WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Bachelor Thesis

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

Daniil Dobriy

Date of Birth: 06.04.2000
Student ID: 11776408

Subject Area: Information Business

Studienkennzahl: UJ033561

Supervisor: Axel Polleres

Date of Submission: 16. August 2023

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK UND OPERATIONS MANAGEMENT DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS MANAGEMENT

Contents

List of Figures

List of Tables

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 work seeks to address the gap by systematically collecting structured data from an extensive corpus of Semantic-MediaWiki-powered portals, in the process establishing a novel tool for automated search engine-mediated discovery of such portals, and providing an in-depth analysis of the ontological diversity (and re-use) amongst these wikis using a variety of ontological metrics. It aims to demonstrate that Semantic Wikis are a valuable and extensive part of Linked Open Data and, in fact, may be considered an active "sub-cloud" within the Linked Open Data ecosystem, which can provide valuable insights into the evolution of small and medium-sized domain-specific Knowledge Graphs.

1 Introduction

The Semantic Web is a visionary field that aims to represent and make available highly interconnected human knowledge both in a machine-readable and human-accessible way [7]. The Semantic Web defines standards used to represent knowledge and provides infrastructure to make knowledge accessible. As such, and considering the foundational value of knowledge to modern society, it is an important field of research that amplifies new knowledge gained in other scientific disciplines.

With the advent of Large Language Models (LLMs), unconventional approaches for extracting structured data from collaborative platforms and knowledge hubs are gaining attention among researchers, including in the Semantic Web community [11]. First introduced in 2005, Semantic MediaWiki[1] (SMW) is an extension that extends the MediaWiki platform, the software underlying most Wikis available on the Web, to enable semantic annotations of Wiki pages [34], making them suitable for collaborative management of structured knowledge. It is also a precursor that majorly inspired Wikidata, the central storage for the structured data of Wikimedia projects like Wikipedia. Therefore, the underlying platform, Wikibase, uses select code from SMW for common tasks [51]. Yet, SMW has been available longer and is more broadly used than Wikibase by various projects to manage their structured data: following the approach described in Section 3, 1458 active SMW instances could be discovered, compared to a lower number of 327 Wikibase instances. Furthermore, the two platforms differ insofar that SMW stores its data as part of its textual page content and has a less complicated data model [51]: SMW's simple subject-predicate-object statement structure known as a *semantic fact* corresponds straightforwardly to the data model underlying the Semantic Web - the Resource Description Framework[2] (RDF), whereby in SMW, subjects are commonly single Wiki pages, properties (predicates) are defined through the use of 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 Knowledge Graphs[4](i.e. collections of facts) along with their

[1] https://www.semantic-mediawiki.org/

[2] The data model is explained in more detail in Section 2.2.

[3] https://www.semantic-mediawiki.org/wiki/Help:Datamodel

[4] This concept is treated in more detail in Section 2.5.

own vocabularies (i.e., the schemas underlying the data).

The SMW platform allows to generate full RDF dumps[5] of its structured data. Additionally, although semantic facts are stored in a relational database by default, RDF statements can also be optionally (via an additional MediaWiki extension) exported/synced with a triplestore, a database specifically designed for storing RDF data, which in turn would provide an endpoint for querying the data.[6] However, in practice, SMW instances rarely publish periodical dumps of their data, nor do they typically make their data available via endpoints that support SPARQL - a query language for querying RDF data.[7] This significantly decreases the effective semantic interoperability and accessibility of KGs provided and maintained through SMW. So far, to the best of the author's knowledge – *an in-depth analysis of the RDF data made available through SMW instances on the Web is missing.*

The re-use of external URLs is possible in semantic facts in SMW, as well as the import of external RDF vocabularies[8] and ontologies[9]; yet, semantic linkage to other KGs is not directly incentivised 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 (i.e., structured data interlinked on the Web[10]) is the lack of a single aggregation point [16, 21]. To this end, the approach proposed herein, SMW Cloud, aims at solving this data accessibility issue by aggregating all publicly available SMWs into a single corpus, addressing the gaps mentioned above by making the following concrete contributions:

- systematically and periodically tracking (for now) 1458 Semantic MediaWiki instances, extracting their page RDF data when technically feasible, and aggregating it to a Linked Data corpus, following a similar approach as the CommonCrawl[11] and WebDataCommons [37] projects,

[5] https://www.semantic-mediawiki.org/wiki/Help:Maintenance_script_dumpRDF.php

[6] In principle, this integration also supports adding additional RDF, cf. https://www.semantic-mediawiki.org/wiki/Help:Using_SPARQL_and_RDF_stores, but herein the focus lies solely on the RDF data exportable from SMW pages directly.

[7] One should note that also there is no "best practice" to detect whether an SMW instance hosts a SPARQL endpoint for querying, as it is not exposed by the SMW API.

[8] https://www.semantic-mediawiki.org/wiki/Help:Import_vocabulary

[9] https://github.com/TIBHannover/ontology2smw

[10] The concept is treated in more detail in Section 2.4.

[11] https://commoncrawl.org

- making this corpus available as an HDT [20] dump (a compressed binary serialization format for RDF), in an accessible, scalable and available, easy to (re-)use and cost-effective manner, following the LOD-a-LOT approach [21] and making calculated metrics for the corpus available in a SPARQL endpoint,
- providing an extensive analysis of the corpus in terms of (a) ontology metrics, following the Neontometrics approach [43] and (b) LOD metrics, following the LODStats approach [15, 16].

As such, the aim is to obtain a comprehensive picture of the current state of structured data stored in SMWs, evaluating the quality and internal structure, thereby tracking the evolution of a significant, previously unexplored part of the LOD ecosystem. It is hypothesised that the domain-specific, small and medium-sized KGs, represented by SMW instances, closely reflect Enterprise Knowledge Graphs (EKGs), for which no publicly available corpus exists. Analysing the corpus, in comparison with the "classic" publicly available collaborative KGs like Wikidata, will therefore gain insights into the possible different parameters to be found in EKGs.

The remainder of this work is structured as follows: Section 2 introduces concepts relevant to this work. Readers may skip portions of this section depending on their familiarity with the Semantic Web. In Section 3, the architecture and the approach to collecting the SMW corpus are described. Section 4 introduces the methods used for corpus analysis. Corpus statistics and results of analyses, including a comparison with similar metrics applied to the "traditional" LOD Cloud and Wikidata, are summarised in Section 5. Section 6 concludes this work with a discussion and an outlook on future work.

2 Background

The following section introduces the basic concepts underlying this work. Its goal is to make the work self-sustained within reasonable limits and to recapitulate those developments in the field of the Semantic Web that gave rise to the current state and challenges of Linked Data and collaborative Knowledge Graphs. Advanced readers familiar with the field may skip this section altogether or pick up at the appropriate level of detail before proceeding to the next sections.

2.1 Semantic Web

As a research area, the Semantic Web "is not primarily driven by certain methods inherent to the field, which distinguishes it from some other areas such as Machine Learning. Rather, it is driven by a shared vision, and as such it borrows from other disciplines as needed" [30]. From the beginning, the vision for the Semantic Web lay in an extension of the existing decentralized hypertext Web through a well-defined informational layer that would both be accessible for humans and enable the intelligent software agents to carry out complex tasks for the users by traversing the Web [7]. This informational layer is enabled by the RDF (Resource Description Framework) data model, which allows interlinking singular concepts rather than documents, as is the case for the hypertext Web, and also introduces typed links [28]. Though the shared vision for the Semantic Web has arguably transformed over time, losing its focus on the intelligent agents and even its characteristic coupling to the Web (Section 2.5), the RDF data model remains indisputably the cornerstone of the field.

2.2 RDF data model

RDF expresses knowledge through the assertion of truthy triples, i.e. subject-predicate-object statements[12] [35]. The model is further characterized by its use of IRIs (Internationalised Resource Identifiers) to uniquely identify concepts used in statements. While IRIs can take any of these three positions in a statement (subject, predicate or object), a typed or a plain literal value can take only the object position and a blank node (which stands for a particular resource or literal not identified by an IRI) can only take subject or object positions in a triple. This arrangement makes it possible to represent data in the form of a directed edge-labelled graph, with subjects and objects being interpreted as nodes and predicates as edges. Recent (most notably, named graphs as part of RDF 1.1[13]) and current developments (e.g., RDF* [4]) concern reification (i.e. statement-level annotations; see [29] for a more detailed and use case supported treatment of the subject). A currently proposed standard, RDF*[14], in particular, would allow referring to the whole statement as the subject or object of another statement.

