An empirical survey of Linked Data conformance

Aidan Hogan [a], Jürgen Umbrich [a], Andreas Harth [b], Richard Cyganiak [a], Axel Polleres [c], Stefan Decker [a]

[a] *Digital Enterprise Research Institute, National University of Ireland, Galway*
[b] *AIFB, Karlsruhe Institute of Technology, Germany*
[c] *Siemens AG Österreich, Siemensstrasse 90, 1210 Vienna, Austria*

Abstract

There has been a recent, tangible growth in RDF published on the Web in accordance with the Linked Data principles and best practices, the result of which has been dubbed the "Web of Data". Linked Data guidelines are designed to facilitate ad hoc re-use and integration of conformant structured data—across the Web—by consumer applications; however, thus far, systems have yet to emerge that convincingly demonstrate the potential applications for consuming currently available Linked Data. Herein, we compile a list of fourteen concrete guidelines as given in the "How to Publish Linked Data on the Web" tutorial. Thereafter, we evaluate conformance of current RDF data providers with respect to these guidelines. Our evaluation is based on quantitative empirical analyses of a crawl of ~4 million RDF/XML documents constituting over 1 billion quadruples, where we also look at the stability of hosted documents for a corpus consisting of nine monthly snapshots from a sample of 151 thousand documents. Backed by our empirical survey, we provide insights into the current level of conformance with respect to various Linked Data guidelines, enumerating lists of the most (non-)conformant data providers. We show that certain guidelines are broadly adhered to (esp. use HTTP URIs, keep URIs stable), whilst others are commonly overlooked (esp. provide licencing and human-readable meta-data). We also compare PageRank scores for the data-providers and their conformance to Linked Data guidelines, showing that both factors negatively correlate for guidelines restricting use of RDF features, while positively correlating for guidelines encouraging external linkage and vocabulary re-use. Finally, we present a summary of conformance for the different guidelines, and present the top-ranked data providers in terms of a combined PageRank and Linked Data conformance score.

Key words: linked data, web of data, semantic web, rdf, web

1 Introduction

As a means of promoting grass-roots adoption of Semantic Web standards, the *Linked Data* community [16] has advocated a set of best principles for collaboratively publishing and interlinking structured data over the Web, as follows (here paraphrasing [7]):

(i) *use URIs* as names for things;
(ii) *use HTTP URIs* so those names can be looked up (aka. *dereferencing*);
(iii) *return useful information* upon lookup of those URIs (esp. RDF);
(iv) *include links* by using URIs that dereference to remote documents.

As such, the Linked Data community encourage those who wish to disseminate structured data on the Web to do so in an interoperable manner using the Semantic Web standards. Thus, Linked Data can be seen as a bottom-up approach to Semantic Web adoption—in particular, this bottom-up philosophy is epitomised by the five-star Linked Data scheme for Web publishing (here paraphrasing [7]):

Email addresses: aidan.hogan@deri.org (Aidan Hogan), juergen.umbrich@deri.org (Jürgen Umbrich), harth@kit.edu (Andreas Harth), richard.cyganiak@deri.org (Richard Cyganiak), axel.polleres@siemens.com (Axel Polleres), stefan.decker@deri.org (Stefan Decker).

★ PUBLISH UNDER AN OPEN LICENSE
★ ★ PUBLISH STRUCTURED DATA
★ ★ ★ USE NON-PROPRIETARY FORMATS
★ ★ ★ ★ USE URIS TO IDENTIFY THINGS
★ ★ ★ ★ ★ LINK YOUR DATA TO OTHER DATA

By promoting an accessible message to the wider Web community, Linked Data has enjoyed increasing adoption over the past four years. In 2007, the W3C "Linking Open Data" project began publishing legacy Web corpora under Linked Data principles. This resulted in rich datasets, most prominently the DBpedia [17] corpus extracted from semi-structured WIKIPEDIA articles. Thereafter, Linked Data adoption spread to various corporate entities, with, e.g., the BBC [66] [1], Thompson Reuters [2], and the New York Times [3] joining the effort and exposing information as Linked Data. More recently, various governmental agencies [8] have begun disseminating various public corpora as Linked Data, beginning with commitments from the US [28] [4] and UK governments [86] [5], and spreading to various other governmental bodies.

Taken together, these (varyingly) interlinked RDF corpora have resulted in a burgeoning, heterogeneous "Web of Data" built using Semantic Web standards and augmented with Linked Data principles. Thereafter, various claims have been made about the potential for new applications that can operate over this "global data space"; Bizer et al. envisage the following scenario(s):

> *"This Web of Data enables new types of applications. There are generic Linked Data browsers which allow users to start browsing in one data source and then navigate along links into related data sources. There are Linked Data search engines that crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data, similar to how a local database is queried today.* [...] *Unlike Web 2.0 mashups which work against a fixed set of data sources, Linked Data applications operate on top of an unbound, global data space. This enables them to deliver more complete answers as new data sources appear on the Web."* **—[16, §1]**

However, although (i) a number of generic web-based browsers have emerged for Linked Data (e.g., Disco [6], Marbles [7], Tabulator [10], Zitgist [8], etc.) and (ii) various warehouses have been proposed to operate over Linked Data from arbitrary domains (e.g., FactForge [13], Falcons [22], Sindice [91], Sig.ma [90], Swoogle [27], SWSE [56], Watson [24], etc.), the above stated vision has *yet* to be entirely realised.

From our experience on the SWSE project [56]—and also in the Pedantic Web group [55] [9] —we have found that (unsurprisingly) RDF data on the Web is of varying *quality*. Aside from concrete issues of noise [55], oftentimes data are modelled in a manner that is not facilitative to generic consumption: for example, common properties for labels are not re-used, properties and classes are invented and not defined, insufficient links are given to enable data discovery, etc. Such issues are something which, according to the above quote, the Linked Data guidelines aim to address.

In general, we currently see a lack of work addressing the issue of *quality* for Linked Data on the Web. Although a quantitative, objective, consumer-agnostic and universal measure of quality for Linked Data is probably unachievable, in this paper, we focus on the conformance of data providers with respect to Linked Data guidelines.

Along these lines, we extract and present a list of fourteen concrete recommendations from the "How to Publish Linked Data on the Web" [15] tutorial, prominently featured on the central `linkeddata.org` site. We discuss the importance of these recommendations, particularly in the light of consumer applications that intend to operate over the data. In particular, we currently focus on recommendations (i) that are targeted at the provision and maintenance of "instance data" as opposed to the provision and maintenance of vocabularies, and (ii) for which we can design straightforward, quantitative analyses.

In terms of experimentation and analysis, we propose a set of measures that can be used to quantify conformance with respect to each guideline highlighted. We then take a large corpus of RDF Web data, consisting of 1.1 billion facts collected from ~4 million RDF/XML documents by means of a breadth-first, open-domain crawl conducted in May 2010, and apply our measures to the data providers involved. We also contrast and

[1] http://www.bbc.co.uk/blogs/bbcinternet/2010/02/case_study_use_of_semantic_web.html; retr. 2011/02/21
[2] http://www.opencalais.com/; retr. 2011/02/21
[3] http://data.nytimes.com/; retr. 2011/02/21
[4] http://www.data.gov/; retr. 2011/09/01
[5] http://data.gov.uk/; retr. 2011/09/01
[6] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/; retr. 2011/09/01
[7] http://marbles.sourceforge.net/; retr. 2011/09/01
[8] http://dataviewer.zitgist.com/; retr. 2011/09/01
[9] http://pedantic-web.org/; retr. 2011/09/01

compare the PageRank scores of data providers and their conformance with respect to different guidelines. Finally, we compare results for different guidelines, and look at aggregate scores that give insights into the overall landscape of Linked Data (non-)conformance.

The rest of the paper is structured as follows:

(i) we discuss related literature in the area of analyses of RDF Web data (§ 2);
(ii) we present core preliminaries and notation (§ 3);
(iii) we describe the corpora used for our study (§ 4);
(iv) we enumerate fourteen Linked Data guidelines (§ 5), where for each we present:
  (a) description of the guideline,
  (b) motivation for the guideline,
  (c) analysis of conformance in our corpora, and
  (d) critical discussion of the analysis;
(v) we present analysis of the PageRank scores of providers, and contrast with conformance (§ 6);
(vi) we summarise, aggregate and discuss our empirical analyses for all issues (§ 7);
(vii) we discuss our results and conclude (§§ 8–9).

2 Background and Related Work

Empirical surveys of RDF Web data are important to generate feedback on current developments and to guide future developments for the Semantic Web and Linked Data. This paper's contribution to the area centres around analysis of conformance with respect to Linked Data publishing guidelines. However, in this section, we provide a comprehensive survey of the research literature concerning empirical analyses of RDF Web data that goes beyond our specific focus and covers the broader background, including:

(i) analyses that generally *characterise* the Semantic Web/Linked Data (§ 2.1);
(ii) works that focus on the *link structure* inherent in published RDF (§ 2.2);
(iii) works that analyse the *semantics* of such data (§ 2.3);
(iv) *miscellaneous*/specialised studies (§ 2.4);
(v) studies of *coverage and use* (§ 2.5); and
(vi) analyses focused on concrete issues of RDF *data quality* (§ 2.6).

2.1 General analyses

Various authors have tried to broadly characterise the Semantic Web down through the years. These analyses now constitute chronological snapshots of the nature of RDF data on the Web at different times.

In 2005, Ding et al. [31] presented one of the earliest analyses of RDF data published on the Web. They collected over 1.5 million RDF/XML documents from the Web and reported about the prevalence of use of various namespaces and properties therein, where the bulk of data were described in the Friend of a Friend (FOAF) and Dublin Core (DC) vocabularies. As opposed to low-volume auto-biographical FOAF profiles, they found that the most prevalent source of RDF data on the Web at that time was given by "FOAF social networks", such as `livejournal.com`, `deadjournal.com`, `ecademy.com`, etc. Further experiments included calculating connected components in the FOAF social network and detecting groups on a subset of 7 thousand FOAF files. They detected various forms of Zipf distributions, such as the number of persons described in each Web document, the number of aliases found per person (found through sharing key values for `foaf:mbox_sha1sum` property), etc.

Roughly a year later, Ding and Finin [26] again looked to characterise the amount of Semantic Web data on the Web. Using search results sizes reported by Google, they estimated the number of RDF documents on the Web at that time to be in the range of 10^7–10^9. Analysing a dataset of 1.4 million RDF sources from 2006—again mostly consisting of FOAF with some RSS 1.0 feeds—they presented various statistics relating to the largest providers, document sizes, last-modified dates, etc. Furthermore, they identified that, e.g., a large fraction of defined classes (>97%) had no instances in their data, and that the majority (more than >70%) of properties are never used. In many cases, the derived statistics follow power law distributions. Towards the end of the paper, they provide some prescient discussion about whether or not the traditional monolithic ontology makes sense for the Web, noting:

> *"Recent work* [...] *argues against large, monolithic ontologies in favor of having many interconnected components. We might even eliminate namespaces as boundaries. For example, the Dublin Core Element ontology has been widely used together with terms from many other semantic web ontologies."*

As we will see later, the emerging trends they remark upon (interconnected, lightweight vocabularies and mixing namespaces) are now central aspects of Linked Data. They also note issues relating to identity and accessibility, which are currently core themes in Linked Data.

Skipping forward to 2008, when Linked Data was gaining a strong foothold on the Web, Hausenblas et al. [48] attempted to empirically gauge the size of the Semantic Web. They differentiate data into schema level

data, and instance level data, where the latter is further split into single-point-of-access datasets (e.g., high-volume publishers) and distributed datasets (e.g., decentralised FOAF files, SIOC descriptions [18], etc.). For the single-point-of-access datasets, they report on the number of triples made available, the level of external linkage, and build a (directed, labelled, weighted) graph of interconnections between the different data providers. They also crawled documents from the decentralised datasets, showing that, e.g., FOAF data was well interlinked within itself, but poorly linked to external datasets. Although no definitive results are given on the effective size of the Semantic Web at that time (informally, a lower bound of two billion triples was established, which we believe to be very conservative), the main conclusion was that more emphasis should be placed on interlinking the datasets.

Various authors have presented unpublished works analysing different Billion Triple Challenge (BTC) corpora down through the years. Most recently, Grimnes [38] provided a detailed analysis of the BTC-2010 [10] dataset. The dataset consists of 3.172 billion statements, representing 1.441 billion unique triples, collected from 8.1 million sources. Grimnes presents a variety of statistics, including triple distributions for subject, predicate and object terms, documents, class memberships, etc. Analysing namespaces, he found that FOAF-related data still contributed the bulk of the corpus, but where new Linked Data domains (particularly `data.gov` related initiatives), and new vocabularies (e.g., GoodRelations data [51]) were also contributing heavily.

2.2 *Link-structure analyses*

As per the previous remarks of Hausenblas et al., an important aspect of the Semantic Web—and of the Web in general—is the interlinkage of content, and the graph structure embodied by those links.

The work of Ge et al. [36] reports experimental results on analysing the complex network structure of the object link graphs (i.e., the RDF data graph) constructed from two large data sets: 11.7 million RDF documents crawled in 2008 and 21.6 million RDF documents crawled in 2009. Both datasets are crawled by the FALCON-S search engine. Their statistics contain the distribution of the number of hosts versus number of documents, and graph invariants such as degree distribution and connectivity. For the 2009 dataset, excluding single non-linked vertices, they find 813 thousand strongly connected components in the graph, where 88.1% of resources are contained in the largest thereof. They estimated the effective diameter of the graph to be ~11.5, indicating the longest (shortest) path between two vertices: i.e., 11.5 is roughly the maximum length walk needed to get from one node to another, taking the shortest route. Comparison between the 2008 and 2009 datasets shows that interlinkage improves slightly.

Josyln et al. look again at the BTC-2010 corpus [63]. Although their work focuses on scalable processing of data using a Cray XMT supercomputing platform, they derive and present an array of useful statistics from the corpus that include: top subjects, predicates and objects; top edge types and node types they connect; top link types; link type bi-grams and tri-grams (i.e., predicate paths of length two and three), as well as connected components and typed paths. They found 208.3 thousand connected components in the RDF graph constituted by the corpus, with the largest component containing 99.8% of the total number of vertices; the discrepancy with Ge et al.'s result is probably due to different sampling techniques for the empirical corpora.

Guéret et al. [39] also look at BTC-2010, but instead measure what they call "robustness" via infrastructure analysis and semantic network analysis, and propose measures for improving the Web of Data; their notion of robustness relates to the reachability of macro-components on the Semantic Web in the case of domains going offline. They derive a hostname graph and a namespace graph from the BTC corpus, and calculate several network measures over those graphs, such as degree distribution and betweenness centrality (which they see as a proxy for robustness). They devise methods to improve the robustness of the Web of Data that aim to minimise the graph's centrality index with the fewest links possible; for this, they propose using a Jaccard distance measure based on vocabulary overlap as a cost function. A qualitative analysis revealed that as much as 80% of the triples do not link to external URIs but refer to either site-internal links, blank nodes or literals. The analysed networks show an extreme distribution and have a brittle structure; much of the connectivity is provided via three central domains (`xmlns.com`, `dbpedia.org` and `purl.org`).

2.3 *Semantic analyses*

Another important aspect of the Semantic Web is, of course, semantics. Various empirical studies down through the years have looked at the use of the RDFS and OWL standards in RDF Web data. On an instance-

[10] http://km.aifb.kit.edu/projects/btc-2010/; retr. 2011/08/21

level, some studies have specifically investigated the use of the owl:sameAs relation on the Web, which is used to relate two *coreferent* resources that talk about the same real-world thing (also known as URI aliases [15,50]). On a schema-level, other studies have looked at how RDFS and OWL are used to define the semantics of classes and properties appearing in various Web vocabularies and ontologies.

Recent works by Ding et al. [29] discuss the use of owl:sameAs for linking URI aliases and retrieving additional data during crawling. They also discuss quality issues arising when using owl:sameAs statements from the Web indiscriminately; in particular, they raise concerns about the symmetric semantics of owl:sameAs links across domains, and about the relaxed use of owl:sameAs. In another 2010 paper, Ding et al. [30] return to this issue, providing quantitative analysis of the owl:sameAs graph extracted from the BTC-2010 dataset. They found that URIs with at least one alias had an average of 2.4 aliases (i.e., the average does not include URIs not in the owl:sameAs graph). The average path length was 1.07, indicating that few transitive aliases are given. They also summarise owl:sameAs linkage between different publishers of Linked Data.

In a similar vein, Halpin and Hayes [42] and Halpin et al. [43] investigate the incorrect use of owl:sameAs. Taking an initial set of 58 million owl:sameAs triples extracted from 1,202 Linked Data domains, they present the top providers of such links, and a distribution of links-per-domain. They then employ four human judges to manually inspect 500 links sampled (using logarithmic weights for each domain) from the full corpus. Their experiments found that approximately 51% (±21%) of owl:sameAs relations were deemed correct: the level of disagreement observed amongst the human judges indicates that coreference between URI aliases is inherently subjective [43]; the authors also note that the RDF descriptions of the aliases were deemed insufficient to make a meaningful judgement in 27% (±19%) of cases.

In a recent paper, we looked more generally at the issue of equality for Linked Data, analysing not only owl:sameAs, but also use of OWL features that allow for inference thereof [58], including inverse-functional properties, functional properties, etc. Surveying the (closure of explicit) owl:sameAs relations in the same corpus used herein, we found that URIs with at least one alias had an average of 2.65 aliases, with the largest set containing over eight thousand aliases (due to incorrect owl:sameAs linkage of online drugs data). We also found that 57% of alias groups contained URIs from more than one domain. Based on our manual evaluation of a sample of one thousand alias pairs, we estimated the accuracy to be ~97.2% (much more encouraging than the results of Halpin et al. [43], although many of our results were deemed "trivially correct" if there was not enough information to suggest otherwise). We also investigated implicit owl:sameAs relations, where most were found through reasoning over inverse-functional properties, but where the vast bulk of additional aliases involved blank nodes within the same domain (accuracy remained stable at ~97.7%).

A recent paper by Mallea at al. [70] discusses the semantics of blank nodes and presents an empirical study of their use in RDF Web data. Although some high-volume publishers export huge amounts of blank nodes in absolute terms, the average use of blank nodes (vs. unique literal or URI terms) across domains was measured as 7.5%, which decreased to 6.1% when only considering domains appearing in the LOD Cloud diagram.[11] Further empirical analysis of the graph-structures formed by blank nodes (where two blank nodes are linked by appearing in the subject and object of the same triple) indicates that 98% of the time, such graphs form trees: the implication is that simple entailment [49] over blank nodes is often tractable in practice despite being NP-complete in theory.

On a schema-level, various works have looked at the expressivity of ontologies on the Web [5,98,23]; these results are somewhat tangential to the focus of this paper, but show that restrictions laid out in the OWL standard (specifically for the OWL Lite and OWL DL dialects) are not well-followed by Web ontologies, but that such ontologies are typically relatively inexpressive. In previous works, we analysed the use of RDFS and OWL in top-ranked vocabularies extracted from an RDF Web crawl (the same as used later); we found that RDFS features were the most prominently used, with OWL (1) features not requiring blank nodes to serialise in RDF also finding use in prominent vocabularies [53].

More recently, Cheng et al. [21] performed a study of 2,996 Web vocabularies, spanning 261 pay-level-domains, finding 396,023 classes and 59,868 properties. Approximately 72% of vocabularies were found to contain no more than 25 terms. Taking a further 15 million "instance" documents, the authors investigate indicators of relatedness between vocabularies, measured with respect to how terms are defined, textual content of vocabularies, explicit interlinkage, and co-occurrence in instance documents. Some resulting high-level conclusions note that various related vocabularies are not interlinked, but that interlinked vocabularies often tend to be co-instantiated in the same documents.

[11] http://lod-cloud.net/; retr. 2012/01/11

2.4 Miscellaneous analyses

Motivated by certain observations or use-cases, a number of specialised analyses of the Semantic Web (esp. Linked Data) have been presented in the literature.

Hartig [46] discusses various issues relating to notions of provenance and the provision of document meta-data. In particular, he provides discussion on current provenance-related properties appearing in popular Linked Data vocabularies. Thereafter—and based on information extracted from Ping-the-Semantic-Web and Sindice—he presents a survey of the approximate number of documents using each of the provenance-related properties, where, of the vocabularies and properties surveyed, DC (esp. `dcterms:created` indicating the date of creation) and FOAF terms (esp. `foaf:maker` indicating an author of the document) were the most prevalently encountered.

Umbrich et al. [93] studied the changes in content of a total of 550 thousand RDF/XML documents crawled from the Web over a period of 24 weeks in 2008. They showed that the `Etag` and `Last-modified` HTTP header fields—which typically indicate the date the document being served was last updated—were not provided in 67.95% of the documents. Thereafter, surveying the content of the documents, they ascertained that 62% of documents remained static over the 24 weeks, whereas 69% of entity descriptions also remained static. Of the documents that did change, 59% were estimated to change at a rate of 12–24 weeks, 23% at a rate of 4–12 weeks, 9% at a rate of 1–4 weeks, and 9% at a rate of <1 week. However, the authors admit that the empirical corpus used was insufficient to derive any concrete, fine-grained conclusions on the dynamicity of RDF Web data, but instead could only offer insights into the approximate level of change.

2.5 Usage-based analyses

We have seen that—starting from at least 2004/05 with FOAF, RSS 1.0 and DC data—there has long been a large base of interlinked RDF documents on the Web. Early use-cases for search over Semantic Web data typically centred around domain-specific ontologies created by academia, and/or hand-crafted FOAF files created by hobbyists, and/or bespoke RDF converted from dumps of legacy structured data [27,24,56,19]. In more recent years, the diversity and volume of RDF Web data has expanded significantly under the banner of Linked Data publishing. Given all of this openly available, interlinked, semantic RDF content, the pertinent question is then: *what can it be used for?*

The overall goal of a 2009 study by Halpin [40,41] was to determine if content on the Semantic Web is potentially of interest to the average Web user, and what the coverage of general-interest topics is like; he identifies two categories, "first generation" content characterised by FOAF social networks, RSS 1.0, etc., and "second generation" content characterised by Linked Data publishing. Halpin acquired a set of 15 million (6.6 million unique) real-world keyword queries from the Microsoft Live search engine, from which, 7.8 thousand unique entity queries (e.g., entity names) and 5.3 thousand unique concept queries (e.g., class names) with more than ten occurrences were extracted. These keyword queries were run against the FALCON-S search engine [22]—which indexes large crawls of RDF Web data—where the results showed that searches either returned a great many results, or none at all. Interestingly, he found that there was no correlation between results sizes and the popularity of the input keyword query, suggesting that a mismatch between the transient nature of Web search and the static nature of RDF data was a possible cause. Many results came from the (second generation) DBpedia domain [17]. Analysis suggested that the spread of results over the different domains ostensibly followed a power-law, but the distribution was ultimately found to have an insignificant fit, possibly because the amount of RDF data published by very large domains greatly outweighs smaller, bespoke publishing. Similar observations of poorly-fitting power laws in the analysis suggest (to us) that RDF data on the Web is still not mature enough to exhibit true power-law distributions. Other results presented in the paper looked at the RDF(S) and OWL features, also discussing the issue of identity where concern were raised about the lack of `owl:sameAs` links to model URI aliases.

