SMW Cloud: A Corpus of Domain-Specific Knowledge Graphs from Semantic MediaWikis

Daniil Dobriy[1(✉)], Martin Beno[1], and Axel Polleres[1,2]

[1] Vienna University of Economics and Business, Vienna, Austria
{daniil.dobriy,martin.beno,axel.polleres}@wu.ac.at
[2] Complexity Science Hub, Vienna, Austria

Abstract. Semantic wikis have become an increasingly popular means of collaboratively managing Knowledge Graphs. They are powered by platforms such as Semantic MediaWiki and Wikibase, both of which enable MediaWiki to store and publish structured data. While there are many semantic wikis currently in use, there has been little effort to collect and analyse their structured data, nor to make it available for the research community. This paper seeks to address this gap by systematically collecting structured data from an extensive corpus of Semantic-MediaWiki-powered portals and providing an in-depth analysis of the ontological diversity (and re-use) amongst these wikis using a variety of ontological metrics. Our paper aims to demonstrate that semantic wikis are a valuable and extensive part of Linked Open Data (LOD), and in fact may be considered an own active "sub-cloud" within the LOD ecosystem, which can provide useful insights into the evolution of small and medium-sized domain-specific Knowledge Graphs.

Keywords: Semantic MediaWiki · Linked Open Data · Ontology · Ontology Metrics · Ontology Links

Resource accessibility

The resource is made openly available with the necessary access information provided below. Section 5.3 provides a sustainability plan, and details further implemented access modalities for the resource.

Resource Type: Dataset
DOI: 10.5281/zenodo.7920174

URL: https://semantic-data.cluster.ai.wu.ac.at/smwcloud/

[1] For more information on the Creative Commons Attribution 4.0 International License, visit https://creativecommons.org/licenses/by/4.0/.

A. Meroño Peñuela et al. (Eds.): ESWC 2024, LNCS 14665, pp. 145–161, 2024.
https://doi.org/10.1007/978-3-031-60635-9_9

1 Introduction

With the advent of Large Language Models (LLMs), unconventional approaches for extracting rich Linked Data from collaborative platforms and knowledge hubs are gaining attention among researchers, including in the Semantic Web community [5]. First introduced in 2005, Semantic MediaWiki[2] (SMW) is an extension for the MediaWiki platform that enables semantic annotations of wiki pages [19], making it a collaborative Knowledge Graph management platform. It is also a precursor that majorly inspired Wikidata, and therefore Wikibase (WB), which uses select code from SMW for common tasks [28]. Yet, SMW has been available for a longer time and is indeed more broadly used than WB by various projects to manage their structured data. Indeed, following the approach described in Sect. 2, we could discover 1458 active Semantic MediaWiki instances, compared to a lower number of 327 Wikibase instances. The two platforms differ insofar that SMW stores its data as part of its textual page content and has a less complicated data model [28]: SMW's simple subject-predicate-object statement structure known as a *semantic facts* corresponds straightforwardly to RDF triples, where subjects are commonly single wiki pages, properties (predicates) are defined by special syntax in pages or via templates and forms enabled through additional extensions, and objects can be of different datatypes[3] (e.g. numbers, dates, pages etc.). Therefore, SMW serves as a flexible tool to collaboratively create and maintain domain-specific KGs, along with their own vocabularies.

RDF/SPARQL Access. The SMW platform allows to generate full RDF dumps[4] of its structured data. Additionally, although semantic facts are stored in a relational database by default, RDF triples can also be optionally exported/synced with a triplestore or a SPARQL endpoint via a special extension.[5] However, in practice, SMW instances rarely publish periodical dumps of their data, nor do they typically make their data available via SPARQL endpoints.[6] This significantly decreases the effective semantic interoperability and accessibility of KGs provided and maintained through SMW. So far, to the best of our knowledge – *an in-depth analysis of the RDF data made available through SMW instances on the Web is missing.*

Linkage and Ontology Re-use. The re-use of external URLs is in principle possible in semantic facts in SMW, as well as the import of external RDF vocabularies[7] and ontologies[8]; yet, semantic linkage to other KGs is not directly incen-

[2] https://www.semantic-mediawiki.org/.
[3] https://www.semantic-mediawiki.org/wiki/Help:Datamodel.
[4] https://www.semantic-mediawiki.org/wiki/Help:Maintenance_script_dumpRDF.php.
[5] In principle, this integration also supports adding additional RDF with SPARQL Updates, cf. https://www.semantic-mediawiki.org/wiki/Help:Using_SPARQL_and_RDF_stores, but herein we focus solely on the RDF data exportable from SMW pages directly.
[6] We note that also there is no "best practice" to detect whether an SMW instance hosts a SPARQL endpoint, as it is not exposed by the SMW API.
[7] https://www.semantic-mediawiki.org/wiki/Help:Import_vocabulary.
[8] https://github.com/TIBHannover/ontology2smw.

tivised within SMW, or respectively, *it is an open question in how far these features are being used*, i.e., in how far the KGs provided by SMW instances *(a) re-use existing ontologies* and *(b) create links to related RDF from other authorities*.

A commonly cited shortcoming of the Linked Open Data (LOD) infrastructure is the lack of a single aggregation point [8,13]. To this end, our approach, SMW Cloud, aims at solving this data accessibility issue by aggregating all publicly available SMWs into a single corpus, addressing the above-mentioned gaps by making the following concrete contributions:

- we systematically track (for now) 1458 Semantic MediaWiki instances, extract their page RDF data when technically feasible, and aggregate it to a Linked Data corpus, following a similar approach as the CommonCrawl[9] and WebDataCommons [21] projects,
- we make this corpus available as an HDT [12] dump, in an accessible, scalable and available, easy to (re-)use and cost-effective manner, following the LOD-a-LOT approach [13] and make calculated metrics for the corpus available in a SPARQL endpoint.
- we provide an extensive analysis of our corpus in terms of (a) ontology metrics, following the Neontometrics approach [24] and (b) LOD metrics, following the LODStats approach [7,8].