Because there is no need for ex-ante schema definition in RDF, the data model makes an ad-hoc extension of data straightforward and has the advantage of integrating separate datasets by concatenation of their statements.

[12] Further referred to as *statements*

[13] `https://www.w3.org/TR/2014/NOTE-rdf11-new-20140225/`

[14] Also referred to as RDF 1.2, see `https://www.w3.org/TR/rdf12-concepts/`

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<https://dobriy.org/foaf.rdf#me> rdf:type foaf:Person.
<https://dobriy.org/foaf.rdf#me> foaf:knows <http://www.polleres.net/foaf.rdf#me>.
```

Figure 1: RDF excerpt from the author's Web page[15]

RDF data can be queried and manipulated with SPARQL[16] (SPARQL Protocol and RDF Query Language), a pendant to SQL designed specifically for the RDF data model.

While a number of textual syntaxes exist for representing RDF data, including in convenient and compact human-readable formats, in the following, the Turtle syntax[17] will be used because of its widespread use and the ability to define prefixed names instead of using full IRIs[18]. In Turtle, "a" is used as a short form for rdf:type, IRIs are enclosed in "<" and ">", a semicolon ";" is used to indicate that the next statement has the same subject and a comma "," indicates the same subject and predicate.

2.3 Ontologies

The RDF data model allows the creation of increasingly complex descriptions (e.g., descriptions in Figure 1) which (informally) include not only the so-called A-Box (Assertion Box) statements that concern individuals (e.g., the author of this work represented by the IRI *<https://dobriy.org/foaf.rdf#me>*), but also T-Box (Terminology Box) statements comprising statements that define classes (e.g., *foaf:Person*) or properties (e.g., *foaf:knows*) respectively. This way of distinguishing statements is inspired by DLs (Description Logics) [31]. In this work, T-Box statements are considered to comprise an ontology, i.e. the set of all statements not describing individuals directly, but rather describing classes and properties that are in turn used to describe individuals and in conjunction with individuals.

The term has, however, no ubiquitously accepted definition and an RDF dataset considered to be an ontology can nevertheless and indeed often does include A-Box statements. Ontologies are commonly defining concepts and relations of a particular limited domain. Such ontologies are referred to as

[15] https://dobriy.org/foaf.rdf

[16] https://www.w3.org/TR/sparql11-overview/

[17] See https://www.w3.org/TR/turtle/

[18] Turtle allows writing *rdfs:label* instead of the full IRI of *http://www.w3.org/2000/01/rdf-schema#label* given the respective prefix *rdfs* has been defined earlier, which is implied for all prefixes in this work.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

foaf:Person rdf:type rdfs:Class,
              owl:Class;
            rdfs:comment "A person.";
            rdfs:isDefinedBy foaf:;
            rdfs:label "Person";
            rdfs:subClassOf <http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>,
                            foaf:Agent;
            owl:disjointWith foaf:Organization,
                             foaf:Project;
            owl:equivalentClass <http://schema.org/Person>,
                                <http://www.w3.org/2000/10/swap/pim/contact#Person>;
            vs:term_status "stable".

foaf:knows rdf:type rdf:Property,
             owl:ObjectProperty;
           rdfs:comment "A person known by this person (indicating some level of reciprocated
                         interaction between the parties).";
           rdfs:domain foaf:Person;
           rdfs:isDefinedBy foaf:;
           rdfs:label "knows";
           rdfs:range foaf:Person;
           vs:term_status "stable".
```

Figure 2: Selected definitions from the FOAF ontology[19]

lower ontologies, e.g., FOAF (Friend of a Friend[20]), whose selected definitions are presented in Figure 2. A *higher ontology*, in turn, describes more general concepts that can be used to model lower ontologies. The standardized nature of RDF, the ability to conveniently join RDF datasets, and the usage of global identifiers promote the reuse of ontologies inside and across their respective domains.

Two broadly reused higher ontologies are RDFS (RDF Schema) and OWL (Web Ontology Language). RDFS provides an ontology for modelling class and property hierarchies through the use of *rdfs:subClassOf* and *rdfs:subPropertyOf* properties respectively and allows to impose classes as domains and ranges of given properties through *rdfs:domain* and *rdfs:range*. OWL further extends RDFS by more complex semantics, including establishing equality (*owl:sameAs*), equivalence (*owl:equivalentClass*, *owl:equivalentProperty*) and disjointness (*owl:disjointWith*, *owl:propertyDisjointWith*) of properties and classes [41]. Other well-known ontologies are DCAT (Data Catalog Vocabulary), FIBO (Financial Industry Business Ontology), Schema.org (rich page metadata) etc.

[19] http://xmlns.com/foaf/0.1/

[20] Ibid.

2.4 Linked Data and Linked Open Data

With time, various authorities have published numerous RDF datasets that became openly available on the Web. The data published as RDF is commonly referred to as *Linked Data* or *LOD (Linked Open Data)* when emphasizing its free and open accessibility. Linked Data follows 4 basic principles [6]:

LD1 Identifying resources through HTTP URIs[21]
(Universal Resource Identifiers)
LD2 Making these URIs dereferenceable[22]
LD3 Dereferencing them to RDF data about a resource
LD4 Linking to related RDF from other authorities

LOD can also be assessed according to the five-star deployment scheme, which includes the following five incremental steps, each adding a star to the total rating [32]:

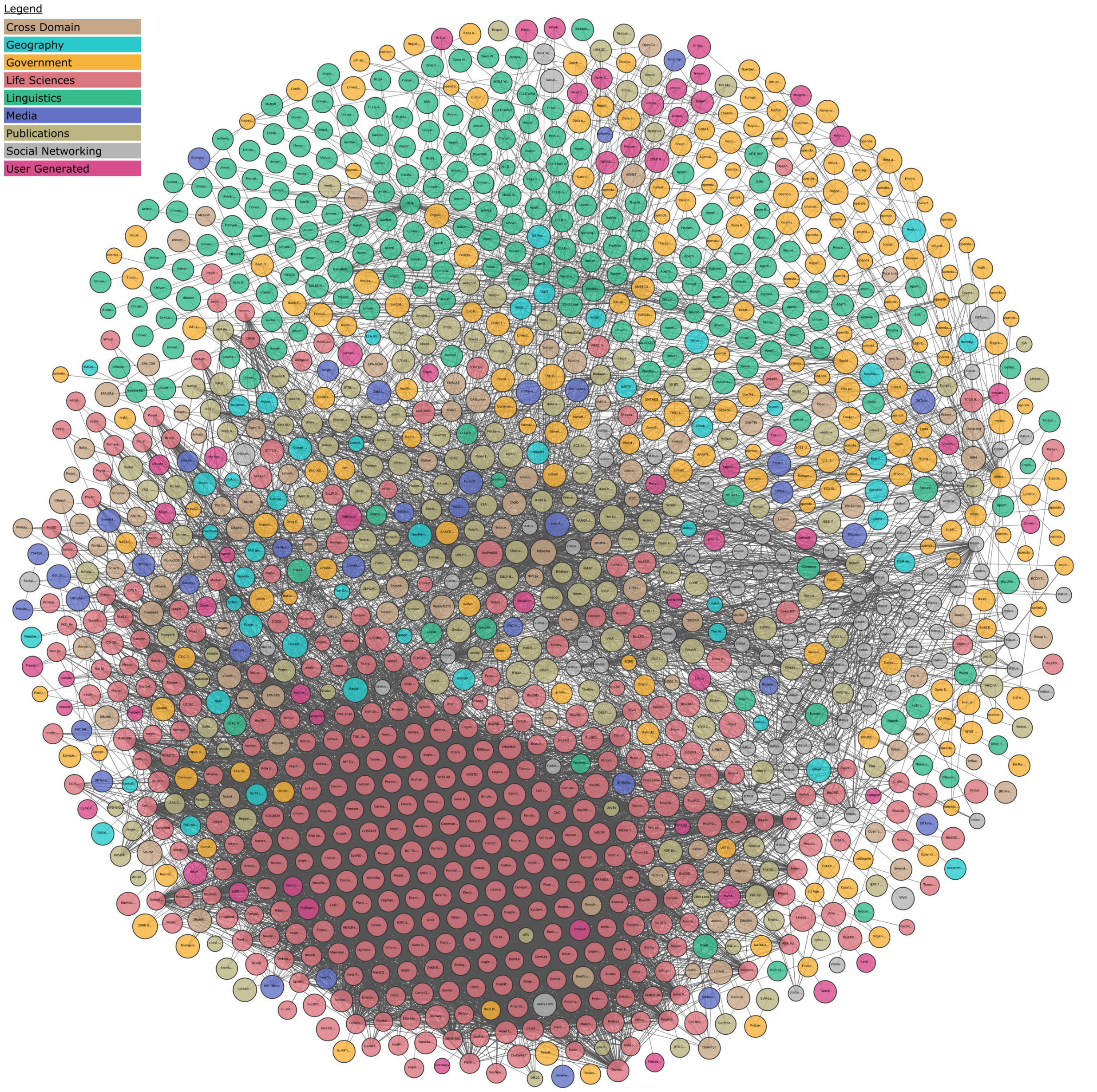
$*$ Publish data on the Web in any format (e.g. PDF, JPEG) accompanied by an explicit Open License (expression of rights).
$**$ Publish structured data on the Web in a machine-readable format (e.g. XML).
$***$ Publish structured data on the Web in a documented, non-proprietary data format (e.g. CSV, KML).
$****$ Publish structured data on the Web as RDF (e.g. Turtle, RDFa, JSON-LD, SPARQL)
$*****$ In your RDF, have the identifiers be links (URLs) to useful data sources.

