BTC-2019: The 2019 Billion Triple Challenge Dataset

José-Miguel Herrera[1], Aidan Hogan[1], and Tobias Käfer[2]

[1] IMFD; DCC, Universidad de Chile
[2] Karlsruhe Institute of Technology (KIT), Germany
{jherrera,ahogan}@dcc.uchile.cl, tobias.kaefer@kit.edu

Abstract. Six datasets have been published under the title of Billion Triple Challenge (BTC) since 2008. Each such dataset contains billions of triples extracted from millions of documents crawed from hundreds of domains. While these datasets were originally motivated by the annual ISWC competition from which they take their name, they would become widely used in other contexts, forming a key resource for a variety of research works concerned with managing and/or analysing diverse, real-world RDF data as found natively on the Web. Given that the last BTC dataset was published in 2014, we prepare and publish a new version – BTC-2019 – containing 2.2 billion quads parsed from 2.6 million documents on 394 pay-level-domains. This paper first motivates the BTC datasets with a survey of research works using these datasets. Next we provide details of how the BTC-2019 crawl was configured. We then present and discuss a variety of statistics that aim to gain insights into the content of BTC-2019. We discuss the hosting of the dataset and the ways in which it can be accessed, remixed and used.
Resource DOI: http://doi.org/10.5281/zenodo.2634588
Resource type: Dataset

1 Introduction

The Billion Triple Challenge (BTC) began at ISWC in 2008 [44], where a dataset of approximately one billion RDF triples crawled from millions of documents on the Web was published. As a demonstration of contemporary Semantic Web technologies, contestants were then asked to submit descriptions of systems capable of handling and extracting value from this dataset, be it in terms of data management techniques, analyses, visualisations, or end-user applications. The challenge was motivated by the need for research on consuming RDF data in a Web setting, where the dataset provided not only a large scale, diverse collection of RDF graphs, but also a snapshot of how real-world RDF data were being published at that time.

A BTC dataset would be published each year from 2008–2012 for the purposes of organising the eponymous challenge at ISWC [44,7,5,6,30], with another BTC dataset published in 2014 [3]. These datasets would become used in a wide variety of contexts unrelated to challenge submissions, not only for evaluating

the performance, scalability and robustness of a variety of systems, but also for analysing Semantic Web adoption in the wild; our survey of how previous BTC datasets have been used (described in more detail in Section 2) reveals :

- **Evaluation**: the BTC datasets have been used for evaluating works on a variety of topics relating to querying [46,25,26,57,35,62,64,65,28], graph analytics [14,15,60,11,33], search [17,8,40,47], linking and matching [49,10,32], reasoning [58,52,42], compression [21,59], provenance [1,61], schemas [9,39], visualisation [22,66], high performance computing [24], information extraction [41], ranking [45], services [53], amongst others.
- **Analysis:** The BTC datasets have further been used for works that aim to analyse the adoption of Semantic Web standards on the Web, including analyses of ontologies and vocabularies [48,23,54], links [20,27], temporal information [51], publishing practices [50], amongst others.

We also found that BTC datasets have been used not only for the eponymous challenges [44,7,5,6,30,29,3], but also for other contests including the TREC Entity Track [2], and the SemSearch Challenge [55].

In summary, the BTC datasets have become a key resource used not only within the Semantic Web community, but also by other communities [14,15,60,11]. Noting that the last BTC dataset was published in 2014 (five years ago at the time of writing), we thus argue that it is nigh time for the release of another BTC dataset (even if not associated with a challenge of the same name).

In this paper, we thus announce the Billion Triple Challenge 2019 dataset. We first provide a survey of how BTC datasets have been used in research works down through the years as both evaluation and analysis datasets. We then describe other similar collections of RDF data crawled from the Web. We provide details on the crawl used to achieve the BTC-2019 dataset, including parameters, seed list, duration, etc.; we also provide statistics collected during the crawl in terms of response codes, triples crawled per hour, etc. Next we provide detailed statistics of the content of the dataset, analysing various distributions relating to triples, documents, domains, predicates, classes, etc., including a high-level comparison with the BTC-2012 and BTC-2014 predecessors; these results further provide insights as to the current state of adoption of the Semantic Web standards on the Web. We then discuss how the data are published and how they can be accessed. We conclude with a summary and outlook for the future.

2 BTC Dataset Adoption

As previously discussed, we found two main types of usage of BTC datasets: for evaluation of systems, and for analysis of the adoption of Semantic Web technologies in the wild. In order to have a clearer picture of precisely how the BTC datasets have been used in the past for research purposes, we performed a number of searches on Google Scholar for the keywords `btc dataset` and `billion triple challenge` (the latter with a phrase search). Given the large

number of results returned, for each search we surveyed the first 50 results, looking for papers that used a BTC dataset for either evaluation or analysis, filtering papers that are later or earlier versions of papers previously found; while this method is incomplete, we already gathered more than enough papers in this sample to get an idea of the past impact of these datasets. We note that Google Scholar uses the number of citations as a ranking measure, such that by considering the first 50 results, we consider the papers with the most impact, but may also bias the sample towards older papers.

In Table 1, we list the research papers found that use a BTC dataset for evaluation purposes; we list a key for the paper, the abbreviation of the venue where it was published, the year it was published, the system, the topic, the year of the BTC dataset used, and the scale of data reported; regarding the latter metric, we consider the figure as reported by the paper itself, where in some cases, samples of a BTC dataset were used, or the BTC dataset was augmented with other sources (the latter cases are marked with '*'). Considering that this is just a sample of papers, we see that BTC datasets have become widely used for evaluation purposes in a diverse range of research topics, in order of popularity: querying (9), graph analytics (5), search (4), linking and matching (3), reasoning (3), compression (2), provenance (2), schemas (2), visualisation (2), high-performance computing (1), information extraction (1), ranking (1), and services (1). While most works consider a Semantic Web setting (dealing with a standard like RDF, RDFS, OWL, SPARQL, etc.), we note that many of the works in the area of graph analytics have no direct connection to the Semantic Web, and rather use the link structure of the dataset to test the performance of network analyses and/or graph algorithms [14,15,60,11]. Furthermore, looking at the venues, we can see that the datasets have been used in works published not only in core Semantic Web venues, but also venues focused on Databases, Information Retrieval, Artificial Intelligence, and so forth. We also remark that many (though not all works) prefer to select a more recent BTC dataset (e.g., from the same year or the year previous).