As such, we strive to obtain a comprehensive picture of the current state of Linked Data stored in SMWs, evaluating the quality and internal structure, thereby tracking the evolution of a significant, previously unexplored part of the LOD ecosystem. Note that we may hypothesise that the domain-specific, small and medium-sized KGs, represented by SMW instances, closely reflect Enterprise Knowledge Graphs (EKGs), for which there is no publicly available corpus. Analysing our corpus, in comparison with the "classic" KGs making up the LOD Cloud[10], but also in comparison with Wikidata, will therefore gain insights into the possible different parameters to be found in EKGs.

The remainder of this paper is structured as follows: In Sect. 2, we describe the architecture and our approach to collecting the SMW corpus. Section 3 introduces the methods used for further corpus analysis. Corpus statistics and results of our analyses, including a comparison with similar metrics applied to the "traditional" LOD Cloud and Wikidata, are summarized in Sect. 4, before we close with a discussion and outlook to future work in Sect. 5.

2 Methods for Collecting the Corpus

Our overall corpus collection approach is illustrated in Fig. 2. In first step, we discover Semantic MediaWiki instances in a three-fold manner:

[9] https://commoncrawl.org.
[10] http://lod-cloud.net.

- We filter the *BuiltWith* MediaWiki collection[11] for SMW instances.
- We query *WikiApiary*[12] – a"meta-wiki" collecting information about public MediaWiki instances – for instances that have the SMW extension installed.
- Lastly, we follow the approach utilized by the web platform discovery tool Crawley[13] and use Search Engine APIs (such as the BingAPI, but also by manual usage of other Search Engines) to discover SMW instances by specific text excerpts or HTML elements commonly found on SMW-powered websites: e.g., "powered by SMW" text snippets).

Following this approach, we collect an extensive list of 1458 SMW instances, for details on the numbers of instances per source, we refer to Fig. 1.

For each found SMW instance, we further retrieve basic statistics directly available via SMW, such as numbers of pages, numbers of users, creation and last modified dates, in order to assess how long the wikis have been operational and how active they are, and the pagelist. Next, we attempt to crawl RDF by querying the pagelist, using the MediaWiki API.[14] Frequent problems encountered in the process of collecting the corpus, such as non-standard behaviours of MediaWiki and SMW platforms, access restrictions, as well as a number of parsing errors, limit technical feasibility of collecting RDF from each instance. Table 1 gives an overview of such issues. Apart from the RDF representation we also crawl the versioning history to identify all changes per page for future analysis.[15] In the last step, we aggregate[16] the RDF by wiki instance to create our corpus. Apart from creating a single HDT file per crawled instance containing all RDF triples, we add metadata per wiki KGs using the VoID [2] and DataCube [10] vocabularies to include more complex statistics and metrics discussed in detail in Sect. 3 below, in the HDT headers.

As an additional item in this metadata aggregation, we collect and classify wikis per "topics", similar to the LOD Cloud topics[17]. Our approach to classify wikis into "LOD Cloud topics" works by, whenever available, fetching metainformation collected by WikiApiary (manually assigned topics) and BuiltWith (SEO-related keywords), plus the textual information from the respective wiki's main page, a random sample of up to 100 page titles and the name of the wiki. We then feed GPT-4 with this textual information[18] to assign it one of the LOD Cloud topics – this current naive approach serves mainly for illustration, cf. Figure 4 below.

[11] https://trends.builtwith.com/websitelist/MediaWiki.
[12] https://wikiapiary.com.
[13] http://purl.org/crawley.
[14] https://www.mediawiki.org/wiki/API:Query.
[15] Cf. Sect. 5, we plan to also extract and analyse the *evolution* of the RDF KGs per SMW instances as future work.
[16] Given the absence of Blank Nodes within instances, skolemization is unnecessary.
[17] https://lod-cloud.net/.
[18] Roughly, our GPT prompt is: *"Given the text "[WIKINAME+MAINPAGETEXT+PAGETITLES+METAINFO]", tell me the best fitting topic among [LODCLOUD-Topiclist]".*

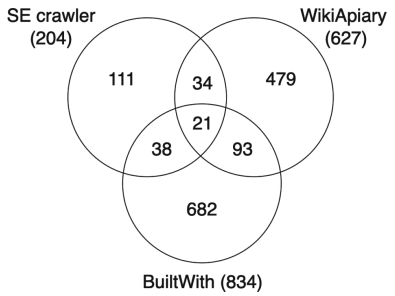

Fig. 1. Venn Diagram of Collected SMWs (1458 in Total) and Their Sources

Table 1. Breakdown of Collection and Processing Losses

SMWs	Description
1458	all active SMW instances
-108	malformed API response
-107	API endpoint unavailable
-55	server-terminated connection
-31	non-standard encoding scheme
1157	instances for which the full pagelist and page RDF could be collected
-51	XML wrongly declared
-36	malformed XML/mismatched tags
-5	non-compliant IRIs
-36	other processing errors
1029	SMW instances for which HDTs could be aggregated