In another 2009 paper, Mika et al. [73] tackle a similar issue that they call the "Semantic Gap", viz., the divide between the supply of data on the Semantic Web and the demand of typical Web users. However, given that their study is centred around the Yahoo! search engine index (which does not index RDF/XML), they focus on analysing structured data embedded in *HTML documents, such as RDFa, eRDF and Microformats. Although they show that RDFa is growing in popularity, the overall percentage of indexed documents containing RDFa was 0.6%, much less than the equivalent percentage for various forms of Microformats (e.g., `tag` was in ~2.6% of indexed documents). They also looked at the ratio of top-ten results pages with embedded meta-data for 7.6 thousand unique, real-world keyword queries. Their results found that 59% of queries had at least one

result with embedded meta-data, but where the equivalent figure considering only RDFa was 2.5%. Although coverage was poor, one conclusion was that Semantic Web technologies could play a more significant role for mainstream web-search if it was adopted by particular sites, or better targeted particular categories of queries.

Looking at the issue of Linked Data usage from another perspective, in 2010, Möller et al. [75] analysed the server access logs of four prominent Linked Data hosts, viz., Semantic Web Dog Food [76], DBpedia [17], DBTune [84] and RKBExplorer [37]. The available logs covered periods from one month (RKBExplorer) to two years (Dog Food), all falling somewhere between 2008–2010. They distinguish "semantic agents" accessing the servers as those issuing SPARQL queries [82] or requesting RDF-specific content-types. Their findings indicate that for the different domains, "semantic traffic" represented 9–19% of the total traffic observed. They then look at whether real-world events affected demand for resources from the different sites (e.g., monitoring access to `dbpedia:Michael_Jackson` around the time of his death), where some topical resources did encounter peaks in demand.

2.6 *Data-quality analyses*

The previous literature gives a somewhat "lukewarm" impression of the use and usefulness of RDF data published on the Web. Aside from issues relating to coverage, interlinkage, semantic expressivity, dynamicity, use-cases, and so on, one possible reason for the (arguably) slow emergence of applications operating over the Semantic Web and Linked Data is that the data being published on the Web is simply not of high-enough *quality*. Publishing problems with respect to accessibility, syntax, semantics, etc., may greatly diminish the potential for applications over the data and/or introduce increased overhead for adoption. More opaque or subjective notions of quality—relating to resource identity, conceptual modelling, competency, etc.—may be more difficult to formalise and quantify. Such matters are further complicated by the inherent decoupling of publishing and modelling from applications, where, aside from coverage, data might be considered of excellent quality for one use-case and of poor quality for another due to, e.g., specific modelling choices.

Indeed, the notion of quality (in this context and in general) is quite a nebulous one. In his dissertation, Vrandečíc [97] asks "*how to assess the quality of an ontology for the Web?*". He refers to an ontology as a formal, explicit specification of a shared conceptualisation that may include classes, properties but *also* instances. Based on previous proposals, he discusses various criteria that a good Web ontology should meet, and then proposes concrete measures relating to accuracy, adaptability, clarity, completeness, computational efficiency, conciseness, consistency and organisational fitness. His conclusion is that a single measure to assess the overall quality of an ontology is elusive, and deriving concrete measures to identify shortcomings in ontologies is a more useful approach; he also states:

> *"[...] instead of aiming for evaluation methods that tell us if an ontology is good, we settled for the goal of finding ontology evaluation methods that tell us if an ontology is bad, and if so, in which way."*

Along these lines he analyses issues relating to naming conventions (i.e., checking if the local part of a URI coincides with a label given to that entity), erroneously introduced punning, "superfluous blank nodes" (i.e., blank nodes that are not used for encoding RDF collections and OWL constructs), to name but a few checks.

In previous works [55], we listed and discussed common errors made by RDF publishers on the Web based on experiments conducted on a dataset acquired from 150k URIs mentioned in a previous RDF dataset. We classified errors into the following four categories: (i) accessibility and derefenceability, (ii) syntax errors, (iii) reasoning: noise and inconsistency and (iv) non-authoritative contributions. Most URIs returned without error, but less than half returned RDF/XML. We found that 8.1% of triples in the resulting dataset used undeclared class URIs, and 14.3% used undeclared property URIs. With respect to noise, the most prevalent issues related to reasoning, where we found many invalid values for inverse-functional properties, and various forms of inconsistency, particularly memberships of disjoint classes. We also identified the issue of "ontology hijacking", where third-parties redefine the meaning of popular vocabulary terms. We argued against application-side workarounds for certain frequently observed problems and buggy datasets, as those would have to be replicated across all applications that use the data, entailing a large barrier-to-entry for potential consumers. Rather, we provided a prototypical validator [12] and initiated an online community—the "Pedantic Web" Group [13] —which aims to educate publishers on issues of data-quality and contacts publishers with bug-reports.

[12] http://swse.deri.org/RDFAlerts/; retr. 2011/08/21; also http://inspector.sindice.com/; retr. 2011/08/21
[13] http://pedantic-web.org/; retr. 2011/08/21

As discussed in the introduction, Linked Data guidelines are designed to enable new types of applications over data published in a conformant manner. With respect to empirical studies of Linked Data conformance, although there is some overlap between some of the more generic guidelines and some results previously discussed, we are not aware of any published work focusing specifically on this topic. At the time of writing, Bizer et al. [14] are currently drafting an online document (currently version 0.3) which—similarly to our contribution—enumerates nine Linked Data guidelines and characterises the conformance of publishers thereto. However, their study of conformance is based on self-reported statistics provided by the publishers themselves in CKAN. [14] We view our work herein as complementary, effectively constituting an empirical, consumer-side study of the issue of Linked Data conformance.

3 Preliminaries

We now move towards presenting our primary contribution, but first we cover some necessary preliminaries relating to RDF and Linked Data principles. We also very briefly discuss the implementation and methods used to conduct our experiments and extract our results.

3.1 RDF

We briefly give some necessary notation relating to RDF constants and RDF triples; cf. [49].

RDF constants

Given the set of URI references $\mathbf{U}$, the set of blank nodes $\mathbf{B}$, [15] and the set of literals $\mathbf{L}$, the set of *RDF constants* is denoted by $\mathbf{C} := \mathbf{U} \cup \mathbf{B} \cup \mathbf{L}$.

Herein, we use CURIEs [12] to denote URIs: we refer the reader to the service at `http://prefix.cc/` (retr. 2011/09/01), where the namespace prefixes used in this paper can be looked up. Following Turtle syntax [6], we may use `a` as a convenient shortcut for `rdf:type`.

RDF triples

The set of all RDF triples is given as $\mathbf{G} := (\mathbf{U} \cup \mathbf{B}) \times \mathbf{U} \times (\mathbf{U} \cup \mathbf{B} \cup \mathbf{L})$. A triple $t := (s, p, o) \in \mathbf{G}$ is called an *RDF triple*, where s is called subject, p predicate, and o object. We call a finite set of triples $G \subset \mathbf{G}$ an RDF *graph*.

Data-level position

We define two *data-level positions* in a triple:

(i) the subject of a triple; and

(ii) the object of a triple *iff* the predicate is *not* `rdf:type`.

Given an RDF graph G, we use the function $\mathsf{dlc}(G)$ to denote the set of RDF constants appearing in the data-level position of some triple in that graph. In particular, we distinguish the data-level positions of a triple from those that are typically occupied by schema terms such as properties appearing in the predicate position or classes appearing as the value of `rdf:type`. (Many of the guidelines we will look at focus on data-level terms and do not naturally apply to these latter schema-level terms, where we would not expect, for example, all members of a class to be given in its dereferenced document, and so forth.)

3.2 Linked Data principles and data sources

Linked Data principles [7] and associated best practices [15] offer clear guidelines for publishing RDF on the Web. We briefly discuss Linked Data principles and notions relating to provenance. [16]

Linked Data principles

Throughout the rest of this paper, we denote the four best practices of Linked Data as follows [7]:

LDP1 use URIs to name things;

LDP2 use HTTP URIs so that those names can be looked up;

LDP3 provide useful structured information when a look-up on a URI is made, called *dereferencing*;

LDP4 include links using external URIs.

Data source

We define the *http-download* function $\mathsf{get} : \mathbf{U} \rightarrow 2^{\mathbf{G}}$ as the mapping from a URI to an RDF graph (set of facts) it may provide by means of a given HTTP lookup [34] that directly returns status code `200 OK` and data in a suitable RDF format; this function also performs a rewriting of blank-node labels (based on the input URI) to ensure uniqueness when merging RDF graphs [49]. We define the set of (RDF) *data sources*

[14] `http://ckan.net/group/lodcloud`; retr. 2011/08/21

[15] We interpret blank nodes as skolem constants, as opposed to existential variables. Also, we rewrite blank-node labels to ensure uniqueness per document, as prescribed in [49].

[16] In a practical sense, all HTTP-level functions $\{\mathsf{get}, \mathsf{redir}, \mathsf{redirs}, \mathsf{deref}\}$ are set at the time of the crawl, and are bounded by the knowledge of our crawl.

($\mathbf{S} \subset \mathbf{U}$) as the set $\mathbf{S} := \{s \in \mathbf{U} : \mathsf{get}(s) \neq \emptyset\}$. In this paper, sources refer to individual RDF/XML documents retrievable over the Web from location s.

RDF triple in context/RDF quadruple

An ordered pair (t, c) with a triple $t = (s, p, o)$, $c \in \mathbf{S}$ and $t \in \mathsf{get}(c)$ is called a *triple in context* c. We may also refer to (s, p, o, c) as an *RDF quadruple* or quad q with context c.

HTTP Redirects/Dereferencing

A URI may provide a HTTP redirect to another URI using a 30x response code [34]; we denote this function as $\mathsf{redir} : \mathbf{U} \to \mathbf{U}$ that may map a URI to itself in the case of failure (e.g., where no redirect exists)—this function would implicitly involve, e.g., stripping the fragment identifier of a URI [11]. We denote the fixpoint of redir as redirs, denoting traversal of a number of redirects (a limit may be set on this traversal to avoid [very rare, possibly malicious] redirect cycles and artificially long redirect paths). We define *dereferencing* as the function $\mathsf{deref} := \mathsf{get} \circ \mathsf{redirs}$ that maps a URI to an RDF graph retrieved with status code 200 OK *after* following redirects, or that maps a URI to the empty set in the case of failure.

Pay-level domains/Data providers

Herein, we use *pay-level domains (PLDs)* [68,30] to distinguish individual *data providers*. A pay-level domain is a direct sub-domain of a top-level domain (TLD) or a reserved second-level country domain (ccSLD); examples of PLDs include `dbpedia.org` and `bbc.co.uk`.

We do not consider general fully-qualified domain names (FQDNs) as indicating different data providers since PLDs such as `livejournal.com` publish data under many FQDNs, assigning a third-level domain to each user (e.g., `danbri.livejournal.com`). Also, we acknowledge that multiple "datasets" may intuitively operate within a given PLD, but note that a pay-level domain is typically under the control of a single person or organisation, which we herein consider to be the granularity of our data providers. We may interchangeably use the terms "data provider", "PLD" or "domain" to refer to such sites that host a set of data sources (in our case, a set of RDF/XML documents).

For convenience, herein we represent PLDs as a set of HTTP URIs—e.g.: `http://dbpedia.org/`—given by the set $\mathbf{P} \subset \mathbf{U}$. We define the function $\mathsf{pld} : \mathbf{U} \to \mathbf{P}$ that maps a HTTP URI to its PLD. Letting $\mathsf{sources}(p) := \{s \in \mathbf{S} : \mathsf{pld}(s) = p\}$ denote the sources under the (direct) control of PLD p, we use the function $\mathsf{data} : \mathbf{P} \to 2^{\mathbf{G}}$ to denote the RDF merge of the set of triples given the set of sources in that PLD ($\mathsf{sources}(p)$); more specifically $\mathsf{data}(p) := \bigcup_{s \in \mathsf{sources}(p)} \mathsf{get}(s)$. We also define the function $\mathsf{local} : \mathsf{pld} \to \mathbf{B} \cup \mathbf{U}$ that maps a PLD p to the union of

(i) the set of blank nodes appearing in a triple of $\mathsf{data}(p)$;
(ii) URIs appearing in a triple of $\mathsf{data}(p)$ such that $\mathsf{pld}(\mathsf{redirs}(u)) = p$.

Intuitively, $\mathsf{local}(p)$ refers to the locally minted non-literal terms under the control of the PLD p [45].

3.3 Extraction of statistics

In this paper, we are more interested in the results of the empirical analysis rather than the implementation thereof, where we consider issues relating to performance, etc., as out of scope. However, we now *briefly* give an insight into the batch-processing techniques used to extract the statistics.

We assume that the dataset to be analysed is given as a flat file of N-Quads, optionally compressed with GZip; further, we assume that knowledge of the crawl is given in a structured input file, including information about response codes, content-types and redirect locations.

In preparation for the analysis, the dataset is sorted according to subject–predicate–object–context order and object–predicate–subject–context order in two separate files: the sorts are performed using standard on-disk external merge-sorts. Thereafter, we perform a merge-join over the two files, joining on the sorted subject/object position, effectively producing batches corresponding to all of the inlinks and outlinks for each resource in the data. Statistics about the various data-providers are accumulated in memory during the scan of resources.

Our implementation is Java based, where we use the Java Statistical Classes library [17] to compute Kendall's τ measure (introduced later for comparing PageRank and conformance). Further, we use a simple RMI infrastructure to perform distributed sorts and scans—details of the infrastructure are available in [56,53]. Analyses are performed using nine machines with 2.2GHz Opteron x86-64 CPU, 4GB main memory, 160GB SATA hard-disks, running Java 1.6.0_12 on Debian 5.0.4.

4 Empirical Corpora

In this section, we give pertinent descriptions of the two corpora we acquired for the purposes of our empir-

[17] `http://www.jsc.nildram.co.uk/`; retr. 2011/09/01

ical analysis. Our primary corpus (§ 4.1) comprises of 1.1 billion quadruples crawled from just under 4 million RDF/XML documents in May 2010, spanning content hosted by 778 data providers (PLDs). We also briefly describe our secondary corpus (§ 4.2), which we use to analyse the stability of documents hosted by individual data providers, and which consists of nine monthly snapshots, accessing a static set of 155 thousand RDF/XML documents and covering the period of March 2010 to November 2010.

4.1 Billion quadruple crawl

To provide insights into current conformance with respect to Linked Data best-practices, the first corpus over which we apply our analyses consists of 1.118 billion quadruples, crawled in mid-May 2010 from 3.985 million RDF/XML documents spanning 778 pay-level domains (data providers). Of the 1.118 billion raw quadruples parsed, 1.106 billion (98.9%) are unique, and 947m (84.7%) are unique triples.

4.1.1 Corpus acquisition

We conducted the crawl in a breadth-first manner over a cluster of nine machines. The crawl was seeded with 42.5 thousand URIs extracted from an older Linked Data crawl conducted in 2009: URIs were randomly sampled from all positions of the RDF triples. To ensure a broad sample of data-providers during the crawl (and to ensure ~polite crawling), we assign each pay-level domain (PLD) an individual priority queue. The PLD queues are sampled in a round-robin fashion during the crawl, with the highest linked URIs *for each domain* being returned first. (For more details on the implementation of our distributed crawler, we refer the interested reader to [53].)

Furthermore, we only access RDF/XML documents using an accept-header `application/rdf+xml` and do not consider documents in other syntaxes, such as RDFa, N-Triples or Turtle. Linked Data guidelines have traditionally suggested that RDF/XML data should be provided as a minimum:

> *"'There are various ways to serialize RDF descriptions. Your data source should at least provide RDF descriptions as RDF/XML which is the only official syntax for RDF"'* —**[15, §5]**

However, more recent Linked Data guidelines [50] and trends suggest that other RDF syntaxes can be used as an alternative to RDF/XML, where in particular, RDFa has been standardised [1] and is growing in popularity.

Along these lines, our empirical corpus is only a sample of Linked Data, and like any non-trivial sample of open Web data (where the nature of the entire population cannot be feasibly known), it has inherent biases that may affect our analysis and results [40]. We identify the following known biases for our corpus:

- given that our crawl was run in May 2010, our corpus does not reflect newer publishers or published data;
- our crawl only samples RDF/XML and does not cover data in other syntaxes;
- given that our crawl does not follow all possible URIs, our corpus is particularly incomplete for domains that host a large amount of documents—towards the end of the crawl, there were still ~50 PLDs whose RDF/XML content was (almost surely) known not to be exhausted;
- given that the crawl is breadth-first and that URIs with higher inlinks are prioritised, our corpus is biased towards containing the most well-linked documents in each domain.

To help counter-act the effects of sampling bias, we focus on presenting conformance measures on a per-PLD basis. Considering the RDF/XML data provided by each PLD as an independent population, we have a varying degree of coverage for each sampling frame. Our results will generally be more accurate for smaller PLDs (for which we have a higher relative coverage), and less accurate for larger PLDs (for which we have a lower relative coverage). Where possible, we present ratio-based conformance measures and other forms of measures that we argue are *less* sensitive to the level of coverage in the sample for a given PLD. Where pertinent, we later discuss possible biases given by our sampling for specific conformance measures.

A more difficult question relates to how the statistics for individual data-providers should be aggregated into an overall conformance score for each guideline. One option is to take the average (i.e., arithmetic mean) of conformance scores for all providers; however, this would assign equal weight to the conformance of, e.g., the high-volume and highly-prominent `dbpedia.org` domain, and the low-volume, obscure `phoenixproductions.org.uk` domain, which publishes a single triple. [18] There is perhaps no single ideal aggregation. For the purposes of Section 5, we only present statistics relating to those 188 data-providers (24.2%) contributing more than 1,000 quadruples to our sample, which is the same cut-off used for datasets to be

[18] `http://www.phoenixproductions.org.uk/newsbomb/index.rdf`; displays hacked notice, 2011/08/15.

included in the LOD cloud.[19] When presenting aggregate scores, we use the arithmetic mean and population standard deviation of conformance across only these 188 providers. Later in Section 7, we present and compare other methods of aggregating the per-PLD conformance scores into an overall score for each guideline.

Again, our corpus constitutes a large collection of documents sampled from a wide variety of Linked Data publishers, and so should yield interesting insights into conformance; it is the base dataset indexed by the SWSE system at the time of writing [56]. We now present statistics that further characterise the particular contents of our corpus.

4.1.2 *Corpus summary*

In Table 1, we present the top twenty-five data providers contributing to our corpus, with respect to the number of quadruples and documents—we extracted the PLDs from the source documents (contexts) and summated occurrences. We see that a large portion of the data is sourced from social networking sites—such as `hi5.com` and `livejournal.com`—that host FOAF exports for millions of users. Notably, the `hi5.com` domain provides 595 million (53.2%) of all quadruples in the data: although the number of documents crawled from this domain was comparable with other high yield domains, the high ratio of triples per document meant that in terms of quadruples, `hi5.com` provides the majority of data. Other providers in the top-five include the `opiumfield.com` domain, which offers LastFM exports; as well as `linkedlifedata.com` and `bio2rdf.org`, which publish data from the life-science domain.

With respect to the nature of the data that these providers contribute to our corpus, we now look at usage of properties and classes in the data. The dominance of `foaf:*` terms for raw triple counts is attributable (in large part) to the high-percentage of data from the `hi5.com` domain.

For properties, we analysed the frequency of occurrence of terms in the predicate position, and for classes, we analysed the occurrences of terms in the object position of `rdf:type` quads. We found 23,155 unique predicates, translating into an average 48,367 quads per predicate; Table 2 gives the listing of the top 25 predicates, where (unsurprisingly) `rdf:type` heads the list (18.5% of all quads), and where `foaf:*` properties also feature prominently.

[19] http://lod-cloud.net/; retr. 2011/09/01

№	PLD	quads (m)	docs (k)	quads/doc
1	hi5.com	595.1	255.7	2,327
2	livejournal.com	77.7	56.0	1,387
3	opiumfield.com	66.1	272.2	243
4	linkedlifedata.com	54.9	253.4	217
5	bio2rdf.org	50.6	227.3	223
6	rdfize.com	38.1	161.9	235
7	appspot.com	28.7	49.9	576
8	identi.ca	22.9	65.2	351
9	freebase.com	18.6	181.6	102
10	rdfabout.com	16.5	135.3	122
11	ontologycentral.com	15.0	1.1	14,798
12	opera.com	14.0	82.8	170
13	dbpedia.org	13.1	144.9	91
14	qdos.com	11.2	14.4	782
15	l3s.de	8.3	163.2	51
16	dbtropes.org	7.4	34.0	217
17	uniprot.org	7.3	11.7	625
18	dbtune.org	6.2	181.0	34
19	vox.com	5.3	44.4	120
20	bbc.co.uk	4.2	262.0	16
21	geonames.org	4.0	213.1	19
22	ontologyportal.org	3.5	0.002	1,741,740
23	ordnancesurvey.co.uk	2.9	43.7	66
24	loc.gov	2.5	166.7	15
25	fu-berlin.de	2.5	135.5	18

Table 1
Top twenty-five PLDs and number of quads and documents they provide.

Analogously, we found 104,596 unique values for `rdf:type`, translating into an average of 1,977 `rdf:type` quadruples per class term; Table 2 gives the listing of the top twenty-five classes, where again FOAF—and in particular `foaf:Person` (79.2% of all `rdf:type` quads)—features prominently.

In order to get an insight into the most instantiated vocabularies, we extracted the "namespace" from predicates and URI-values for `rdf:type`: we simply strip the URI upto the last hash or slash. Table 2 also gives the top twenty-five occurring namespaces for a cumulative count, where FOAF, RDFS, and RDF dominate; in contrast, Table 2 also gives the top twenty-five namespaces for unique URIs appearing as predicate or value of `rdf:type`, where in particular namespaces relating to DBPedia, Yago and Freebase offer a diverse set of *instantiated* terms; note that (i) the terms need not be defined in that namespace (e.g., `foaf:tagLine` used by LiveJournal) or may be misspelt versions of defined terms (e.g., `foaf:image` used by LiveJournal instead of `foaf:img` [55]), and (ii) 460 of the 489 terms in the `rdf:` namespace are predicates of the form `rdf:_`*n*.

№	predicate	triples	class	triples	namespace	triples	namespace	terms
1	rdf:type	206,799,100	foaf:Person	163,699,161	foaf:	615,110,022	yagor:	41,483
2	rdfs:seeAlso	199,957,728	foaf:Agent	8,165,989	rdfs:	219,205,911	yago:	33,499
3	foaf:knows	168,512,114	skos:Concept	4,402,201	rdf:	213,652,227	dbtropes:	16,401
4	foaf:nick	163,318,560	mo:MusicArtist	4,050,837	b2r:/b2rr:	43,182,736	fb:	14,884
5	b2rr:linkedToFrom	31,100,922	foaf:PersonalProfileDocument	2,029,533	lldpubmed:	27,944,794	dbp:	7,176
6	lldegene:pubmed	18,776,328	foaf:OnlineAccount	1,985,390	lldegene:	22,228,436	own16:	1,149
7	rdfs:label	14,736,014	foaf:Image	1,951,773	skos:	19,870,999	semwebid:	1,024
8	owl:sameAs	11,928,308	opiumfield:Neighbour	1,920,992	fb:	17,500,405	opencyc:	937
9	foaf:name	10,192,187	geonames:Feature	983,800	owl:	13,140,895	estoc:	927
10	foaf:weblog	10,061,003	foaf:Document	745,393	opiumfield:	11,594,699	dbo:	843
11	foaf:homepage	9,522,912	owl:Thing	679,520	mo:	11,322,417	sumo:	581
12	lldpubmed:chemical	8,910,937	estoc:cphi_m	421,193	dc:	9,238,140	rdf:	489
13	foaf:member_name	8,780,863	gr:ProductOrServiceModel	407,327	estoc:	9,175,574	wn:	303
14	foaf:tagLine	8,780,817	mo:Performance	392,416	dct:	6,400,202	b2r:/b2rr:	366
15	foaf:depiction	8,475,063	fb:film.performance	300,608	b2rns:	5,839,771	movie:	251
16	foaf:image	8,383,510	fb:tv.tv_guest_role	290,246	sioc:	5,411,725	uniprot:	207
17	foaf:maker	7,457,837	fb:common.webpage	288,537	vote:	4,057,450	aifb:	196
18	lldpubmed:journal	6,917,754	dcmit:Text	262,517	geonames:	3,985,276	factbook:	191
19	foaf:topic	6,163,769	estoc:irt_h_euryld_d	243,074	skipinions:	3,466,560	foaf:	150
20	lldpubmed:keyword	5,560,144	owl:Class	217,334	dbo:	3,299,442	geospecies:	143
21	dc:title	5,346,271	opwn:WordSense	206,941	uniprot:	2,964,084	b2rns:	133
22	foaf:page	4,923,026	bill:LegislativeAction	193,202	estoc:	2,630,198	ludopinions:	132
23	b2r:linkedToFrom	4,510,169	fb:common.topic	190,434	lldlifeskim:	2,603,123	oplweb:	130
24	skos:subject	4,158,905	rdf:Statement	169,376	ptime:	2,519,543	esns:	127
25	skos:prefLabel	4,140,048	mu:Venue	166,374	dbp:	2,371,396	drugbank:	119

Table 2. Top twenty-five (i) predicates by number of triples they appear in; (ii) values for `rdf:type` by number of triples they appear in; and (iii) namespaces by number of triples the contained predicates or values for `rdf:type` appear in; (iv) namespaces by unique predicates or values for `rdf:type` that they contain.