In Table 2, we instead look at papers that have performed analyses of Semantic Web adoption on the Web based on a BTC dataset. In terms of the types of analysis conducted, most relate to analysis of ontologies/vocabularies (3) or links (2), with temporal meta-data (1) and publishing practices relating to SPARQL endpoint (1) also having been analysed. Though fewer in number, these papers play an important role in terms of guiding Semantic Web research and practice.

Most of the papers discussed were not associated with a challenge (perhaps due to how we conducted our survey). For more information on the challenges using the BTC dataset, we refer to the corresponding descriptions for the TREC [2], SemSearch [55], and Billion Triple Challenges [44,7,5,6,30,29,3].

We reiterate that this is only a sample of the works that have used these datasets, where a deeper search of papers would likely reveal further research depending on the BTC dataset. Likewise, we have only considered published works, and not other applications that may have benefited from or otherwise

Table 1. Use of BTC datasets as evaluation datasets

Paper	Venue	Year	System	Topic	BTC	Max. scale
Neumann & W. [46]	SIGMOD	2009	RDF-3X*	Querying	2008	562,469,278
Urbani et al. [58]	ISWC	2009	–	Reasoning	2008	864,800,000
Delbru et al. [17]	ESWC	2010	SIREN	Search	–	*10,000,000,000
Papadakis et al. [49]	iiWAS	2010	–	Linking	2009	1,150,000,000
Fang et al. [63]	TREC	2010	Purdue	Search	2010	–
Arias et al. [22]	SemSearch	2011	–	Visualisation	2010	13,000,000
Blanco et al. [8]	ISWC	2011	–	Search	2009	1,140,000,000
Böhm et al. [9]	JWS	2011	–	Schema	2010	3,170,000,000
Cheng et al. [14]	ICDE	2011	–	Analytics	2009	673,300,000
Goodman et al. [24]	ESWC	2011	–	HPC	2009	–
Groppe & G. [26]	SAC	2011	–	Queries	2009	830,000,000
Mulay & K. [45]	COMAD	2011	SPRING	Ranking	2010	10,000,000
Ladwig & T. [40]	CIKM	2011	–	Search	2009	10,000,000
Speiser & H. [53]	ESWC	2011	LIDS	Services	2010	3,162,149,151
Cheng et al. [15]	KDD	2012	–	Analytics	2009	773,000,000
Bohm et al. [10]	CIKM	2012	LINDA	Linking	2010	566,200,000
Görlitz et al. [25]	ISWC	2012	SPLODGE	Querying	2011	2,000,000
Neumayer et al. [47]	ECIR	2012	–	Search	2009	1,140,000,000
Wang & C. [60]	VLDB	2012	–	Analytics	2010	773,000,000
Shaw et al. [52]	Datalog	2012	–	Reasoning	2010	3,200,000,000
Umbrich et al. [57]	ISWC	2012	–	Querying	2011	2,145,000,000
Fernandez et al. [21]	JWS	2013	HDT	Compression	2010	232,542,405
Hose & S. [35]	DESWEB	2013	WARP	Querying	2008	*562,469,278
Urbani et al. [59]	Concurrency	2013	–	Compression	–	3,180,000,000
Yang et al. [62]	DASFAA	2013	–	Querying	2010	1,280,000,000
Yuan et al. [64]	VLDB	2013	TripleBit	Querying	2012	1,048,920,108
Zeng et al. [65]	VLDB	2013	Trinity	Querying	2010	3,171,793,030
Bu et al. [11]	VLDB	2014	Pregelix	Analytics	2014	*6,177,086,016
Zhang et al. [66]	JWS	2014	–	Visualisation	2014	1,436,500,000
Liu et al. [42]	Cybernetics	2015	IDIM	Reasoning	2012	1,436,545,555
Avgoustaki et al. [1]	ESWC	2016	–	Provenance	2009	500,000
Gurajada et al. [28]	SIGMOD	2014	TriAD	Querying	2012	1,048,920,108
Konrath et al. [39]	JWS	2012	SchemEx	Schema	2011	2,100,000,000
Lehmerg et al. [41]	JWS	2015	MSJ Engine	Inf. Ex.	2014	4,000,000
Heflin & S. [32]	AAAI	2016	–	Linking	2012	1,400,000
Hogan [33]	TWEB	2017	BLabel	Analytics	2014	4,000,000
Wylot et al. [61]	TKDE	2017	TripleProv	Provenance	2009	42,944,553

leveraged these datasets. Still however, our survey reveals the considerable impact that BTC datasets have had on research in the Semantic Web community, and indeed in other communities. Though the BTC-2019 dataset has only recently been published, we believe that this analysis indicates the potential impact that the newest edition of the BTC dataset should have.

3 Related Work

The BTC datasets are not the only RDF corpora to have been collected from the Web. In this section we cover some of the other initiatives found in the literature for acquiring such corpora.

Predating the release of the first BTC dataset in 2008 were the corpora collected by a variety of search engines operating over Semantic Web data, including Swoogle [19], SWSE [31], Watson [16], Falcons [13], and Sindice [56]. These works described methods for crawling large volumes of RDF data from the Web. Also predating the first BTC dataset, Ding and Finin [18] collected one of the first

Table 2. Use of BTC datasets as analysis datasets

Paper	Venue	Year	Analysis	BTC	Max. scale
Rula et al. [51]	ISWC	2012	Temporal	2011	2,100,000
Ding et al. [20]	ISWC	2010	Linking	2010	9,358,227
Gueret et al. [27]	ISWC	2010	Linking	2009	3,200,000,000
Nikolov & M. [48]	COLD	2010	Ontologies	2009	1,140,000,000
Glimm et al.[23]	LDOW	2012	Ontologies	2011	2,145,000,000
Stadtmüller et al. [54]	CSWS	2012	Ontologies	2011	2,100,000,000
Paulheim & H. [50]	ISWC	2013	Publishing	2012	10,000

large corpora of RDF data from the Web, containing 279,461,895 triples from 1,448,504 documents. They proceeded to analyse a number of aspects of the resulting dataset, including the domains on which RDF documents were found, the age and size of documents, how resources are described, as well as an initial analysis of quality issues relating to `rdfs:domain`. Though these works serve as an important precedent to the BTC datasets, to the best of our knowledge, the corpora were not published and/or were not reused.