The Semantic Web standards have created a foundation for Linked Data where anyone can publish potentially conflicting statements about any resource identified by any particular IRI. By design, the Semantic Web and, consequently, LOD support the AAA principle (Anyone can say Anything about Any topic) [3].

The manually curated LOD Cloud (illustrated in Figure 3) collects many published datasets that conform to the Linked Data Principles [38]. Studies analysing Linked Data have shown that LOD Cloud datasets have significant issues with availability [49] (precluding LD2, LD3). Furthermore, it has been found that links between datasets in the LOD Cloud are very sparse on the level of individuals and although ontology reuse is common, many links are broken [38], pointing to issues with LD4 as well as LD2 and LD3 conformance.

[21] A subset of IRIs

[22] When a given URI can be used to fetch information about the resource that the URI represents.



The Linked Open Data Cloud from lod-cloud.net

Figure 3: LOD Cloud, illustration from `http://lod-cloud.net`

2.5 Knowledge Graphs

Besides LOD, which is openly published on the Web, *KGs (Knowledge Graphs)* represent a related development that similarly had its roots in the Semantic Web. A KG can be defined as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities" [31]. In contrast to LOD, the first KGs were maintained by large technology companies and have, therefore, been predominantly closed to the public, albeit with notable exceptions. The prominent KGs published online are Wikidata, DBpedia, YAGO and BabelNet [31].

2.6 Wikis

Wikis, stemming from “Wikiwiki” (Hawaiian for “quick”), is another development that significantly affected the collaborative aspect of the Web, allowing their users to collaboratively edit the content of informational pages. They are characterized by 1) a non-linear hypertext structure (pages linking to each other), 2) easy and extensive access, 3) not requiring any special client software and 4) not requiring any additional training for usage [14].

These characteristics make Wikis especially well suited for collaborative Knowledge Management in organizations, where they pose as "antidotes to many barriers in information sharing" [25]. Today, most publicly available Wikis are powered by the open-source MediaWiki software.[23] Evaluating all the MediaWiki-powered portals collected by methods described in Section 3.2, we can attest to their broad geographic adoption (whenever geographic data is available) both worldwide (as illustrated by Figure 4) and in Europe (as illustrated by Figure 5).

By utilizing LLMs to freely classify the topics of known Wikis,[24] we can also observe a broad variety of underlying subjects (see Figure 6), especially noting the significant portion of Wikis used in business and industry.

2.7 Semantic MediaWiki

First available in 2005, SMW (Semantic MediaWiki) is likewise an extension for the MediaWiki platform that adds semantic annotations to pages.[25] It

[23] https://www.mediawiki.org/wiki/MediaWiki

[24] To this end, text extracted from the main page of each Wiki instance is passed in a prompt to GPT-4: *"Given the following text from the main page of a Wiki, propose a likely category for this Wiki: [WIKITEXT]".*

[25] https://www.semantic-mediawiki.org/wiki/Help:Introduction_to_Semantic_MediaWiki#User_manual

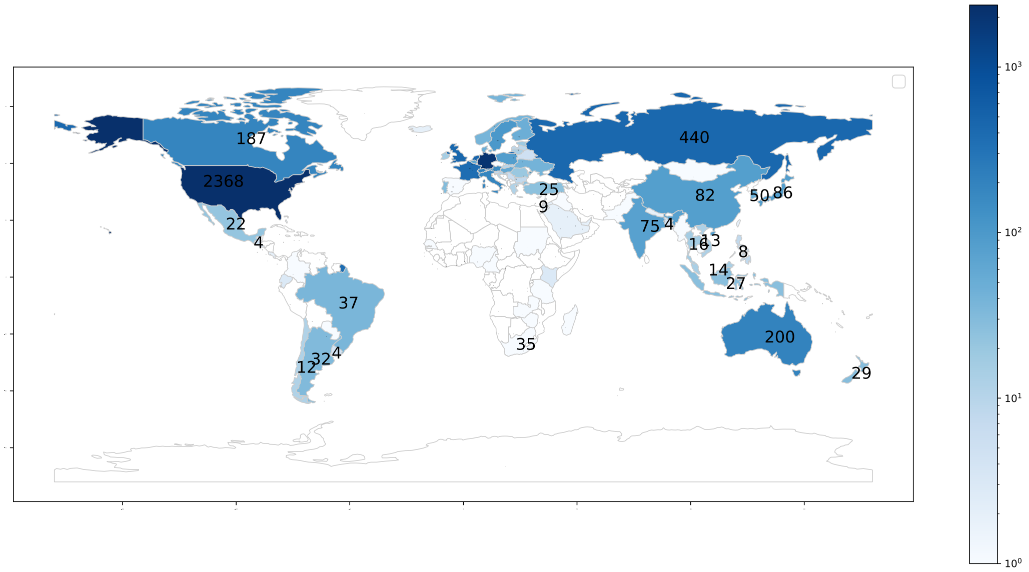


Figure 4: Worldwide adoption of Wikis (small values and Europe excluded)

is also a precursor that majorly inspired Wikidata (and therefore Wikibase) which uses select code from SMW for common tasks [50]. It has been available for a longer time and is more broadly used than Wikibase by various projects to manage their structured data. SMW differs from another related platform, Wikibase,[26] in that it stores data as part of its textual page content and has a less complicated data model [50].

Because of its broad adoption and availability of tools for importing and exporting ontologies, SMW portals are the focus of this work. The data is stored in a relational database by default but can also be reflected in potentially any triple store that accepts SPARQL updates, consequently making it queryable with SPARQL.[27] The SPARQL querying is made available with special extensions.[28] The platform allows to import external RDF vocabularies[29] and has available dedicated tools for specifically importing ontologies.[30]