4.2 Nine monthly snapshots

We now briefly discuss the nature of our nine monthly snapshots, which we later analyse in order to determine the stability with which the individual data-providers host their documents; in particular, we focus on the parameters of the crawl and the crawler.

Each month, we accessed a static set of the URIs of 155 thousand RDF/XML documents—these URIs were randomly sampled from a large crawl conducted in January 2010, and so contain a similar sample bias to that of the larger crawl. The nine monthly snapshots contain an average of 51 million quadruples each. The accessed documents are served by 850 data providers (PLDs). Of these, 457 data providers coincide with our larger crawl.

Given that we will present the stability with which different providers host data, it is perhaps important to note the specific times for crawling and the timeouts used. Each snapshot was crawled starting at 00:01 a.m. GMT on the first Sunday of each month. We carried out the crawl with the LDSpider framework. [20] The crawler uses a 128-second socket timeout, and a 64-second timeout for establishing a connection. Further, we (i) enabled Nagle's algorithm [21] , which tries to conserve bandwidth by minimising the number of segments that are sent; and (ii) enabled GZip compression (as available).

5 Best Practices for Data Providers

Linked Data principles and publishing guidelines are designed to make structured data more amenable to ad hoc consumption on the Web. However, it is currently unclear how closely RDF publishers follow these best-practices.

From [15]—up until recently, the definitive guide to publishing Linked Data on the Web as prominently promoted on the http://linkeddata.org/ (retr. 2011/09/01) site—we derive a list of fourteen concrete guidelines, and empirically evaluate their uptake with respect to our corpora. Going through this list, we first quote the advice from [15] verbatim, and discuss the rationale, feasibility and repercussions thereof. We design and briefly formalise some (typically) straightforward metrics that aim to quantify the level of conformance with respect to the given guidelines, and then present the results of some empirical analyses over our corpora that indicate how closely data-providers follow the given best practice.

We aim to have a good coverage of the broad-range of recommendations and topics covered in [15]. However, we note that we omit a couple of issues for which we found it difficult to design some quantitative experiments; for example, we do not look at the use of unique keys in URIs, or at the dereferenceability of information vs. non-information resources. Also, again we focus on issues relating to data providers, and not to vocabulary providers. Otherwise, we believe that our guidelines have good competency with respect to the discussion in [15].

Importantly, we also acknowledge that many of the best-practices we outline may not be applicable to all scenarios, and that reasonable exceptions may often apply—for example, although best-practices discourage the use of blank nodes, they may be useful for representing highly-transient resources, or perhaps for n-ary predicate constructs. [22] However, we believe that the presented recommendations apply in the general case—since we look at a significant spectrum of issues, we necessarily need to apply straightforward, objective, quantitative analysis.

We also present lists of the top five and bottom five most/least conformant domains for each guideline; full versions of all tables are available online at http://aidanhogan.com/ldstudy/; retr. 2012/01/12. When presenting statistics about specific data providers (to avoid connotations of "pointing the finger") we mark data providers with which at least one of the authors is directly *involved* (†) or with which at least one of the authors is directly *affiliated* (‡). We (humbly) note that these domains often appear towards the bottom of our conformance rankings.

Moving forward, we organise issues into categories, presenting them together in the following subsections:

(i) *naming* resources (§ 5.1);
(ii) *linking* to external data providers (§ 5.2);
(iii) *describing resources* (§ 5.3);
(iv) *dereferenced representations* (§ 5.4).

5.1 Naming

In this section, we look at best-practices relating to the naming of resources as discussed in [15,85].

[20] http://code.google.com/p/ldspider/; retr. 2011/09/01
[21] See RFC 896: http://tools.ietf.org/html/rfc896; retr. 2011/09/01
[22] See http://richard.cyganiak.de/blog/2011/03/blank-nodes-considered-harmful/ (retr. 2011/09/01.) for some informal discussion.

ISSUE I: AVOID BLANK NODES

"We discourage use of blank nodes. It is impossible to set external RDF links to a blank node, and merging data from different sources becomes much more difficult when blank nodes are used." —**[15, § 2.2]**

What? Blank-node identifiers are local to a given document, and thus cannot be externally referenced. The above quote recommends minimal usage of blank nodes in Linked Data publishing.

Why? Primarily, Linked Data best-practices emphasise the importance of interlinking and re-using names across domains, whereby use of blank nodes would pose obvious problems.

Further, classical RDF semantics mandates an existential interpretation of blank nodes [49] not well-supported by Linked Data tools (or arguably even understood by adopters), where, for example, the current RDF semantics of blank nodes does not align well with SPARQL, which interprets them as names [70].

In addition, when merging documents, local blank-node labels must be mapped to globally unique labels. [23]

Conformance? We see minimal (or no) use of blank nodes as a general indicator of Linked Data conformance. Along these lines, we use the following metric to determine conformance for a PLD p, where a higher percentage is interpreted as having higher conformance (here recalling notation from § 3):

$$\neg\mathsf{bn}(p) := \frac{|\mathsf{dlc}(p) \cap \mathbf{U}|}{|\mathsf{dlc}(p) \cap (\mathbf{U} \cup \mathbf{B})|}$$

where $\mathsf{dlc}(p)$ is a shortcut denoting the set of data-level constants appearing in the data hosted by the PLD p. Here, $\neg\mathsf{bn}(p)$ (*not blank node*) gives the ratio (expressed as a quotient) of the set of unique data-level URIs vs. the set of unique data-level URIs and blank nodes in data hosted by the PLD p—here we exclude literals.

In Table 3, we present the top-five and bottom-five data providers with respect to $\neg\mathsf{bn}$ (expressed as a percentage). We found that 64 data-providers (34% of 188 offering more than 1,000 quads) did not use any blank nodes, where Table 3 only enumerates the five largest such providers with respect to the total quads in our sample. Further, 86 providers (45.7%) used less than <1% blank nodes. The average score for $\neg\mathsf{bn}$ across the 188 data-providers included was 84.3% (±24.2 pp.) [24]

№	PLD	¬bn [%]
1	linkedlifedata.com	100
...	appspot.com	100
...	dbpedia.org‡	100
...	l3s.de	100
...	dbtropes.org	100
184	okkam.org	20.2
185	opencalais.com	17.8
186	hi5.com	9.6
187	ontologycentral.com†	2.0
188	prefix.cc†	0.2

Table 3
Top five PLDs and bottom five PLDs ordered by percentage use of URIs vs. blank nodes (ties ordered by number of quads in sample)

Bias? Regarding the above results, the high-level sampling biases discussed in Section 4.1.1 again apply (e.g., not considering RDFa data, etc.). Regarding sampling biases specific to the above measures, we note that documents that contain a high percentage of blank nodes may be less likely to be crawled since they contain less dereferenceable URIs (and thus, less opportunities to be linked). Perhaps more importantly, given that local URIs can be re-used across documents whereas blank nodes cannot, the ratio of blank nodes vs. local URIs may increase when more documents are available for analysis; for example, if a domain publishes a single consistent local URI and a single unique blank-node in each document it hosts, the ratio of blank nodes will increase linearly as the number of documents considered increases. Thus, for the very large domains that we only partially sample, the ratio of blank nodes may be under-represented assuming significant re-use of local URIs across documents.

Conclusion? Given the high standard deviation (24.2 pp), we can still see that conformance to this guideline varies widely across domains. However, although a number of high-volume publishers still make heavy use of blank nodes—thus ensuring their prevalence in absolute terms—we have seen that most domains make relatively sparse (or no) use of blank nodes.

There are various possible valid reasons for using blank nodes, including use for transient items that only

[23] This is not necessarily an expensive process: in our case, we use an escaped concatenation of context and the local blank-node label to generate a global ID.

[24] We use '±' to indicate popuation standard deviation. 'pp' indicates percentage point units.

exist at request time, or for resources that should not be externally referenced, or as shortcuts for representing *n*-ary predicates in RDF, or use for serialising certain OWL axioms, etc. However, in the general case, avoiding blank nodes makes RDF Web data better subject to interlinking and re-use. Traditionally, blank nodes were heavily used to identify non-information resources, particularly by high-volume publishers of FOAF data [54], where we can still see the effects of this practice in RDF published today (e.g., `hi5.com`).

In fact, the recently reconvened RDF W3C Working Group has been discussing the possibility of specifying an informative, agreed-upon mechanism for converting (aka. *Skolemising*) blank nodes into unique URIs [70]. [25] This would allow for legacy blank nodes to be converted, serialised and consumed as URIs by software agents. Current proposals centre around the use of `.well-known` URIs with a reserved path prefix [78].

ISSUE II: USE HTTP URIS

> *"In the context of Linked Data, we restrict ourselves to using HTTP URIs"* —**[15, §2.1]**

What? The above quote recommends only using URIs with the `http://` or `https://` schemes, and thus avoiding other URI schemes, such as `ftp:`, `file:`, `mailto:`, `urn:`, `info:`, etc.

Why? Unlike blank nodes, URIs give a direct mechanism for globally identifying a given resource. In addition, HTTP URIs are compatible with the identification of resources with respect to Web Architecture principles [69], such that (related) representations of the referent can be returned by means of a HTTP lookup [34]. (We will see more in the next issue.)

Conformance? We see a high percentage use of HTTP URIs as a general indicator of Linked Data conformance, where we use the following metric to quantify this conformance:

$$\mathsf{hu}(p) := \frac{\left|\mathsf{dlc}(p) \cap \{u \in \mathbf{U} : \mathsf{sch}(u) \in \{\texttt{http}, \texttt{https}\}\}\right|}{|\mathsf{dlc}(p) \cap \mathbf{U}|}$$

where $\mathsf{sch}(u)$ denotes the URI scheme of u, and where $\mathsf{hu}(p)$ (*HTTP URIs*) represents the ratio of unique URIs appearing in a data-level position of a triple in $\mathsf{data}(p)$ that have the `http` or `https` scheme. Note that we do not count blank nodes here since they are accounted for by the previous metric.

In Table 4, we present the top-five and bottom-five data providers with respect to hu (represented as a percentage). We found that 112 data-providers (60% of 188 providers hosting more than 1,000 quads) did not use any non-HTTP URIs, where Table 4 only enumerates the five largest such providers (with respect to total quads hosted). Further, we note that 162 data-providers (86.2%) used >99% HTTP URIs. The average percentage use of HTTP URIs was 98.8% (±4.8 pp).

№	PLD	hu [%]
1	`hi5.com`	100
...	`linkedlifedata.com`	100
...	`rdfize.com`	100
...	`identi.ca`	100
...	`freebase.com`	100
184	`code4lib.org`	87.4
185	`gregheartsfield.com`	81.4
186	`fluffyandmervin.com`	70.4
187	`smhowell.net`	68.8
188	`loc.gov`	56.7

Table 4
Top five PLDs and bottom five PLDs ordered by percentage use of HTTP URIs (when tied, ordered by number of quads)

Bias? Aside from the high-level biases already discussed—and as per the use of blank nodes in the previous guideline—documents with high percentages of non-HTTP URIs are perhaps less likely to be crawled due to a lack of dereferenceable names.

Conclusion? We have seen that most domains surveyed are highly-conformant with this guideline. The most significant counter-example to this trend is the `loc.gov` domain, which assigns each locally minted URI an alias with the `info:` scheme; in fact, this case is not problematic given that a `http:` alias is available for all resources.

Despite the guideline, there are valid reasons to use non-HTTP URIs, esp. for identifying legacy resources, where schemes like `mailto:` and `tel:` can be used to directly indicate email and telephone numbers respectively, and where many information resources are identified/accessible through an `ftp:` scheme URI. Instead, the guideline is implicitly encouraging new identifiers to be minted with the `http:` scheme (as opposed to, e.g., using URN schemes), which, in particular, enables

[25] cf. `http://www.w3.org/2011/rdf-wg/track/issues/40`; retr. 2011/08/10

information about the resource to be dereferenced, as per the next guideline.

Issue III: Mint dereferenceable URIs

> *"Define your URIs in an HTTP namespace under your control, where you actually can make them dereferenceable."* —**[15, §3]**

> *"When publishing Linked Data on the Web, we represent information about resources using the Resource Description Framework (RDF)."* —**[15, §2.2]**

> *"Your data source should at least provide RDF descriptions as RDF/XML which is the only official syntax for RDF."* —**[15, §5]**

What? As alluded to by the previous guideline, HTTP URIs can be dereferenced by means of a HTTP lookup. In the context of Linked Data, we would expect some RDF representation to be returned as discussed in the second quote above. Given the third quote, we would also expect data to be returned in RDF/XML format (and then optionally in Turtle or TriX, etc.[15, §5]).

Why? When a URI identifying some resource is looked up, a consumer should reasonably respect a (related) representation thereof to be returned. Given that applications are consuming Linked Data in an ad hoc manner, such applications require structured data to be provided in a known, standardised fashion. Again, in the context of Linked Data, RDF provides the core, interoperable data-model. RDF/XML is traditionally the most widely supported RDF syntax, although we again acknowledge that RDFa has enjoyed recent growth in adoption.

We note that many Linked Data systems rely on dereferenceable URIs being used in the data. First, use of dereferenceable URIs is important for locating information about resources, useful for processing SPARQL queries over a priori unknown data sources (e.g., see [47]), for "Linked Data browsers", which allow for navigating the Web of Data through dereferenceable URIs (e.g., see [10]), etc. [26] Similarly, dereferenceability establishes an important relationship between resources and their authoritative representations, often used as an indicator of provenance or trustworthiness of the information in a specific source with respect to a specific resources, used in applications such as ranking (e.g., see [45]) or reasoning (e.g., see [57,20]). Finally, performing HTTP lookups on non-derefereneable URIs can cause significant wasted computation-time for agents, especially Web crawlers used in warehousing approaches, which may perform many millions of lookups, and live-querying and browsing systems, which must retrieve sources at query-time.

[26] In the context of Linked Data browsers, a non-dereferenceable URI equates to a dead-link on the traditional HTML Web.

Conformance? We consider providers that mint a high ratio of local URIs that dereference to RDF/XML content (using `Accept: application/rdf+xml`) as highly conformant. Note however that our crawl is incomplete, where we do not perform lookups on all URIs in the corpus. Thus, herein we restrict our analyses to look at the percentage of URIs that were confirmed *not* to be dereferenceable, where we would expect conformant data providers to mint fewer non-dereferenceable URIs; more formally, for a PLD p, we measure:

$$\mathsf{du}(p) := 1 - \frac{|\mathsf{ldlc}(p) \cap \{u \in U : \mathsf{deref}(u) = \emptyset\}|}{|\mathsf{ldlc}(p) \cap \mathbf{U}|}$$

where $U \subset \mathbf{U}$ is the set of HTTP URIs looked up during the crawl of our corpus, and $\mathsf{ldlc}(p) := \mathsf{dlc}(p) \cap \mathsf{local}(p)$ denotes the set of local, data-level constants in the data hosted by p. For a PLD p, a lower ratio of confirmed non-dereferenceable URIs results in a higher value for $\mathsf{du}(p)$ indicating better conformance.

We found three domains that did not mint any local URIs ($\mathsf{ldlc}(p) \cap \mathbf{U} = \emptyset$): `hopcroft.name`, `lehigh.edu`, and `unitn.it`. We exclude these three domains from tables that have local URIs as a denominator, and consider them as having a score of zero when calculating averages or orderings.

Along these lines, Table 5 presents the top five and bottom five data-providers with respect to du conformance (represented as a percentage). We note that no non-dereferenceable URIs were found for 14 providers (7.4%); in total, 36 PLDs (19.1%) have less than 1% of their local URIs confirmed as non-dereferenceable. The average score for du was 70.3% (±26.8 pp) across the 188 data-providers.

However, we note that this metric and these results do not consider the amount of data returned about the given URI in the dereferenceable document; hence, we also look at another metric, as follows:

$$\mathsf{dt}(p) := \frac{\sum_{u \in DU_p} |\{t \in \mathsf{deref}(u) : u \in \mathsf{dlc}(\{t\})\}|}{|DU_p|}$$

where

$$DU_p := \{u \in U : \mathsf{deref}(u) \neq \emptyset \wedge u \in \mathsf{ldlc}(p)\}$$

№	PLD	du [%]
1	ontologycentral.com†	100
...	zitgist.com	100
...	zbw.eu	100
...	ebusiness-unibw.org	100
...	174.129.12.140 (open-biomed.org.uk)	100
181	br3nda.com	21.4
182	ajft.org	19.6
183	smhowell.net	18.1
184	snell-pym.org.uk	16.7
185	typepad.com	7.8

Table 5
Top five PLDs and bottom five PLDs ordered by percentage of locally used URIs that were not found to be non-dereferenceable (when tied, ordered by number of quads)

№	PLD	dt [triple]
1	prefix.cc†	417.0
2	bio2rdf.org	247.7
3	linkedlifedata.com	214.8
4	br3nda.com	180.3
5	dbpedia.org‡	69.5
181	bestbuy.com	2.1
182	lingvoj.org	2.0
183	livejournal.com	2.0
184	opiumfield.com	1.1
185	hi5.com	0

Table 6
Top five and bottom five PLDs ordered by average number of triples mentioning dereferenceable URI in resp. dereferenced documents

denotes the set of URIs for a PLD p that (i) are mentioned in the data of p and (ii) were looked up and found to dereference to RDF/XML data during our crawl. Thus, for each PLD p, the dt metric takes the average across all $u \in DU_p$, of the number of triples mentioning u (in a data-level position) in the dereferenced document of u.

In Table 6, we give the top five and bottom five PLDs for the dt measure. We note that the prefix.cc domain only had two dereferenceable (information resource) URIs, denoting the two documents found on that domain, where each document had a large set of foaf:topic outlinks. [27] Documents in the bottom half of the table typically only had dereferenceable information resources (the documents themselves). For example, the hi5.com domain only hosts dereference document URIs, where every document has links to external RDF/XML documents, but has no mention of itself. Similarly, each document on livejournal.com only speaks about itself in two triples. [28]

The average value for dt across all documents was 17.5 (± 40.3) triples: the high standard deviation tells us that the outliers at the top of the table have a strong effect on the average.

Note that we further analyse the information dereferenced by different domains later in Section 5.4.

Bias? Besides the high-level sampling biases already discussed, it is important to note that we would consider URIs that dereference (only) to RDFa as non-dereferenceable: we do not detect RDF embedded in HTML documents. Further still, since we only partially crawl the local URIs of large data-providers, the ratio of confirmed non-dereferenceable URIs would be under-represented for these domains. It is also worth noting that documents with few or no dereferenceable URIs are less likely to be well-linked and thus, again, less likely to be crawled.

Conclusions? We have seen that although many publishers largely abide by the dereferenceable-URIs guideline (average of 70.3%), there is still some notable variability in conformance (standard deviation of 26.8 pp).

Again, legacy information resources on the Web are most naturally identified using their native URL. For certain local resources referenced in RDF data—e.g., online spreadsheets, images, etc.—it is often infeasible to make their URIs dereference to a valid RDF description; similarly, embedding RDFa into certain HTML documents may not be feasible or currently cost-effective. Thus, despite the guideline, it is often infeasible to make all (local) URIs dereference to RDF. Indeed, the prohibitive cost involved in, e.g., embedding RDFa metadata into the legacy HTML content of established web-sites, or maintaining content-negotiation and redirect schemes, etc., might discourage potential adopters if the guideline were enforced more rigorously. In addition to overhead, prior to Linked Data principles, making "RDF URIs" dereferenceable was not a priority. The earliest recommendations relating to dereferenceability were specific to class and property terms published by vocabularies [74], where older RDF Web data may still feature sparse use of dereferenceable URIs.

[27] cf. http://prefix.cc/popular/all.file.vann and http://prefix.cc/rdf,owl,foaf,dc.file.vann; retr. 2010/08/22.
[28] This has since changed; LiveJournal FOAF files now do not host *any* information about themselves, although they offer links to external documents; cf. http://danbri.livejournal.com/data/foaf; retr. 2011/08/23.

Issue IV: Keep URIs short

"Keep implementation cruft out of your URIs. Short, mnemonic names are better." —**[15, §3]**

What? The above recommendation recommends avoiding, for example, URIs that contain query parameters, or that are very long, instead preferring short, human readable URIs. Many Web server solutions offer URI rewriting engines that enable mapping from longer, low-level implementational URIs to short, mnemonic URIs.

Why? On the Web, humans must often deal directly with URLs, keying them into browser address bars, memorising the locations of commonly accessed pages, advertising the web-site of a company, etc. Thus, URLs are not solely designed to be computer-processable addresses, but are also purposefully designed to be human-cognisable—for example, mnemonic domain names are used to represent numerical IP addresses. Following the same rationale, the use of mnemonic URIs in Linked Data is explicitly encouraged in [15].

Further, the prevalent use of shorter URI strings offers some obvious benefits for large-scale and/or frequent processing of RDF data; for example, short URIs allow for (i) smaller on-disk indexes, allowing for shorter disk reads; (ii) storing more data in main memory, translating into larger caches (e.g., see [71]) and more scalable in-memory applications; (iii) efficient compression techniques for further reducing memory-footprint (e.g., see [72,33]); (iv) faster serialisation of RDF data, requiring less bandwidth and reducing latency; etc.

Conformance? We deem data providers that locally mint (on average) shorter URIs as being generally more compliant with Linked Data best practices. Along these lines, we use the following measure to quantify conformance:

$$\overline{\mathsf{ul}}(p) := \frac{\sum_{u \in \mathsf{ldlc}(p) \cap \mathbf{U}} \mathsf{len}(u)}{|\mathsf{ldlc}(p) \cap \mathbf{U}|}$$

where $\mathsf{len}(u)$ is the character length of the URI u; i.e., $\overline{\mathsf{ul}}(p)$ gives the average length of URIs local to a PLD p.

Table 7 presents the top five and bottom five data-providers with respect to $\overline{\mathsf{ul}}$ as observed in our large corpus (excluding the three that did not define any local URIs). The mean average-length of local URIs across the 185 PLDs surveyed was 52.4 (±16.4) characters.