On the other hand, since the first BTC dataset, a number of collections of RDF Web data have been published. The Sindice 2011 [12] contains 11 billion statements from 231 million documents, collecting not only RDF but also Microformats, and was used in 2011 for the TREC Entity Track; unfortunately the dataset is no longer available from its original location. The Dynamic Linked Data Observatory (DyLDO) [37] has been collecting RDF data from the Web each week since 2013; compared with the BTC datasets (which are yearly, at best), the DyLDO dataset are much smaller, crawling in the order of 16–100 million quads per week. LOD Laudromat [4] is an initiative to collect, clean, archive and republish Linked Datasets, offering a range of services from descriptive metadata to SPARQL endpoints and visualisations; unlike the BTC datasets, the focus is rather on collecting and republishing datasets in bulk. Meusel et al. [43] have published the WebDataCommons, extracting RDFa, Microdata and Microformats from the massive Common Crawl dataset; the result is a collection of 17,241,313,916 RDF triples, which, to the best of our knowledge, is the largest collection of crawled RDF data to have published to-date; however, the nature of the WebDataCommons dataset is different from a typical BTC instance since it collects a lot of relatively shallow metadata from HTML pages, where the most common properties instantiated by the data are, for example, Open Graph metadata such as `ogp:type`, `ogp:title`, `ogp:url`, `ogp:site_name`, `ogp:image`, etc.; hence while WebDataCommons is an important resource, it is somewhat orthogonal to the BTC series of datasets.

4 Crawl

We follow a similar procedure for crawling the BTC-2019 dataset as in the most recent years. Our crawl uses the most recent version of the LDspider [36] (version

1.3[3]), which offers a variety of features for configuring crawls of native RDF content, including support for various RDF syntaxes, various traversal strategies, various ways to scope the crawl, and most importantly, components to ensure a "polite" crawl that respects the `robots.txt` exclusion protocol and implements a minimal delay between requests to the same server to avoid DoS-like patterns.

The crawl was executed on a single virtual machine running Ubuntu 18.04 on an Intel Xeon Silver 4110 CPU@2.10GHz, with 30G of RAM. The machine was hosted in the University of Chile. Following previous configurations for BTC datasets, LDspider is configured to crawl RDF/XML, Turtle and N-Triples following a breadth-first strategy; the crawler does not yet support JSON-LD, while enabling RDFa currently tends to gather a lot of shallow disconnected metadata from webpages, which we interpret as counter to the goals of BTC datasets. IRIs ending in `.html`, `.xhtml`, `.json`, `.jpg`, `.pdf` are not visited with the assumption that they are unlikely to yield content in one of the desired formats. To enable higher levels of scale, the crawler is configured to use the hard-disk to manage the frontier list (the list of unvisited URLs). Based on initial experiments with the available hardware, 64 threads were chosen for the crawl (adding more threads did not increase performance); implementing a delay between subsequent requests to the same (pay-level) domain is then important to avoid DoS-style polling, where we allow a one second delay. The crawler respects the `robots.txt` exclusion protocol[4] and will not crawl domains or documents that are blacklisted by the respective file. All HTTP(S) IRIs from an RDF document without a blacklisted extension – irrespective of the subject/predicate/object position – are considered candidates for crawling. In each round, IRIs are prioritised in terms of the number of links found, meaning that unvisited IRIs mentioned in more visited documents will be prioritised for crawling. We store the data collected as an N-Quads file, where we use the graph term to indicate the location of the document in which the triple is found; a separate file indicating the redirects encountered, as well as various logs, are also maintained.[5] A diverse list of 442 URLs taken from DyDLO [37] were given as input to the crawl.[6]

We ran the crawl with this configuration continuously for one month from 2018/12/12 until 2019/01/11, during which we collect 2,162,129,316 quads. Since we apply streaming parsers to be able to handle large RDF documents, in cases where a document contains duplicated triples, the initial output will contain duplicate quads; when later removed, we were left with 2,155,856,225 unique

[3] `https://github.com/ldspider/ldspider`

[4] One exception is the `Crawl delay` definition, where all websites are configured for a one second delay only irrespective of the `robots.txt` file.

[5] The script used to run the call – including all arguments passed to LDspider – is available at `https://github.com/jotixh/RDFLiteralDefinitions/blob/master/ldspider-runner/bin/crawl.sh`.

[6] `https://github.com/jotixh/RDFLiteralDefinitions/blob/master/ldspider-runner/seed.txt`

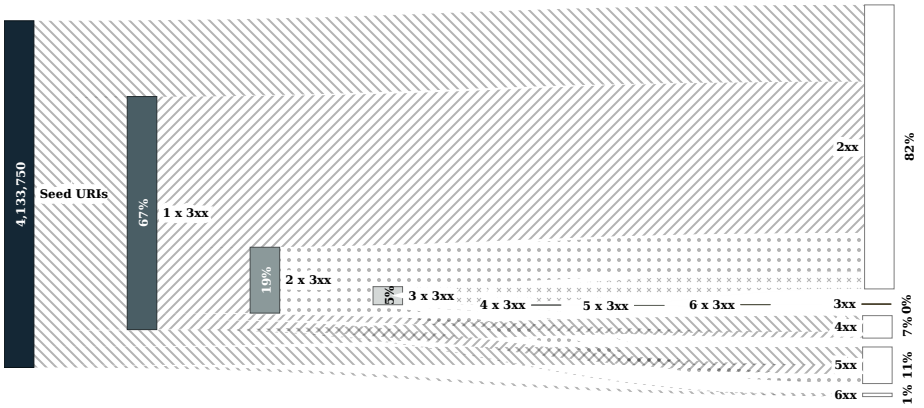

Fig. 1. Sankey diagram showing response codes for the crawled URIs. n x 3xx indicates n-th redirection.

quads in the dataset from a total of 2,641,253 documents on 394 pay-level-domains (PLDs).[7]

In Figure 1, we show the crawling behaviour on the HTTP level. As the HTTP status does not cover issues on the networking level, we added a class (`6xx`) for networking issues, which allows us to present the findings on HTTP and networking level in a uniform manner. We assigned exceptions that we encountered during crawling to the status code classes, according to whether we consider them a server problem (eg. `SSLException`) or a networking issue (eg. `ConnectionTimeoutException`) as in [38]. The number of seed URIs is composed of all URIs we ever tried to dereference during the crawling, where in total we tried to dereference 4,133,750 URIs. We see that about two thirds of seed URIs responded with an HTTP status code of the Redirection class (`3xx`), which are about three times as many as URIs that directly provided a successful response (`2xx`). A total of 6% of requests immediately fail due to server or network issues (`5xx`/`6xx`). In total, 82% of seed URIs eventually yielded a successful response, i.e., about 3.3M seed URIs, which is considerable more than documents in the final crawl (2.6 M); reasons for this difference include the fact that many seed URIs redirect to the same document, that multiple hash URIs from the same documents are in the seed list, etc.