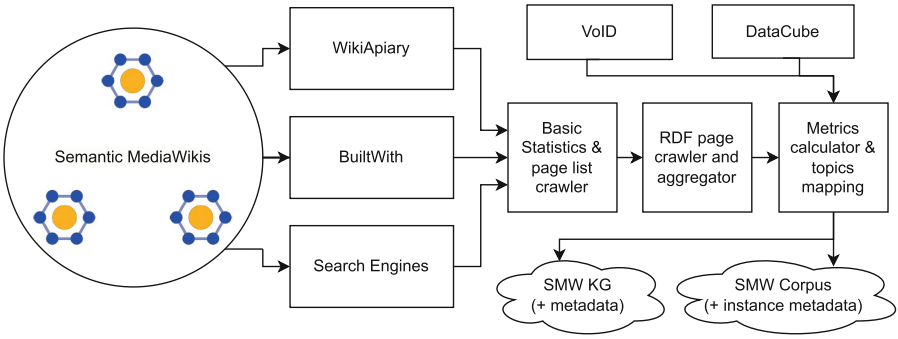

Fig. 2. Architecture of the SMW Crawler and Metrics Processor

2.1 Corpus Provision

Our SMW corpus is registered on Zenodo[19] and is available under a permanent URL[20] and directly on the open data repository of the institute[21] we provide both separate RDF HDT [12] dumps per SMW instance, as well as in a single HDT file for the whole corpus. A SPARQL endpoint, serving the VoID & DataCube metadata in a queryable form, is available at: https://smwcloud-sparql.cluster.ai.wu.ac.at/. The resource including all calculated metrics are provided under the Creative Commons Attribution 4.0 International License.[22]. The canonical citation for the SMW Cloud is:

Dobriy, D., Beno, M., & Polleres, A. (2023). SMW Cloud: A Corpus of Domain-Specific Knowledge Graphs from Semantic MediaWikis. Retrieved from http://purl.org/SMWCloud.

[19] 10.5281/zenodo.7920175.
[20] http://purl.org/SMWCloud.
[21] https://semantic-data.cluster.ai.wu.ac.at/smwcloud/.
[22] https://creativecommons.org/licenses/by/4.0/.

Comprehensive documentation for the resource is available at: https://github.com/semantisch/smwcloud. For illustration, we provide the VoID metadata and selected DataCube entries available through our endpoint (and as part of the respective HDT header) for "Wien Geschichte Wiki", an SMW instance providing historical information about Vienna [18] in Turtle syntax in Fig. 3.

```
@prefix smwcloud: <http://purl.org/smwcloud/> .

smwcloud:7f5cb281-76f8-4d16-aee1-a4ad7c660eec void:inDataset <http://purl.org/smwcloud/> .

smwcloud:7f5cb281-76f8-4d16-aee1-a4ad7c660eec a void:Dataset ;
  foaf:homepage <https://www.geschichtewiki.wien.gv.at/Wien_Geschichte_Wiki>;
  foaf:page <https://www.geschichtewiki.wien.gv.at/api.php>;

  dcterms:title "Wien Geschichte Wiki";
  dcterms:source <https://www.geschichtewiki.wien.gv.at>;
  dcterms:modified "2023-05-05"^^xsd:date;
  dcterms:license <https://www.geschichtewiki.wien.gv.at/Impressum>;

  # The dcterms:description was generated by summarising the wikis main page text using GPT:
  dcterms:description
    "\"Wien Geschichte Wiki\" is the historical knowledge platform of the city of Vienna, based
   on the "Historical Lexicon of Vienna" by Felix Czeike, which brings together expertise from
   city administration and the public and currently has over 48,000 contributions,
      279,000 addresses, and 15,000 images.";

  # The dcterms:subject topic was assigned one of the LODCloud categories using using GPT:
  dcterms:subject "Government";
  void:feature <http://purl.org/HDT/hdt#HDTv1>;
  void:dataDump <ACTUAL_URL_FOR_SINGLE_WIKI_HDT_DUMP> ;
  void:uriSpace "https://www.geschichtewiki.wien.gv.at/";

  # Void observations:
  void:triples 5236436;
  void:entities 1038817;
  void:classes 245;
  void:properties 256;
  void:distinctSubjects 436147;
  void:distinctObjects 401053;
  void:documents 328141;

# A DataCube observation:
smwcloud:7f5cb281-76f8-4d16-aee1-a4ad7c660eec/sameIndividualsAxioms/2023-05-05 a qb:Observation ;
    qb:dataSet smwcloud:7f5cb281-76f8-4d16-aee1-a4ad7c660eec ;
    smwcloud:referenceDate "2023-05-05"^^xsd:date ;
    smwcloud:sameIndividualsAxioms 7491 .
```

Fig. 3. Metadata for "Wien Geschichte Wiki"

3 Methods for Analysing the Corpus

Due to their topical diversity, data and schemas differ considerably among SMW instances, but also we hypothesize that the RDF data from SMW instances has fundamentally different characteristics than other Linked Data corpora, such as the LOD Cloud datasets, or Wikidata. In order to verify these assumptions, a comprehensive characterization of our corpus requires two things: (1) a fundamental understanding of each dataset, and (2) an overview of all available data

Table 2. Related Work and Implemented Metrics

Use case \ Paper	Fernandez et al.[13]	Ermilov et al. [8][9][7]	Nogalez et al. [22]	Reiz et al. [24]	Abedjan et al. [1]	Haller et al. [17]	Rietveld et al. [26]
Basic graph metrics	X	X	X	X	X	X	X
Basic ont. metrics			X	X		X	X
Quality analysis		X		X	X		X
Vocabulary re-use		X	X			X	
Dataset interlinkage		X				X	
Languages analysis		X	X				

Paper	Eval. dimensions
Ermilov et al. [7]	Basic graph metrics
Reiz et al. [24]	Basic ontology metrics[a]
Haller et al. [17]	Instance, ontology links
Yao et al. [31]	Cohesion
Yang et al. [30]	Complexity
Fernandez et al. [14]	Coverage, structure
Duque-Ramos et al. [6][25]	SQuaRE[b] -based quality
Gangemi et al. [15]	Structure, functionality, usability
Tartir et al. [27]	Populated ontology (instances, schema)
Orme et al. [23]	Quality, completeness, and stability

[a] Basic ontology metrics can be used a building block for calculation of quality frameworks.