№	PLD	$\overline{\mathsf{ul}}$ [char.]
1	gromgull.net	25.6
2	4july.me	26.3
3	urmf.net	28.3
4	chirup.com	31.1
5	waka.me	32.6
181	idehen.name	95.1
182	rkbexplorer.com	96.3
183	dividing.com	101.0
184	uniba.it	104.3
185	nuigalway.ie‡	113.6

Table 7
Top five PLDs and bottom five PLDs ordered by average length of local URIs

Conclusions? The average length of local URIs varies by a notable factor of ~4× in the sample of data-providers analysed (i.e., between `nuigalway.ie` and `gromgull.net`).

Although shorter URIs do enable more efficient indexing and serialisation, longer URIs composed of recognisable patterns or words—e.g., a well-structured directory scheme or a full resource label—may often be mnemonically preferable, or better indicate the resource they identify, than shorter URIs—e.g., an authority followed by a trailing nine-digit number as commonly produced by URL shorteners. Of course, the guideline is more concerned with avoiding excessive URI length [29] than making URIs as short as possible.

Issue V: Host stable URIs

"Try to keep your URIs stable and persistent. Changing your URIs later will break any already-established links, so it is advisable to devote some extra thought to them at an early stage." —**[15, §3]**

What? The above quote advises against the use of transiently/intermittently dereferenceable URIs. Once dereferenceable URIs are minted, they should be kept dereferenceable over time (even if the underlying redirects and/or RDF content are dynamic).

Why? As per the importance of dereferenceability, URIs should also be stable over time: URIs that are only temporarily or intermittently dereferenceable—or that identify different non-information resources over

[29] As exemplified by the following document on our home university server: http://rss.library.nuigalway.ie/rdf/Medicine-new-books.rdf; retr. 2011/08/16.

time—damage previous efforts at external linking and mapping. As such, unstable URIs can seriously harm the performance of agents and applications [81] and the reliability of query answers from search engines or live query processors. Similarly, the connectivity of the Web of Data (and the reachability of its various subsets) is heavily dependant on the stability of resources with a high (betweenness) *centrality* [39].

Conformance? We deem data providers that maintain a higher percentage of stable URIs (minted locally) as generally being more conformant to Linked Data best practices. Along these lines, we use the following metric of conformance for stability of documents:

$$\overline{\mathsf{st}}(p, \mathcal{S}) := \frac{|\{(s,i) : s \in S_i \in \mathcal{S} \wedge \mathsf{pld}(s) = p\}|}{|\mathcal{S}| \times |\{s \in S \in \mathcal{S} : \mathsf{pld}(s) = p\}|}$$

where $\mathcal{S} \in 2^{\mathbf{S}}$ denotes a collection of sets of sources from which RDF graphs were successfully retrieved—as such, $\mathcal{S}$ represents our monthly snapshots of documents. Intuitively, $\overline{\mathsf{st}}$ represents the average number of appearances of local sources for p in the snapshots (i.e., sources that appeared at least once in one of the snapshots). [30]

Along these lines, Table 7 presents the top five and bottom five data-providers with respect to the $\overline{\mathsf{st}}$ measure for the respective domain for our nine-month crawl (when tied, ordered thereafter by number of quads hosted). We only consider documents that appeared at least once in the snapshot. Note again that we only have monthly information available for 141 (75%) of the 188 providers hosting more than 1,000 quads. We found that 65 of these providers (46%) had an average availability of 100%, and that 75 (53%) had an average availability in excess of 99%. The lowest data provider hosted one document in one snapshot. The mean availability of documents was 88.8% (±19.4 pp) across all data-providers.

Bias? Besides the high-level bias, one possible concern with this analysis is the low number of observations available for each document (i.e., nine) and the large interval between observations (i.e., per month). For example, a domain could regularly experience capacity problems one whole day each month (~96% uptime), which we would have only a $\frac{9}{30}$ probability of encountering in one of our monthly snapshots. The presented figures thus serve as an informative indicator of medium-term stability.

[30] Note again that the list of URIs we attempt to retrieve in each snapshot is static.

№	PLD	$\overline{\mathsf{st}}$ [%]
1	ontologyportal.org	100
...	ordnancesurvey.co.uk	100
...	fao.org	100
...	kit.edu‡	100
...	nytimes.com	100
137	4july.me	36.4
138	reshouts.com	32.5
139	deri.ie‡	26.5
140	kaufkauf.net	11.8
141	ourcoffs.org.au	11.1

Table 8
Top five and bottom five PLDs ordered by average percentage availability of documents for our nine monthly snapshots (ordered thereafter by number of quads)

Conclusions? The average stability of documents being hosted across the nine snapshots was relatively high, at 88.8%. One may note however that a more granular analysis with more frequent snapshots may yield different results (we plan on gathering such a corpus as future work).

There are few if any good reasons to host unstable URIs. However, in reality there are currently few (if any) revenue streams available through Linked Data publishing, leading to less server resources and inevitably less emphasis on quality-of-service. In addition, Linked Data consumers are sometimes naïve/impolite with respect to their demands on data providers, where prominent publishers such as `dbpedia.org` (with a stability of 86% in our analysis) receive high levels of traffic, and must carefully implement triple-limits for SPARQL queries and dereferenced documents to keep services running. Conversely, many Linked Data sites are hosted on stable, high-bandwidth, university servers.

In any case, instable URIs are to be expected in Linked Data, especially as it expands and diversifies. Link monitoring and maintenance frameworks such as DSNotify [81] should help attenuate the problem of URI instability by monitoring when remote resources are created, removed, changed, updated or moved, and revising links to these resources accordingly.

5.2 Linking

Herein, we now discuss conformance with respect to how data providers provide external links to other data providers (note that we do not examine internal interlinkage). Again, we continue to follow best practices extracted from [15].

ISSUE VI: USE EXTERNAL URIS

> *"[...] the most valuable RDF links are those that connect a resource to external data published by other data sources, because they link up different islands of data into a Web. Technically, such an external RDF link is a RDF triple which has a subject URI from one data source and an object URI from another data source."* —**[15, §2.2]**

What? Data providers are encouraged to provide a diverse set of URIs that dereference to external Linked Data domains, effectively providing links to remote data.

Why? Defining RDF links to external providers allows data consumers to serendipitously *discover* related information on the Web, be it in a (semi-)automated manner as performed by crawlers, or in a direct manner as performed by users of Linked Data browsers.

In fact, the principle aesthetic of Linked Data—as its name suggests—is the importance of well-interlinked data. Not only do links connect together islands of information, but self-organising phenomena—such as preferential attachment [3]—bring an inherent structure to the resulting network, where the most in-demand nodes become the most heavily connected, etc. The resulting structure is then amenable to various analyses—such as those discussed in § 6—that allow for identifying the "importance" of various nodes in the graph.

Conformance? We deem data providers that offer a higher outdegree of RDF links to external (RDF) data providers as being, in general, more conformant with respect to Linked Data best-practices. We count the number of external links as follows:

$$\mathsf{el}(p) := |P \cap \mathsf{plds}(\mathsf{dlc}(p) \cap \mathbf{U}) \setminus \{p\}|$$

where P is the set of 778 PLDs providing RDF to our corpus. In other words, for a PLD p, $\mathsf{el}(p)$ counts the number of unique external PLDs linked from a data-level position in the data hosted by p; only links to PLDs found to host RDF are counted. We see a higher value of el as denoting better conformance with respect to the stated guideline.

In Table 9, we present the top five and bottom five PLDs with respect to el; again, we only consider links to providers that were confirmed to host RDF in our crawl. In total, in our corpus, we found five PLDs that did not provide links to any external PLD. The 188 data-providers analysed linked to an average of 20.4 (±38.2) external PLDs.

№	PLD	el [PLD]
1	identi.ca	300
2	status.net	191
3	soton.ac.uk	167
4	semanticweb.org	147
5	appspot.com	129
184	semantic-web-grundlagen.de	0
...	prefix.cc†	0
...	opiumfield.com	0
...	hi5.com	0
...	fgiasson.com	0

Table 9
Top five PLDs and bottom five PLDs ordered according to the number of external PLDs they link to

Bias? Since this measure is not based on a ratio or other form of quotient—but instead an absolute count—our results would under-represent the level of external links for domains that are only partially sampled, particularly those with very diverse link-sets. In other words, the more documents analysed, the more external links are likely to be found. However, by counting links on the level of RDF domains, we believe that the absolute count would plateau more quickly than counting, e.g., the number of external documents linked. Also, since we omit links to PLDs for which we did not find RDF/XML data, we may under-represent the level of links to external domains providing RDFa.

Conclusions? Although the absolute figures here are difficult to interpret (how many externally linked domains are "enough"?), we can see that there is very high variability in terms of the level of external linking on different domains, highlighted by a standard deviation (38.2 PLDs) which is greater than the average (20.4 PLDs). Those domains featuring diverse links to external RDF domains are typically collaborative platforms (e.g., `semanticweb.org` hosts a Semantic MediaWiki platform [67]) or offer some form of centralised service (e.g., `identi.ca` and `status.net` act as central hubs for an open-source micro-blogging platform, linking to installations on other sites).

High-quality links between remote data providers are crucial to realising the "Linked Data vision". However, in the general case, creating high-quality links to external RDF providers is often a challenging task for publishers. Along these links, link-generation frameworks

and tools are important to see conformance to this guideline realised in practice. One such proposal is SILK [96], which allows publishers to specify declarative criteria by which two datasets should be linked; once the criteria have been defined, they can be re-executed intermittently to refresh the interlinkage between the two providers.

ISSUE VII: PROVIDE owl:sameAs LINKS

> *"It is common practice to use the owl:sameAs property for stating that another data source also provides information about a specific non-information resource. An owl:sameAs link indicates that two URI references actually refer to the same thing. Therefore, owl:sameAs is used to map between different URI aliases [...]"* **—[15, §6]**

What? The `owl:sameAs` property is used to directly relate two URIs aliases: i.e., URIs that are coreferent. As such, `owl:sameAs` denotes a form of equality between resources, and has a corresponding transitive, symmetric, and reflexive semantics [52]. Linked Data best-practices encourage publishers to specify `owl:sameAs` relations between local resources and known URI aliases, particularly to URIs minted in another domain.

Why? Linked Data principles mandate use of dereferenceable URIs to identify resources (ISSUE III); now, if two different data providers wish to contribute information about the same resource, they must mint separate URIs to ensure this dereferenceability. Thereafter, `owl:sameAs` links can be used between the two URIs to specify that they denote the same resource. From another perspective (and assuming correct usage) an `owl:sameAs` link states that an agent can find more information about the given resource under the given URI alias by dereferencing that URI alias.

These relations can be used by live Linked Data browsers to (possibly semi-automatically) pull in additional remote information about a given resource. Additionally, various warehousing systems use (and/or generate) `owl:sameAs` relations to *consolidate* or *smush* local data [94,88,87,37,59,53,58], unifying the information about a given resource—specified by different data providers under different URI aliases—under a canonical identifier, thus effectively integrating the different data contributions about that resource.

Conformance? We deem data providers that offer a higher outdegree of `owl:sameAs` links to external providers as being more conformant to Linked Data principles. Along these lines, we use a similar metric as for the previous issue, but restricted to `owl:sameAs` links:

$$\mathsf{el}^{=}(p) := |P \cap \mathsf{plds}(\mathsf{dlc}(\mathsf{sa}(p)) \cap \mathbf{U}) \setminus \{p\}|$$

where $\mathsf{sa}(p)$ is the set of triples with the predicate `owl:sameAs` hosted by p. This is equivalent to the previous measure, but restricted to the set of `owl:sameAs` links (external URIs in the subject and object of such triples are counted).

Theraftер, Table 10 enumerates the top five and bottom five providers in terms of hosting `owl:sameAs` links to external providers. Firstly, of the 188 PLDs analysed, 56 PLDs (29.8%) had an `owl:sameAs` link to some external PLD also contributing RDF data to our corpus. [31] Each provider offered `owl:sameAs` links to an average of 1.79 (±5.19) external PLDs (the high population standard deviation indicates that a small number of PLDs dominate the average).

№	PLD	$\mathsf{el}^{=}$ [PLD]
1	harth.org†	41
2	uriburner.com	39
3	revyu.com	21
4	deri.org‡	20
5	semanticweb.org	18
57...	appspot.com	0
...	linkedlifedata.com	0
...	opiumfield.com	0
...	livejournal.com	0
...	hi5.com	0

Table 10
Top five PLDs and bottom five PLDs ordered according to the number of external PLDs they link to using `owl:sameAs` (and thereafter, by number of quads)

Bias? As per the previous guideline, the $\mathsf{el}^{=}(p)$ measure is not a quotient, but an absolute count. Thus again, our results may under-represent the level of external links for domains that are only partially sampled and that offer diverse `owl:sameAs` links.

Conclusions? The level of `owl:sameAs` interlinkage across domains is seemingly quite low, with 29.8%

[31] Notably, `uriburner.com` had `owl:sameAs` links to a rather impressive 1,274 external providers, but only 39 were to RDF PLDs contributing to our corpus.

of the domains considered offering such links to an external RDF domain appearing in our corpus.

Compared to the previous guideline recommending the provision of generic (RDF) links, generating `owl:sameAs` links to remote domains is even more challenging given the definitive semantics of the `owl:sameAs` relation. Again, tools such as SILK [96] can be used to generate `owl:sameAs` links to remote domains; various works have explored domain-specific techniques for interlinking the URI aliases of RDF datasets (e.g., see [84,62]); other authors present best-effort mechanisms for mining URI aliases from large RDF corpora in a generic and automatic manner (e.g., see [60,58]).

Conversely, herein we have not looked at the accuracy of such `owl:sameAs` links, which is difficult to determine by automatic means. (We refer the reader to the works already mentioned in § 2.3 for more detail on this topic.)

5.3 Describing resources

In this section, we look at Linked Data best practices that discuss how the local resources of interest should be described.

Issue VIII: Avoid prolix RDF features

> *"We discourage the use of RDF reification as the semantics of reification are unclear and as reified statements are rather cumbersome to query with the SPARQL query language.* [...] *You should think twice before using RDF collections or RDF containers as they do not work well together with SPARQL.* [...] *can the information also be expressed using multiple triples having the same predicate?"* —**[15, §2.2]**

What? Various RDF primitives are discouraged in Linked Data publishing, including those that relate to (i) RDF reification, viz., the properties `rdf:subject`, `rdf:predicate`, `rdf:object` and the class `rdf:Statement`; (ii) RDF containers, viz., properties of the form `rdf:_`n ($n \in \mathbb{N}$), the property `rdfs:member` and the classes `rdf:Alt`, `rdf:Bag`, `rdf:Seq` and `rdfs:Container`; and (iii) RDF collections, viz., the properties `rdf:first`, `rdf:rest`, and the class `rdf:List`.

Why? With respect to RDF reification—speaking about triples themselves within the RDF data-model—few systems support or use this feature, and it is widely considered as cumbersome where requests have been made for its deprecation [9,32].

Similarly, RDF containers have enjoyed little uptake in the wild, with sparse support from tools. For example, the lightweight semantics of RDF containers encoded in RDFS mandates an infinite number of axiomatic triples of the form ($\forall n \in \mathbb{N}$):

```
rdf:_n a rdfs:ContainerMembershipProperty ;
       rdfs:domain rdfs:Resource ;
       rdfs:range rdfs:Resource .
```

where scalable RDFS materialisation engines are typically forced to omit such inferences [100,95,77] (for discussion, see [99]). Similarly, three classes of containers have been defined in RDF—`rdf:Bag`, `rdf:Seq` and `rdf:Alt`—but the semantics thereof have not been adequately specified. Again, tool support is sparse, and calls have been made for deprecation [9,32].

Finally, collections are perhaps the most widely adopted of the three discouraged above—most notably, various OWL axioms rely on RDF collections [52], which, importantly, can be terminated to indicate that the given set of elements is "closed". [32] However, collections require a nested structure containing linked sublists, which is cumbersome to represent in triples, and can be expensive to support in performance- or data-intensive environments.

Further, as noted in the above quote, no explicit support for any of the three features have been provided in SPARQL (other than Turtle shortcuts for RDF collections)—for example, there is no support for returning the members of arbitrary length collections, etc. [33]

Finally, we note that such primitives are typically expressed using blank nodes—which are generated from RDF/XML and Turtle shortcuts thereof—where, as per Issue I, blank nodes are expressly discouraged in Linked Data best practices.

Conformance? We deem data providers that avoid use of RDF reification, containers and collections to be more conformant to Linked Data best practices. Along these lines, we use the following metric to measure conformance with respect to this best practice:

$$\neg\mathsf{rcc}(p) := \frac{|\mathsf{data}(p) \setminus \mathbf{RCC}|}{|\mathsf{data}(p)|}$$

[32] However, Linked Data best practices implicitly discourage use of the OWL constructs that require collections [15].

[33] We note that support for such queries are indirectly covered by SPARQL 1.1 proposals pertaining to property paths [44].

where **RCC** denotes the set of all RDF triples relating to **r**eification, **c**ontainers and **c**ollections as discussed at the outset, with:

(i) a predicate from the set { `rdf:subject`, `rdf:predicate`, `rdf:object`, `rdfs:member`, `rdf:first`, `rdf:rest` } or of the form `rdf:_`n ($n \in \mathbb{N}$); *or*

(ii) the predicate `rdf:type` and an object from the set { `rdf:Statement`, `rdf:Alt`, `rdf:Bag`, `rdf:Seq`, `rdfs:Container`, `rdf:List`}.

Thereafter, ¬rcc represents the ratio of triples hosted by p (in our corpus) that are *not* of this form.

Table 11 enumerates the top five and bottom five providers in terms of not using the discouraged RDF primitives (presented in ascending order of ¬rcc, and thereafter by quadruple count). Of the 188 PLDs analysed, we note that 148 PLDs (78.7%) had no use of reification/containers/collections, whereas 167 PLDs (88.8%) had less than 1% use thereof. Each provider hosted 99.1% (±4.7 pp) of non-**RCC** triples.

№	PLD	¬rcc [%]
1	hi5.com	100
...	livejournal.com	100
...	opiumfield.com	100
...	linkedlifedata.com	100
...	rdfize.com	100
184	nuigalway.ie‡	91.3
185	sourceforge.net	91.2
186	ivan-herman.net	90.9
187	okkam.org	67.3
188	uniprot.org	47.5

Table 11
Top five PLDs and bottom five PLDs ordered according to the total percentage of triples that do not relate to RDF reification, containers or collections (ordered thereafter by number of quads)

Conclusions? Most publishers (78.7%) do not use the features of RDF discouraged by the guidelines; many of those that do only make sparse use of such features.

However, the guideline may be considered somewhat simplistic. In particular, collections offer a standardised means of specifying ordered, closed lists of items in RDF, and as such, form an important part of serialising certain OWL axioms in RDF, including union classes, intersection classes, enumerations, property chains, compound keys, pair-wise disjoint sets, etc. Although most of these OWL features are rarely used in prominent Linked Data vocabularies [50, § 4.4.3], [53], union classes and intersection classes are used in the formal definition of, e.g., the SKOS vocabulary and the Music Ontology [83], amongst others.

Issue IX: Re-use existing terms

> *"In order to make it as easy as possible for client applications to process your data, you should reuse terms from well-known vocabularies wherever possible."* —**[15, §3]**

What? Another important aspect of Linked Data is the (re-)use of declarative, extensible, shared vocabularies across the Web. The above best-practice encourages re-use of existing class and property terms—used prominently by other data-providers—as defined in de-facto agreed-upon vocabularies.

Why? Re-using well-known terms to describe resources in a uniform manner increases the interoperability of data published in this manner. Indeed, the re-use of well-known vocabularies supports not only data integration and management tasks, but is also important for Linked Data consumer applications that have tailored support for the most common vocabularies (e.g., [90]), as well as for applications that offer domain agnostic user-interfaces for browsing and querying the data (e.g., [27,56,22,91,24,13]). Otherwise, given complete disagreement on the use of vocabularies between different data-providers, consumers are faced with the crippling problem of heterogeneity with respect to how the data can be interpreted, queried and displayed [61].

Conformance? We deem data-providers that exhibit a higher-overlap (with respect to external providers) in the vocabularies used to describe their data as being more compliant with Linked Data best practices. Along these lines, we first quantify the level of overlap of class-membership terms for the local data of a provider p as:

$$\mathsf{olc}(p) := \sum_{x \in \mathsf{cmem}(p)} \left|\{p' \in P \setminus \{p\} : x \in \mathsf{cmem}(p')\}\right|$$

where P denotes the known set of PLDs, $\mathsf{cmem}(p')$ denotes the set of terms appearing in the object of a triple $t \in \mathsf{data}(p)$, where $t.obj \notin \mathbf{B}$ and $t.pred =$ `rdf:type` (class membership terms). Intuitively, olc denotes the sum of the number of external PLDs also using the local class membership term, for each such term. Analogously, we quantify the level of the overlap of local predicate-terms with external providers as follows:

$$\mathsf{olp}(p) := \sum_{x \in \mathsf{pred}(p)} \left| \{p' \in P \setminus \{p\} : x \in \mathsf{pred}(p')\} \right|$$

where $\mathsf{pred}(p')$ denotes the set of terms appearing in the predicate of some triple $t \in \mathsf{data}(p)$, but where we exclude `rdf:type` (which was used by 187/188 PLDs, the exception being `lehigh.edu`). Finally, to aggregate these two values, we use a simple summation:

$$\mathsf{olt}(p) := \mathsf{olc}(p) + \mathsf{olp}(p)$$

denoting the total overlap of vocabulary terms used to describe local data, with respect to external providers.

In Table 12, we present the top five and bottom five data-providers with respect to olt. The 188 data-providers analysed had an average overlap of 6,607 (±3,667) shared uses of a term.

№	PLD	olt [PLD × term]
1	w3.org	15,980
2	mit.edu	13,864
3	qdos.com	13,637
4	kanzaki.com	13,445
5	kasei.us	13,252
184	unitn.it	301
185	rkbexplorer.com	270
186	prefix.cc†	214
187	freebase.com	134
188	lehigh.edu	120

Table 12
Top five PLDs and bottom five PLDs ordered according to total overlap of all classes and properties with respect to use by external data providers

Bias? As per the previous measures relating to linking, the given olt metric is not based on a quotient, but on an absolute count. Thus, again, the smaller the relative sample of data we have for a PLD, the more likely its score is to be under-represented. This would be particularly true of domains with very heterogeneous documents: i.e., with high variability in the class and property terms used across individual local documents.

Conclusions? Although it is difficult to determine an optimal figure for overlap, we do see that there are non-trivial levels of re-use across different data-providers. Some exceptions do exist, however. For example, `freebase.com` is a prominent publisher of RDF, providing general-interest resource descriptions on a broad range of topics, but was found to have relatively low overlap with other domains; most of the class and property terms are minted in a local namespace, with few external properties used (viz. `owl:sameAs`, `cc:attributionURL` and RDFS terms).

To improve the amount of vocabulary overlap between different domains, it seems that two things are required: (i) the continuous proposal and promotion of new vocabularies to expand coverage; (ii) tools and search engines that enable publishers to find the correct, most widely-adopted terms for their needs. With respect to (i), web-based vocabulary editors such as Neologism [4] are an important development, promoting vocabulary development and maintenance as a community-driven process. With respect to (ii), initiatives such as "Linked Open Vocabularies" [34] that study and promote legacy vocabularies are of vital importance.