In Figure 2, we show the number of (non-distinct) quads crawled as the days progress, where we see that half of the data are crawled after about 1.6 days; the rate at which quads are crawled decays markedly over time. This decay in

[7] A pay-level domain (PLD) is one that must be paid for to be registered; examples would be `dbpedia.org`, `data.gov`, `bbc.co.uk`, but not `en.dbpedia.org`, `news.bbc.co.uk`, etc. Oftentimes datasets will rather report fully-qualified domain names (FQDNs), which we argue is not a good practice since, for example, sub-domains can be used for individual user accounts (as was the case for sites like Livejournal, which had millions of sub-domains: one for each user).

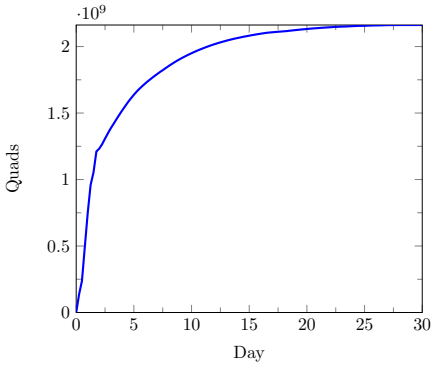

Fig. 2. Quads crawled after each day

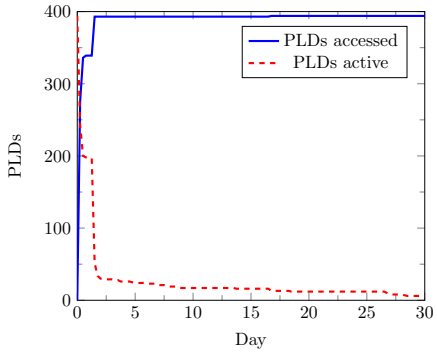

Fig. 3. PLDs included after each day

performance occurs because at the start of the crawl there are more domains to crawl from, where smaller domains are exhausted early in the crawl; this leaves fewer active domains at the end of the crawl. Figure 3 then shows the number of PLDs contributing quads to the crawl as the days progress (accessed), where all but one domain is found after 1.5 days. Figure 3 also shows the number of active PLDs: the PLDs that will contribute quads to the crawl in the future, where for example we see based on the data for day 15 that that the last 15 days of the crawl will retrieve RDF successeful from 16 PLDs. By the end of the crawl, there are only 6 PLDs active from which the crawler can continue to retrieve RDF data. These results explain the trend in Figure 2 of the crawl slowing as it progresses: the crawl enters a phase of incrementally crawling a few larger domains, where the crawl delay becomes the limit to performance. For example, at the end of the crawl, with 6 domains active, a delay limit of 1 second means that 6 documents can be crawled per second. Similiar crawls of RDF documents on the Web have encountered this same phenomenon of "PLD starvation" [34].

In summary, we crawl for 30 days collecting a total of 2,155,856,033 unique quads from 2,641,253 RDF documents on 394 pay-level domains. Per Figure 2, running the crawl for more time would have limited effect on the volume of data.

5 Dataset Statistics

The data are collected from 2,641,253 RDF documents collected from 394 pay-level domains containing a total of 2,155,856,033 unique quads. Surprisingly, the number of unique triples in the dataset is much lower: 256,059,356. This means that on average, each triples is repeated in approximately 8.4 different documents; we will discuss this issue again later. In terms of schema, the data contain 38,156 predicates and instances of 120,037 unique classes; these terms are defined in a total of 1,746 vocabularies (counting unique namespaces).

Next we look at the sources of data for the crawl. RDF content was successfully crawled from a total of 394 different PLDs. In Table 3, we show the

top 25 PLDs with respect to the number of documents crawled and the overall percentage of documents sourced from that site; the largest provider of documents is `dbpedia.org` (6.14%), followed by `loc.gov` (5.68%), etc. We remark that amongst these top PLDs, the distribution is relatively equal. This is because documents are crawled from each domain at a maximum rate of 1 per second, meaning that typically a document will be polled from each active domain with the same interval. To counter the phenomenon of PLD starvation, we stop the polling of active domains when the number of active domains is below a certain threshold and move to the next hop (the documents in the queues of the domains are ranked by in-links as a measure of importance). The result is that large domains are often among the last active domains, where the polling is stopped before the domain is crawled exhaustively and for all domains after downloading almost the same number of documents. However, looking at Table 4, which displays the top 25 PLDs in terms of unique triples, we start to see some skew, where 52.15% of all unique triples come from Wikidata (despite it accounting for only 5.35%) of documents. Even more noticeably, if we look at Table 5, which displays the top-25 PLDs by number of quads, we see that Wikidata accounts for 93.06% of all quads; in fact, if we divide the number of quads for Wikidata by the number of documents, we find that it contains, on average, approximately 14,208 triples per document! By way of comparison, DBpedia contains 226 triples per document. Hence given that the crawl, by its nature, balances the number of documents polled from each domain, and that Wikidata's RDF documents are orders of magnitude larger than those of other domains, we see why the skew in quads occurs. Further cross referencing quads with unique triples, we see a lot of redundancy in how Wikidata exports RDF, repeating each triple in (on average) 15 documents; by way of comparison, DBpedia repeats each unique triple in (on average) 1.11 documents. This skew occurs as a result of how Wikidata chooses to export its data; while representing how real-world data are published, consumers of the BTC-2019 dataset should keep this skew in mind when using the data, particularly if conducting analyses of adoption; for example, analysing the most popularly-used predicates by counting the number of quads using each predicate would be disproportionately affected by Wikidata.

Turning towards the use of vocabularies in the data, Table 6 presents the most popular vocabularies (extracted from predicate and class terms) in terms of the number of PLDs on which they are used (and the percentage of PLDs). Unsurprisingly core Semantic Webs standards head the list, followed by Friend of a Friend (FOAF), Dublin Core (DC) vocabularies, etc.; almost all of these vocabularies have been established for over a decade, with the exception of the Linked Data Platform (LDP) vocabulary which appears in 21st place. On the other hand, Table 7 presents the number of PLDs per predicate, while Table 8 presents the number of PLDs per class, where again there are few surprises at the top of the list, with most terms corresponding to the most popular namespaces. We conclude that BTC 2019 is a highly diverse dataset featuring hundreds of thousands of vocabulary terms from thousands of vocabularies.

