, i.e., while Jena derived mostly *Type1* information from the skewed LODL sample, in the non-skewed sample *Type1* information form only a small portion of all information the Jena derived. Furthermore, in the non-skewed sample, *Type4* information accounts for almost *20%* of all derivations of both reasoners, while in the skewed sample *Type4* derivations are almost non-existent. This considerable contrast between the results of skewed and non-skewed sample stems from the fact that skewed sample is populated with a large number of small documents with similar structure that belong to archives that Lodlaundromat crawled from *wright.edu*. These documents alone account for roughly *70%* of the whole skewed sample from LODL. Thus, it is not surprising that the results are biased toward the inference results of these documents. The biased conclusion of the skewed LODL sample shows the necessity for non-skewed sample to conclude unbiased results.

Failures *Type2* and *Type3* derivations should be easy to calculate since they can be typically derived with a single pass on the data. Unfortunately, we still witnessed a number of failures with both reasoners. These failures were rare in $\mathtt{WDC}_1$ (i.e., less than 0.1% for both reasoners, and all these cases were *exceptions* caused by syntax errors). With $\mathtt{LODL}_1$, Jena successfully finished for more than 99.9% of the input documents. When it did not, *timeout* was the primary cause of failure. With Pellet we witnessed a higher percentage of failures (about 12% of the inputs). In more than 72.5% of these cases, Pellet threw an *exception*, while the rest of the cases the reasoner *timed out*. Interestingly, more than 92% of exceptions were raised by inconsistencies while the rest were raised due to other internal reasons (e.g. *Unknown concept type exception*).

4 IRI De-referencing

We will now present the results of some experiments to investigate whether the inclusion of additional remote content obtained by de-referencing IRIs in the documents leads to more derivations. To this end, we considered all documents of $\mathtt{LODL}_1$ and $\mathtt{WDC}_1$ for which local reasoning succeeded. Given the low failure rate, these samples are roughly equivalent to the original $\mathtt{LODL}_1$ and $\mathtt{WDC}_1$ datasets. In this section, we refer to these two subsets as $\mathtt{LODL}_2$ and $\mathtt{WDC}_2$ respectively.

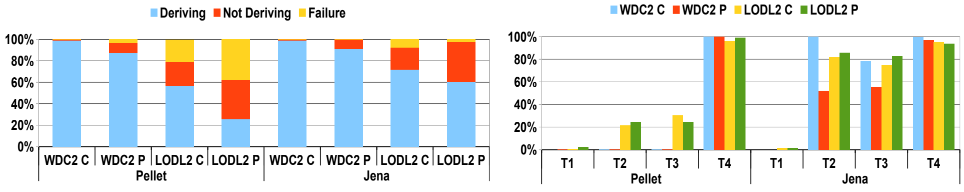
Unfortunately, de-referencing every IRI in each document is not technically feasible due to high latencies and limited bandwidth. To reduce the workload, first we avoided de-referencing IRIs that were part of the standard vocabularies (RDF, RDFS, OWL, XSD) since that content is typically already known by the reasoner. Second, we limited de-referencing to only two subsets of IRIs: *predicate IRIs* and *Class IRIs*. The firsts are IRIs that appear as predicates of triples. These IRIs (excluding those from standard vocabularies) appear in 99.7% of the $\mathtt{LODL}_2$ documents and 83.4% of the $\mathtt{WDC}_2$ documents. The seconds are IRIs that were either explicitly defined as instances of *rdfs:Class* or appeared as subjects or objects of predicates that we knew their domains or ranges were instances of *rdfs:Class* (e.g., the object of the *rdf:type* predicate). De-referencing class IRIs was not always possible: in fact, only 67.63% of documents in $\mathtt{LODL}_2$ and 71.79% of documents in $\mathtt{WDC}_2$ contain class IRIs. Tab. 2 reports statistics about the number of distinct predicates and classes in the $\mathtt{LODL}_2$ and $\mathtt{WDC}_2$ datasets.