Issue X: Cherry-pick vocabularies

> *"It is common practice to mix terms from different vocabularies."* —**[15, §4]**

What? Related to the previous issue, in re-using class and property terms from legacy vocabularies (which are themselves likely to be re-used by external datasets), data providers may often have to "cherry pick" appropriate terms defined in different vocabularies and namespaces: mixing terms from different vocabularies is endorsed by Linked Data best practices.

Why? For example, many data providers need to describe information about people, where FOAF is the vocabulary of choice. Similarly, many providers may wish to describe information about online presence or users, where SIOC is the de facto agreed-upon vocabulary. For describing metadata about documents, terms from the DC vocabulary are often used. Taxonomies and tagging schemes are often represented in SKOS, etc.

Following the same rationale as for the previous issue—where Linked Data best practices encourage data providers to use the same terms to specify similar information about resources in an agreed-upon manner—publishers are herein encouraged to re-use terms wherever available. In other words, publishers are (implicitly) discouraged from unnecessarily remodelling "insular" vocabularies from scratch in their own namespace.

[34] http://labs.mondeca.com/dataset/lov/index.html; retr. 2011/08/17.

Conformance? We deem the use of a larger number of vocabularies—for class and property terms—by a given data provider to be an indicator of conformance with respect to Linked Data best practices. Along these lines, we use the following simple metric to quantify "conformance":

$$\mathsf{nss}(p) := |\{\mathsf{ns}(u) : u \in \mathsf{pred}(p) \cup \mathsf{cmem}(p)\}|$$

where $\mathsf{ns}(u)$ denotes the namespace of a URI, which we compute as the set of characters up until the last hash or slash. Note that we only consider the namespaces of HTTP URIs and that we only count namespaces that appear for at least one other PLD. Intuitively, nss denotes the number of unique namespaces given by the vocabulary terms used to describe the local data of p.

Along these lines, in Table 13 we present the top five and bottom five data-providers in terms of the number of unique namespaces used in their respective contributions to our corpus. Each of the 188 providers used predicates/classes from an average of 8.6 (±7.1) namespaces.

№	PLD	nss [URIs]
1	w3.org	53
2	uriburner.com	34
...	openlinksw.com	34
4	b4mad.net	32
5	wasab.dk	28
159...	hi5.com	3
185	appspot.com	2
...	ontologyportal.org	2
...	unitn.it	2
188	lehigh.edu	1

Table 13
Top five PLDs and bottom five PLDs ordered according to number of unique namespaces for class and property terms (and thereafter, by number of quads)

Bias? Again, the nss metric is based on an absolute count. Domains for which we have smaller samples, and that have high variability in the class and property namespaces used across documents, may thus be under-represented by the measure's score.

Conclusions? We see that it is indeed common practice to mix vocabularies and select class and property terms from different namespaces when describing Linked Data, with an average of 8.6 namespaces used per domain. As per the previous guideline, this indicates that Linked Data publishers (often) partially self-organise by selecting numerous legacy vocabularies as opposed to defining (each time) a novel, insular, local vocabulary. Such agreement helps reduce the heterogeneity of datasets merged from different providers, making them easier to consume.

Interestingly, prominent Linked Data vocabularies often provide mappings to other vocabularies, formalised using the RDF and OWL standards [53]. Thus, even if two publishers choose different (but mapped) vocabularies, the data they provide can oftentimes be semantically integrated using reasoning techniques. We note however, that reasoning over Linked Data is a relatively new and challenging topic [53], where efforts to avoid or minimise the need for reasoning in the first place (i.e., by fostering agreement on vocabulary use as discussed for the previous issue) obviously have immediate practical benefits.

5.4 *Dereferencing resources*

Herein, we look at those Linked Data best practices—as introduced in [15]—that discuss what information about a resource should be returned when its respective URI is dereferenced.

Issue XI: Give human-readable meta-data

> *"We especially recommend the use of rdfs:label and foaf:depiction properties whenever possible as these terms are well-supported by client applications."* —**[15, §4]**

What? Publishers are encouraged to assign human-consumable information to their resources in a standard way; in particular, use of the property `rdfs:label` (used for attaching human readable "labels" or "'names" to resources) and the property `foaf:depiction` (used for attaching digital images to resources) are explicitly encouraged.

Why? Although RDF focuses on the provision of computer-readable resource descriptions, end-user applications often need to render a human-consumable description of the resources. Indeed, with respect to non-information resources, computers have no knowledge of the referents (the entities) being described—no way of mapping from a URI to the actual thing it identifies—and thus human-consumable meta-information is a fundamental requirement for applications to directly and effectively convey to users *what* is being talked about. The usability of such applications is thus dependent on

the provision of some core information for a high percentage of resources in the corpus: in particular, a label or "title" for the resource being described, and/or an image depicting the resource.

Conformance? We deem data-providers that provide a high percentage of their dereferenceable resources with some value for the `rdfs:label` and/or `foaf:depiction` properties to be more conformant with respect to Linked Data best practices. Along these lines, for each PLD, we check the percentage of such resources (appearing in a data-level position of a triple) that are provided a value for `rdfs:label`:

$$\mathsf{hrl}(p) := \frac{|\{u \in DU_p : \mathsf{lab}(u) \cap \mathsf{data}(p) \neq \emptyset\}|}{|DU_p|}$$

where $\mathsf{lab}(u)$ denotes the (infinite) set of triples given by $\{u\} \times \{\texttt{rdfs:label}\} \times \mathbf{L}$ and recalling that DU_p denotes the set of URIs from PLD p that were looked up and found to dereference to RDF/XML content during our crawl. We also compute the analogous measure hrp, but for $\mathsf{pic}(u)$, given similarly as the set of triples $\{u\} \times \{\texttt{foaf:depiction}\} \times \mathbf{U}$. Finally, we compute an average of the hrl and hrp to give our final measure of conformance:

$$\mathsf{hr}(p) := \frac{\mathsf{hrl}(p) + \mathsf{hrp}(p)}{2}$$

In Table 14, we present the top five PLDs with respect to this hr measure. Notably, of the 188 PLDs, 71 (37.8%) did not have a value for either property for any locally dereferenceable resources. Each PLD provided a label/depiction (hr) for, on average, 10.2% (±16.0) of locally dereferenceable resources; taking just labels (hrl), the analogous figure was 19.1% (±31.1 pp); taking depictions (hrp), the analogous figure was 1.2% (±5.2 pp) respectively.

Given the prolific use of sub-properties of `rdfs:label` on the Web of Data—properties such as `foaf:name`, `doap:name` or the various SKOS label properties—we checked to see whether reasoning would be able to automatically find more human-readable metainformation for the above two properties. For `rdfs:label`, we found 24 (possibly indirect) sub-properties in our corpus. For `foaf:depiction`, we found a sub-property within FOAF itself (`foaf:img`), an inverse-property also in FOAF (`foaf:depicts`), and four further sub-properties in remote vocabalaries (`sioc:avatar`, `ov:houseColor`, `mo:image`, `swid:Property-3AFoaf-3Aimg`). We then extend the previous measures to include these additional human-readable properties that can be found through reasoning

№	PLD	hr [%]
1	rdfize.com	66.1
2	kit.edu‡	51.1
3	ebusiness-unibw.org	50
4	l3s.de	49.9
5	ontologydesignpatterns.org	49.6
109...	freebase.com	0
...	identi.ca	0
...	opiumfield.com	0
...	livejournal.com	0
...	hi5.com	0

Table 14
Top five PLDs and bottom five PLDs ordered according to the percentage coverage of labels and depictions defined for local dereferenceable resources (and thereafter, by number of quads)

(hrl^+, hrp^+, and their average: hr^+)—we also include the original properties (thus, e.g., $\mathsf{hr}^+ \geq \mathsf{hr}$).

Thereafter, Table 15 gives analogous results, but for hr^+. This time, of the 188 PLDs, 20 (10.6%) still did not provide a (possibly implicit) value for either property for any dereferenceable resources (a reduction of 27.2 pp over hr). The average coverage of possibly implicit labels and depictions (i.e., hr^+) was 20.2% (±16.5 pp), an overall increase of 10 pp with reasoning. Considering just labels (hrl^+), the analogous figures were 32.8% (±30.4 pp) average, an increase of 13.7 pp with reasoning. Considering just depiction (hrp^+), the figures were 7.5% (±10.8 pp) average, a 6.3 pp increase with reasoning.

№	PLD	hr^+ [%]
1	rdfize.com	66.1
2	dbtune.org	51.5
3	kit.edu‡	51.1
4	advogato.org	50
5	robots.net	50
168...	vox.com	0
...	freebase.com	0
...	opiumfield.com	0
...	livejournal.com	0
...	hi5.com	0

Table 15
Top five PLDs and bottom five PLDs ordered according to the percentage coverage of possibly implicit labels and depictions defined for local dereferenceable resources (and thereafter, by number of quads)

Conclusions? Publishers frequently do not conform to this guideline, particularly for providing images.

After reasoning, the average coverage of dereferenceable resources with at least one `rdfs:label` value roughly doubles, and the number of resources with a `foaf:depiction` value increases by a factor of roughly 6×, albeit still remaining low at 7.5%.

There are a number of possible factors for this. Firstly, the prominent use of alternative naming properties that are not formally mapped to `rdfs:label`—such as `foaf:nick`, `dc:title`, `rss:title`, and `dct:title` (cf. Table 2)—leads to a lack of agreement on how human readable labels should be assigned to resources, which cannot be resolved automatically by reasoning. Consumers typically must "hard-code" the most popular labelling alternatives into their applications. Secondly, information resources and other "auxiliary resources", such as those used to represent n-ary predicates, may not have natural human-readable labels, or may not have any suitable images available, etc. Another possible explanation is that publishers do not appreciate the importance of making human-consumable metadata available due to a lack of tangible applications using their data.

Providing human-consumable meta-information for resources is important for allowing users to visualise, browse, and understand RDF data, where providing labels and depictions establishes a baseline. Further textual information about the resource, preferably given as a value for `rdfs:comment`, can also be valuable. Extending the concept further, different vocabularies may have different combinations of properties whose values are interesting to human users; work like Fresnel [80], which allows for specifying how data from different vocabularies should be rendered, go in this direction, with the potential to extend human-readability beyond just labels and images.

ISSUE XII: DEREFERENCE FORWARD-LINKS

> "The description: *The representation should include all triples from your dataset that have the resource's URI as the subject. This is the immediate description of the resource.*" —**[15, §5]**

What? As discussed in ISSUE III, the URIs assigned to resources should dereference to (related) representations thereof. This representation should contain all locally available triples where the given URI of the resource is in the subject position.

Why? Many Linked Data applications rely on the assumption that content *relevant* to the resource will be returned, in RDF, as a response to a HTTP lookup on its URI.[35] Intuitively, the above recommendation encourages data providers to return as complete an "immediate description" of a given local resource as possible to those requesting agents, allowing applications to achieve a higher recall with respect to query answering, or to render a more complete description of the resource of interest to the user.

Conformance? We deem data-providers that return a high percentage of triples with locally minted subject URIs—triples that appear in a local document—to also be given in the dereferenced document of that subject; for brevity, we call such triples "local outlinks". Along these lines, we give the following quantification of conformance:

$$\mathsf{do}(p) := \frac{|\{t \in \mathsf{deref}(t.sub) : t.sub \in \mathsf{ldlc}(p) \cap U\}|}{|\{t \in \mathsf{data}(p) : t.sub \in \mathsf{ldlc}(p) \cap U\}|}$$

where we only include URIs ($t.sub$) that were looked up during our crawl (the set U). Intuitively, for each PLD p, do denotes the average number of outlinks of local URIs found in the respectively dereferenced documents.

In Table 16, we present the top and bottom five PLDs with respect to the given do measure. Of the 188 PLDs—where again, three did not mint any local URIs (ISSUE III)—36 PLDs (19.1%) gave all known local outlinks in all known dereferenced documents; 60 PLDs (31.9%) provided more than 99% of local outlinks in the respectively dereferenced documents. We encountered no local outlinks for any URI local to the `hi5.com`: for this domain, all such local URIs are documents that are only given `rdfs:seeAlso` *inlinks*. Across the 188 data providers surveyed, the average percentage of local outlinks returned in the respective dereferenceable documents was 83.6% (±20.1 pp).

Bias? Again, we do not consider the case where URIs dereference to RDFa embedded in HTML documents. Also, documents that contain very few (or no) dereferenceable URIs are less likely to be picked up by our crawler. Finally, for data providers with partial samples, we may underestimate the number of local outlinks, which may cause the do measure to be over-represented. (It's worth noting that this analysis does not consider blank nodes—covered already by ISSUE I—which are naturally not dereferenceable.)

[35] We have previously made proposals for a priori checks that determine a degree of likelihood as to whether dereferencing a URI is likely to contain relevant content for a certain mime-type [92].

№	PLD	do [%]
1	livejournal.com	100
...	opiumfield.com	100
...	ontologycentral.com†	100
...	vox.com	100
...	ontologyportal.org	100
181	twoozer.com	37.7
182	xmlns.com	33.5
183	nickshanks.com	26.9
184	gregheartsfield.com	26.5
185	hi5.com	–

Table 16
Top five PLDs and bottom five PLDs ordered according to percentage of local outlinks given in the dereferenced document (and thereafter by number of quads)

Conclusions? This guideline is core to the Linked Data principles themselves, allowing consumers to "follow their nose" when looking for information about a resource of interest. As previously discussed, the dereferenceability of data is a key assumption for many Linked Data applications (cf. [47,10]).

Indeed, it seems that there is relatively high conformance to this guideline, where the domains surveyed provide, on average, 83.6% of local outlinks in the dereferenced document for the given subject URI. Thus, a Linked Data consumer can expect a high yield of triples where a resource appears in the subject position by dereferencing its URI. However, as we will see for the next issue, this yield does not hold to the same extent for triples where the resource appears in other positions.

Issue XIII: Dereference back-links

> "Backlinks: *The representation should also include all triples from your dataset that have the resource's URI as the object* [allowing] *browsers and crawlers to traverse links in either direction."* —**[15, §5]**

What? Closely related to the previous issue, Linked Data best practices recommend the provision of all "local inlinks" or backlinks—locally available triples in which the resource URI appears as an object—in the dereferenced document returned for the given resource.

Why? Arguably the distinction between inlinks and outlinks in terms of descriptiveness is a trivial one for RDF, where a resource appearing in an object position is equally being described. For example, consider:

```
ex:page foaf:maker ex:Joan .
ex:Joan foaf:made ex:page .
```

Both triples describe the resources `ex:page` and `ex:Joan` in an equivalent manner, irrespective of the positioning thereof. Similarly, let's say that we only know the second triple above, and that `ex:page` is dereferenced: by only returning outlinks, the consuming agent will not know of any relation between `ex:page` and `ex:Joan`, and, taking the example of a live Linked Data browser, users will not be able to navigate between these nodes. In other words, inlinks can themselves implicitly represent outlinks,[36] and inlinks allow for navigating from a given resource to those resources that are related to it.

Conformance? For locally minted and dereferenceable URIs, we deem data-providers that return a high percentage of local triples with the dereferenced term as object ("local inlinks"), in the respectively dereferenced document, to be highly conformant. We give a similar quantification of conformance as for local outlinks:

$$\mathsf{di}(p) := \frac{|\{t \in \mathsf{deref}(t.obj) : t.obj \in \mathsf{ldlc}(p) \cap U\}|}{|\{t \in \mathsf{data}(p) : t.obj \in \mathsf{ldlc}(p) \cap U\}|}$$

Note that by the definition of $\mathsf{ldlc}(p)$, as a special case, we do not count objects of `rdf:type` triples in the analysis, where, e.g., it would be unrealistic (and probably undesirable) to expect FOAF to provide a list of all `foaf:Person` members in the FOAF vocabulary dereferenced by that class term.

In Table 17, we present the top and bottom five PLDs with respect to the given di measure. Of the 188 PLDs, 14 PLDs (7.4%) gave all known local inlinks in all known dereferenced documents; 20 PLDs (10.6%) provided more than 99% of local inlinks in the respectively dereferenced documents. Conversely, 11 PLDs (5.9%) offered no dereferenceable outlinks, with 17 domains (9%) providing less than 1% of locally available inlinks through dereferencing. We also encountered two domains—`unitn.it` and `prefix.cc`—that had no local inlinks for any local URI (we consider these scores as zero). Across the 188 data-providers, the average percentage of local outlinks returned in the respectively dereferenced documents was 55.2% (±32.9 pp).

Bias? Similar biases exist as per the previous issue.

[36] Such implicit knowledge can be formally represented using an `owl:inverseOf` relation.

№	PLD	di [%]
1	ontologyportal.org	100
...	ebusiness-unibw.org	100
...	174.129.12.140 (open-biomed.org.uk)	100
...	skipforward.net	100
...	semantic-web-grundlagen.de	100
177...	umbel.org	0
...	lexvo.org	0
...	loc.gov	0
...	geonames.org	0
...	hi5.com	0

Table 17
Top five PLDs and bottom five PLDs ordered according to percentage of local inlinks given in the dereferenced document (and thereafter, by number of quads)

Conclusions? Compared to dereferencing "outlinks", publishers are much less conformant when it comes to dereferencing inlinks: compared with an average 83.6% of outlinks being dereferenced across the domains surveyed, the analogous figure for inlinks was 55.2%. Similarly, a relatively high standard deviation of 32.9 pp indicates significant variability in conformance across publishers. Some of the domains not providing any dereferenceable inlinks are quite prominent in the Linked Data community, where in particular, the dereferenceable RDF hosted by `geonames.org` consists only of the local outlinks of the geographical resource in question, and meta-data about the document. [37]

Many of the local URIs appearing in the object position are information resources, often HTML pages. These are associated with a given subject resource with typed links, e.g., `foaf:page`, `foaf:weblog`, `gn:locationMap`, etc. Such information resources are typically assigned no further meta-data other than the aforementioned inlink(s), and do not dereference to RDF (albeit, we do not check for RDFa). Further still, certain resources may feature a high indegree, which makes the inclusion of all inlinks in the dereferenced document somewhat impractical. For example, the URI `http://identi.ca` had 1.66 million inlinks through the `foaf:accountServiceHomepage` property in our corpus, many of which were local; making all of these inlinks dereferenceable—e.g., by embedding RDFa into the main `identi.ca` webpage—would obviously be impractical.

[37] See, e.g., `http://sws.geonames.org/2964179/about.rdf`; retr. 2011/08/18.

Issue XIV: Describe & licence documents

> "Metadata: *The representation should contain any metadata you want to attach to your published data, such as a URI identifying the author and licensing information. These should be recorded as RDF descriptions of the information resource that describes a non-information resource; that is, the subject of the RDF triples should be the URI of the information resource.* [...] *In order to enable information consumers to use your data under clear legal terms, each RDF document should contain a license under which the content can be used."* —**[15, §5]**

What? The information resources (possibly) returned through dereferencing non-information resources are, of course, themselves dereferenceable resources. Thus, by implication, the previous two premises of Linked Data best practices again apply: locally known inlinks and outlinks relating to the information resource should be returned in the dereferenced document. Emphasis is placed on returning licencing information.

Why? Returning meta-information about documents follows the same rationale as before: descriptions of information resources can similarly contain any form of meta-data the provider deems relevant. However, the above stated best practice emphasises that licencing information should be attached, such that consumers are made aware of the legal rights and permissiveness under which the pertinent data are made available.

Conformance? Since this issue is partially covered by the previous two, herein we focus on conformance with respect to (i) providing meta-information about documents and (ii) licensing information. Thus, we deem data providers that return (i) a high percentage of resource descriptions for their documents, and (ii) a high percentage of licencing information for these resource descriptions, as being better conformant to Linked Data best practices. With respect to (i), we quantify conformance as follows:

$$\mathsf{dmr}(p) := \frac{|\{s \in S : \mathsf{pld}(s) = p \wedge s \in \mathsf{dlc}(\mathsf{get}(s))\}|}{|\{s \in S : \mathsf{pld}(s) = p\}|}$$

where S denotes the set of known sources. Here, dmr denotes the percentage of source URIs local to p that themselves appear as a data-level constant in the RDF graph they return.

Along these lines, in Table 18 we provide the top five and bottom five data-providers with respect dmr. We found 20 providers (10.6%) that gave no meta-data for any of their documents, where in Table 18, we show the five largest. Conversely, 77 PLDs (41%) offered some meta-data for all documents, where again we only show the five largest. On average, the 188 data providers offered some meta-data in 75.7% (±36.6 pp) of the documents.

№	PLD	dmr [%]
1	identi.ca	100
...	dbtropes.org	100
...	vox.com	100
...	ontologyportal.org	100
...	twatter.com	100
168...	uniprot.org	0
...	ontologycentral.com†	0
...	rdfabout.com	0
...	freebase.com	0
...	hi5.com	0

Table 18
Top five PLDs and bottom five PLDs ordered according to percentage of local documents with some embedded meta-information (ordered thereafter by number of quads)

With respect to licencing, we note that the guidelines do not mention a specific property to relate a document to its licence. From the set of property terms appearing in the predicate position of a triple in our corpus, we performed a search for the string "licen" to determine a set of candidates that publishers might be using. After filtering out some obviously irrelevant properties (such as fb:common.licensed_object.provenance), we present the top ten such properties in Table 19 according to the number of times they were used as a predicate. As suggested by Bizer et al., we also include the properties dc:rights and dct:rights [14]. We note that some of these properties may not be intended for usage on documents, where we note that the value of the property doap:license should give licencing information with respect to a software project. In any case, we believe that Linked Data guidelines should more explicitly recommend a chosen licencing property for RDF documents published on the Web.

With respect to conformance, we re-use the dmr(p) metric, but where in the numerator, we only consider descriptions of documents that included a value for a property containing the string "licen" or for the properties dc:rights or dct:rights; we denote this value as dmr$'$(p). The top-five and bottom-five providers resulting from this analysis are presented in Table 20,

№	property	quads
1	xhtml:license	179,375
2	dc:licence	176,029
3	cc:license	59,790
4	dc:rights	7,007
5	sz:license_text	2,035
6	dbo:license	1,653
7	dct:licence	1,591
8	dbp:licence	383
9	wrcc:license	151
10	doap:license	92
–	dct:rights	23

Table 19
Top ten licencing properties according to use in our corpus

where we found that only 27 PLDs (14.4%) returned some licencing information for some local document. We found that, on average, providers gave licencing information for 3.4% (±15.4 pp) of local documents.

№	PLD	dmr$'$ [%]
1	fluffyandmervin.com	100
2	l3s.de	99.8
3	geospecies.org	99.7
4	smhowell.net	96
5	mfd-consult.dk	50
28...	rdfize.com	0
...	linkedlifedata.com	0
...	opiumfield.com	0
...	livejournal.com	0
...	hi5.com	0

Table 20
Top five PLDs and bottom five PLDs ordered according to percentage of local documents with embedded licencing meta-information (ordered thereafter by number of quads)

Conclusions? Few documents provide licencing information directly as part of the document meta-data. Further still, there is a palpable need for (i) an agreed-upon licencing property, and (ii) an agreed set of common licence URIs; to avoid consumers again having to hard-code support for all alternatives used by publishers. The most complete proposal along these lines is provided by the Creative Commons vocabulary. [38]

We note that there may be other licencing practices not checked by our analysis. For example, publishers may choose to make licencing meta-data available for an entire dataset—a logical grouping of documents—in a single VoiD description [2]. However, others have

[38] http://creativecommons.org/ns; retr. 2011/08/28.

also observed a worrying lack of licencing information for RDF documents published on the Web, where, e.g., Bizer et al. put the figure at 15% of Linked Data publishers offering document-level licencing [14]. [39]

6 PageRank of Domains

Having analysed various issues relating to Linked Data conformance, we now briefly look at measures that use links-based analysis to rank the different domains hosting RDF in our corpus. In particular, we are interested in whether or not there is a correlation between the PageRank scores of the different domains and their conformance to the guidelines measured in the previous section. Later, we will also use the PageRank scores to present a weighted aggregation of conformance measures (as opposed to the arithmetic-mean aggregation introduced thus far).