Table 3. PLDs by docs.

№	PLD	Docs.	%
1	dbpedia.org	162117	6.14%
2	loc.gov	150091	5.68%
3	bnf.fr	146186	5.53%
4	sudoc.fr	144877	5.49%
5	theses.fr	141228	5.35%
6	wikidata.org	141207	5.35%
7	linkeddata.es	130459	4.94%
8	getty.edu	130398	4.94%
9	fao.org	92838	3.51%
10	ontobee.org	92812	3.51%
11	dbtune.org	91755	3.47%
12	wals.info	88786	3.36%
13	lexvo.org	87584	3.32%
14	ordnancesurvey.co.uk	86801	3.29%
15	idref.fr	83670	3.17%
16	glottolog.org	79365	3.00%
17	l3s.de	77650	2.94%
18	uba.de	73648	2.79%
19	uni-mannheim.de	71883	2.72%
20	pokepedia.fr	70300	2.66%
21	ontologycentral.com	64407	2.44%
22	bl.uk	55951	2.12%
23	d-nb.info	55731	2.11%
24	cnr.it	47955	1.82%
25	bne.es	39437	1.49%

Table 4. PLDs by triples

№	PLD	Triples	%
1	wikidata.org	133535555	52.15%
2	dbpedia.org	32981420	12.88%
3	idref.fr	16820681	6.57%
4	bnf.fr	11769268	4.60%
5	getty.edu	6571525	2.57%
6	linkeddata.es	5898762	2.30%
7	loc.gov	5362064	2.09%
8	sudoc.fr	4972647	1.94%
9	ontologycentral.com	4471962	1.75%
10	theses.fr	4095897	1.60%
11	dbtune.org	3697811	1.44%
12	l3s.de	2747392	1.07%
13	bl.uk	2575875	1.01%
14	glottolog.org	1913034	0.75%
15	d-nb.info	1501742	0.59%
16	wals.info	1441392	0.56%
17	uba.de	1400424	0.55%
18	fao.org	1170742	0.46%
19	pokepedia.fr	1117102	0.44%
20	ordnancesurvey.co.uk	822175	0.32%
21	myexperiment.org	815221	0.32%
22	bne.es	788499	0.31%
23	lexvo.org	774028	0.30%
24	githubusercontent.com	683901	0.27%
25	kit.edu	641578	0.25%

Table 5. PLDs by quads

№	PLD	Quads	%
1	wikidata.org	2006338975	93.06%
2	dbpedia.org	36686161	1.70%
3	idref.fr	22013225	1.02%
4	bnf.fr	12618155	0.59%
5	getty.edu	7453134	0.35%
6	sudoc.fr	7176301	0.33%
7	loc.gov	6725390	0.31%
8	linkeddata.es	6485114	0.30%
9	theses.fr	4820874	0.22%
10	ontologycentral.com	4633947	0.21%
11	dbtune.org	3943928	0.18%
12	bl.uk	3348410	0.16%
13	l3s.de	3084744	0.14%
14	pokepedia.fr	3039193	0.14%
15	myexperiment.org	2401693	0.11%
16	kit.edu	2361368	0.11%
17	glottolog.org	1936776	0.09%
18	d-nb.info	1719665	0.08%
19	uba.de	1474952	0.07%
20	wals.info	1459402	0.07%
21	ontobee.org	1332477	0.06%
22	uni-mannheim.de	1316328	0.06%
23	fao.org	1170742	0.05%
24	ordnancesurvey.co.uk	1165124	0.05%
25	githubusercontent.com	1015635	0.05%

Table 6. PLDs per vocab.

№	Vocab.	PLDs	%
1	rdf:	389	98.73%
2	rdfs:	224	56.85%
3	foaf:	218	55.33%
4	owl:	170	43.15%
5	dce:	145	36.80%
6	dct:	138	35.03%
7	skos:	76	19.29%
8	geo:	58	14.72%
9	admin:	52	13.20%
10	schema:	43	10.91%
11	rss:	36	9.14%
12	con:	34	8.63%
13	bibo:	33	8.38%
14	cc:	31	7.87%
15	void:	28	7.11%
16	cert:	28	7.11%
17	atom:	26	6.60%
18	vann:	23	5.84%
19	sioc:	23	5.84%
20	vcard:	23	5.84%
21	ldp:	23	5.84%
22	doap:	22	5.58%
23	content:	21	5.33%
24	bio:	20	5.08%
25	wot:	19	4.82%

Table 7. PLDs per pred.

№	Predicate	PLDs	%
1	rdf:type	389	98.73%
2	foaf:name	168	42.64%
3	rdfs:label	165	41.88%
4	foaf:homepage	151	38.32%
5	rdfs:seeAlso	146	37.06%
6	foaf:primaryTopic	134	34.01%
7	owl:sameAs	117	29.70%
8	foaf:knows	102	25.89%
9	foaf:maker	102	25.89%
10	dce:title	99	25.13%
11	rdfs:comment	98	24.87%
12	foaf:mbox_sha1sum	98	24.87%
13	foaf:nick	87	22.08%
14	foaf:workplaceHomepage	87	22.08%
15	foaf:depiction	86	21.83%
16	dct:title	79	20.05%
17	rdfs:subClassOf	76	19.29%
18	foaf:title	75	19.04%
19	dct:modified	72	18.27%
20	foaf:mbox	72	18.27%
21	dce:creator	67	17.01%
22	rdfs:range	67	17.01%
23	rdfs:subPropertyOf	67	17.01%
24	rdfs:domain	65	16.50%
25	foaf:family_name	64	16.24%