[b] `https://www.iso.org/standard/64764.html`

[3]. Therefore, we perform our analyses both on the single wiki datasets as well as on the corpus as a whole in terms of commonly used metrics.

In order to establish a foundation for comparative analysis, we review related LOD analysis studies in Table 2. This allows us to discern and categorize the prevailing analytical themes. The most common analyses within these studies are performed on *Basic graph metrics*, *Basic ontology metrics*, and *Quality*. In addition to the aforementioned themes, other forthcoming analyses encompass *Vocabulary re-use*, *Dataset interlinkage* and *Language usage*.

General graph metrics give a comprehensive and comparable characterization of the corpus [7]. Additionally, we calculate *basic ontology metrics* to assess the used schemata/ontologies of the individual SMW instances in a comparable manner, and evaluate the corpus on the basis of established *quality analysis* frameworks to gain a better understanding of quality characteristics. Explicitly including ontology metrics and common ontology quality frameworks in our analyses through the implementation of the Neontometrics calculation engine [24], we address the lack of metric validation for real-world data [24], especially in *small KGs*: we perceive SMW instances as a publicly available pendant/proxy for enterprise KGs with their own characteristics.

Table 2 briefly summarizes the various metrics and quality frameworks we have calculated for the SMW corpus: as for calculating the basic graph and ontology metrics, we use the Neontometrics OPI (Ontology Programming Interface).

Afterwards, we apply these metrics to calculate common quality frameworks: Cohesion Metrics [31], Complexity Metrics [30], Fernandez et al. [14], OQuaRE [6,25], Gangemi et al. [15], OntoQA [27] and Orme et al. [23]. Please refer Table 2 for a summary of their respective dimensions.

4 Results and Corpus Statistics

In this section, we provide a brief overview of the insights into the corpus based on the collected statistics. We excluded the datasets causing processing errors as described in Table 1, resulting in 1029 datasets which collectively form the corpus and the basis for analyses.

4.1 Basic RDF Metrics

Table 3 gives a distilled overview of of the corpus dimensions. Notably, the absolute numbers we report for SMW Cloud fall short of the statistics for the number of statement reported by the instances themselves: totalling 1,012,521,773 statements and 206,997 unique properties, calculated by aggregating statistics reported by individual SMWs. At the same time, the SMW Cloud dimensions are comparable to that of LODStats [8], which totals 192,230,648 triples. Though Wikidata has grown considerable in the last 5 years, from 3b triples in 2018 to more than 17b in 2022, exceeding SMW Cloud in size considerably [11]. Nevertheless, despite its comparatively smaller size in terms of total number of statements, the SMW Cloud exhibits a significantly broader range of unique properties, exceeding both Wikidata and LODStats, as well as suggesting limited vocabulary re-use in SMW Cloud.

For a better overview of the characteristics of individual datasets comprising the corpus, we refer to Table 4. There, we compare their parameters to the individual LOD Cloud datasets [8]. Though LOD Cloud datasets are significantly larger on average (2,180,651 triples to 186,813 for SMWs), the majority of LOD datasets are smaller (median 2,486 vs 12,596 for SMWs) implying more uniform sizes of the SMWs. Another key characteristic of SMWs is the higher number of properties and classes per datasets as well as the higher number of properties per entity, suggesting a more granular and detailed data modeling approach in SMWs and a user-centric, bottom-up nature of ontology creation. The lack of class and property depth suggests a flat ontology structure in the SMW instances. We attribute this to the a) user communities with rich domain knowledge and limited expertise in ontology management, b) ad-hoc nature of ontology creation and c) decentralized ontology development in SMWs. Other characteristics regarding high numbers of labeled subjects and large median typed and untyped string lengths for Literals further differentiate user-centric SMWs from LOD datasets.

Table 3. SMW Cloud summary statistics

Dataset	#Triples	#Subjects	#Predicates	#Objects	#Literals
LODStats [8]	192,230,648	Not reported	49,916	Not reported	90,261,655
SMW Cloud	236,505,705	24,010,566	52,670	66,052,823	160,108,216
Wikidata 2021[a]	17,662,800,665	1,625,057,179	38,867	Not reported	Not reported
LOD-a-lot [13]	28,362,198,927	3,214,347,198	1,168,932	3,178,409,386	1,302,285,394

[a] http://gaia.infor.uva.es/hdt/wikidata/wikidata20210305.hdt.gz

Table 4. SMW Cloud and LOD Cloud comparison

	SMW Cloud				LOD Cloud [8]			
Metric	Mean	Min	Max	Median	Mean	Min	Max	Median
Triples p. dataset	186,813	0	31,582,870	12,595.6	2,180,651	2	247,620,294	2,486
Entities	14,828	0	5,036,913	1,281.0	390,325.95	0	63,494,920	106.5
Literals	54,076	0	18,514,040	3,304.0	790,000.57	0	88,315,872	127.0
Blanks					484,540.68	0	202,745,495	0.0
Blanks (subjects)	0.04	0	14	0	399,680.75	0	166,901,812	0.0
Blanks (objects)	0.02	0	6	0	143,005.6	0	50,803,539	0.0
Subclasses	0.14	0	46	0	14.07	0	2,000	0.0
Typed subjects	13,068.83	0	44,374.58	1,047.0	109,790.35	0	25,848,850	22.0
Labeled subjects	2,241.23	0	760,619	267.0	28,652.13	0	11,588,129	0.0
Properties p. entity	6.22	0.88	12.32	4.10	2.86	0	27.27	2.54
String length (typed)	9.32	0	476.05	9.51	9.5	0	1,854.0	0.0
String length (untyped)	72.43	0	369.77	73.12	38.24	0	2,688.0	20.0
Class hierarchy depth	0.009	0	2	0.0	1.63	0	6	0.0
Property hierarchy depth	0	0	0	0.0	1.04	0	5	0.0
Classes	356.52	0	113,270	27.0	20.09	1	1,328	5.0
Properties	155.01	0	45,209	38.0	30.36	1	885	20.0