[26] See https://wikiba.se

[27] https://www.semantic-mediawiki.org/wiki/Help:Using_SPARQL_and_RDF_stores

[28] https://www.mediawiki.org/wiki/Manual:Managing_data_in_MediaWiki

[29] https://www.semantic-mediawiki.org/wiki/Help:Import_vocabulary

[30] https://github.com/TIBHannover/ontology2smw

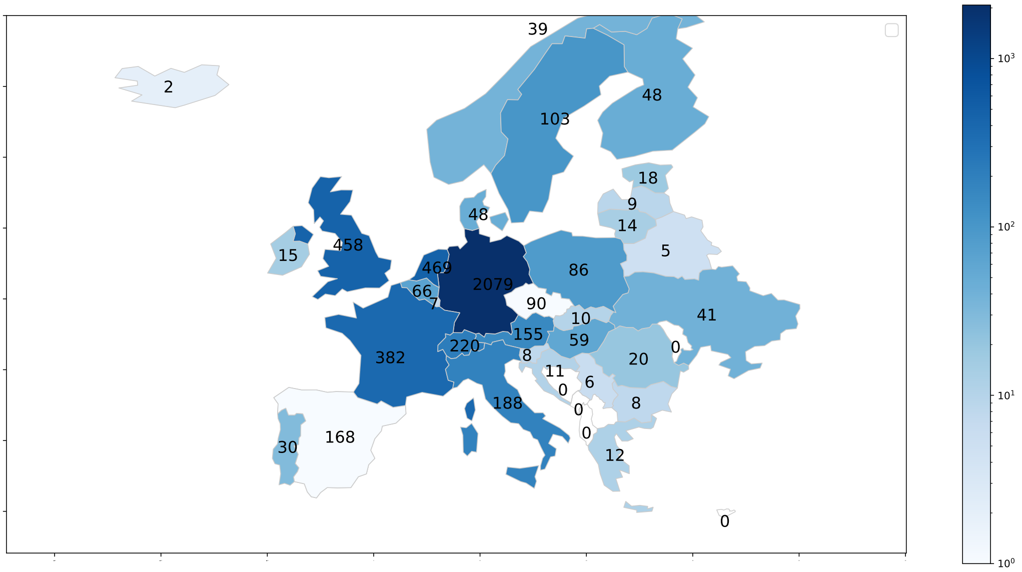


Figure 5: Adoption of Wikis in Europe

MediaWiki platform uses *categories* to classify its articles. While categories are merely tags and, therefore, do not allow typed links, their notation is extended to capture typed links in SMW. The platform thus views properties as "categories for values".[31]

The data model of SMW also fundamentally follows the subject-predicate-object statement structure known as a *semantic fact* in SMW, where subjects are commonly single pages, and objects can be of different datatypes (e.g. numbers, dates, pages).[32] Properties (predicates) are normally defined by typing Wikitext in pages or with templates and forms enabled through additional extensions. The default property datatype at creation is *Page*, but can be changed anytime. Properties with datatype *Reference* can be used to define both value and provenance information.[33]

SMW annotation data can be exported only partially or as an RDF dump of all existing annotations. The category assigned to a regular page is mapped to RDF as the value of rfd:type property of the page. Similarly, when a

[31] https://www.semantic-mediawiki.org/wiki/Help:Properties_and_types#User_manual

[32] https://www.semantic-mediawiki.org/wiki/Help:Datamodel

[33] https://github.com/SemanticMediaWiki/SemanticMediaWiki/issues/1808

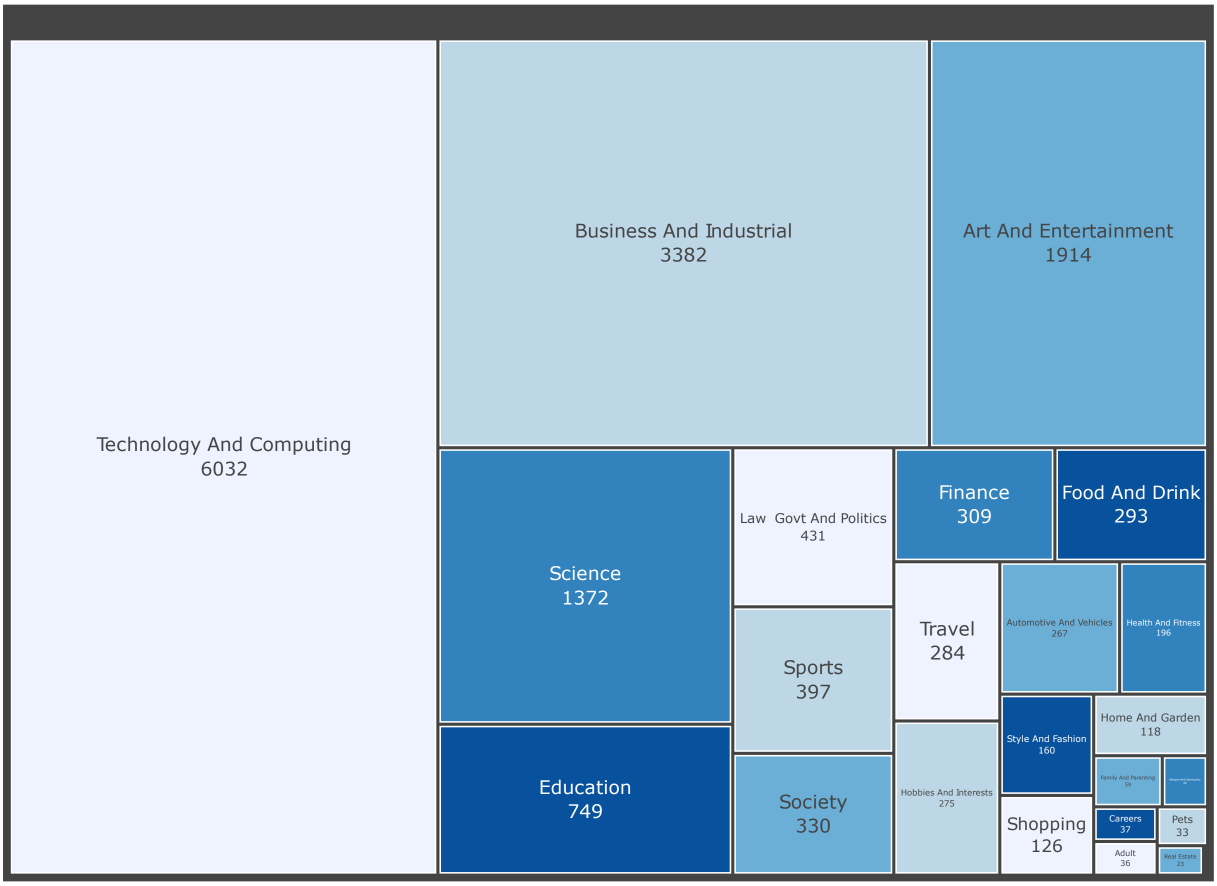


Figure 6: Topical diversity of Wikis

category is assigned to itself a category page, it is mapped as the value of rdfs:subClassOf of the category page.[34]

3 Methods for collecting the corpus

As noted in Section 1, the initial contribution of this work lies in establishing a process of discovering and extracting the RDF contents from available Semantic MediaWiki instances. The overall corpus collection approach is illustrated in Figure 9. In the first step, Semantic MediaWiki instances are discovered in a three-fold manner:

[34] https://www.semantic-mediawiki.org/wiki/Help:RDF_export

- filtering the *BuiltWith* MediaWiki collection[35] for SMW instances;
- querying *WikiApiary*[36] – a "meta-wiki" collecting information about public MediaWiki instances – for instances that have the SMW extension installed;
- lastly, using search engine APIs to discover SMW instances by specific text excerpts or HTML elements commonly found on SMW-powered websites: e.g., "powered by Semantic MediaWiki" text snippets. This approach allows the discovery of previously unknown instances and is further described in Section 3.1.

3.1 Discovering Wikis with search engines[37]

The field of Web platform discovery, which involves the systematic identification of websites, is a research priority for discovering Linked Open Data (LOD) [8] and accessing the factual extent of the Semantic Web. This subject intersects with Web crawling, an automated process concerned with the traversal and extraction of Web content and search engine scraping.

Investigations in the field [48] have presented scalable algorithms for pattern mining, significantly enhancing the efficiency of media-type focused crawling. Additionally, efforts like MultiCrawler have proposed pipeline architectures for more effective crawling and indexing of the Semantic Web data [27]. Other notable tools, such as Apache Any23,[38] offer extraction libraries and Web services that transform structured data from HTML and other Web documents to more useful formats. The relevance of the application of such tools is illustrated by services like Portalwatch [47] and WikiApiary[39], which monitor the deployment and usage of specific Open Data and Wiki platforms on the Web. Finally, due to the inherent cost of the platform and dataset discovery, services like LOD Laundromat [5] and LOD Cloud[40] exist to provide an entry point and catalogue linked datasets.