Furthermore, not all IRIs could be accessed: Only 4.7% of predicates and 35.9% of the class IRIs in $\mathtt{WDC}_2$ were de-referencable. The $\mathtt{LODL}_2$ dataset presented a significantly different situation: There, roughly 73% of predicates and 74.5% of class IRIs were accessible. We analyzed the inaccessible predicate IRIs in $\mathtt{WDC}_2$ and found that more than 84.6% of them pointed to non-existent resources on *schema.org*. We reported the full list of accessible and inaccessible IRIs in the public repository of this study.

Experimental Procedure We proceeded as follows: First we performed reasoning only on the single document (see Sec. 3). Then, we repeated the process only considering the remotely-fetched triples, and finally considering the document plus its remotely-

	$\mathtt{LODL}_2$				$\mathtt{WDC}_2$			
	Max	Average	Median	Min	Max	Average	Median	Min
Predicate	432	14.3	6	0	67	3.6	3	0
Predicate Domains	11	2.5	2	0	5	1.4	1	0
Class	496	7.9	1	0	14	1.3	2	0
Class Domains	11	1.5	1	0	4	0.7	1	0

Table 2. Statistics of predicate/classes IRIs per document in each sample.



(a) (b)

Fig. 5. After de-referencing predicate (P) and class (C) IRIs: Ratio of *Deriving*, *Not Deriving* and *Failed* reasoning processes (a). Ratio of documents that derive each derivation type (b).

fetched triples. We counted as *new* only those derivations that could have been derived in this last step (document plus the remote triples). In other words, we only count the derivations that were impossible to derive without adding external content to the input.

4.1 Experimental Results

Based on the number of new derivations, we divided the input documents into three groups: Those that yielded new derivations (*Deriving*), those that produced no new derivations (*Not-Deriving*), and those for which the reasoning process failed (*Failure*). Fig. 5a shows the percentage of documents in each group. The figure shows that a relatively large percentage of documents in $\mathtt{LODL}_2$ derived no additional information after remote triples were added. Furthermore, documents in $\mathtt{WDC}_2$ are more likely to yield new derivation after de-referencing IRIs than documents in $\mathtt{LODL}_2$. Moreover, the figure also suggests that de-referencing class IRIs is more likely to produce additional derivations than de-referencing predicate IRIs. This is interesting because documents often contain more predicate IRIs than class IRIs.

***Deriving* documents** To study how the de-referencing of IRIs affects the number of derivations, Fig. 6 shows a comparison between the size of the input documents and the number of new derivations. The figure shows that for $\mathtt{WDC}_2$, regardless the type of IRI that is de-referenced and irrespective of the reasoner, the number of new derived triples is proportional to the size of input document. This is similar to the local reasoning results (see Fig. 1). On the contrary, the reasoning for $\mathtt{LODL}_2$ is different from local reasoning results, especially with Jena.

In order to gain more insights, we classified the newly derived triples into the four categories defined in Sec. 2. Fig. 5b reports the ratio of documents that derive each specific derivation type (*T1*-*T4*) after de-referencing predicate (P) or class (C) IRIs.

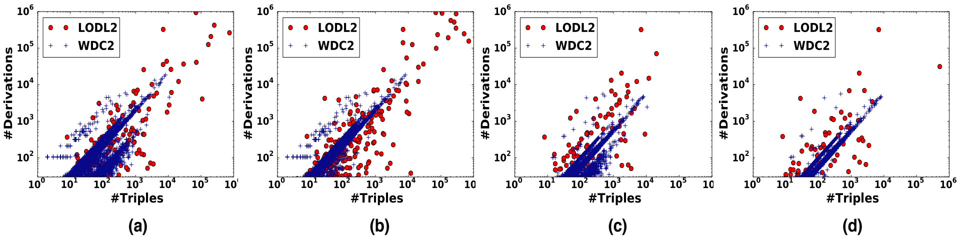


Fig. 6. Number of derivations vs input size after de-referencing: classes with Jena (a), predicates with Jena (b), classes with Pellet (c), predicates with Pellet (d).