4.2 Topical Analysis

Semantic MediaWikis capture a variety of highly-specialized domain knowledge, as visualized in Fig. 4 reusing the LOD Cloud categories:[23] 16% wikis specialize on publishing/annotating *media*, 10% are *life science* wikis, 6% *geography*-oriented, 5% *government*-related, 4.3% *language*-related. Still, about half of all wikis (48%) could not be classified under one of the categories (i.e. summarized under *user generated other*), hinting at the topic diversity and high degree of specialization of SMWs.

Manual coding of a randomly selected sample of 100 wikis (as partially represented in Table 5) has shown that the chosen topics represent the wiki well, overlap and are similar in meaning (87% similarity) with WikiApiary tags and BuildWith verticals. Therefore, as a result, we characterize each wiki by 3 independently created tag/domain collections which promotes discoverability and an automatically generated description.

We further analyse the distribution of specialized domains in the corpus with freely annotated tags, see Fig. 5. The big components of the hitherto unclassified wikis are therefore: *gaming* with 134 instances in total, *technology* (115), *education* and *community*. The 101 unclassified wikis indeed can hardly be classified

[23] https://lod-cloud.net.

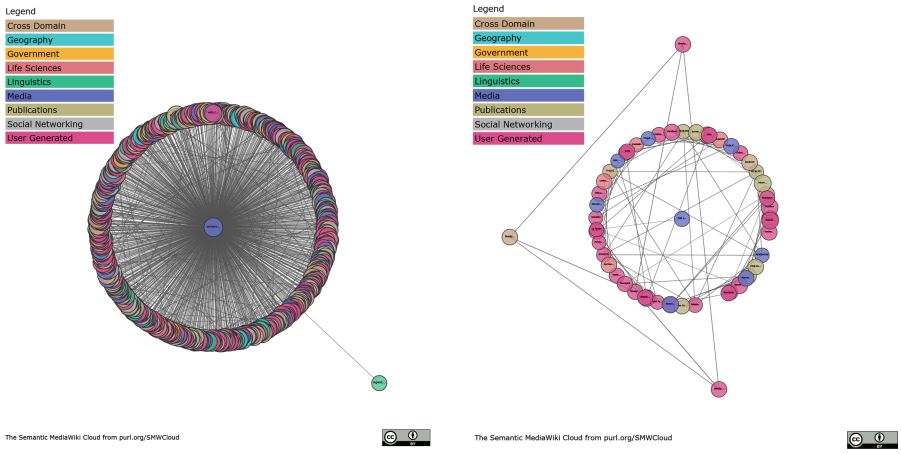

Fig. 4. Connected SMW Cloud instances with and without semantic-mediawiki.org

Table 5. Examples of SMW domain annotations

SMW	LOD Cloud classification[a]	Free classification	Description	Topic (WikiApiary)	Topic (BuiltWith)
geschichtewiki .wien.gv.at	history (government, geography)	Vienna, Austria	*The Wien Geschichte Wiki is an encyclopedia of historical, geographic, and cultural information related to the city of Vienna and its surrounding regions.*	city wiki, history, vienna	Art And Entertain-ment
bacid.eu	government	public administra-tion	*BACID Wiki contains information about decentralized governance, capacity building, and public administration initiatives in the Danube region.*	–	Business And Industrial
korrekt.org	publications (media, user generated)	knowledge-based systems	*Korrekt.org is a wiki focused on the research and publications of Professor Markus Krötzsch, covering topics such as description logic, semantic wikis, and knowledge-based systems.*	homepage, semantic mediawiki	Science
www. gardenology. org	life sciences	plants, gardening, encyclopedia	*A comprehensive wiki encyclopedia covering plants and gardening, featuring detailed entries and photographs.*	–	–

[a] Classifications enclosed within parentheses are also produced by the model, serving as alternatives.

because of the lack of investigated content (see Sect. 2). While we see that common LOD Cloud topics are well suited for classifying about half of SMW Cloud, other significant topics emerge: SMW Cloud has a rich number of technology, education and community wikis not prominently featured in LOD Cloud.

4.3 Ontological Analysis

A number of metrics has been been proposed for the analysis of KGs and ontologies. Since a KG comprises both A-Box as well as T-Box statements, we consider