Table 8. PLDs per class

№	Class	PLDs	%
1	foaf:Person	167	42.39%
2	foaf:PersonalProfileDocument	88	22.34%
3	owl:Class	76	19.29%
4	owl:Ontology	65	16.50%
5	owl:ObjectProperty	61	15.48%
6	foaf:Document	60	15.23%
7	owl:DatatypeProperty	57	14.47%
8	skos:Concept	50	12.69%
9	foaf:Organization	38	9.64%
10	rss:channel	34	8.63%
11	owl:Restriction	34	8.63%
12	rdf:Property	32	8.12%
13	foaf:OnlineAccount	31	7.87%
14	owl:AnnotationProperty	30	7.61%
15	rdf:Seq	27	6.85%
16	rdfs:Class	27	6.85%
17	atom:feed	26	6.60%
18	skos:ConceptScheme	25	6.35%
19	rss:item	24	6.09%
20	geo:Point	24	6.09%
21	foaf:Project	24	6.09%
22	cert:RSAPublicKey	23	5.84%
23	schema:Person	22	5.58%
24	owl:TransitiveProperty	22	5.58%
25	owl:FunctionalProperty	21	5.33%

6 Comparison with BTC-2012 and BTC-2014

We now provide a statistical comparison between BTC-2019 and its two most recent predecessors: BTC-2014 and BTC-2012. We downloaded these two datasets from their corresponding webpages and ran the same statistical code over both as run for the BTC-2019. Noting that BTC-2014 and BTC-2012 included HTTP header meta-data as part of their RDF dump, for the purposes of comparability, we filtered such information from these crawls as they were not part of the native RDF document (and thus were not included in the BTC-2019 files).

We begin in Table 9 with a comparison of high-level statistics between the three datasets, where we see that in terms of quads, BTC-2019 is larger than BTC-2012 but smaller than BTC-2014; as previously discussed, BTC-2014 extracted a lot of shallow HTML-based metadata from small RDFa documents, which we decided to exclude from BTC-2019: as can be seen by cross referencing the quads and documents column, where BTC-2019 had on average 816 quads per document, while BTC-2012 had on average 147 quads per document and BTC-2014 had on average 91 quads per document. Of note is the relatively vast quantity of predicates, classes and vocabularies appearing in the BTC-2014 dataset; upon further analysis, most of this is noise relating to a bug in the exporter of a single site – `gorodskoyportal.ru` – which linked to recursively nested namespaces of the form:

`http://gorodskoyportal.ru/moskva/rss/channel/.../channel/*`

where "..." indicates repetitions of the `channel` sub-path.

We see that BTC-2019 also comes from fewer domains than BTC-2012 and much fewer than BTC-2014; this is largely attributable not only to our decision to not include data embedded in HTML pages, but also to a variety of domains who have ceased publishing RDF data. Regarding the largest contributors of data in terms of PLDs, Table 10 provides a comparison of the domains contributing the most documents to each of the three versions of the BTC datasets, where we see some domains in common across both (e.g., `dbpedia.org`, `loc.gov`), some domains appearing in older versions but not in BTC-2019 that have gone offline (`freebase.com`, `kasabi.com`, `opera.com`, etc.), as well as some new domains appearing only in the more recent BTC-2019 version (e.g., `wikidata.org`).

7 Publication

We publish the files on the Zenodo service, which provides hosting in CERN's data centre and also assigns resources published with DOIs. The DOI of the BTC 2019 dataset is `http://doi.org/10.5281/zenodo.2634588`. The data are published in N-Triples format using GZip compression. Due to the size of the dataset, rather than publish the data as one large file, we publish the following:

Unique triples (1 file: 3.1GB) this file stores only the unique triples of the BTC-2019 dataset.

Table 9. Comparison of BTC 2012, 2014, 2019: High-level Statistics

Statistic	BTC 2012	BTC 2014	BTC 2019
Quads	1,230,391,773	3,974,427,819	2,155,856,033
Unique Triples	974,810,809	3,168,111,983	256,059,356
PLDs	829	47,634	394
Documents	8,373,075	43,598,858	2,641,253
Predicates	57,235	2,192,434	38,156
Classes	296,605	2,700,640	120,037
Vocabularies	1,775	977,606	1,746

Table 10. Comparison of BTC 2012, 2014, 2019: Top PLDs per Documents

№	BTC-2012		BTC-2014		BTC-2019	
	PLD	Docs	PLD	Docs	PLD	Docs
1	dbpedia.org	2,714,588	openlinksw.com	1,885,141	dbpedia.org	162,117
2	freebase.com	1,849,859	crossref.org	1,388,354	loc.gov	150,091
3	data.gov.uk	1,328,918	b3kat.de	1,189,744	bnf.fr	146,186
4	kasabi.com	324,769	legislation.gov.uk	1,153,601	sudoc.fr	144,877
5	opera.com	297,657	sysoon.com	1,142,464	theses.fr	141,228
6	loc.gov	192,125	bibsonomy.org	1,118,619	wikidata.org	141,207
7	fu-berlin.de	162,455	dbpedia.org	1,107,836	linkeddata.es	130,459
8	vu.nl	149,920	loc.gov	1,099,278	getty.edu	130,398
9	europa.eu	145,351	linkedct.org	1,052,459	fao.org	92,838
10	lexvo.org	127,924	rdfize.com	1,049,708	ontobee.org	92,812

Quads (114 files: 26.1GB total) given the large volume of quads, we split the data up, creating a separate file for the quads collected from each of the top 100 PLDs, and an additional file containing the quads for the remaining 294 PLDs. Given the size of Wikidata, we split its file into 14 segments, each containing at most 150 million quads each and taking 1.8GB of space.

Hence we offer consumers a number of options for how they wish to use the BTC-2019 dataset. Consumers who are mostly interested in the graph structure (e.g., for testing graph analytics or queries on a single graph) may choose to download the unique triples file. On the other hand, other consumers can select smaller files from the PLDs of interest, potentially remixing the BTC-2019 into various samples; another possibility, for example, would be to take one file from each PLD (including Wikidata), thus potentially reducing the skew in quads previously discussed. Aside from the data themselves, we also publish a VoID file describing metadata about the crawl, and offer documentation on how to download all of the files at once, potential parsers that can be used, etc.

8 Conclusion

In this paper, we have provided a survey indicating how the BTC datasets have been used down through the years, providing a strong motivation for continuing

the tradition of publishing these datasets. Observing that the last BTC crawl was conducted 5 years ago in 2014, we have thus crawled and published the newest edition to the BTC series: BTC 2019. We have provided various details on the crawl used to acquire the dataset, various statistics regarding the resulting dataset, as well as discussion on how the data are published in a sustainable way.