Fig. 5b shows a different situation than in the local case (Fig. 2b). If we perform only local reasoning, then only a rather small percentage of $\texttt{LODL}_1$ documents derived *Type4* triples. Instead, after we de-reference IRIs, the majority of documents in both datasets did derive *Type4* triples.

Also, while every document in the local reasoning experiments derived *Type1* triples, such new derivation is almost non-existent after IRIs are de-referenced. The absence of *Type1* new derivations was expected because *Type1* triples are most often RDFS and OWL axioms that reasoners can derive anyway, thus they are not considered as *new* derivation. Aside from that, Fig. 5b shows that Jena derived many more *Type2* and *Type3* derivations than Pellet. Our manual inspection of these new *Type2* and *Type3* triples revealed that these derivations are mostly basic statements such as “an entity is of type *owl:Thing*”, or “resource is different from another resource”. Pellet is usually capable of concluding such derivation without additional data; hence, for this reasoner these statements are not counted as new derivations. This is not the case for Jena, and therefore it can derive them after external triples are included.

Finally, we observed that with $\texttt{WDC}_2$ documents the number of all *Type2*, *Type3*, and *Type4* derivations tends to be proportional to the size of input document. This situation is different for larger $\texttt{LODL}_2$ documents because these documents tend to use richer OWL ontologies which trigger more reasoning. Consequently, the number of derived triples is no longer proportional to the input size.

***Not deriving* documents** Fig. 5a also shows that there are cases where both reasoners did not derive any new triples. We scrutinized each *Not Deriving* document and the corresponding remotely-fetched triples, and found two main reasons for this: First, in some cases these triples only stated comments, labels, and descriptions intended for human interpretation. Reasoners can only conclude a limited number of derivations from such data. Second, the remote ontologies are dependent on yet more external ontologies, and so the inclusion of the remote data without its dependencies leads to no new derivation. Fig. 5a also shows a larger number of *Not Deriving* and *Failure* cases with $\texttt{LODL}_2$ than with $\texttt{WDC}_2$. This was surprising to us since we expected that IRI de-referencing was more effective in native RDF datasets than in datasets embedded in HTML pages.

Failures Fig. 5a reports a non-negligible number of cases where the inclusion of remote triples led to a failure of the reasoning process. Note that in this experiment the input samples only contain documents for which local reasoning had succeeded.

Therefore, the failures we refer to are caused by the inclusion of external ontologies. Pellet had the largest proportion of failed cases over $\mathtt{LODL}_2$ documents (20-40%). From our execution traces, we noticed that Pellet almost always failed due to inconsistencies (this accounts for 99% of the cases with predicates, and almost 94% with classes). Sometimes these inconsistencies were caused by conflicts introduced between triples fetched from different sources, while sometimes the conflict was between the external knowledge and the input document. Jena failed less times, but this is due to the fact that it is less stringent about consistency. Whenever Jena failed, it was because it timed out.

Further inspections indicated that inconsistencies are exacerbated when triples are included from more sources. When Pellet failed due to inconsistencies over $\mathtt{LODL}_2$, on average we de-referenced predicates from more than 18 sources (median 9), and classes from more than 11 sources (median 6). The average is significantly lower if we consider the cases where Pellet did not fail: predicates were from 5 sources (median 2), and classes from around 7 sources (median 2). This indicates that an excessive linking to multiple sources increases the chances of stumbling into inconsistencies.

5 OWL imports

Data Collection The directive *owl:imports* is another standard mechanism to link the document to external ontologies. In this section, we study how such inclusion affects the outcome of the reasoning process. This directive is used in less than 0.2% (939 documents) of the whole LODL dataset and in only 121 documents of the WDC dataset. Therefore, we do not sample them but instead use all of them. First, we executed local reasoning on them and filtered out all the documents for which this process failed. This reduced our input to 554 LODL documents (83 sources) while the size of the WDC documents remained unchanged: 121 documents from 16 different sources. In this section, we refer to these subsets of documents as $\mathtt{LODL}_3$ and $\mathtt{WDC}_3$ respectively.