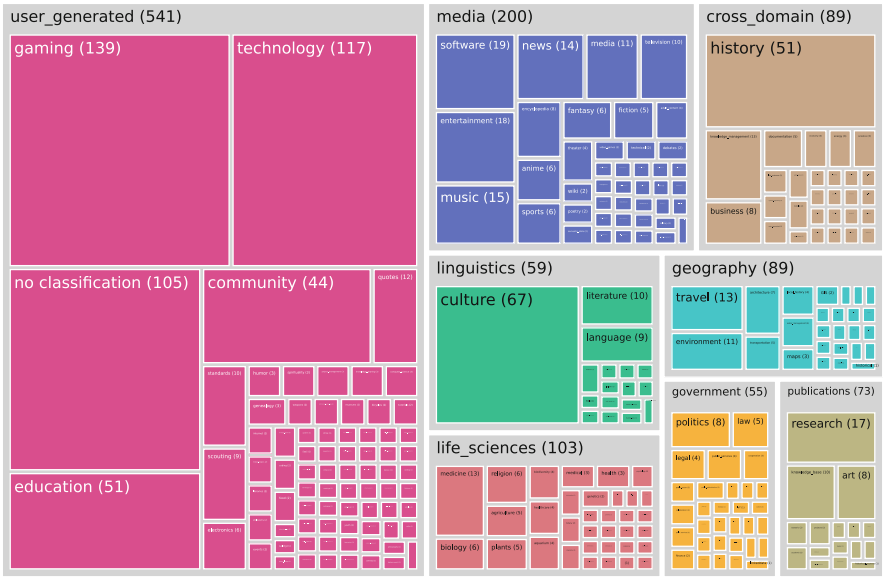

Fig. 5. Prevalent Topics Across SMW Cloud Instances

both metrics for characterizing KGs and ontologies as suitable for analysing KG with notable limitations applying to ontology metrics discussed separately.

Reiz et al. [24] proposes an ontology describing common metrics used to characterize Knowledge Graphs and also introduces an open source tool automating the calculation of a broad variety of metric and quality analyses calculations[24]. The website of the tool also includes a Metric Explorer with metric overviews and descriptions. For SMW Cloud, Table 6 summarizes the basic ontology metrics for SMWs. We have also calculated all quality analyses based on these basic ontology metrics suggested by the Neontometrics engine (see Table 8).

Ontologically, SMW do contain a large number of class and property assertion axioms, same individual assertions as well as individual, class and property annotations. Notably, although SMW technically allows the use of RDFS and OWL concepts, only a total of 5 SMWs[25] implement class hierarchies (via *rdfs:subClassOf*) and no SMW instance implements property hierarchies (via *rdfs:subPropertyOf*) in practice, while 5 SMW instances use *owl:equivalentProperty* definitions, some of which seem redundant.[26] A closer analysis of the RDF(S) and OWL vocabulary used in 1029 crawled SMW instances in terms of Description Logics expressivity (testable again through the Neontometrics tool) is illustrated in Table 7: here, the concrete RDF(S)+OWL

[24] http://neontometrics.com/.

[25] Specifically: *wiki.spell-plattform.de*, *wiki.fablab.is*, *wiki.attraktor.org*, *spiele.j-crew.de* and *dotawiki.de*.

[26] E.g., a triple *dcterms:isPartOf owl:equivalentProperty dcterms:isPartOf* in *pool.publicdomainproject.org*.

Table 6. Basic ontology metrics

	SMW Cloud			
Metric	Mean	Min	Max	Median
Class assertion axioms	12,782.75	0	540,701	530.0
Object property assertion axioms	21,735.10	0	1,169,337	635.5
Data property assertion axioms	55,397.72	0	3,144,162	1,793.0
Same individuals axioms	346.05	0	18,512	10.0
General annotation axioms	1,759.55	0	49,391	172.0
Annotation assertion axioms	5,065.86	0	374,185	508.0
Data property annotations	6.92	0	178	6.0
Class annotations	3.25	0	374	2.0
Object property annotations	15.58	0	374	5.0
Individual annotations	5,004.62	0	374,116	439.0
Axioms	95,654.59	0	5,236,436	3,951.5
Logical axioms	90,308.68	0	4,861,744	3,297.5
Classes(See footnote 27)	195.57	0	9,082	22.5
Classes with individuals	188.06	0	9,074	14.5
Object properties	30.83	0	1,500	21.0
Data properties	48.80	0	1,249	32.0
Individuals	12153.97	0	728,482	650.5

constructs being used in each DL expressivity class are analysed, which reveals that only a small fraction of RDFS' and OWL's statements are being used in SMW instances. Indeed, for instance, neither *rdfs:domain* and *rdfs:range* definitions, nor *owl:equivalentClass*, with one exception, hardly any multi-triple OWL axioms are being used in SMW instances: under-use of complex OWL constructs, not even mentioning OWL2, can therefore be also observed on the SMW ecosystem, in a similar and even more pronounced form than already observed more than 10 years ago for the LOD Cloud [16]. Also, due to the sparse use of subclassing and sub- or equivalent properties, further analysis does not further focus on ontology metrics and quality framework metrics, which emphasize schema depth/inheritance richness [27], see also Footnote[27] in Table 6.

In our metric processing, we calculate all Quality Frameworks indicated in Table 2. Although it is not feasible to discuss the evaluation of frameworks in full, we demonstrate the OQueRe framework in Table 8, examplified by metrics receiving the best score and the worst score suggesting a more in-depth analysis as subject of future work capitalizing on the resource established in this work.

[27] Due to the observed lack of hierarchical structure, the *number of classes* is equivalent with the number of *root classes*, *paths to leaf classes*, *absolute leaf cardinality* and *absolute depth*, so we do not provide these metrics separately..

Table 7. Use of the RDF(S)+OWL and DL expressivity of SMW instances

Number of SMW instances:	$\mathcal{AL}$ (99)	$\mathcal{AL}$(D) (165)	$\mathcal{ALO}$(D) (708)	$\mathcal{ALEO}$(D) – (50)	$\mathcal{ALHO}$(D) – – (3)	(1029)
rdf:type	10206194	217002	18593506	1807	69973	29088482
rdfs:isDefinedBy	1829566	35365	2987260	665	3336	4856192
rdfs:label	1829657	35365	2988428	666	3336	4857452
rdfs:seeAlso	–	–	2	–	–	2
rdfs:subClassOf	–	51	16	–	–	67
rdfs:comment	–	–	255	–	–	255
owl:imports	1430995	26574	2065990	443	3065	3527067
owl:Ontology	1430995	26574	2065990	443	3065	3527067
owl:Class	136758	3934	193832	39	7175	341738
owl:DatatypeProperty	8043	3027	35345	17	216	46648
owl:ObjectProperty	8377	1232	19169	24	82	28884
owl:sameAs	450445	–	358972	76	164	809657
owl:differentFrom	–	–	16	–	–	16
owl:equivalentProperty	–	–	–	–	5	5
owl:intersectionOf	–	–	–	3	–	3
owl:Restriction	–	–	–	3	–	3
owl:onProperty	–	–	–	3	–	3
owl:hasValue	–	–	–	3	–	3
rdf:first	–	–	–	6	–	6
rdf:rest	–	–	–	6	–	6
rdf:nil	–	–	–	3	–	3