In terms of the statistics, we noted two problematic aspects: a relatively low number of PLDs contributing to the crawl, leading to exhausting the available PLDs relatively quickly, and a large skew in the number of quads sourced from Wikidata. These observations are based on how the data are published on the Web rather than being a particular artifact of the crawl. Still, the resulting dataset is highly diverse, and can be used for evaluating methods on real-world data; furthermore, with appropriately designed metrics taking into account the skew on Wikidata, the BTC 2019 dataset contains valuable insights on how data are being published on the Web today.

References

1. Avgoustaki, A., Flouris, G., Fundulaki, I., Plexousakis, D.: Provenance Management for Evolving RDF Datasets. In: European Semantic Web Conferece (ESWC). pp. 575–592. Springer (2016)
2. Balog, K., Serdyukov, P., de Vries, A.P.: Overview of the TREC 2011 entity track. In: Text REtrieval Conference (TREC). NIST (2011)
3. Bechhofer, S., Harth, A.: The semantic web challenge 2014. J. Web Semant. **35**, 141 (2015)
4. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data. In: International Semantic Web Conference. pp. 213–228. Springer (2014)
5. Bizer, C., Maynard, D.: The semantic web challenge, 2010. J. Web Semant. **9**(3), 315 (2011)
6. Bizer, C., Maynard, D.: The semantic web challenge, 2011. J. Web Semant. **16**, 32 (2012)
7. Bizer, C., Mika, P.: The semantic web challenge, 2009. J. Web Semant. **8**(4), 341 (2010)
8. Blanco, R., Mika, P., Vigna, S.: Effective and Efficient Entity Search in RDF Data. In: International Semantic Web Conference (ISWC). pp. 83–97. Springer (2011)
9. Böhm, C., Lorey, J., Naumann, F.: Creating void descriptions for web-scale data. J. Web Semant. **9**(3), 339–345 (2011)
10. Böhm, C., de Melo, G., Naumann, F., Weikum, G.: LINDA: distributed web-of-data-scale entity matching. In: ACM International Conference on Information and Knowledge Management (CIKM). pp. 2104–2108. ACM (2012)
11. Bu, Y., Borkar, V.R., Jia, J., Carey, M.J., Condie, T.: Pregelix: Big(ger) Graph Analytics on a Dataflow Engine. PVLDB **8**(2), 161–172 (2014)
12. Campinas, S., Ceccarelli, D., Perry, T.E., Delbru, R., Balog, K., Tummarello, G.: The Sindice-2011 Dataset for Entity-Oriented Search on the Web of Data. In: International Workshop on Entity-Oriented Search (EOS). pp. 26–32 (2011)
13. Cheng, G., Ge, W., Qu, Y.: Falcons: searching and browsing entities on the semantic web. In: International Conference on World Wide Web (WWW). pp. 1101–1102. ACM (2008)

14. Cheng, J., Ke, Y., Chu, S., Özsu, M.T.: Efficient core decomposition in massive networks. In: International Conference on Data Engineering (ICDE). pp. 51–62. IEEE (2011)
15. Cheng, J., Zhu, L., Ke, Y., Chu, S.: Fast algorithms for maximal clique enumeration with limited memory. In: SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 1240–1248. ACM (2012)
16. d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E.: Characterizing Knowledge on the Semantic Web with Watson. In: International Workshop on Evaluation of Ontologies (EON). pp. 1–10. CEUR-WS.org (2007)
17. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G.: A Node Indexing Scheme for Web Entity Retrieval. In: Extended Semantic Web Conference (ESWC). pp. 240–256. Springer (2010)
18. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: International Semantic Web Conference. pp. 242–257. Springer (2006)
19. Ding, L., Finin, T.W., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: International Conference on Information and Knowledge Management (CIKM). pp. 652–659. ACM (2004)
20. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.L.: SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. In: International Semantic Web Conference (ISWC). pp. 145–160. Springer (2010)
21. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). J. Web Semant. **19**, 22–41 (2013)
22. Gallego, M.A., Fernández, J., Martínez-Prieto, M., de la Fuente, P.: RDF visualization using a three-dimensional adjacency matrix. In: Semantic Search Workshop (SEMSEARCH) (2011)
23. Glimm, B., Hogan, A., Krötzsch, M., Polleres, A.: OWL: Yet to arrive on the Web of Data? In: Linked Data on the Web (LDOW). CEUR-WS.org (2012)
24. Goodman, E.L., Jimenez, E., Mizell, D., Al-Saffar, S., Adolf, B., Haglin, D.J.: High-Performance Computing Applied to Semantic Databases. In: Extended Semantic Web Conference. pp. 31–45. Springer (2011)
25. Görlitz, O., Thimm, M., Staab, S.: SPLODGE: Systematic Generation of SPARQL Benchmark Queries for Linked Open Data. In: International Semantic Web Conference (ISWC). pp. 116–132. Springer (2012)
26. Groppe, J., Groppe, S.: Parallelizing join computations of SPARQL queries for large semantic web databases. In: Symposium on Applied Computing (SAC). pp. 1681–1686. ACM (2011)
27. Guéret, C., Groth, P.T., van Harmelen, F., Schlobach, S.: Finding the Achilles Heel of the Web of Data: Using Network Analysis for Link-Recommendation. In: International Semantic Web Conference (ISWC). pp. 289–304. Springer (2010)
28. Gurajada, S., Seufert, S., Miliaraki, I., Theobald, M.: TriAD: a distributed shared-nothing RDF engine based on asynchronous message passing. In: SIGMOD International Conference on Management of Data. pp. 289–300. ACM (2014)
29. Harth, A., Bechhofer, S.: The semantic web challenge 2013. J. Web Semant. **27-28**, 1 (2014)
30. Harth, A., Maynard, D.: The semantic web challenge 2012. J. Web Semant. **24**, 1–2 (2014)