There is a long history of links-based analysis over Web data—and in particular over hypertext documents—where links are seen as a vote for the relevance or importance of a given document. Seminal works exploiting the link structure of the Web for ranking documents include HITS [65] and PageRank [79].

Whilst link-based analysis, such as PageRank, are an established technique when considering the Web of Documents, there are some fundamental differences between the notion of a (hyper)link on the traditional Web of Documents, and the notion of a (RDF) link on the Web of Data. On the Web of Documents, a hyperlink is typically interpretable as a pointer to the content of the target page; when considering PageRank, hyperlinks are often intuited as "votes" from source pages to target pages. On the Web of Data, links can have arbitrary labels (i.e., predicates), can be of various forms, and may serve a variety of purposes, including (but not limited to):

(i) the target domain hosts a description of a class or property used on the host domain (schema links);
(ii) the target domain was involved in the generation of the source data, or provides a centralised service upon which the source domain relies;
(iii) the target domain describes legacy resources that refer to the same real-world entities as the source domain (`owl:sameAs` links);
(iv) the source domain does not wish to describe a particular resource, but instead out-sources the description to the target domain with a link;

Along these lines, there are many possible ways one might consider applying PageRank over Linked Data.

More recent works (e.g., [35,25,45]) have presented various attempts at incorporating links-based analysis techniques for ranking RDF data, with various end-goals in mind: most commonly, prioritisation of informational artefacts in user result-views. Detailed discussion of the different approaches is out of the current scope, where for our purposes, we choose a straightforward approach inspired by the work of Harth et al. [45], who propose (amongst other approaches) a PLD-level ranking of Linked Data. (A similar proposal has been put forward by Delbru et al. [25], who also look at performing ranking on a "dataset-level", the results of which are then propagated to ranks of intra-dataset entities.)

The first step towards ranking PLDs is to construct a directed graph representing the link structure between the different PLDs, which we now discuss.

6.1 PLD-level graph

Recalling the Linked Data principles enumerated in § 3.2, according to LDP4, links should be specified simply by using external URI names in the data. These URI names should dereference to an RDF description of themselves according to LDP2 and LDP3 respectively.

Following these principles, we define our *PLD-level graph* as follows. Let $D := (V, E)$ represent a simple directed graph where $V \subset \mathbf{P}$ is a set of PLDs (vertices), and $E \subset V \times V$ is a set of pairs of vertices (edges). Letting $p_i, p_j \in V$ be two vertices, then $(p_i, p_j) \in E$ iff $p_i \neq p_j$ and there exists some $u \in \mathbf{U}$ such that u appears in $\mathsf{data}(p_i)$, and $\mathsf{pld}(u) = p_j$. In other words, an edge extends from p_i to p_j if p_i hosts a triple that contains a URI under the authority of p_j and/or that redirects to p_j. Notably, the link is forwarded through any redirect such that, e.g., if the URI mentioned in p_i has the authority `purl.org` but redirects to `xmlns.com` (as would be the case for a FOAF URI), the link is given to the latter domain (p_j = `xmlns.com`), not the former redirection domain. [40]

6.2 PageRank algorithm

We now introduce the PageRank algorithm [79], which we apply to the PLD-level graph.

[39] Dodds has also raised similar concerns; cf. http://www.flickr.com/photos/ldodds/4043803502/; retr. 2011/08/12.

[40] In fact, `purl.org` does not appear as a PLD in our analysis at all since it always redirects to an external domain.

Taking $D := (V, E)$, let $E(p)$ denote the set of direct successors (outlinks) of vertex (PLD) p, let $E^-(p)$ denote the set of direct predecessors (inlinks) of p, and let

$$V_\emptyset := \{p \in V : E(p) = \emptyset\}$$

denote the set of vertices with no outlinks (aka. dangling vertices). The PageRank of a vertex p_i in the directed graph $D := (V, E)$ is then given as follows [79]:

$$\mathsf{pr}(p_i) := \frac{1-d}{|V|} + d \sum_{p_\emptyset \in V_\emptyset} \frac{\mathsf{pr}(p_\emptyset)}{|V|} + d \sum_{p_j \in E^-(p_i)} \frac{\mathsf{pr}(p_j)}{|E(p_j)|}$$

where d is a damping constant (typically $d := 0.85$ [79]), which helps ensure convergence in the following iterative calculation, and where the middle component splits the ranks of dangling nodes evenly across all other nodes.

Now let $w := \frac{1-d}{|V|}$ represent the weight of a universal (weak link) given by all non-dangling nodes to all other nodes—dangling nodes split their vote evenly and thus don't require a weak link; we can use a weighted adjacency matrix M as follows to encode the graph $D := (V, E)$:

$$m_{i,j} := \begin{cases} \frac{d}{|E(p_j)|} + w, & \text{if } (p_j, p_i) \in E \\ \frac{1}{|V|}, & \text{if } p_j \in E_\emptyset \\ w, & \text{otherwise} \end{cases}$$

where this stochastic matrix can be thought of as a Markov chain (dubbed the random-surfer model). The ranks of all PLDs can be expressed algebraically as the principal eigenvector of M, which in turn can be estimated using the power iteration method up until some termination criteria (fixed number of iterations, convergence measures, etc.) is reached. We refer the interested reader to [79] for more detail on PageRank, and to [56] for our distributed implementation thereof.

6.3 PLD PageRank results

From our billion-quadruple corpus, we extracted the PLD-level graph, which contained 778 vertices and 7,647 edges, giving an average degree of 9.83 edges per vertex. Four vertexes had no inlinks [41], whereas every vertex had at least one outlink. In Table 21, we present the top-25 scoring PLDs after applying the PageRank analysis over this graph. Note that we italicise domains that contributed less than 1,000 quads to our corpus. We will now discuss the top ten results.

Unsurprisingly, the `w3.org` domain—which hosts the `rdf:`, `rdfs:`, `owl:` and `skos:` namespace documents, amongst others—tops the table. [42]

Otherwise, the top half of the table is dominated by domains that host popular vocabularies: (2) `dublincore.org` hosts the `dc:` and `dct:` namespaces; (3) `xmlns.com` hosts the `foaf:` and `wot:` namespaces; (5) `rdfs.org` hosts the `sioc:` and related namespaces; (6) `resource.org` hosts the `rss:` namespace, as well as an older `cc:` namespace; (8) `vocab.org` hosts a variety of namespaces, including `bio:`, `frbr:`, `ov:`, `rel:`, `vann:`, `whisky:`; and (6) `usefulinc.com` hosts the `doap:` namespace.

Further down the table, we find domains not necessarily associated with popular vocabularies. Notably, (4) `loc.gov` and (9) `vu.nl` rank highly despite having a much smaller in-degree than many PLDs below them: `loc.gov` was one of two domains linked by `dublincore.org`, from which it gained a significant boost in rank, where, in turn, `vu.nl` was one of three domains linked from `loc.gov`, which accounted for most of its rank. This is an example of highly-ranked domains with low out-degree passing on high rank scores to their neighbours. Wrapping up the top ten—and as already mentioned—(10) the `dbpedia.org` domain is a prominently-linked publisher of RDF on the Web.

Thus, we see that top-ranked domains attract inlinks (and thus higher PageRank) for very different reasons. In fact, we believe its unclear whether Linked Data is mature enough, and well-linked enough, to allow for meaningful links-based analysis, particularly on the level of domains: many of the cross-provider links are being generated in automated ways, or are being generated in large batches by a few data providers (to few data providers). We believe that there is little in the way of ad hoc, manual interlinkage being performed by humans on the current Web of Data—using the analogy of links being interpreted as "votes" on the Web, there are relatively few human voters (or, currently, incentives to vote). However, as the Web of Data diversifies, so too will the benefits of PageRank-esque measures over the burgeoning graph. For now, we are interested to see how the presented PageRank scores correlate with the conformance scores for the data providers in our corpus.

7 Synopsis of Analysis

Herein, we present overall summaries and aggregations for conformance with respect to the different

[41] This implies that the links to these domains must only have appeared in the seed list of URIs for the crawl.

[42] Only the `globalnames.org`, `sig.ma` and `unitn.it` domains did not not link to `w3.org` in our data.

№	PLD	rank	in-degree	out-degree
1	w3.org	0.175582	774	71
2	dublincore.org	0.092568	306	2
3	xmlns.com	0.068402	690	2
4	loc.gov	0.043293	19	3
5	*rdfs.org*	0.017477	330	7
6	*resource.org*	0.017409	119	2
7	ldodds.com	0.016112	100	22
8	vocab.org	0.014381	199	20
9	vu.nl	0.013240	22	12
10	dbpedia.org‡	0.010859	118	125
11	*usefulinc.com*	0.010494	71	3
12	identi.ca	0.009672	179	305
13	semanticweb.org	0.009231	86	147
14	rdfweb.org	0.007931	47	57
15	creativecommons.org	0.006679	73	3
16	mit.edu	0.006079	50	58
17	*isi.edu*	0.006066	18	5
18	geonames.org	0.005914	168	4
19	*danbri.org*	0.005886	53	16
20	*wordpress.com*	0.005859	84	3
21	daml.org	0.005656	56	1
22	*stanford.edu*	0.005403	29	7
23	sourceforge.net	0.005315	65	12
24	*umd.edu*	0.005080	31	5
25	mindswap.org	0.004688	26	12

Table 21
Top twenty-five ranked PLDs and number of inlinks and outlinks from/to external PLDs; domains with less than 1,000 quads are italicised

guidelines, also looking at correlation with the PageRank scores of the different domains (§ 7.2). We then look at aggregating conformance scores across the different guidelines and PageRank scores into one overall measure (§ 7.3).

7.1 Kendall's τ coefficient

First, we introduce *Kendall's τ coefficient* [64], which we use to compare the orderings given by conformance scores of each domain and its respective PageRank score. Given that our data may contain outliers and follow non-normal distributions, we favour the non-parametric (rank-based) Kendall's τ over the parametric (value-based) Pearson's coefficient: although some information is lost by "compressing" absolute values into ranks, non-parametric tests are more robust in the face of outliers, which are to be expected in data such as ours. We also favour Kendall's τ over Spearman's ρ (a non-parametric version of Pearson's) since τ is based on simple distances, whereas ρ is based on squared differences, implying that outliers in ordering—i.e., fewer, longer distances—are punished more by ρ than τ when compared with many, shorter distances. Additionally, we are also grateful for the fact that Kendall's τ measure is simpler to present and explain [64], vs. the Spearman's ρ intuition of characterising the monotonicity for a function mapping one ordering to the other. (Informally, we also ran Spearman's ρ measures and found that they correspond closely with Kendall's τ where we choose to only present the latter for brevity.)

Towards defining Kendall's τ, let $\leq_1$ and $\leq_2$ denote two total orderings defined for a set S (where $|S| \geq 2$), intuitively, Kendall's τ quantifies the amount of agreement between the ordered pairs given by the two orderings. In particular, it measures the ratio of (dis)agreement across all possible pairs for the two orderings, represented in an interval $[-1, 1]$.

First let

$$\mathcal{A}gree(\leq_1, \leq_2, S) := S \times S \cap <_1 \cap <_2$$

denote the set of ordered pairs (s_i, s_j) from $S \times S$ for which $s_i <_1 s_j$ and $s_i <_2 s_j$—i.e, $\mathcal{A}gree$ is the set of unique unequal and non-tied pairs such that both orderings agree. Also, let

$$\mathcal{D}isagree(\leq_1, \leq_2, S) := S \times S \cap <_1 \cap >_2$$

similarly denote the unique set of unequal and non-tied pairs such that the orderings disagree. Now, for $\leq_1$, $\leq_2$, and S (omitting the arguments for brevity):

$$\tau' := \frac{(|\mathcal{A}gree| - |\mathcal{D}isagree|)}{\frac{n(n-1)}{2}}$$

where $n = |S \times S|$ denotes the cardinality of the set of all ordered pairs that can be constructed from S, and where $\frac{n(n-1)}{2}$ denotes the cardinality of the set of all unordered, unequal pairs. Thus, Kendall's τ measures the difference between the total number of all pairs for which both orderings agree and those for which they disagree, normalised by the total number of independent non-trivial pairs to compare. *Thus, τ' is a rational number in the interval $[-1, 1]$, where a value of 0 indicates no correlation (agreement) between the orderings, 1 indicates perfect correlation (agreement) between the orderings, and -1 indicates perfect disagreement between the orderings (i.e, $<_1 = >_2$).*

However, ties may often occur in our scenario, where $s_i =_1 s_j$ or $s_i =_2 s_j$ (or both). For Kendall's τ, tied pairs are simply excluded from the analysis, with the denominator modified to only count non-tied pairs.

Finally, we also present the *statistical significance* (p-value) of the τ measure, which denotes the probability of the given observations occurring under the

null hypothesis: i.e., the probability of finding the given (or weaker) correlation if the two orderings were completely independent. As per tradition [89], we interpret a p-value of less than 0.05 to be a *significant* result.

7.2 Aggregating Results for Issues

In Table 22, we summarise all of the results for all of the issues/measures encountered thus far. Note that ISSUE IV (referring to URI length) is the only exception to the rule that a higher value corresponds to higher conformance. We present a number of aggregate scores for each measure. We first present the average and population standard deviation for each guideline across the 188 data providers contributing more than 1,000 quads to our corpus. We then present the analogous figures for all 778 data providers in our corpus. Given that the latter averages are mostly influenced by low-volume publishers, we also present weighted averages based on the size (quad count) of providers, and their PageRank.

First, let $\mathsf{pos}_w(p)$ denote a straightforward (non-parametric) ranking of PLDs prescribed by some ordering—in this case, we use *s*ize and *P*age*R*ank, denoted pos_s and pos_{pr} respectively. For size, we count the number of documents, where for example $\mathsf{pos}_s(p') = 3$ indicates that PLD p' contributed the third-most documents to our corpus (possibly tied with another PLD). Similarly, $\mathsf{pos}_{pr}(p'') = 1$ would indicate that p'' was the highest ranked PLD in our PageRank scores (i.e., p'' =`w3.org`; cf. Table 21). We then use these weights (denoted generically by $\mathsf{pos}_w(.)$) for averaging the conformance scores for an issue x as follows:

$$ {}_w\overline{\mathsf{cs}}_x := \frac{\sum_{p \in P}(|P| + 1 - \mathsf{pos}_w(p)) \times \mathsf{cs}_x(p)}{\sum_{p \in P} |P| + 1 - \mathsf{pos}_w(p)} $$

where P is the set of all PLDs in our corpus. Again, w stands for a generic weight, which we instantiate by $s \mid pr$ such that ${}_s\overline{\mathsf{cs}}_x$ denotes the size-weighted average for issue x and ${}_{pr}\overline{\mathsf{cs}}_x$ denotes the PageRank-weighted average for issue x. We also present the accompanying biased weighted standard deviation (such that distances from the mean are also weighted). Where w is omitted ($\overline{\mathsf{cs}}_x$), we denote a non-weighted arithmetic mean.

Finally, we also present Kendall's τ correlation between PageRank and conformance scores, where again, a positive value indicates positive correlation between the two orderings they prescribe for the PLDs.

On a high level, with respect to conformance, we see that providers current abide by guidelines regarding the use of HTTP URIs (ISSUE II), hosting stable URIs (ISSUE V), avoiding use of verbose RDF features (ISSUE VIII), and making local outlinks dereferenceable (ISSUE XII). On the other hand, other guidelines are not well abided by, particularly the provision of human-readable metadata (ISSUE XI), providing licencing information for documents (ISSUE XIV_b), and dereferencing inlinks (ISSUE XIII). Thus, applications relying on these features of Linked Data—for example, for allowing users to navigating through inlinks, rendering domain-agnostic display of resources, or determining whether the consumer's intended use of the data is legal—are inherently affected.

Regarding the differences between considering only the 188 PLDs with > 1,000 quads and all PLDs, we see that many of the average conformance scores remain relatively stable. However, by including the lower-volume publishers, we (unsurprisingly) see a marked drop in those scores given in absolute terms, such as the level of external linkage, the number of triples dereferenced, and the variety of vocabularies and vocabulary terms used.

When looking at the size-weighted averages, the conformance figures quite often float between the scores considering only the PLDs with >1,000 quads and all PLDs; we see a slight increase again in the level of external linkage, dereferenced triples, and in the variety of vocabulary usage.

When looking at the PageRank-weighted averages, we see a slight drop in some conformance scores, particularly those that restrict use of RDF features such as blank nodes, non-HTTP URIs and reification/containers/collections. We note that there may be valid exceptions to these guidelines, particular when modelling vocabularies in OWL; recalling that many of the highest-ranked domains host vocabularies, this may explain these observations. Further, looking at the significant results given by Kendall's τ for correlation between conformance and PageRank, we see some similar results. Highly-ranked domains tend not to follow the aforementioned guidelines restricting use of RDF features and non-HTTP URIs. In addition, highly-ranked PLDs are less likely to provide all inlinks in the locally dereferenced document, or to provide metadata for the document itself. Conversely, highly-ranked PLDs tend to provide more dereferenceable data, to contain a higher level of external linkage, to use more vocabularies, and to provide more human-readable meta-data (again, vocabularies commonly provide `rdfs:label` scores directly for class and property terms).

	issue			plds > 1,000 q		all plds							
	id (x)	mnemonic	sym.	$\overline{\text{cs}}_x$	$\pm_x$	$\overline{\text{cs}}_x$	$\pm_x$	${}_s\overline{\text{cs}}_x$	${}_s\pm_x$	${}_{pr}\overline{\text{cs}}_x$	${}_{pr}\pm_x$	τ	p
NAMING	I	*ratio: URIs vs. URIs and b'nodes*	¬bn	84.32%	24.21 pp	86.71%	19.83 pp	86.33%	21.25 pp	83.91%	20.21 pp	-0.22	~
	II	*ratio: HTTP-URIs vs. local URIs*	hu	98.84%	4.81 pp	96.40%	8.50 pp	98.18%	5.66 pp	95.34%	9.73 pp	-0.29	~
	III_a	*ratio: local deref. URIs vs. local URIs*	du	70.26%	26.8 pp	52.52%	28.63 pp	63.93%	26.45 pp	54.07%	30.30 pp	0.04	0.12
	III_b	*avg. deref triples*	dt	17.51 t	40.29 t	11.48 t	21.82 t	13.24 t	27.55 t	12.41 t	18.74 t	0.11	~
	IV	*avg. URI string length*	$\overline{\text{ul}}$	52.41 c	16.41 c	44.51 c	14.12 c	49.41 c	16.03 c	45.41 c	15.19 c	0.04	0.10
	V	*avg. uptime of docs over 9 months*	$\overline{\text{st}}$	88.81%	19.45 pp	92.15%	16.75 pp	89.87%	18.34 pp	92.47%	15.97 pp	-0.04	0.41
LINK	VI	*external PLDs for RDF links*	el	20.40 p	38.16 p	7.02 p	20.56 p	13.95 p	31.89 p	11.01 p	27.43 p	0.36	~
	VII	*external PLDs for* `owl:sameAs` *links*	$\text{el}^{=}$	1.79 p	5.19 p	.56 p	2.74 p	1.13 p	3.96 p	.89 p	3.46 p	0.22	~
DATA	VIII	*ratio: non-reif./coll/cont. triples*	¬rcc	99.15%	4.68 pp	99.00%	4.48 pp	99.11%	4.51 pp	98.84%	5.19 pp	-0.14	0.01
	IX	*PLD–vocabulary-term overlap measure*	olt	6,607 pu	3,667 pu	5,860 pu	2,787 pu	6,553 pu	3,290 pu	5,936 pu	3,025 pu	0.03	0.17
	X	*distinct vocabulary namespaces used*	nss	8.61 u	7.12 u	5.08 u	4.52 u	7.21 u	6.04 u	6.07 u	5.53 u	0.25	~
DEREFERENCE	XI_a	*avg. entities w/ labels, pics*	hr	10.17%	15.96 pp	7.28%	14.21 pp	7.70%	14.09 pp	9.50%	15.04 pp	0.27	~
	XI_b	*avg. entities w/ labels, pics (reasoning)*	hr^{+}	20.17%	16.51 pp	22.39%	19.61 pp	19.96%	16.20 pp	23.62%	21.31 pp	0.07	0.01
	XII	*ratio: deref. outlinks vs. local outlinks*	do	83.60%	20.09 pp	86.10%	28.63 pp	86.00%	18.88 pp	85.70%	29.44 pp	-0.01	0.69
	XIII	*ratio: deref. inlinks vs. local inlinks*	di	55.19%	32.94 pp	65.10%	40.80 pp	57.64%	33.33 pp	61.14%	41.86 pp	-0.11	~
	XIV_a	*ratio: local docs w/ deref. meta-data*	dmr	75.73%	36.57 pp	78.19%	38.37 pp	79.61%	33.54 pp	72.61%	40.47 pp	-0.23	~
	XIV_b	*ratio: local docs w/ deref. licence*	dmr^{l}	3.40%	15.42 pp	4.26%	19.14 pp	4.08%	17.47 pp	4.67%	19.76 pp	0.11	0.09

Table 22. For each issue, we summarise the (i) average values of conformance ($\overline{\text{cs}}_x$), and the resulting population standard deviation ($\pm_x$) for PLDs with greater than 1,000 quads and for all PLDs, (ii) size weighted average (${}_s\overline{\text{cs}}_x$) and weighted standard deviation (${}_s\pm_x$) for all PLDs, (iii) PageRank weighted average (${}_{pr}\overline{\text{cs}}_x$) and weighted standard deviation (${}_{pr}\pm_x$) for all PLDs, and (iv) Kendall's τ correlation between PageRank and average conformance, with corresponding p-value. The non-standard units are c: *c*haracters, p: *P*LDs, q: *q*uads, t: *t*riples, u: *U*RIs, pu: *P*LDs × *U*RIs; as before, pp indicates *p*ercentage *p*oints. Finally, we use '~' to indicate a value <0.005.

7.3 Aggregating Results for PLDs

In the previous section, we looked at aggregating scores for each guideline by taking various forms of mean across the PLDs surveyed. In this section, we conversely look at aggregating scores for each PLD across the guidelines presented. As such, we formulate a high-level conformance metric that aggregates scores across Issue I–XIV, as follows:

(i) each individual issue is given an equal weight with respect to the overall conformance measure;
(ii) wherever possible, each PLD should gain/lose conformance score in a manner appropriate with their absolute conformance to each issue.

With respect to Item (ii) above, we considered using a purely non-parametric (position-based) aggregation of overall conformance, but found this to often be unrepresentative and overly simplistic—for example, `ordnancesurvey.co.uk` uses one blank node and 371 thousand URIs, but would be in position 65/188 for the ¬bn metric.

Along these lines, for each issue we score each data-provider on a scale from [0–100], where 100 denotes the highest level of conformance. Metrics that are in the interval [0–1] are directly converted to percentages. For metrics that are not percentage- or ratio-based—viz., those discussed for Issue IV, VI, VII, IX, X—we resort to a positional based ranking that we then linearly bin into the [0–100] interval using:

$$\mathsf{cs}_x(p) := \frac{(|P| + 1 - \mathsf{pos}_x(p)) \times 100}{|P|}$$

where x denotes an issue in { IV, VI, VII, IX, X }, $|P|$ is the number of data-providers under analysis, and $\mathsf{pos}_x(p)$ is the position assigned to that data-provider by the associated metric of conformance. Thus, for example, all providers tied for first (most conformant) for issue x will be assigned $\mathsf{cs}_x(p) = 100$; all tied for third will be assigned $\mathsf{cs}_x(p) = \frac{(188+1-3)\times 100}{188} = 98.9$, etc.