In the case of WikiApiary, the service provides a comprehensive repository that tracks and catalogues Wikis and their respective metadata on the Web. Most notably, WikiApiary also collects Semantic Wikis. Despite its extensive coverage and reliance on bots ("bees") to keep the metadata up-to-date, the catalogue is manually curated through community submissions, which could

[35] https://trends.builtwith.com/websitelist/MediaWiki

[36] https://wikiapiary.com

[37] This section has been accepted to the Posters, Demos and Industry Track of the 22$^{\text{nd}}$ International Semantic Web Conference as a standalone submission [12].

[38] https://any23.apache.org/

[39] https://wikiapiary.com

[40] http://lod-cloud.net

potentially introduce gaps in data collection. The proposed tool aims to enhance and ease Web platform discovery in this area.

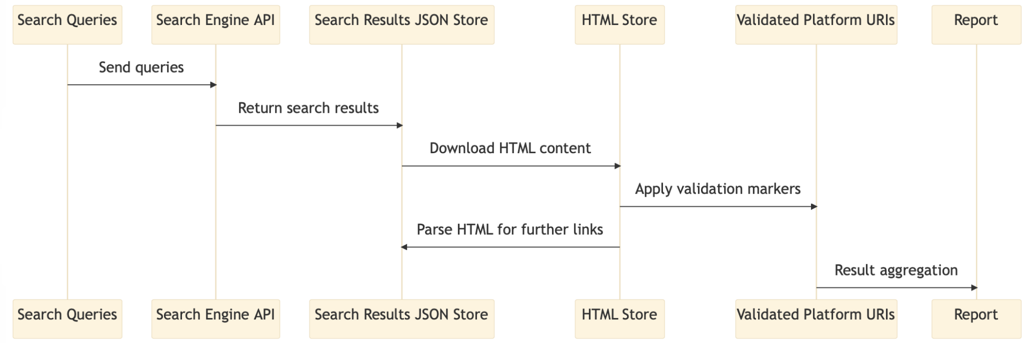


Figure 7: Architecture diagram of the search engine crawler [12]

Crawley is an open-source Python-engineered command-line tool designed to streamline the discovery and validation of specific technological platforms. It is currently available together with documentation on GitHub[41] under a CC-BY 4.0 license.[42] Figure 7 illustrates the high-level architecture of the tool. The tool extends various search engine APIs (SERP API, BING API) as a reliable solution to search engine querying. Thus, the search is performed with Google, Bing, Yandex, Yahoo, DuckDuckGo, Baidu and Naver.

The user can initiate a search event, which is defined by a search engine (i.e., Google, Bing, Yandex, Yahoo, DuckDuckGo, Baidu, Naver) and the query itself.[43] The tool then queries the search engines, performing result pagination until all the query results are exhausted and prints the actual number of unique sites, giving the user a heuristic estimation of how prodigious a certain query-search engine combination is, and aggregates the search results in the ./results folder. Although the queries can be formulated freely, we recommend using a subset of markers defined in the paragraph below that have a probability of being indexed by search engines (i.e., text snippets and image annotations, but not code excerpts). A trade-off pattern is observed whereby more general queries lead to more results, but fewer validation hits at the end, and more specific queries lead to fewer results, but a larger proportion of hits, which gives merit to formulating both general and specific queries.

[41] https://github.com/semantisch/crawley

[42] http://creativecommons.org/licenses/by/4.0/

[43] Cf. documentation for the tool on https://github.com/semantisch/crawley/README

The results/platform validation process with Crawley begins with the user identifying text/code snippets commonly found on sites using a particular technology of interest: "Powered by Semantic MediaWiki", "CKAN API", "Socrata API" as well as components of URL commonly used by a specific platform (e.g., *.../dataset*). We designate these as *markers*. Having identified possible markers and defined them in the configuration,[44] the user can initiate a validation phase, whereby the tool requests HTML contents for the collected search results and then matches them against the markers, returning the total number of validation hits for each platform type and producing a validation report.

Finally, the tool is able to recursively extract further links from validated sites. This is a useful feature which relies on the fact that similar platforms often contain hyperlinks to each other. The extracted links are then treated as search results in the pipeline and can be validated further, whereby previous HTML collection and validation events, as well as results, are cached for efficiency.

In the case of discovering Semantic MediaWikis, a search (without recursive link collection) and validation have been performed with Crawley using the Bing search engine.

A set of custom markers has been identified in association with Semantic Wikis:

```
<meta name="generator" content="MediaWiki"
<link rel="ExportRDF"
Powered by MediaWiki
```

However, as noted before, only "Powered by MediaWiki" was then used for the queries, as other snippets are not indexed by search engines. Additional queries were therefore devised: "MediaWiki", "Semantic MediaWiki", and "Semantic Wiki".

3.2 Methods for collecting statistics

Combining all three approaches, an extensive list of 1458 SMW instances is collected, for details on the numbers of instances per source, refer to Figure 8. For each found SMW instance, basic statistics directly available via SMW are retrieved, such as the number of pages, number 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 page list. Next, a crawl of RDF is attempted

[44] Cf. documentation for the tool on https://github.com/semantisch/crawley/README

by querying the page list, using the MediaWiki API.[45] 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, the versioning history is crawled to identify all changes per page for future analysis.[46] In the last step, the RDF is aggregated[47] by Wiki instance to create the corpus. Apart from creating a single HDT file per crawled instance containing all RDF triples, metadata per Wiki KG is added using the VoID [2] and DataCube [18] vocabularies to include more complex statistics and metrics discussed in detail in Section 4 below, in the HDT headers.

As an additional item in this metadata aggregation, Wikis are collected and classified per "topics", similar to the LOD Cloud topics.[48] The 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 sample of page titles and the name of the Wiki. The GPT-4 is then fed with this textual information[49] to assign it one of the LOD Cloud topics – this current naive approach serves mainly for illustration, cf. Figure 11 below.

3.3 Corpus availability

The SMW corpus is available under a permanent URL[50] and directly on the open data repository of the institute.[51] Both separate RDF HDT [20] dumps per SMW instance, as well as a single HDT file for the whole aggregated corpus are provided. 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 is provided under the Creative Commons Attribution 4.0 International License.[52]

[45] `https://www.mediawiki.org/wiki/API:Query`

[46] Cf. Section 6: It is planned to also extract and analyse the *evolution* of the RDF KGs per SMW instances as future work.

[47] Given the absence of Blank Nodes within instances, skolemization is unnecessary.

[48] `https://lod-cloud.net/`

[49] The respective prompt is: *"Given the text "[WIKI-NAME+MAINPAGETEXT+PAGETITLES+METAINFO]", tell me the best fitting topic among [LODCloudTopiclist]".*

[50] `http://purl.org/SMWCloud`

[51] `https://semantic-data.cluster.ai.wu.ac.at/smwcloud/`

[52] `https://creativecommons.org/licenses/by/4.0/`

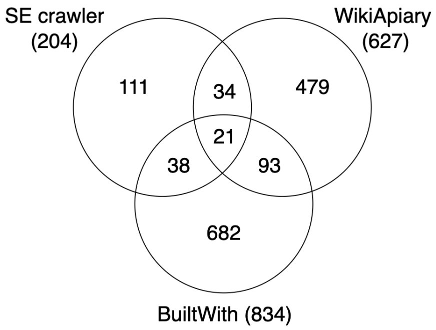


Figure 8: Venn diagram of collected SMWs and their sources

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 page list 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

Table 1: Breakdown of collection and processing losses

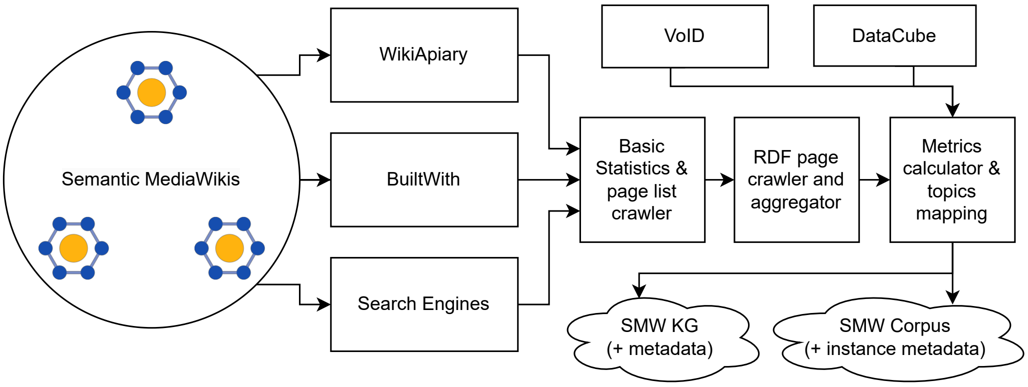

Semantic MediaWikis
WikiApiary
BuiltWith
Search Engines
VoID
DataCube
Basic Statistics & page list crawler
RDF page crawler and aggregator
Metrics calculator & topics mapping
SMW KG (+ metadata)
SMW Corpus (+ instance metadata)

Figure 9: Architecture of the SMW crawler and metrics processor

For illustration, the VoID metadata and selected DataCube entries available through an endpoint (and as part of the respective HDT header) are provided for "Wien Geschichte Wiki," an SMW instance providing historical information about Vienna [33] in Turtle syntax in Figure 10.