31. Harth, A., Umbrich, J., Decker, S.: MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. In: International Semantic Web Conference (ISWC). pp. 258–271. Springer (2006)
32. Heflin, J., Song, D.: Ontology Instance Linking: Towards Interlinked Knowledge Graphs. In: AAAI Conference on Artificial Intelligence. pp. 4163–4169. AAAI (2016)
33. Hogan, A.: Canonical forms for isomorphic and equivalent RDF graphs: Algorithms for leaning and labelling blank nodes. TWEB **11**(4), 22:1–22:62 (2017)
34. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. J. Web Semant. **9**(4), 365–401 (2011)
35. Hose, K., Schenkel, R.: WARP: workload-aware replication and partitioning for RDF. In: Workshops Proceedings of the International Conference on Data Engineering (ICDE). pp. 1–6. IEEE (2013)
36. Isele, R., Umbrich, J., Bizer, C., Harth, A.: LDspider: An Open-source Crawling Framework for the Web of Linked Data. In: ISWC Posters & Demonstrations. CEUR-WS (2010)
37. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing linked data dynamics. In: Extended Semantic Web Conference (ESWC). pp. 213–227. Springer (2013)
38. Käfer, T., Wins, A., Acosta, M.: Modelling and analysing dynamic linked data using RDF and SPARQL. In: Workshop on Dataset PROFILing and fEderated Search for Web Data (PROFILES) (2017)
39. Konrath, M., Gottron, T., Staab, S., Scherp, A.: SchemEX - Efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Semant. **16**, 52–58 (2012)
40. Ladwig, G., Tran, T.: Index structures and top-k join algorithms for native keyword search databases. In: Conference on Information and Knowledge Management (CIKM). pp. 1505–1514. ACM (2011)
41. Lehmberg, O., Ritze, D., Ristoski, P., Meusel, R., Paulheim, H., Bizer, C.: The Mannheim Search Join Engine. J. Web Semant. **35**, 159–166 (2015)
42. Liu, B., Huang, K., Li, J., Zhou, M.: An Incremental and Distributed Inference Method for Large-Scale Ontologies Based on MapReduce Paradigm. IEEE Trans. Cybernetics **45**(1), 53–64 (2015)
43. Meusel, R., Petrovski, P., Bizer, C.: The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In: International Semantic Web Conference (ISWC). pp. 277–292. Springer (2014)
44. Mika, P., Hendler, J.: The semantic web challenge, 2008. J. Web Semant. **7**(4), 271 (2009)
45. Mulay, K., Kumar, P.S.: SPRING: ranking the results of SPARQL queries on linked data. In: International Conference on Management of Data (COMAD). pp. 47–56. Allied Publishers (2011)
46. Neumann, T., Weikum, G.: Scalable join processing on very large RDF graphs. In: SIGMOD International Conference on Management of Data. pp. 627–640. ACM (2009)
47. Neumayer, R., Balog, K., Nørvåg, K.: When Simple is (more than) Good Enough: Effective Semantic Search with (almost) no Semantics. In: European Conference on IR Research (ECIR). pp. 540–543. Springer (2012)
48. Nikolov, A., Motta, E.: Capturing emerging relations between schema ontologies on the Web of Data. In: Consuming Linked Data (COLD). CEUR (2010)

49. Papadakis, G., Demartini, G., Fankhauser, P., Kärger, P.: The missing links: Discovering hidden same-as links among a billion of triples. In: International Conference on Information Integration and Web-based Applications and Services ( ). pp. 453–460. ACM (2010)
50. Paulheim, H., Hertling, S.: Discoverability of SPARQL endpoints in linked open data. In: ISWC Posters & Demonstrations. pp. 245–248. CEUR-WS.org (2013)
51. Rula, A., Palmonari, M., Harth, A., Stadtmüller, S., Maurino, A.: On the Diversity and Availability of Temporal Information in Linked Open Data. In: International Semantic Web Conference (ISWC). pp. 492–507. Springer (2012)
52. Shaw, M., Koutris, P., Howe, B., Suciu, D.: Optimizing Large-Scale Semi-Naïve Datalog Evaluation in Hadoop. In: International Workshop on Datalog in Academia and Industry. pp. 165–176. Springer (2012)
53. Speiser, S., Harth, A.: Integrating Linked Data and Services with Linked Data Services. In: Extended Semantic Web Conference (ESWC). pp. 170–184. Springer (2011)
54. Stadtmüller, S., Harth, A., Grobelnik, M.: Accessing information about linked data vocabularies with vocab.cc. In: Semantic Web and Web Science - 6th Chinese Semantic Web Symposium and 1st Chinese Web Science Conference, CSWS 2012, Shenzhen, China, November 28-30, 2012. pp. 391–396 (2012)
55. Tran, T., Mika, P., Wang, H., Grobelnik, M.: Semsearch'11: the 4th semantic search workshop. In: International Conference on World Wide Web (Companion Volume). pp. 315–316. ACM (2011)
56. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. In: International Semantic Web Conference and Asian Semantic Web Conference (ISWC/ASWC). pp. 552–565. Springer (2007)
57. Umbrich, J., Karnstedt, M., Hogan, A., Parreira, J.X.: Hybrid SPARQL queries: Fresh vs. fast results. In: International Semantic Web Conference (ISWC). pp. 608–624. Springer (2012)
58. Urbani, J., Kotoulas, S., Oren, E., van Harmelen, F.: Scalable Distributed Reasoning Using MapReduce. In: International Semantic Web Conference (ISWC). pp. 634–649. Springer (2009)
59. Urbani, J., Maassen, J., Drost, N., Seinstra, F.J., Bal, H.E.: Scalable RDF data compression with MapReduce. Concurrency and Computation: Practice and Experience **25**(1), 24–39 (2013)
60. Wang, J., Cheng, J.: Truss decomposition in massive networks. PVLDB **5**(9), 812–823 (2012)
61. Wylot, M., Cudré-Mauroux, P., Hauswirth, M., Groth, P.T.: Storing, tracking, and querying provenance in linked data. IEEE Trans. Knowl. Data Eng. **29**(8), 1751–1764 (2017)
62. Yang, T., Chen, J., Wang, X., Chen, Y., Du, X.: Efficient SPARQL Query Evaluation via Automatic Data Partitioning. In: Database Systems for Advanced Applications (DASFAA). pp. 244–258. Springer (2013)
63. Yi, Si, L., Somasundaram, N., Al-Ansari, S., Yu, Z., Xian, Y.: Purdue at trec 2010 entity track: A probabilistic framework for matching types between candidate and target entities
64. Yuan, P., Liu, P., Wu, B., Jin, H., Zhang, W., Liu, L.: TripleBit: a Fast and Compact System for Large Scale RDF Data. PVLDB **6**(7), 517–528 (2013)
65. Zeng, K., Yang, J., Wang, H., Shao, B., Wang, Z.: A Distributed Graph Engine for Web Scale RDF Data. PVLDB **6**(4), 265–276 (2013)
66. Zhang, X., Song, D., Priya, S., Daniels, Z., Reynolds, K., Heflin, J.: Exploring linked data with contextual tag clouds. J. Web Semant. **24**, 33–39 (2014)