The *owl:imports* directive defines a transitive process, i.e., an imported ontology may itself import additional ontologies [1]. The documents in $\mathtt{WDC}_3$ only import the *goodrelations*[4] ontology, which is accessible and does not contain links to any other ontology. On the other hand, the documents in $\mathtt{LODL}_3$ import 221 distinct ontologies from 62 different domain names. 76.9% of such imported ontologies were accessible, and only 52 of the documents imported ontologies with nested *owl:imports* statements. We found that the maximum length of transitive *owl:imports* chain is 4. Tab. 3 provides more information about the documents and the imported ontologies they mention. In the public repository, we report also the list of all inaccessible ontologies and more details on the ones that we fetched.

Experimental Procedure We proceeded in a similar way to Sec. 4, namely, we performed three reasoning processes: one over the documents without the imported ontologies, one over only the set of imported ontologies, and one over the document and its imported ontologies combined. Also in this case, we count as *new* derivation only those triples that are exclusively present in the last step(i.e., triples that are impossible to derive without importing external ontologies).

[4] http://purl.org/goodrelations/v1

	# Triples				# Imported ontologies			
	Max	Average	Median	Min	Max	Average	Median	Min
LODL3	4.3M	51.7K	397	2	48	4.5	4	1
WDC3	281	31.1	29	22	1	1	1	1

Table 3. Number of triples and number of imported links per document.

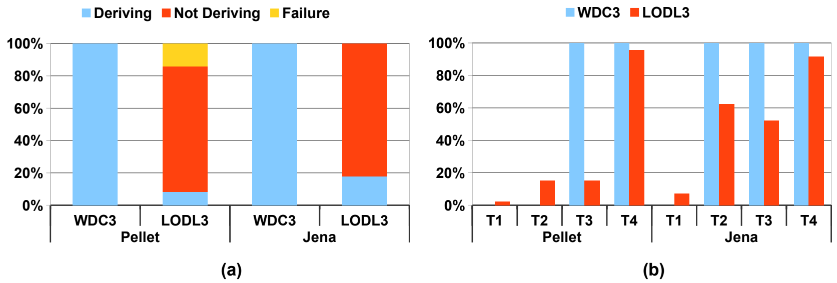


Fig. 7. (a) Ratio of documents in Derived/Not Derived/Failed groups. (b) Ratio of documents that derived each type of derivations.

5.1 Experimental Results

Similarly to Sec. 4.1, we categorized documents into the three groups of *Deriving*, *Not-Deriving*, and *Failure*, and present the collected statistics in Fig. 7a. The figure shows that both reasoners derived new triples from every document in WDC3. However, we also see that for a significant number of documents in LODL3 both reasoners were not able to derive any new triple. This was surprising to us since these documents were explicitly pointing to the external ontologies so we assumed that the import process would lead to at least some new derivations.

***Deriving* documents** Fig. 8 reports the number of new derived triples against the number of triples in the input document. We observe no proportional relation between

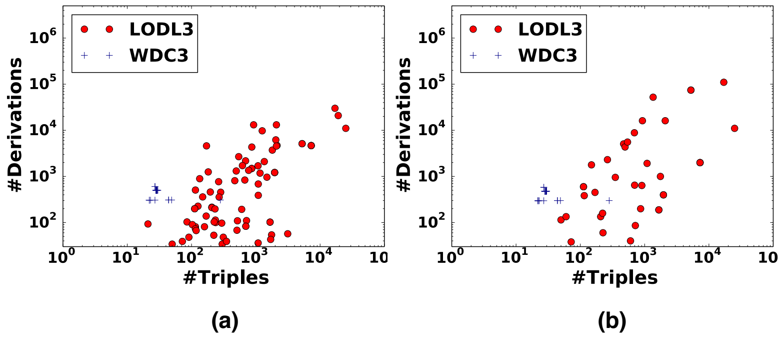


Fig. 8. Number of new derivations vs the document size after importing ontologies using Jena (a), and Pellet (b).