4 Methods for analysing the corpus

Due to their topical diversity, data and schemas differ considerably among SMW instances, but it is also hypothesized that the RDF data from SMW instances has fundamentally different characteristics than other Linked Open Data corpora, such as the LOD Cloud datasets or Wikidata.

```
@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 .
```

Figure 10: Metadata for “Wien Geschichte Wiki”

4.1 Analytical framework

In order to verify these assumptions, a comprehensive characterization of the corpus requires two things: (1) a fundamental understanding of each dataset, and (2) an overview of all available data [9]. Therefore, the analyses are performed both on the single Wiki datasets as well as on the corpus as a whole in terms of commonly used metrics.

Use case \ **Paper**	Fernandez et al. [21]	Ermilov et al. [16] [17] [15]	Nogalez et al. [39]	Reiz et al. [44]	Abedjan et al. [1]	Haller et al. [26]	Rietveld et al. [45]
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. [15]	Basic graph metrics
Reiz et al. [43]	Basic ontology metrics[53]
Haller et al. [26]	Instance, ontology links
Yao et al. [54]	Cohesion
Yang et al. [53]	Complexity
Fernandez et al. [22]	Coverage, structure
Duque-Ramos et al. [13] [42]	SQuaRE[54]-based quality
Gangemi et al. [23]	Structure, functionality, usability
Tartir et al. [46]	Populated ontology (instances, schema)
Orme et al. [40]	Quality, completeness, and stability

Table 2: Related work and implemented metrics

In order to establish a foundation for comparative analysis, related LOD analysis studies are reviewed 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*.

4.2 Processed metrics

General graph metrics give a comprehensive and comparable characterization of the corpus [15]. Additionally, *basic ontology metrics* are calculated 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 the analyses through the implementation of the Neontometrics calculation engine [44], the lack of metric validation for real-world data [43]

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

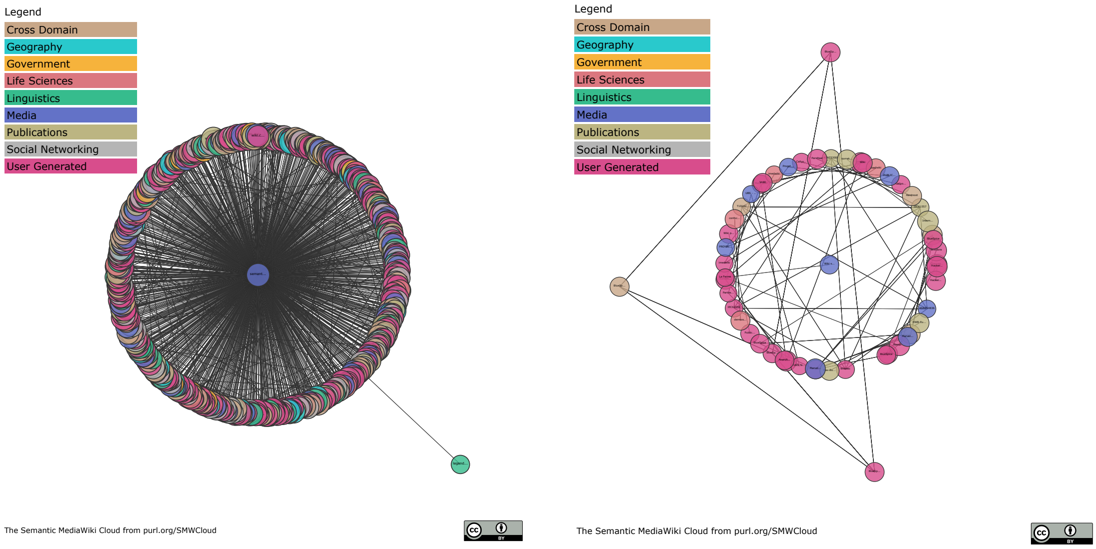
[54] `https://www.iso.org/standard/64764.html`

is addressed, especially in *small KGs*: SMW instances can be viewed as a publicly available pendant/proxy for enterprise KGs with their own characteristics.

Table 2 briefly summarizes the various metrics and quality frameworks calculated for the SMW corpus. As for calculating the basic graph and ontology metrics, the Neontometrics OPI (Ontology Programming Interface) is used. Afterwards, these metrics are applied to calculate common quality frameworks: Cohesion Metrics [54], Complexity Metrics [53], Fernandez et al. [22], OQuaRE [13][42], Gangemi et al. [23], OntoQA [46] and Orme et al. [40]. Please refer to Table 2 for a summary of their respective dimensions.

5 Results and corpus statistics

In this section, a brief overview of the insights into the corpus based on the collected statistics is provided. The datasets causing processing errors as described in Table 1 are excluded, resulting in 1029 datasets that collectively form the corpus and the basis for analyses.



The Semantic MediaWiki Cloud from purl.org/SMWCloud

The Semantic MediaWiki Cloud from purl.org/SMWCloud

Figure 11: SMW Cloud with and without `semantic-mediawiki.org`

5.1 Basic RDF metrics

Table 3 gives a distilled overview of the corpus dimensions. Notably, the absolute numbers reported for SMW Cloud in this work fall short of the statistics for the number of statements reported by the instances themselves,

Dataset	#Triples	#Subjects	#Predicates	#Objects	#Literals
LODStats [16]	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[55]	17,662,800,665	1,625,057,179	38,867	Not reported	Not reported
LOD-a-lot [21]	28,362,198,927	3,214,347,198	1,168,932	3,178,409,386	1,302,285,394

Table 3: SMW Cloud summary statistics

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 [16], which totals 192,230,648 triples. Wikidata has grown considerably in the last 5 years, from 3b triples in 2018 to more than 17b in 2022, exceeding SMW Cloud in size [19]. 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, refer to Table 4. There, their parameters are compared to the individual LOD Cloud datasets [16]. 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 12,595.6 for SMWs), implying more uniform sizes of SMWs. Another key characteristic of SMWs is a higher number of properties and classes per dataset as well as a higher number of properties per entity, suggesting a more granular and detailed data modelling approach in SMWs and a user-centric, bottom-up nature of ontology creation. The lack of class and property depth is indicative of a flat ontology structure in the SMW instances. This is attributed 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 labelled subjects and large average typed and untyped string lengths for Literals further differentiate user-centric SMWs from LOD datasets.