Next, for Issue III, XI, & XIV where we presented two measures each, we first take the local average of these measures as the final score for cs_{III}, cs_{XI} and cs_{XIV}: for example, cs_{XI} is given the average score for providing human-readable meta-information with and without reasoning enabled.

Finally, we take the overall per-PLD aggregated score as the average of the scores for the individual issues:

$$\overline{\mathsf{cs}}(p) = \frac{\sum_{x \in \{\text{I}\ldots\text{XIV}\}} \mathsf{cs}_x(p)}{|\{\text{I}\ldots\text{XIV}\}|}$$

giving us our final overall conformance measure for provider p. We exclude the conformance measure $\mathsf{cs}_{\text{V}}(p)$ from the average for the 47 providers for which we had no information about the stability of URIs from our nine snapshots.

Thereafter, for reference, we present the results for the 188 PLDs with >1,000 quads in Tables A & B (at the end of the paper), along with their individual score for each issue presented. *We do not claim that lower ranked providers definitively host data of less "quality" or "worth", but rather, we instead claim that they host data in a manner that is less conformant to Linked Data guidelines.* The average score for the providers was 64.7% (±8.0 pp).[43] In the Table, we also present the ranking position of each provider in the rightmost column (under pr). Looking for correlation between the orderings given by a higher $\overline{\mathsf{cs}}$ score and a higher PageRank score, we computed Kendall's $\tau = 0.17$ with $p = 0.00005$, denoting a significant, weak-to-moderate correlation between the two orderings. The largest distance between the two orderings was given by `loc.gov`, which was ranked 9th in terms of PageRank, but ranked 172nd in terms of overall conformance (recall from § 6.3 that this PLD received much of its PageRank through a single link from `dublincore.org`). Along these lines, Table 23 enumerates the remaining top ten, where all results bar № 8 & № 10 had high PageRank and low conformance scores.

№	PLD	$\lvert\mathsf{pos}_{\mathsf{pr}} - \mathsf{pos}_{\overline{\mathsf{cs}}}\rvert$	$\mathsf{pos}_{\overline{\mathsf{cs}}}$	$\mathsf{pos}_{\mathsf{pr}}$
1	`loc.gov`	178	182	4
2	`unitn.it`	151	188	37
3	`geonames.org`	149	162	13
4	`vu.nl`	142	149	7
5	`okkam.org`	141	181	40
6	`typepad.com`	138	179	41
7	`livejournal.com`	135	155	20
8	`chirub.com`	135	41	176
9	`xmlns.com`	134	137	3
10	`jobsonica.com`	133	17	150

Table 23
Top ten PLDs with the greatest (absolute) difference in position in terms of PageRank (pr) and conformance ($\overline{\mathsf{cs}}$)

Conversely, Table 24 presents the top ten data providers with respect to the highest average position in terms of conformance and PageRank. We note that these providers rank highly for our two distinct "quality" measures, and thus we would consider them—in a generic sense—to be the "highest scoring" providers resulting from our analysis, providing highly-conformant

[43] We acknowledge that with the inclusion of the position-based scores, the absolute values of $\overline{\mathsf{cs}}$ have little by way of *direct* meaning.

data, and being heavily linked from other highly-ranked providers.

№	PLD	$\frac{pos_{\overline{cs}}+pos_{pr}}{2}$	$pos_{\overline{cs}}$	pos_{pr}
1	dbpedia.org‡	7	6	8
2	mit.edu	11	10	12
3	identi.ca	11.5	14	9
4	w3.org	12.5	24	1
5	rdfweb.org	13	15	11
6	qdos.com	15.5	1	30
7	l3s.de	16	5	27
8	sourceforge.net	16.5	18	15
9	bblfish.net	17	8	26
10	fu-berlin.de	20.5	20	21

Table 24
Top ten PLDs with the highest average positions for PageRank (pr) and conformance ($\overline{cs}$)

8 Discussion and Outlook

We have seen that the conformance of data providers varies significantly for the different Linked Data guidelines highlighted, which in turn may have implications for ad hoc consumers operating over the Web of Data. Although publishers may (reasonably) decide to (partially) forego compliance with respect to individual guidelines—and as we have discussed in this paper—each such guideline has, in the general case, a clear rationale. By aggregating a conformance score for a wide range of guidelines, we believe that the result offers a good indication as to the inherent consumability of the resulting data by generic, domain-agnostic, applications—be it live Linked Data browsers, or warehousing engines, or systems operating on similar principles. Along these lines, we presented a comprehensive summary of results for all providers of a non-trivial amount of data found in our empirical corpus, giving a breakdown of their conformance score for fourteen individual guidelines, as well as their aggregated conformance score and independent PageRank score—we hope that this will serve as a useful reference list for publishers as well as developers of consumer applications.

Non-conformance could be explained by a number of factors. First, certain data accessed during our crawl may be old, and possibly pre-date Linked Data publishing. Second, as we have discussed, while all guidelines are well-motivated in the general case, some guidelines are not necessarily definitive or universal where there may be valid reasons for occasional non-conformance. Third, following certain guidelines may be impractical for certain domains, where for example providing images for all entities is often not practical or useful (esp., if the entities described are more conceptual, such as time periods or sensor measurements, etc.). Fourth, we believe that many patterns emerging in Linked Data publishing are down to precedent, where newer publishers follow the example set out by more established publishers; this may explain, for example, the endemic lack of per-document licencing. That said, we note that guidelines that are core to the original Linked Data principles (use HTTP URIs, make them dereferenceable, etc.) are typically well adhered to, with the possible exception of making inlinks dereferenceable (a specialisation of LDP4).

However, aggregated conformance alone is itself insufficient to characterise the quality of a data-provider: for example, we could—with fairly minimal effort—create a new data-provider that would earn the highest possible aggregated conformance score in our analysis (without necessarily having any meaningful content). Indeed, we have seen (e.g., in Tables A–B) that many times, low-volume publishers are the most compliant with the guidelines. Thus, we see PageRank and other links-based analysis measures as complimenting our conformance scores: our intuition here is that the conformance scores give insights as to the "structural" quality of the data provider's contribution, whereas the PageRank scores give insights into the "importance" of their contribution. We saw that highly-ranked providers tended to be non-conformant with respect to certain guidelines, particular those discouraging use of particular RDF features; we argue that these guidelines are more exceptional in nature. Conversely, highly-ranked providers tended to be more conformant for guidelines pertaining to interlinkage.

Returning to the more general topic of Linked Data quality, we have very much followed Vrandečíc's intuition [97] of looking at specific, quantifiable issues, which often tell more about what data providers are doing wrong, as opposed to what they are doing right. Indeed, one of the most useful indicators of data quality is *competency* with respect to a given task: are the data sufficient to enable a particular application? Linked Data guidelines represent a basic, structural form of competence for applications to locate, parse, retrieve, discover, and consume content. However, other than some specific recommendations with respect to providing human-readable meta-data and licencing information, the presented guidelines are quite vague on the topic of how content should be modelled and presented, what granularity of modelling maximises data utility in the general case, how data should be versioned, how

authorship provenance should be specified, how the semantics of the data can be effectively used to increase interoperability, how the coverage and scope of the data should be advertised, etc. Such issues are inherently difficult to study, but also inherently important to study.

With respect to future work, we would next like to do a more specialised empirical study for vocabularies in Linked Data, particularly their use of the RDFS and OWL standards, how they are interlinked and mapped, how they are externally redefined, what kinds of reasoning the defined semantics enables, what modelling patterns exist, what are the prevalent issues, etc. We have already compiled some initial results on the most commonly used features of RDFS and OWL in Linked Data vocabularies [53, Table 5.2], which we would like to expand into a more comprehensive study of the Linked Vocabulary ecosystem. Furthermore, we hope to repeat the experiments presented in this paper for a future, sufficiently different sample of the Web of Data; we would be particular interested in studying data published as RDFa or other embedded formats, how trends change over time, conformance for new guidelines that emerge (e.g., as per [50]), etc. We would also like to investigate a more granular means of identifying individual data-providers than the current rather "catch-all" notion of PLDs used herein. Finally, we are currently setting up some monitoring experiments that will take snapshots of Linked Data from different publishers at regular intervals, and that when studied, we hope will yield insights into the dynamicity and evolution of such datasets over time.

9 Conclusion

With respect to Linked Data—where the provision of data is only loosely coupled with the modus-operandi of consumer applications—universal notions of quality are inherently difficult to pinpoint and measure. Herein, we have focused on two particular quantifiable aspects relating to Linked Data quality for individual data providers: their conformance with respect to Linked Data guidelines and their PageRank score.

We have offered insights into the current level of conformance with respect to the current wisdom on how to publish Linked Data, where we see, for example, that few providers attach human-readable meta-data to their resources (particularly images), or licencing information to their documents. [44] Similarly, providers often do not provide locally-known inlinks in the dereferenced document of a given resource. Thereafter, lack of such conformance has a varying knock-on effect with respect to consumer applications, which must be taken into account by developers.

We also looked at the PageRank scores of data providers as a complimentary analysis to our conformance measures. We found that highly-ranked data providers are more likely to use RDF features discouraged by Linked Data guidelines, but are also more likely to offer a diverse set of links to external domains.

We then proposed a straightforward aggregated conformance measure for data-providers, presenting results for 188 domains in our sample of data; we considered the `qdos.com` domain to be the most compliant across all guidelines, where many "personal domains" also featured highly. For our proposed aggregated conformance, we found a significant, moderate correlation with PageRank; however, we also found highly-ranked providers that had very low conformance with respect to the stated guidelines. In particular, the `loc.gov` domain, which was highly ranked on the basis of a single link from `dublincore.org`, was found to have a low conformance score. A similar result was given for the more established `geonames.org` and `livejournal.com` domains, etc.; loosely speaking, the data provided by these domains were found to be rather uniform (e.g., not using diverse vocabulary) and insular (e.g., linking to few external domains). Conversely, the two most conformant *and* highly ranked providers were the prominent `dbpedia.org` and `mit.edu` domains.

To conclude, empirical analyses of Linked Data adoption are imperative to understand what is working and what is not and to inform future directions for the Semantic Web standards and Linked Data guidelines. Herein, we presented our own contribution to the area, which focuses on Linked Data conformance. We hope to see many more such empirical analyses—particularly those that go beyond raw dataset sizes, and those that focus on Linked Data quality and usage patterns—emerge in the next few years.

Acknowledgements *We would like to thank the anonymous reviewers and the editors for their helpful feedback and comments. We would also like to thank those involved in the Pedantic Web Group. The work presented herein has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), and by an IRCSET postgraduate scholarship.*

[44] A result echoed by Dodds; see http://www.ldodds.com/tmp/iswc-legal-frameworks-overview.pdf (retr. 2011/09/01) for discussion regarding licencing on the Web of Data.

References

[1] Ben Adida and Mark Birbeck. RDFa Primer. W3C Working Group Note, October 2008. http://www.w3.org/TR/xhtml-rdfa-primer/.

[2] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note, March 2011. http://www.w3.org/TR/void/.

[3] Albert L. Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] Cosmin Basca, Stéphane Corlosquet, Richard Cyganiak, Sergio Fernández, and Thomas Schandl. Neologism: Easy Vocabulary Publishing. In *Proceedings of the Workshop on Scripting for the Semantic Web*, June 2008.

[5] Sean Bechhofer and Raphael Volz. Patching Syntax in OWL Ontologies. In *International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, pages 668–682. Springer, November 2004.

[6] David Beckett and Tim Berners-Lee. Turtle – Terse RDF Triple Language. W3C Team Submission, January 2008. http://www.w3.org/TeamSubmission/turtle/.

[7] Tim Berners-Lee. Linked Data. W3C Design Issues, July 2006. From http://www.w3.org/DesignIssues/LinkedData.html; retr. 2010/10/27.

[8] Tim Berners-Lee. Putting Government Data online. W3C Design Issues, 2009. From http://www.w3.org/DesignIssues/GovData.html; retr. 2011/01/21.

[9] Tim Berners-Lee. The Future of RDF. W3C Design Issues, 2010. From http://www.w3.org/DesignIssues/RDF-Future.html; retr. 2010/10/28.

[10] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In *The 3rd International Semantic Web User Interaction Workshop (SWUI06)*, November 2006.

[11] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986, January 2005. http://tools.ietf.org/html/rfc3986.

[12] Mark Birbeck and Shane McCarron. CURIE Syntax 1.0 – A syntax for expressing Compact URIs. W3C Recommendation, January 2009. http://www.w3.org/TR/curie/.

[13] Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. FactForge: A fast track to the web of data, July 2010.

[14] Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the LOD Cloud. Technical Report V. 0.3, Freie Universität Berlin, 2011. http://www4.wiwiss.fu-berlin.de/lodcloud/state/.

[15] Christian Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web. linkeddata.org Tutorial, July 2008. http://linkeddata.org/docs/how-to-publish.

[16] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[17] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – A crystallization point for the Web of Data. *J. Web Sem.*, 7(3):154–165, 2009.

[18] John G. Breslin, Stefan Decker, Andreas Harth, and Uldis Bojars. SIOC: an approach to connect web-based communities. *IJWBC*, 2(2):133–142, 2006.

[19] Gong Cheng, Weiyi Ge, and Yuzhong Qu. Falcons: searching and browsing entities on the semantic web. In *World Wide Web*, pages 1101–1102, April 2008.

[20] Gong Cheng, Weiyi Ge, Honghan Wu, and Yuzhong Qu. Searching Semantic Web Objects Based on Class Hierarchies. In *Proceedings of Linked Data on the Web Workshop*, 2008.

[21] Gong Cheng, Saisai Gong, and Yuzhong Qu. An empirical study of vocabulary relatedness and its application to recommender systems. In *International Semantic Web Conference (1)*, pages 98–113, 2011.

[22] Gong Cheng and Yuzhong Qu. Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. *Int. J. Semantic Web Inf. Syst.*, 5(3):49–70, 2009.

[23] Mathieu d'Aquin, Claudio Baldassarre, Laurian Gridinoc, Sofia Angeletou, Marta Sabou, and Enrico Motta. Characterizing Knowledge on the Semantic Web with Watson. In *5th International Workshop on Evaluation of Ontologies and Ontology-based Tools*, pages 1–10, November 2007.

[24] Mathieu d'Aquin and Enrico Motta. Watson, more than a semantic web search engine. *Semantic Web*, 2(1):55–63, 2011.

[25] Renaud Delbru, Nickolai Toupikov, Michele Catasta, Giovanni Tummarello, and Stefan Decker. Hierarchical Link Analysis for Ranking Web Data. In *ESWC (2)*, pages 225–239, 2010.

[26] Li Ding and Tim Finin. Characterizing the Semantic Web on the Web. In *Proceedings of the 5th International Semantic Web Conference*, pages 242–257, November 2006.

[27] Li Ding, Timothy W. Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM*, pages 652–659, 2004.

[28] Li Ding, Timothy Lebo, John S. Erickson, Dominic DiFranzo, Gregory Todd Williams, Xian Li, James Michaelis, Alvaro Graves, Jinguang Zheng, Zhenning Shangguan, Johanna Flores, Deborah L. McGuinness, and James A. Hendler. TWC LOGD: A portal for linked open government data ecosystems. *J. Web Sem.*, 9(3):325–333, 2011.

[29] Li Ding, Joshua Shinavier, Tim Finin, and Deborah L. McGuinness. owl:sameAs and Linked Data: An Empirical Study. In *WebSci10: Extending the Frontiers of Society On-Line*, April 2010.

[30] Li Ding, Joshua Shinavier, Zhenning Shangguan, and Deborah L. McGuinness. SameAs networks and beyond: analyzing deployment status and implications of owl:sameAs in linked data. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, pages 145–160, Berlin, Heidelberg, 2010. Springer-Verlag.

[31] Li Ding, Lina Zhou, Tim Finin, and Anupam Joshi. How the Semantic Web is Being Used: An Analysis of FOAF Documents. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04*, pages 113.3–, Washington, DC, USA, 2005. IEEE Computer Society.

[32] Lee Feigenbaum. Cambridge Semantics Position. In *W3C Workshop on RDF Next Steps*, Stanford, Palo Alto, CA, USA, June 2010.

[33] Javier D. Fernández, Claudio Gutierrez, and Miguel A. Martínez-Prieto. RDF compression: basic approaches. In *WWW*, pages 1091–1092, 2010.

[34] Roy T. Fielding, James Gettys, Jeffrey C. Mogul, Henrik Frystyk, Larry Masinter, Paul J. Leach, and Tim Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616, June 1999. http://www.ietf.org/rfc/rfc2616.txt.

[35] Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. TripleRank: Ranking Semantic Web Data by Tensor Decomposition. In *International Semantic Web Conference*, pages 213–228, 2009.

[36] Weiyi Ge, Jianfeng Chen, Wei Hu, and Yuzhong Qu. Object Link Structure in the Semantic Web. In *ESWC (2)*, volume 6089 of *Lecture Notes in Computer Science*, pages 257–271. Springer, 2010.

[37] Hugh Glaser, Ian Millard, and Afraz Jaffri. RKBExplorer.com: a knowledge driven infrastructure for linked data providers. In *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, ESWC'08, pages 797–801, Berlin, Heidelberg, 2008. Springer-Verlag.

[38] Gunnar Aastrand Grimnes. (still) Nothing Clever. Personal Weblog. http://gromgull.net/blog/category/semantic-web/billion-triple-challenge/; retr. 2012/01/12.

[39] Christophe Guéret, Paul T. Groth, Frank van Harmelen, and Stefan Schlobach. Finding the Achilles Heel of the Web of Data: Using Network Analysis for Link-Recommendation. In *9th International Semantic Web Conference*, pages 289–304, November 2010.

[40] Harry Halpin. A Query-Driven Characterization of Linked Data. In *3rd International Workshop on Linked Data on the Web (LDOW2009)*, April 2009.

[41] Harry Halpin. Is there anything worth finding on the semantic web? In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1065–1066, New York, NY, USA, 2009. ACM.

[42] Harry Halpin and Patrick J. Hayes. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *3rd International Workshop on Linked Data on the Web (LDOW2010)*, April 2010.

[43] Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In *International Semantic Web Conference*, pages 305–320, November 2010.

[44] Steve Harris, Andy Seaborne, and Eric Prud'hommeaux. SPARQL 1.1 Query Language. W3C Working Draft, October 2010. http://www.w3.org/TR/sparql11-query/.

[45] Andreas Harth, Sheila Kinsella, and Stefan Decker. Using Naming Authority to Rank Data and Ontologies for Web Search. In *International Semantic Web Conference*, pages 277–292, 2009.

[46] Olaf Hartig. Provenance Information in the Web of Data. In *3rd International Workshop on Linked Data on the Web (LDOW2009)*, April 2009.

[47] Olaf Hartig, Christian Bizer, and Johann-Christoph Freytag. Executing SPARQL queries over the Web of Linked Data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 293–309, Berlin, Heidelberg, 2009. Springer-Verlag.

[48] Michael Hausenblas, Wolfgang Halb, Yves Raimond, and Tom Heath. What is the Size of the Semantic Web? In *I-Semantics 2008: International Conference on Semantic Systems*, Graz, Austria, 2008.

[49] Patrick Hayes. RDF Semantics. W3C Recommendation, February 2004. http://www.w3.org/TR/rdf-mt/.

[50] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st Edition)*, volume 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011. Available from http://linkeddatabook.com/editions/1.0/.

[51] Martin Hepp. Product Variety, Consumer Preferences, and Web Technology: Can the Web of Data Reduce Price Competition and Increase Customer Satisfaction? In *EC-Web*, page 144, 2009.

[52] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. OWL 2 Web Ontology Language Primer. W3C Recommendation, October 2009. http://www.w3.org/TR/owl2-primer/.

[53] Aidan Hogan. *Exploiting RDFS and OWL for Integrating Heterogeneous, Large-Scale, Linked Data Corpora*. PhD thesis, Digital Enterprise Research Institute, National University of Ireland, Galway, 2011. Available from http://aidanhogan.com/docs/thesis/.

[54] Aidan Hogan, Andreas Harth, and Stefan Decker. Performing Object Consolidation on the Semantic Web Data Graph. In *1st I3 Workshop: Identity, Identifiers, Identification Workshop*, 2007.

[55] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the Pedantic Web. In *3rd International Workshop on Linked Data on the Web (LDOW2010)*, April 2010.

[56] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing Linked Data with SWSE: the Semantic Web Search Engine. *J. Web Sem.*, 9(4):365–401, 2011.

[57] Aidan Hogan, Jeff Z. Pan, Axel Polleres, and Stefan Decker. SAOR: Template Rule Optimisations for Distributed Reasoning over 1 Billion Linked Data Triples. In *International Semantic Web Conference*, 2010.

[58] Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan Decker. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *J. of Web Sem.*, 2012. In press.

[59] Wei Hu, Jianfeng Chen, Gong Cheng, and Yuzhong Qu. ObjectCoref and Falcon-AO: Results for OAEI 2010. In *Fifth International Workshop on Ontology Matching*, November 2010.

[60] Wei Hu, Jianfeng Chen, and Yuzhong Qu. A self-training approach for resolving object coreference on the semantic web. In *Proc. of WWW 2012*, pages 87–96. ACM, 2011.

[61] Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P. Sheth. Linked Data is Merely More Data. In *In: AAAI Spring Symposium "Linked Data Meets Artificial Intelligence", AAAI*, pages 82–86. AAAI Press, 2010.

[62] Anja Jentzsch, Jun Zhao, O. Hassanzadeh, Kei-Hoi Cheung, Matthias Samwal, and Bosse Andersson. Linking Open Drug Data (Triplification Challenge Report). In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS'09)*, 2009.

[63] Cliff Joslyn, Bob Adolf, Sinan al Saffar, John Feo, Eric Goodman, David Haglin, Greg Mackey, and David Mizell. High Performance Semantic Factoring of Giga-Scale Semantic Graph Databases, November 2010. Billion Triple Challenge 2010.

[64] Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938.

[65] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[66] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets Semantic Web – how the BBC uses DBpedia and Linked Data to make connections. In *ESWC*, pages 723–737, 2009.

[67] Markus Krötzsch, Denny Vrandecic, Max Völkel, Heiko Haller, and Rudi Studer. Semantic wikipedia. *J. Web Sem.*, 5(4):251–261, 2007.

[68] Hsin-Tsang Lee, Derek Leonard, Xiaoming Wang, and Dmitri Loguinov. IRLbot: scaling to 6 billion pages and beyond. In *WWW*, pages 427–436, 2008.

[69] Rhys Lewis. Dereferencing HTTP URIs. Draft Tag Finding, May 2007. http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14; retr. 2011/02/25.

[70] Alejandro Mallea, Marcelo Arenas, Aidan Hogan, and Axel Polleres. On blank nodes. In *International Semantic Web Conference*, pages 421–437, 2011.

[71] Michael Martin, Jörg Unbehauen, and Sören Auer. Improving the Performance of Semantic Web Applications with SPARQL Query Caching. In *ESWC (2)*, pages 304–318, 2010.

[72] B. Scott Michel, Konstantinos Nikoloudakis, Peter L. Reiher, and Lixia Zhang. URL Forwarding and Compression in Adaptive Web Caching. In *INFOCOM*, pages 670–678, 2000.

[73] Peter Mika, Edgar Meij, and Hugo Zaragoza. Investigating the semantic gap through query log analysis. In *International Semantic Web Conference*, pages 441–455, 2009.