the number of derived triples and the size on input document in $\texttt{LODL}_3$ and the outcome with the two reasoners is different. This is in contrast with $\texttt{WDC}_3$ because here both reasoners derived roughly an equal number of derivations. Furthermore, each reasoner in $\texttt{WDC}_3$ derived almost the same number of triples per document (dots overlay each other in the figure). There are two reasons behind such regularity on $\texttt{WDC}_3$: *First*, as Tab. 3 shows, documents in $\texttt{WDC}_3$ tend to be of similar size; *Second*, all documents in $\texttt{WDC}_3$ import the same ontology (goodrelations).

Similarly as before, we classified the newly derived triples into our four categories and report the results in Fig. 7b. We notice that the type of new derivations is akin to what reasoners derived when IRIs were de-referenced (see Fig. 5b). In both cases, new *Type1* triples are almost nonexistent and almost all documents lead to the derivation of *Type3* and *Type4* triples. Additionally, we also observe that with Jena more documents derive *Type2* and *Type3* triples than with Pellet. As we explained in Sec. 4, this is because *Type2* and *Type3* triples usually include information that Pellet can derive without external ontological data (and hence are not counted as new derivations).

***Not Deriving* documents** While there is no *Not Deriving* document in $\texttt{WDC}_3$, as Fig. 7a shows, the percentage of *Not Deriving* documents in $\texttt{LODL}_3$ is remarkably higher than when IRIs are de-referenced (see Figure 5a). To find the cause, we studied the connections between the documents and the ontologies they import. In some cases, we found that the ontological information included from external sources was wither already in the document or reasoners were able to derive it from the triples in the document itself. In other cases (which were the majority), we found that the *owl:imports* statement was the only link between the document and the imported ontology. In other words, no term from the directly or indirectly imported ontologies was used in triples of the input document.

We can only speculate on the possible reasons behind the lack of links between the documents and the imported ontologies. One possible explanation could be that publishers put the *owl:imports* statements at the beginning of a large file (as a sort of "header") even though the remote knowledge was relevant for triples that were serialized much later on. Then, the large file was split in smaller ones without replicating the *owl:imports* statement on each file. In such a case, the only file that would contain the *owl:imports* statement is the first split, but this split does not contain any relevant triple for the remote ontology and hence no new derivation is produced (and the ones that could benefit from the remote content do not contain a link to the ontology).

Similarly, another case could occur if the publisher stores the *TBox* and *ABox* triples into different files and the *owl:imports* statement is put in the TBox file even though it points to relevant information for the ABox triples. In this case, if the ABox files do not import the TBox file, then the *owl:imports* statement will appear in a file (the TBox one) where it is not needed while files which might need it are not properly linked.

Failures In about 18% of the cases, Pellet failed and threw an exception about inconsistency. There were no failures with $\texttt{WDC}_3$. Jena timed out in only ∼0.3% of the cases. Pellet never timed out.

6 Related Work

Various aspects of Linked Open Data have been extensively studied in the last decade. Studies span a wide range of subjects including the quality of data [23, 20, 4], inconsistencies in the schema [2], the utilization of the standard vocabularies, and the depth and quality of the ontologies [30, 10, 11]. In [11], the authors provide some statistics about the utilization of ontologies and vocabularies. Bechhofer et al. [3] analyze a number of ontologies on the Web and find that the majority are OWL Full, mostly because of the syntactic errors or misuse of the vocabulary. Wang et al. [30] present similar finding and also report the frequency of the OWL language constructs and the shape of class hierarchies in the ontologies. Authors of [15] processed a large number of ontologies with various reasoners and show that most OWL reasoners are robust against the web.

As part of their research, authors of [9] report that only a small percentage of graphs on the Web uses *owl:imports*, a claim that our results confirm. The authors of [16] introduce ϵ-Connections to provide modelers with suitable means for developing Web ontologies in a modular way, and to provide an alternative to *owl:imports*.