5.2 Topical analysis

Semantic MediaWikis capture a variety of highly-specialized domain knowledge, as visualized in Figure 11 reusing the LOD Cloud categories:[56] 16% Wikis specialize on publishing/annotating *media*, 10% are *life science* Wikis,

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

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

	SMW Cloud				LOD Cloud [16]			
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

Table 4: SMW Cloud and LOD Cloud comparison

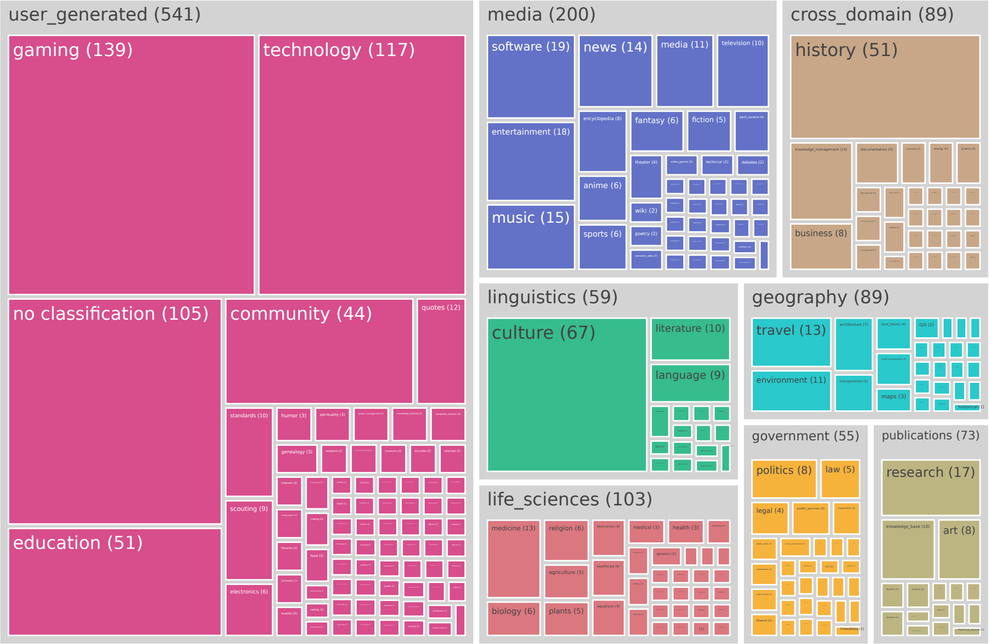


Figure 12: Prevalent topics across SMW Cloud instances

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, each Wiki is characterized by 3 independently created tag/domain collections, which promote discoverability and an automatically generated description.

Further, the distribution of specialized domains in the corpus with freely annotated tags is analyzed, see Figure 12. 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 because of the lack of investigated content (see Section 3). While 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.

SMW	LOD Cloud classification[57]	Free classification	Description	Topic (WA)	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 Entertainment
bacid.eu	government	public administration	*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 Kroetzsch, 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.*	-	-

Table 5: Examples of SMW domain annotations

5.3 Ontological analysis

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

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

metrics for characterizing KGs and ontologies as suitable for analysing KGs are proposed, with notable limitations applying to ontology metrics discussed separately.

Reiz et al. [43] 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 metrics and quality analyses calculations.[58] 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. All quality analyses are calculated based on these basic ontology metrics suggested by the Neontometrics engine (see Table 8).

Ontologically SMWs 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[59] 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.[60] 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 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 [24]. 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 [46], see also Footnote 61 in Table 6.

[58] http://neontometrics.com/

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

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

[61] 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 these metrics are not provided separately.

	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[61]	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

Table 6: Basic ontology metrics

	$\mathcal{AL}$	$\mathcal{AL}$(D)	$\mathcal{ALO}$(D)	$\mathcal{ALEO}$(D) -	$\mathcal{ALHO}$(D) -	
Number of SMW instances:	(99)	(165)	(708)	(50)	(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 7: Use of the RDF(S)+OWL and DL expressivity of SMW instances

Metric	Mean	Min	Max	Median	Score
Mean number of annotations per class [13, 42]	37,96	0,65	710,16	14,29	1 (excellent)
Mean number of attributes per class [13, 42]	1,05	0,01	2,67	0,77	1 (excellent)
Weighted Method Count: Mean number of properties per class [13, 42]	345,37	0,70	55439,4	35,19	5 (unsatisfactory)

Table 8: OQueRE metrics and scores

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

6 Conclusion and future work

This work presented and characterized the SMW Cloud corpus, an extensive collection of RDF data collected from Semantic MediaWiki instances. 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 [52]. It is planned 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, it is recommended 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-endpoint is available, make it discoverable through SMW API and 3) consider adopting HDT as a compact format for publishing regular dumps as it is a highly compact format for SMW data (achieving a Data Compression Ratio of 17,5 for SMW Cloud compared to NTriples, more efficiently than for other benchmarks evaluated [36]).

In terms of evaluation and benchmarking as a field of interest of the Semantic Web community [10], SMW Cloud provides a novel and distinct dataset with unique characteristics that introduce variety into the field of LOD sources investigated so far; these unique characteristics have been demonstrated 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. it is expected that the SMW Corpus will enable previously unexplored approaches in LOD and EKG research.

Finally, this work also presented a novel approach for web platform discovery, Crawley. The tool could be broadly useful both for standalone web platform discovery as well as for the extension of existing manually curated catalogues.

6.1 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.

6.2 Future work

As future work, it is theorized that the SMW corpus can also provide a basis for longitudinal analysis, link analysis [26], etc. This will enable a better understanding of the dynamics and evolution of vocabularies. It is the goal to create profiling tools and resources *enabling users to create an assessment* *of the data at hand* [9]. As noted in Section 3, 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.

With respect to the new method for platform discovery, the promising areas for future work include 1) parallelizing requests, 2) implementing standalone Search Engine crawling and 3) automatic marker discovery, which could greatly increase the efficiency of the discovery process, positioning Crawley as an increasingly valuable asset for comprehensive web platform discovery. Finally, it is planned to applying the method to discover and monitor more Semantic Web resources, such as public Wikibase instances and SPARQL endpoints.

6.3 Resource Availability Statement

The corpus is hosted through the Institute for Data, Process and Knowledge Management. The institute has already hosted various widely adopted Semantic Web resources for several years now and promotes the sustainability strategy within ongoing community activities such as the "Distributed Knowledge Graphs" COST Action,[62] which as one of its activities aims at aligning and sustaining community services and tools.

1. The aggregated SMW Cloud dataset is made available via the institutional repository.[63]
2. The SMW Cloud corpus containing individual SMW datasets is also made available via the institutional repository.[64]
3. The calculated metrics for the corpus are available via a public SPARQL endpoint.[65]

6.4 Acknowledgements

The author expresses profound gratitude to Bernhard Krabina, Amr Azzam, Martin Beno and Alexandra Hager for their invaluable support and expert advice and gives special acknowledgement to Thomas Seyffertitz for his consultation in the area of research data management as well as Gary Brewer, BuiltWith, for providing an extensive MediaWiki collection.

[62] https://cost-dkg.eu/

[63] https://semantic-data.cluster.ai.wu.ac.at/smwcloud/

[64] https://semantic-data.cluster.ai.wu.ac.at/smwcloud/corpus/

[65] https://smwcloud-sparql.cluster.ai.wu.ac.at/

References

[1] Ziawasch Abedjan, Toni Gruetze, Anja Jentzsch, and Felix Naumann. Profiling and mining RDF data with ProLOD++. In *2014 IEEE 30th International Conference on Data Engineering*, pages 1198–1201, Chicago, IL, USA, March 2014. IEEE.

[2] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets. CEUR Workshop Proceedings, 2009.

[3] Dean Allemang, James A Hendler, and Fabien Gandon. *Semantic web for the working ontologist*. ACM Press, 2020.

[4] Dörthe Arndt, Jeen Broekstra, Bob DuCharme, Ora Lassila, Peter F. Patel-Schneider, Eric Prud'hommeaux, Ted Jr. Thibodeau, and Bryan Thompson. Rdf-star and sparql-star, final community group report 17 december 2021, Dec 2021. https://w3c.github.io/rdf-star/cg-spec (Last accessed 23 May 2023).

[5] Wouter Beek, Laurens Rietveld, Hamid R Bazoobandi, Jan Wielemaker, and Stefan Schlobach. Lod laundromat: a uniform way of publishing other people's dirty data. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*, pages 213–228. Springer, 2014.

[6] Tim Berners-Lee. Linked data—design issues. w3c, 2006. https://www.w3.org/DesignIssues/LinkedData (Last accessed 23 May 2023).

[7] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[8] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global, 2011.

[9] Christoph Bohm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grutze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Profiling linked open data with ProLOD. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 175–178, Long Beach, CA, March 2010. IEEE.

[10] Peter Boncz, Irini Fundulaki, Andrey Gubichev, Josep Larriba-Pey, and Thomas Neumann. The Linked Data Benchmark Council Project. *Datenbank-Spektrum*, 13(2):121–129, July 2013.

[11] Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. A suite of generative tasks for multi-level multimodal webpage understanding, 2023.