[74] Alistair Miles, Thomas Baker, and Ralph Swick. Best Practice Recipes for Publishing RDF Vocabularies, March 2006. Version available from: http://www.w3.org/TR/2006/WD-swbp-vocab-pub-20060314/l. Superseded by Berrueta & Phipps: http://www.w3.org/TR/swbp-vocab-pub/.

[75] Knud Möller, Michael Hausenblas, Richard Cyganiak, Gunnar Aastrand Grimnes, and Siegfried Handschuh. Learning from Linked Open Data usage: Patterns & metrics. In *WebScience 2010*, 2010.

[76] Knud Möller, Tom Heath, Siegfried Handschuh, and John Domingue. Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects. In *ISWC/ASWC*, pages 802–815, 2007.

[77] Sergio Muñoz, Jorge Pérez, and Claudio Gutiérrez. Minimal Deductive Systems for RDF. In *ESWC*, pages 53–67, 2007.

[78] Mark Nottingham and Eran Hammer-Lahav. Defining Well-Known Uniform Resource Identifiers (URIs). RFC 5785, April 2010. http://www.ietf.org/rfc/rfc5785.txt.

[79] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[80] Emmanuel Pietriga, Christian Bizer, David R. Karger, and Ryan Lee. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *International Semantic Web Conference*, pages 158–171, 2006.

[81] Niko P. Popitsch and Bernhard Haslhofer. DSNotify: handling broken links in the web of data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 761–770, New York, NY, USA, 2010. ACM.

[82] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C Recommendation, January 2008. http://www.w3.org/TR/rdf-sparql-query/.

[83] Yves Raimond and Mark B. Sandler. A web of musical information. In *ISMIR*, pages 263–268, 2008.

[84] Yves Raimond, Christopher Sutton, and Mark B. Sandler. Interlinking Music-Related Data on the Web. *IEEE MultiMedia*, 16(2):52–63, 2009.

[85] Leo Sauermann and Richard Cyganiak. Cool URIs for the Semantic Web. W3C Interest Group Note, December 2008. http://www.w3.org/TR/cooluris/.

[86] John Sheridan and Jeni Tennison. Linking uk government data. In *LDOW*, 2010.

[87] Lian Shi, Diego Berrueta, Sergio Fernández, Luis Polo, and Silvino Fernández. Smushing RDF instances: are Alice and Bob the same open source developer? In *PICKME Workshop*, 2008.

[88] Jennifer Sleeman and Tim Finin. Learning Co-reference Relations for FOAF Instances. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[89] Stephen Stigler. Fisher and the 5% level. *Chance*, 21(4):12, 2008.

[90] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, and Stefan Decker. Sig.ma: Live views on the Web of Data. In *Semantic Web Challenge (ISWC2009)*, 2009.

[91] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice.com: Weaving the Open Linked Data. In *ISWC/ASWC*, pages 552–565, 2007.

[92] Jürgen Umbrich, Andreas Harth, Aidan Hogan, and Stefan Decker. Four heuristics to guide structured content crawling. In *Proceedings of the 2008 Eighth International Conference on Web Engineering-Volume 00*, pages 196–202. IEEE Computer Society, 2008.

[93] Jürgen Umbrich, Michael Hausenblas, Aidan Hogan, Axel Polleres, and Stefan Decker. Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In *3rd International Workshop on Linked Data on the Web (LDOW2010) at WWW2010*, Raleigh, USA, April 2010.

[94] Jacopo Urbani, Spyros Kotoulas, Jason Maassen, Frank van Harmelen, and Henri E. Bal. OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples. In *ESWC (1)*, pages 213–227, 2010.

[95] Jacopo Urbani, Spyros Kotoulas, Eyal Oren, and Frank van Harmelen. Scalable Distributed Reasoning Using MapReduce. In *International Semantic Web Conference*, pages 634–649, 2009.

[96] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and Maintaining Links on the Web of Data. In *8th International Semantic Web Conference*, pages 650–665, November 2009.

[97] Denny Vrandečíc. *Ontology Evaluation*. PhD thesis, Karlsruhe Institute of Technology, June 2010.

[98] Taowei David Wang, Bijan Parsia, and James A. Hendler. A Survey of the Web Ontology Landscape. In *International Semantic Web Conference*, pages 682–694, 2006.

[99] Jesse Weaver. Redefining the RDFS Closure to be Decidable. In *W3C Workshop on RDF Next Steps*, Stanford, Palo Alto, CA, USA, June 2010.

[100] Jesse Weaver and James A. Hendler. Parallel Materialization of the Finite RDFS Closure for Hundreds of Millions of Triples. In *International Semantic Web Conference (ISWC2009)*, pages 682–697, 2009.

№	pld	cs	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	pr
1	qdos.com	83	99	100	93	84	89	93	97	100	99	96	1	70	93	50	30
2	harth.org†	81	97	97	77	85	100	94	100	100	54	78	9	98	100	50	43
3	wikier.org	81	94	96	71	81	89	86	96	100	89	92	25	97	68	50	58
4	iandavis.com	80	99	99	52	87	100	81	82	99	89	90	14	100	83	50	44
5	l3s.de	79	100	100	94	8	98	94	79	100	28	57	50	100	100	100	27
6	dbpedia.org‡	79	100	100	98	52	86	97	96	100	30	81	45	100	100	18	8
7	tommorris.org	78	65	99	83	88	100	73	79	100	60	73	25	100	100	50	78
8	bblfish.net	78	74	97	64	45	100	85	70	98	88	82	39	95	98	50	26
9	gromgull.net	77	58	99	73	100	100	62	79	100	90	78	19	100	91	33	81
10	mit.edu	75	83	94	51	49	92	91	94	99	99	97	19	95	71	22	12
11	deri.org‡	75	99	100	55	12	42	96	98	100	97	92	32	99	95	36	57
12	kanzaki.com	75	76	100	45	48	100	71	70	94	98	93	48	97	90	22	46
13	brainbird.net	75	100	100	78	79	89	79	70	100	76	52	11	83	83	50	73
14	identi.ca	75	100	100	76	89	98	100	70	100	76	52	13	59	65	50	9
15	rdfweb.org	74	93	94	58	27	100	90	70	100	61	87	23	100	86	49	11
16	openlinksw.com	74	100	100	83	5	80	96	90	100	93	99	28	88	44	28	52
17	jobsonica.com	74	100	100	88	97	78	26	70	100	81	52	10	92	91	50	150
18	sourceforge.net	73	63	100	62	59	100	69	70	91	84	73	29	98	92	39	15
19	geospecies.org	73	100	88	74	48	–	78	70	100	55	96	1	100	45	100	68
20	fu-berlin.de	73	100	99	57	6	95	82	94	100	82	87	23	98	75	30	21
21	urmf.net	73	100	100	67	99	100	44	70	100	76	52	14	74	82	50	144
22	dbtune.org	73	100	100	90	9	61	73	88	100	29	78	50	100	99	47	51
23	status.net	73	100	100	55	64	79	99	70	100	76	52	11	86	83	50	22
24	w3.org	73	73	99	46	22	100	91	95	98	100	100	21	99	24	50	1
25	ivan-herman.net	72	59	99	63	68	67	78	90	91	77	89	5	100	100	28	24
26	iberliner.de	72	100	100	55	77	100	55	70	100	76	52	12	86	80	50	108
27	ontologi.es	72	68	93	69	71	89	72	90	98	45	89	7	93	83	43	67
28	walsh.name	72	66	99	50	55	89	75	70	100	83	73	0	100	100	50	32
29	b4mad.net	72	54	98	72	11	98	70	70	100	96	98	19	99	84	41	94
30	bleeper.de	72	100	100	69	88	85	65	70	100	76	52	12	71	70	50	103
31	bio2rdf.org	72	100	100	100	95	77	69	93	100	27	81	16	100	0	50	76
32	budez.com	72	100	100	83	94	–	35	70	100	76	52	13	79	83	50	134
33	linkedmdb.org	72	100	100	76	19	91	58	86	100	23	57	48	100	98	48	136
34	rdfize.com	72	100	100	95	61	99	35	86	100	26	73	66	99	14	50	93
35	soton.ac.uk	72	78	100	56	10	99	99	93	100	97	84	7	78	56	48	29
36	floss.pro	72	100	100	58	89	100	58	70	100	76	52	13	73	65	50	131
37	deri.ie‡	71	93	99	37	53	26	93	96	96	95	88	31	94	94	5	34
38	3kbo.com	71	95	100	50	33	100	84	79	100	92	84	3	84	45	45	100
39	semanticweb.org	71	84	100	48	36	97	98	98	98	85	93	32	59	44	22	10
40	nickshanks.com	71	27	97	82	76	–	78	86	100	87	84	35	27	92	50	135
41	chirup.com	71	100	100	80	98	–	11	70	100	81	52	11	83	84	50	176
42	plasr.com	71	100	100	54	91	98	26	70	100	81	52	11	75	82	50	163
43	0nl1ne.at	70	100	100	56	86	100	35	70	100	76	52	11	81	71	50	144
44	android-pt.net	70	100	100	67	93	–	44	70	100	76	52	13	74	78	50	131
45	tomheath.com	70	93	100	55	31	96	80	94	100	86	73	1	89	40	45	63
46	kit.edu‡	70	99	100	58	50	100	84	70	100	20	52	51	100	99	0	64
47	br3nda.com	70	100	100	60	70	–	62	70	100	61	21	17	100	100	50	102
48	androidpt.com	70	100	100	61	86	–	44	70	100	76	52	13	79	79	50	116
49	myxavier.com	70	95	100	68	96	100	11	70	100	59	52	9	87	81	50	163
50	microblogs.org	70	100	100	62	96	–	49	70	100	76	52	10	80	63	50	119
51	kjernsmo.net	70	34	98	68	54	100	78	93	100	90	92	23	46	66	33	31
52	mkbergman.com	70	94	100	54	21	100	97	70	100	85	78	11	81	36	48	83
53	tweet.ie	69	100	100	48	93	–	49	70	100	81	52	9	78	73	50	143
54	recit.org	69	100	100	62	70	94	55	70	100	76	52	13	67	63	50	113
55	opera.com	69	91	100	86	87	99	75	70	100	44	52	11	97	9	50	28
56	nuigalway.ie‡	69	97	99	64	2	100	62	82	91	63	78	23	100	100	7	84
57	uriburner.com	69	91	100	88	6	61	95	99	99	56	99	0	97	77	0	107
58	advogato.org	69	100	100	63	57	100	91	97	100	32	16	38	98	27	50	23
59	9001700.com	69	100	100	47	81	–	49	70	100	76	52	9	83	79	50	112
60	twitterpile.com	69	100	100	47	79	–	35	70	100	81	52	11	88	81	50	176
61	kasei.us	69	46	100	66	34	100	75	93	100	98	97	6	95	19	35	55
62	dagoneye.it	68	90	100	61	15	100	92	70	100	95	81	4	75	25	49	85
63	milaro.net	68	93	100	55	82	100	35	70	100	59	52	11	80	71	50	140
64	mindswap.org	68	76	96	40	46	81	63	70	98	81	78	24	100	74	28	16
65	naijapulse.com	68	100	100	61	66	73	49	70	100	76	52	10	76	72	50	150
66	bbc.co.uk	68	89	100	58	19	89	83	79	100	45	92	33	96	50	21	48
67	pipian.com	68	95	100	56	32	89	71	70	100	86	82	0	82	42	46	104
68	taxonconcept.org	68	100	98	59	30	–	69	70	100	22	78	36	93	78	50	159
69	4july.me	68	100	100	66	99	36	35	70	100	76	52	16	77	74	50	163
70	sodetailed.org	68	100	100	43	54	–	44	70	100	76	52	16	92	87	50	127
71	tuitui.info	68	100	100	45	72	–	55	70	100	76	52	7	83	72	50	114
72	torrez.us	68	84	99	53	44	100	89	70	100	91	81	0	64	26	49	59
73	nytimes.com	68	100	100	70	41	100	49	88	100	15	68	0	97	53	67	66
74	mulestable.net	68	100	100	56	82	–	44	70	100	76	52	11	74	65	50	131
75	mfd-consult.dk	67	63	96	29	58	–	80	70	100	96	96	0	92	24	70	47
76	lingvoj.org	67	98	100	52	95	100	58	86	99	12	52	2	99	65	25	71
77	bentio.com	67	100	100	45	80	–	35	70	100	76	52	13	83	70	50	144
78	semantic-web.at	67	93	100	52	16	100	95	70	100	84	78	1	77	26	48	33
79	revyu.com	67	100	100	55	43	99	89	99	100	29	69	22	62	40	33	61
80	snell-pym.org.uk	67	57	99	18	24	100	69	90	98	94	95	0	94	33	67	117
81	tflearning.com	67	100	100	77	45	100	65	70	100	42	66	0	91	36	45	188
82	twit.tv	67	41	100	77	94	100	69	70	100	52	16	12	74	82	50	70
83	plik.com.br	67	100	100	66	85	–	26	70	100	81	52	13	63	64	50	176
84	larkc.eu	67	94	100	57	60	100	87	70	100	42	66	1	80	28	48	111
85	petstatus.com	67	100	100	46	61	–	44	70	100	76	52	15	80	73	50	127
86	heppnetz.de	66	81	100	73	52	100	35	70	95	19	57	46	100	100	0	49
87	idehen.name	66	99	100	88	4	–	35	79	100	55	90	34	79	56	41	96
88	smhowell.net	66	99	69	55	63	–	49	79	99	43	81	1	83	42	96	137
89	leuksman.com	66	98	100	49	62	–	44	70	100	62	52	11	84	76	50	110
90	tllts.org	66	98	100	59	46	100	11	70	100	59	52	9	85	82	50	106
91	robots.net	66	100	100	45	64	100	62	70	100	32	16	38	94	49	50	124
92	googlecode.com	66	100	100	81	37	100	26	70	100	5	52	48	100	100	0	17
93	twatter.com	66	90	100	62	74	–	55	70	100	59	52	14	65	61	50	150
94	linkedct.org	65	100	100	61	42	66	35	79	100	19	16	50	100	100	49	90

Table A
Aggregated Table: Top Half

№	pld	cs	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	pr
95	ldodds.com	65	71	99	20	76	100	78	70	100	91	87	0	77	0	45	5
96	ordnancesurvey.co.uk	65	100	100	65	21	100	26	70	100	26	78	19	100	59	50	25
97	squeak.im	65	90	100	48	90	44	35	70	100	59	52	12	87	76	50	142
98	metafora.it	65	100	100	58	37	100	44	70	100	35	66	12	96	43	48	138
99	waka.me	65	62	100	50	98	100	26	70	100	52	16	17	85	82	50	126
100	haktweet.com	64	100	100	59	74	–	11	70	100	81	52	14	66	61	50	175
101	us.es	64	100	100	53	63	–	26	70	100	76	52	16	72	61	50	88
102	ajft.org	64	39	100	14	67	100	69	70	100	94	95	0	100	0	50	97
103	ontologyd.p..org	64	98	100	65	13	100	44	70	99	13	52	50	97	96	0	130
104	scoffoni.net	64	100	100	37	56	–	63	70	100	63	21	6	90	75	50	89
105	eye48.com	64	98	100	81	5	100	49	88	100	54	88	1	93	14	23	105
106	linkedlifedata.com	64	100	100	99	31	–	49	70	100	7	21	2	100	99	50	155
107	dublincore.org	64	100	100	51	59	100	26	70	100	12	52	44	100	78	0	2
108	dowhatimean.net†	64	84	100	63	20	89	88	70	100	53	73	0	73	28	49	72
109	appspot.com	63	100	100	97	55	54	98	70	100	24	2	36	99	2	50	35
110	titticimmino.com	63	95	100	60	14	100	72	70	100	42	66	4	83	32	46	157
111	ebusiness-unibw.org	63	100	100	61	16	100	26	70	100	9	52	50	100	100	0	75
112	hii.co.il	63	100	100	49	73	–	35	70	100	53	21	15	78	75	50	144
113	umbel.org	63	100	100	60	71	99	63	70	100	11	57	0	100	0	50	53
114	shef.ac.uk	63	35	98	60	53	–	44	79	100	82	52	15	100	69	31	60
115	alandix.com	63	94	100	54	18	100	81	70	100	42	66	1	78	28	48	95
116	svenbilt.com	63	62	100	49	65	100	49	70	100	52	16	12	82	70	50	91
117	semanticschool.com	63	93	100	50	25	99	88	70	100	42	66	0	71	22	49	139
118	bluereek.com	63	96	100	47	32	100	58	70	100	42	66	1	83	31	48	184
119	miggi.fi	63	49	100	41	91	100	55	70	100	52	16	16	65	70	50	163
120	talis.com	63	96	99	41	15	44	80	88	100	43	94	28	99	35	13	18
121	ontologyportal.org	62	100	100	60	23	100	11	79	100	13	2	37	100	100	50	38
122	data.gov.uk	62	100	100	66	29	–	35	79	100	21	78	17	100	37	50	87
123	wasab.dk	62	53	100	40	11	74	87	70	97	93	98	11	87	17	35	39
124	secondintegral.com	62	93	100	57	26	89	70	70	100	42	66	2	81	27	48	162
125	dbtropes.org	62	100	100	94	27	–	26	79	100	10	21	13	98	90	50	185
126	uniprot.org	62	100	100	88	65	82	58	70	47	11	57	2	99	87	0	86
127	dataincubator.org	62	100	100	48	38	–	26	79	100	24	82	15	100	43	50	174
128	yeebok.com	62	97	100	54	80	–	26	70	100	62	52	7	65	41	50	163
129	bestbuy.com	62	100	100	52	14	93	26	70	100	16	57	1	100	90	46	118
130	thesecretlair.com	62	100	100	63	39	–	26	70	100	81	52	15	57	49	50	144
131	diving-robot.de	61	96	100	62	12	100	62	70	100	42	66	3	81	21	44	160
132	uniba.it	61	94	100	58	3	100	65	70	100	60	66	0	74	19	49	82
133	daml.org	61	98	100	53	39	100	11	70	100	6	52	0	100	96	33	14
134	vocab.org	61	32	100	54	68	53	78	70	99	30	87	36	64	66	21	6
135	ourcoffs.org.au	61	100	100	57	40	11	35	70	100	76	52	16	79	71	50	144
136	tdwg.org	61	100	100	78	43	–	26	70	100	5	21	24	79	100	50	45
137	xmlns.com	61	100	100	71	90	100	26	70	100	14	52	24	34	23	49	3
138	gregheartsfield.com	61	68	81	39	73	100	55	79	100	87	87	16	27	16	25	120
139	zbw.eu	61	97	99	94	83	–	26	70	100	18	57	3	98	21	25	185
140	fluffyandmervin.com	61	98	70	57	62	–	11	70	98	25	73	0	82	41	100	163
141	spip.org	60	34	100	54	92	100	49	70	100	52	16	16	51	61	50	101
142	wingerz.com	60	88	100	50	28	89	65	70	100	34	66	1	74	29	49	141
143	dajobe.org	60	58	100	21	23	50	85	70	99	88	69	8	87	50	33	19
144	freebase.com	60	79	100	98	66	99	49	88	100	1	16	0	100	44	0	62
145	twoozer.com	60	82	100	65	77	–	44	70	100	59	52	10	38	33	50	163
146	sebastiankruk.com	60	90	100	54	20	100	55	70	100	42	66	0	74	19	48	122
147	feebleforce.com	60	39	100	55	47	100	26	70	100	52	16	23	68	92	50	163
148	pbj.ca	60	48	100	41	97	78	26	70	100	52	16	17	66	76	50	98
149	vu.nl	60	64	100	86	18	89	62	70	100	76	57	0	100	14	0	7
150	fancygonzo.net	60	98	100	39	35	44	70	70	100	42	66	0	88	33	49	161
151	daviding.com	60	98	100	50	3	100	58	70	100	42	66	0	85	11	49	180
152	vaenl1.com	59	56	100	37	78	100	11	70	100	52	16	15	75	73	50	163
153	aregar.it	59	91	100	54	26	78	44	70	100	42	66	0	75	29	47	183
154	rkbexplorer.com	59	100	100	55	4	89	26	82	100	2	16	0	100	100	50	74
155	livejournal.com	59	24	100	50	60	96	86	70	100	33	52	0	100	1	50	20
156	designmills.com	59	89	100	60	17	100	26	70	100	34	66	0	78	31	48	121
157	zitgist.com	58	100	100	69	22	67	83	70	100	27	52	22	100	0	8	54
158	telegraphis.net	58	60	100	76	36	100	62	86	100	14	52	14	99	18	0	179
159	s.-w.-grundlagen.de	58	100	100	40	7	100	3	70	100	4	16	25	100	100	50	115
160	korrekt.org	58	100	100	40	30	100	11	70	100	4	16	49	99	94	0	92
161	rdfabout.com	58	33	100	93	24	87	55	82	100	28	73	31	71	32	0	77
162	geonames.org	58	100	100	45	78	96	44	82	100	17	57	0	59	0	28	13
163	fgiasson.com	57	100	100	70	34	66	3	70	100	36	66	0	100	9	49	69
164	eswc2006.org	57	100	100	35	38	100	44	70	100	18	52	18	100	24	2	80
165	europa.eu	57	100	100	97	57	–	44	70	100	15	57	1	100	0	0	36
166	vox.com	57	31	100	58	69	77	35	70	100	44	52	0	100	7	50	65
167	174.129.12.140	56	45	100	88	13	44	35	82	100	23	16	37	100	100	0	99
168	code4lib.org	56	100	87	96	28	97	26	70	100	16	52	7	100	0	2	42
169	outestoi.com	56	51	100	29	69	89	35	70	100	52	16	14	63	42	50	129
170	skipforward.net	55	100	100	59	10	–	11	70	100	6	21	42	100	100	0	156
171	fffff.at	55	52	100	40	51	–	26	70	100	52	16	21	66	70	50	150
172	todaysmama.com	55	33	100	57	47	–	44	70	100	52	16	16	57	69	50	150
173	ahvoice.com	55	31	100	39	84	100	26	70	100	52	16	14	42	39	50	163
174	lexvo.org	54	100	100	84	49	–	55	86	100	8	16	5	100	0	0	125
175	opiumfield.com	54	80	99	51	40	100	3	70	100	31	21	0	100	9	50	158
176	opencalais.com	54	18	100	95	9	96	55	82	100	20	52	1	100	2	21	109
177	fao.org	53	100	100	74	41	100	26	70	100	21	57	0	51	1	0	50
178	reshouts.com	53	22	100	71	72	33	11	70	100	52	16	16	56	71	50	163
179	typepad.com	52	85	96	10	7	100	44	70	91	35	68	0	100	0	25	41
180	kaufkauf.net	52	100	100	60	44	12	26	70	100	10	52	47	100	3	1	185
181	okkam.org	50	20	100	77	35	–	11	70	67	7	16	46	94	61	50	40
182	loc.gov	49	100	57	85	51	100	11	70	100	3	16	0	100	0	0	4
183	prefix.cc†	47	0	100	100	75	–	3	70	100	2	16	0	100	0	50	182
184	ontologycentral.com†	44	2	100	54	29	44	26	79	100	9	21	49	100	0	0	181
185	hi5.com	38	10	100	50	56	99	3	70	100	22	16	0	0	0	0	79
186	lehigh.edu	35	100	100	1	2	100	11	79	100	1	1	0	0	0	0	56
187	hopcroft.name	31	75	100	1	2	–	11	70	100	31	21	0	0	0	0	123
188	unitn.it	31	100	100	1	2	–	26	70	100	3	2	0	0	0	0	37

Table B
Aggregated Table: Bottom Half