More recently, Glimm et al. [14] discuss the current availability of OWL data on the Web. They report a detailed analysis on the number of used RDFS/OWL terms and highlight that the *owl:sameAs* triples are very popular. Similarly, Matentzoglu et al. [21] present another evaluation of the OWL landscape on the Web and a method to build an OWL DL corpus for evaluation of OWL engines. There have also been extensive studies on quality assessments and consistency of graphs on the Web. For instance, Zaveri et al. [31] provide a framework for linked data assessment. Feeney et al. [13] found string interdependencies between vocabularies and provide a tool to combine common linked data vocabularies into a single local logical model. Furthermore, they suggest a set of recommendation for linked data ontology design. None of these methods evaluate the interplay between data distribution and reasoning as we do. Therefore, we believe our results are a natural complement to all the above works.

7 Conclusions

The goal of this paper was to better understand how the distribution and reusage of ontologies affect reasoning on the Web of data. To this end, we analyzed several samples from LODLaundromat, which is a large crawl of RDF documents, and from Web Data Commons, which contains knowledge graphs that are embedded in HTML pages. We selected samples from hundreds of different domains in order to be as representative as possible. We compared the derivations produced by Pellet and Jena with and without remote external ontologies to understand, both from a quantitative and qualitative perspective, which are the major changes in terms of new derivations.

What have we learned? If we do not include any remote ontology, then reasoning tends to be rather trivial in the sense that it mainly returns RDFS and OWL axioms or description of the terms used in the document (e.g. that a property is an instance of *rdfs:Property*). However, if we do include remote ontologies, either by IRI de-referencing or *owl:imports*, then reasoners are able to derive many more non-trivial derivations.

Next to these positive findings, our analysis highlights some important problems:

- Reasoning on single documents is not always possible. In fact, we observed a number of failures (0.1-12%) during the reasoning process with both reasoners. These failures are due to either syntax errors, timeouts or inconsistencies;
- There are a non-negligible number of cases where the inclusion of the remote ontologies did not lead to any new derivation. Also, there are cases where the inclusion of remote ontologies breaks the reasoning process since it causes inconsistencies;
- The *owl:imports* directive is rarely used. Furthermore, it seems in many cases it is not used correctly (e.g., if the dataset is split in multiple files, the *owl:imports* statement is not replicated on each file) and this greatly reduces its potential;
- A significant number of IRIs are not accessible anymore. This is an important problem because the Semantic Web encourages ontological reuse as a basic principle, and if an ontology becomes unavailable then all documents that link to it will be unable to access its knowledge.

Some of these issues are already being studied in the community (for instance the rare usage of *owl:imports* is shown in [9], and the problem of non-accessible IRIs is well-known [18]) while others are not well-studied yet. Possible directions for future work could aim at researching techniques to selectively pick the "best" remote ontologies to avoid stumbling in errors. Also, it would be interesting to design methods to try to recover from situations where the documents do not point to any remote ontology by considering, for instance, ontologies that were linked for similar data. All these techniques could be potentially useful to make the Semantic Web more resilient to adverse situations.

With this paper, we provided a first snapshot of the current state of reasoning on the Web of Data. Our findings are encouraging, and our hope is that they stimulate the community to reflect on the adoption of current semantic technologies.

References

1. Antoniou, G., Van Harmelen, F.: Web Ontology Language: OWL. In: Handbook on ontologies, pp. 67–92 (2004)
2. Baclawski, K., Kokar, M.M., Waldinger, R., Kogut, P.A.: Consistency Checking of Semantic Web Ontologies. In: Proceedings of ISWC, pp. 454–459 (2002)
3. Bechhofer, S., Volz, R.: Patching syntax in OWL ontologies. In: Proceedings of ISWC, pp. 668–682 (2004)
4. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: A Uniform Way of Publishing Other People's Dirty Data. In: Proceedings of ISWC, pp. 213–228 (2014)
5. Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A metrics-driven approach for quality assessment of linked open data. Journal of theoretical and applied electronic commerce research 9(2), 64–79 (2014)
6. Berners-Lee, T.: Linked data-design issues (2006), http://www.w3.org/DesignIssues/LinkedData.html
7. Brickley, D., Guha, R.V.: RDF Schema 1.1. W3C Recommendation (2014)