Table 8. OQueRE Metrics and Scores

Metric	Mean	Min	Max	Median	Score
Mean number of annotations per class	37.96	0.65	710.16	14.29	1 (excellent)
Mean number of attributes per class	1.05	0.01	2.67	0.77	1 (excellent)
Weighted Method Count: Mean number of properties per class [6]	345.37	0.70	55439.4	35.19	5 (unsatisfactory)

5 Conclusion and Future Work

This paper presented and characterized the SMW Cloud corpus, an extensive collection of RDF data collected from Semantic MediaWiki instances. We demonstrate, with a focused crawling pipeline, that we can identify and collect data from over 1000 SMW instances, some of which have been in existence and active use for over 15 years, which demonstrates the considerable interest of relevant communities and SW technology users of this largely unobserved part of the Semantic Web.

To promote interoperability and ease of use, the SMW corpus is made available as HDT and the corpus' metadata is queryable via a SPARQL endpoint, in line with the FAIR data principles [29]. We plan to update the SMW Cloud regularly and extend it by discovering and crawling RDF from new SMW instances as they appear.

Following the same approach, we recommend that Semantic MediaWiki developers 1) enable RDF dump generation by default rather than requiring administrators to manually make use of a maintenance script to create a dump (or us to crawl the RDF data per page) 2) when a SPARQL-enpoint is available, make it discoverable through SMW API and 3) consider adopting HDT as a compact format for publishing regular dumps as we have demonstrated that it is highly compact format for SMW data (achieving a Data Compression Ratio of 17,5 for SMW Cloud compared to NTriples, more efficiently that for other benchmarks evaluated [20]).

In terms of evaluation and benchmarking as a field of interest of the Semantic Web Community [4], SMW Cloud provides a novel and distinct dataset with unique characteristics that introduces variety into the field of LOD sources investigated so far; we have demonstrated these unique characteristics in terms of a variety of common basic graph and ontology metrics, that illustrate significant differences of RDF usage within SMW instances and the rest of the LOD Cloud. We expect the SMW Corpus to enable previously unexplored approaches in LOD and EKG research.

5.1 Limitations

Statistics calculated by us can not be directly integrated back into individual SMWs, creating a discoverability problem. To this end, it is planned to introduce an SMW extension that will 1) schedule regular RDF dump generation, 2) notify the proposed architecture of the Wiki, and 3) fetch calculated statistics from the SMW Corpus and integrate them into the Wiki.

5.2 Future Work

As future work we think that the SMW corpus can also provide a basis for longitudinal analysis, link analysis [17], etc. This will enable a better understanding of dynamics and evolution of vocabularies. It is our goal to create profiling tools and resources *enabling users to create an assessment of the data at hand* [3]. Finally, as noted in Sect. 1, Wikibases are the second most widely used semantic wiki platform to date. Therefore, crawling efforts and analyses of Wikibase instances following a similar methodology are a prioritised part of future work.

5.3 Sustainability Plan

To account for the updates in the SMW Cloud, which particularly concerns the introduction of new SMW instances, but also general changes to already

accounted SMWs, we plan to release new versions of SMW Cloud at least annualy. The sustainability plan includes further extension of the automatic metric calculation in the frame of future work and other continuing research as well as tight collaboration with other researchers and users with regards to its further development and maintenance, as well as use case documentation.

We are commited to sustainably host and maintain the corpus thought our institute that already hosts various widely adopted Semantic Web resources for several years now and promote the sustainability strategy within ongoing community activities such as the "Distributed Knowledge Graphs" COST Action[28], which as one of its activities aims at aligning and sustaining community services and tools.

The Resource is made Accessible in Following Ways:

1. The aggregated SMW Cloud dataset is made available via the institutional repository.[29]
2. The SMW Cloud corpus containing individual SMW datasets is also made available via the institutional repository.[30]
3. The calculated metrics for the corpus are available via a public SPARQL endpoint.[31]

Acknowledgements. The authors would like to express their profound gratitude and give special acknowledgment to Alexandra Hager for her invaluable, expert and time-critical support, Thomas Seyffertitz for expert advice in the area of research data management and Gary Brewer, BuiltWith, for providing an extensive MediaWiki collection. This work is part of a project funded by the *WU Anniversary Fund of the City of Vienna*.

References

1. Abedjan, Z., Gruetze, T., Jentzsch, A., Naumann, F.: Profiling and mining RDF data with ProLOD++. In: 2014 IEEE 30th International Conference on Data Engineering, pp. 1198–1201. IEEE, Chicago, IL, USA, March 2014. https://doi.org/10.1109/ICDE.2014.6816740, http://ieeexplore.ieee.org/document/6816740/
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: LDOW (2009)
3. Bohm, C., et al.: Profiling linked open data with ProLOD. In: 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010), pp. 175–178. IEEE, Long Beach, CA, March 2010. https://doi.org/10.1109/ICDEW.2010.5452762, http://ieeexplore.ieee.org/document/5452762/
4. Boncz, P., Fundulaki, I., Gubichev, A., Larriba-Pey, J., Neumann, T.: The linked data benchmark council project. Datenbank-Spektrum **13**(2), 121–129 (2013)

[28] https://cost-dkg.eu/.
[29] https://semantic-data.cluster.ai.wu.ac.at/smwcloud/.
[30] https://semantic-data.cluster.ai.wu.ac.at/smwcloud/corpus/.
[31] https://smwcloud-sparql.cluster.ai.wu.ac.at/.