8. Cochran, W.G.: Sampling techniques. John Wiley & Sons (2007)
9. Delbru, R., Tummarello, G., Polleres, A.: Context-Dependent OWL Reasoning in Sindice-Experiences and Lessons Learnt. In: Web Reasoning and Rule Systems. pp. 46–60 (2011)
10. Ding, L., Kolari, P., Ding, Z., Avancha, S.: Using Ontologies in The Semantic Web: A Survey. In: Ontologies, pp. 79–113 (2007)
11. Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and Ranking Knowledge on the Semantic Web. In: Proceedings of ISWC, pp. 156–170 (2005)
12. Fallside, D.C., Walmsley, P.: XML schema part 0: Primer. W3C recommendation (2004)
13. Feeney, K., Mendel-Gleason, G., Brennan, R.: Linked data schemata: fixing unsound foundations. Semantic Web Journal-Special Issue on Quality Management of Semantic Web Assets (2015)
14. Glimm, B., Hogan, A., Krötzsch, M., Polleres, A.: OWL: Yet to arrive on the Web of Data? In: WWW2012 Workshop on Linked Data on the Web. vol. 937. CEUR-WS.org (2012)
15. Gonçalves, R.S., Matentzoglu, N., Parsia, B., Sattler, U.: The empirical robustness of description logic classification. In: Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035. pp. 277–280. CEUR-WS. org (2013)
16. Grau, B.C., Parsia, B., Sirin, E.: Combining OWL ontologies using ϵ-connections. Journal of Web Semantics 4(1), 40–59 (2006)
17. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 web ontology language primer. W3C recommendation (2009)
18. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing Linked Data Dynamics. In: The Semantic Web: Semantics and Big Data, pp. 213–227 (2013)
19. Klyne, G., Carroll, J.J., McBride, B.: RDF 1.1 concepts and abstract syntax. W3C Recommendation (2014)
20. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven Evaluation of Linked Data Quality. In: Proceedings of WWW. pp. 747–758 (2014)
21. Matentzoglu, N., Bail, S., Parsia, B.: A Snapshot of the OWL Web. In: Proceedings of ISWC, pp. 331–346 (2013)
22. McBride, B.: Jena: A semantic web toolkit. IEEE Internet computing 6(6), 55–59 (2002)
23. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: Linked Data Quality Assessment and Fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 116–123. EDBT-ICDT '12 (2012)
24. Motik, B., Nenov, Y., Piro, R., Horrocks, I., Olteanu, D.: Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems. In: Proceedings of AAAI. pp. 129–137 (2014)
25. Mühleisen, H., Bizer, C.: Web Data Commons–Extracting Structured Data from Two Large Web Corpora. Proceedings of the Workshop on Linked Data on the Web 937, 133–145 (2012)
26. Parsia, B., Sirin, E., Kalyanpur, A.: Debugging OWL ontologies. In: Proceedings of WWW. pp. 633–640 (2005)
27. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semantics: science, services and agents on the World Wide Web (2007)
28. Urbani, J., Jacobs, C., Krötzsch, M.: Column-Oriented Datalog Materialization for Large Knowledge Graphs. In: Proceedings of AAAI. pp. 258–264 (2016)
29. Urbani, J., Kotoulas, S., Maassen, J., Van Harmelen, F., Bal, H.: WebPIE: A Web-scale Parallel Inference Engine using MapReduce. Journal of Web Semantics 10, 59–75 (2012)
30. Wang, T.D., Parsia, B., Hendler, J.A.: A Survey of the Web Ontology Landscape. In: Proceedings of ISWC. pp. 682–694 (2006)
31. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web 7(1), 63–93 (2015)
32. Zumel, N., Mount, J., Porzak, J.: Practical data science with R. Manning (2014)