5. Burns, A., et al.: A suite of generative tasks for multi-level multimodal webpage understanding. arXiv preprint arXiv:2305.03668 (2023)
6. Duque-Ramos, A., Fernández-Breis, J.T., Stevens, R., Aussenac-Gilles, N.: OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. J. Res. Pract. Inf. Technol. **43**(2) (2011)
7. Ermilov, I., Demter, J., Martin, M., Lehmann, J., Auer, S.: Lodstats–large scale dataset analytics for linked open data. Under review in ISWC (2013)
8. Ermilov, I., Lehmann, J., Martin, M., Auer, S.: LODStats: the data web census dataset. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9982, pp. 38–46. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46547-0_5
9. Ermilov, I., Martin, M., Lehmann, J., Auer, S.: Linked open data statistics: collection and exploitation. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2013. CCIS, vol. 394, pp. 242–249. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41360-5_19
10. Escobar, P., Candela, G., Trujillo, J., Marco-Such, M., Peral, J.: Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. Comput. Stand. Interfaces **68**, 103378 (2020)
11. Fahl, W., Holzheim, T., Westerinen, A., Lange, C., Decker, S.: Getting and hosting your own copy of wikidata. In: Wikidata Workshop @ ISWC 2022 (2022). https://ceur-ws.org/Vol-3262/paper9.pdf
12. Fernández, J.D., Martınez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). J. Web Semant. **19**(2) (2013). https://doi.org/10.1016/j.websem.2013.01.002, http://www.polleres.net/publications/fern-etal-2013-HDT-JWS.pdf
13. Fernández, J.D., Beek, W., Martínez-Prieto, M.A., Arias, M.: LOD-a-lot. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 75–83. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_7
14. Fernández, M., Overbeeke, C., Sabou, M., Motta, E.: What makes a good ontology? A case-study in fine-grained knowledge reuse. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 61–75. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10871-6_5
15. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 140–154. Springer, Heidelberg (2006). https://doi.org/10.1007/11762256_13
16. Glimm, B., Hogan, A., Krötzsch, M., Polleres, A.: OWL: yet to arrive on the web of data? In: WWW2012 Workshop on Linked Data on the Web (LDOW2012). Lyon, France, April 2012. http://www.polleres.net/publications/glim-etal-2012LDOW.pdf
17. Haller, A., Fernández, J.D., Kamdar, M.R., Polleres, A.: What are links in linked open data? A characterization and evaluation of links between knowledge graphs on the web. J. Data Inf. Qual. **12**(2), 1–34 (2020)
18. Krabina, B.: Building a knowledge graph for the history of Vienna with semantic MediaWiki. J. Web Semant. **76**, 100771 (2023)
19. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: Cruz, I., et al. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 935–942. Springer, Heidelberg (2006). https://doi.org/10.1007/11926078_68
20. Martínez-Prieto, M.A., Arias Gallego, M., Fernández, J.D.: Exchange and consumption of huge RDF data. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 437–452. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30284-8_36

21. Meusel, R., Petrovski, P., Bizer, C.: The webdatacommons microdata, RDFa and microformat dataset series. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 277–292. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_18
22. Nogales, A., Angel Sicilia-Urban, M., García-Barriocanal, E.: Measuring vocabulary use in the linked data cloud. Online Inf. Rev. **41**(2), 252–271 (2017)
23. Orme, A.M., Yao, H., Etzkorn, L.H.: Indicating ontology data quality, stability, and completeness throughout ontology evolution. J. Softw. Maint. Evol. Res. Pract. **19**(1), 49–75 (2007)
24. Reiz, A., Sandkuhl, K.: Neontometrics–a public endpoint for calculating ontology metrics. In: Proceedings of Poster and Demo Track and Workshop Track of the 18th International Conference on Semantic Systems co-located with 18th International Conference on Semantic Systems (SEMANTiCS 2022), vol. 13, pp. 22–15. CEUR-WS, Vienna (2022)
25. Reiz, A., Sandkuhl, K.: Harmonizing the OQuaRE quality framework:. In: Proceedings of the 24th International Conference on Enterprise Information Systems, pp. 148–158. SCITEPRESS - Science and Technology Publications, Online Streaming (2022). https://doi.org/10.5220/0011077200003179, https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011077200003179
26. Rietveld, L., Beek, W., Hoekstra, R., Schlobach, S.: Meta-data for a lot of LOD. Semant. Web **8**(6), 1067–1080 (2017)
27. Tartir, S., Arpinar, I.B.: Ontology evaluation and ranking using ontoqa. In: International Conference on Semantic Computing (ICSC 2007), pp. 185–192. IEEE (2007)
28. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85. ACM, New York, NY, USA (2014)
29. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18, https://www.nature.com/articles/sdata201618, number: 1 Publisher: Nature Publishing Group
30. Yang, Z., Zhang, D., Ye, C.: Ontology analysis on complexity and evolution based on conceptual model. In: Leser, U., Naumann, F., Eckman, B. (eds.) DILS 2006. LNCS, vol. 4075, pp. 216–223. Springer, Heidelberg (2006). https://doi.org/10.1007/11799511_19
31. Yao, H., Orme, A.M., Etzkorn, L.: Cohesion metrics for ontology design and application. J. Comput. Sci. **1**(1), 107–113 (2005)