[12] Daniil Dobriy and Axel Polleres. Crawley: A tool for web platform discovery. In *Proceedings of the 22nd International Semantic Web Conference 2023 Posters, Demos and Industry Tracks*, 2023.

[13] Astrid Duque-Ramos, Jesualdo Tomás Fernández-Breis, Robert Stevens, and Nathalie Aussenac-Gilles. OQuaRE: A SQuaRE-based Approach for Evaluating the Quality of Ontologies. *Journal of Research and Practice in Information Technology*, 43(2), 2011.

[14] Anja Ebersbach, Markus Glaser, Richard Heigl, and Alexander Warta. *Wiki: Web Collaboration.* Springer Science & Business Media, 2008.

[15] Ivan Ermilov, Jan Demter, Michael Martin, Jens Lehmann, and Sören Auer. Lodstats–large scale dataset analytics for linked open data. *Under review in ISWC*, 2013.

[16] Ivan Ermilov, Jens Lehmann, Michael Martin, and Sören Auer. LODStats: The Data Web Census Dataset. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, volume 9982, pages 38–46. Springer International Publishing, Cham, 2016. Series Title: Lecture Notes in Computer Science.

[17] Ivan Ermilov, Michael Martin, Jens Lehmann, and Sören Auer. Linked Open Data Statistics: Collection and Exploitation. In Pavel Klinov and Dmitry Mouromtsev, editors, *Knowledge Engineering and the Semantic Web*, volume 394, pages 242–249. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Communications in Computer and Information Science.

[18] Pilar Escobar, Gustavo Candela, Juan Trujillo, Manuel Marco-Such, and Jesús Peral. Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. *Computer Standards & Interfaces*, 68:103378, February 2020.

[19] Wolfgang Fahl, Tim Holzheim, Andrea Westerinen, Christoph Lange, and Stefan Decker. Getting and hosting your own copy of wikidata. In *Proceedings of the 3rd Wikidata Workshop*, 2022.

[20] Javier D. Fernández, Miguel A. Martınez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF Representation for Publication and Exchange (HDT). *Journal of Web Semantcics*, 19(2), 2013.

[21] Javier D. Fernández, Wouter Beek, Miguel A. Martínez-Prieto, and Mario Arias. LOD-a-lot. In Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin, editors, *The Semantic Web – ISWC 2017*, Lecture Notes in Computer Science, pages 75–83, Cham, 2017. Springer International Publishing.

[22] Miriam Fernández, Chwhynny Overbeeke, Marta Sabou, and Enrico Motta. What Makes a Good Ontology? A Case-Study in Fine-Grained Knowledge Reuse. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *The Semantic Web*, volume 5926, pages 61–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. Series Title: Lecture Notes in Computer Science.

[23] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. Ontology evaluation and validation an integrated formal model for the quality diagnostic task. 2005. https://api.semanticscholar.org/CorpusID:3087032.

[24] Birte Glimm, Adian Hogan, Markus Krötzsch, and Axel Polleres. OWL: Yet to arrive on the web of data? In *WWW2012 Workshop on Linked Data on the Web (LDOW2012)*, Lyon, France, April 2012.

[25] Tay Pei Lyn Grace. Wikis as a knowledge management tool. *Journal of knowledge management*, 13(4):64–74, 2009.

[26] Armin Haller, Javier D. Fernández, Maulik R. Kamdar, and Axel Polleres. What Are Links in Linked Open Data? A Characterization and Evaluation of Links between Knowledge Graphs on the Web. *Journal of Data and Information Quality*, 12(2):1–34, June 2020.

[27] Andreas Harth, Jürgen Umbrich, and Stefan Decker. Multicrawler: A pipelined architecture for crawling and indexing semantic web data. In *The Semantic Web-ISWC 2006: 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006. Proceedings 5*, pages 258–271. Springer, 2006.

[28] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[29] Daniel Hernández, Aidan Hogan, and Markus Krötzsch. Reifying rdf: What works well with wikidata? *SSWS@ ISWC*, 1457:32–47, 2015.

[30] Pascal Hitzler. Semantic Web: A Review Of The Field. *Semantic Web*.

[31] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge*, 12(2):1–257, 2021.

[32] Bernadette Hyland, Ghislain Atemezing, Michael Pendleton, and Biplav Srivastava. Linked data glossary. *W3C Government Linked Data Working Group. http://www. w3. org/TR/ld-glossary*, 2013.

[33] Bernhard Krabina. Building a Knowledge Graph for the History of Vienna with Semantic MediaWiki. *Journal of Web Semantics*, 76:100771, April 2023.

[34] Markus Krötzsch, Denny Vrandecic, and Max Völkel. Semantic mediawiki. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors, *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006, Proceedings*, volume 4273 of *Lecture Notes in Computer Science*, pages 935–942. Springer, 2006.

[35] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107):6, 2004. https://www.w3.org/TR/rdf-primer/ (Last accessed 23 May 2023).

[36] Miguel A. Martínez-Prieto, Mario Arias Gallego, and Javier D. Fernández. Exchange and Consumption of Huge RDF Data. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications*, volume 7295, pages 437–452.

Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. Series Title: Lecture Notes in Computer Science.

[37] Robert Meusel, Petar Petrovski, and Christian Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014*, pages 277–292, Cham, 2014. Springer International Publishing.

[38] Sebastian Neumaier, Axel Polleres, Simon Steyskal, and Jürgen Umbrich. Data integration for open data on the web. In *Reasoning Web International Summer School*, pages 1–28. Springer, 2017.

[39] Alberto Nogales, Miguel Angel Sicilia-Urban, and Elena García-Barriocanal. Measuring vocabulary use in the Linked Data Cloud. *Online Information Review*, 41(2):252–271, April 2017.

[40] Anthony M. Orme, Haining Yao, and Letha H. Etzkorn. Indicating ontology data quality, stability, and completeness throughout ontology evolution. *Journal of Software Maintenance and Evolution: Research and Practice*, 19(1):49–75, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smr.341.

[41] Axel Polleres, Aidan Hogan, Renaud Delbru, and Jürgen Umbrich. Rdfs and owl reasoning for linked data. In *Reasoning Web International Summer School*, pages 91–149. Springer, 2013.

[42] Achim Reiz and Kurt Sandkuhl. Harmonizing the OQuaRE Quality Framework:. In *Proceedings of the 24th International Conference on Enterprise Information Systems*, pages 148–158, Online Streaming, — Select a Country —, 2022. SCITEPRESS - Science and Technology Publications.

[43] Achim Reiz and Kurt Sandkuhl. Neontometrics: A flexible and scalable software for calculating ontology metrics. In *18th International Conference on Semantic Systems, SEMPDW 2022, 13 September 2022 through 15 September 2022*. CEUR-WS, 2022.

[44] Achim Reiza and Kurt Sandkuhla. 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)*. CEUR-WS, Vienna, 2022.

[45] Laurens Rietveld, Wouter Beek, Rinke Hoekstra, and Stefan Schlobach. Meta-data for a lot of LOD. *Semantic Web*, 8(6):1067–1080, August 2017.

[46] Samir Tartir and I. Budak Arpinar. Ontology Evaluation and Ranking using OntoQA. In *International Conference on Semantic Computing (ICSC 2007)*, pages 185–192, Irvine, CA, USA, September 2007. IEEE.

[47] Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Towards assessing the quality evolution of open data portals. In *Proceedings of ODQ2015: Open Data Quality: from Theory to Practice Workshop, Munich, Germany*, 2015.

[48] Jürgen Umbrich, Marcel Karnstedt, and Andreas Harth. Fast and scalable pattern mining for media-type focused crawling. *KDML*, page 119, 2009.

[49] Ruben Verborgh, Tobias Kuhn, and Tim Berners-Lee. Proceedings of the 2nd workshop on decentralizing the semantic web co-located with the 17th international semantic web conference (iswc 2018). In *2nd Workshop on Decentralizing the Semantic Web co-located with the 17th International Semantic Web Conference (ISWC 2018)*, volume 2165, 2018.

[50] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[51] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. Publisher: ACM New York, NY, USA.

[52] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao,

and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. Number: 1 Publisher: Nature Publishing Group.

[53] Zhe Yang, Dalu Zhang, and Chuan Ye. Ontology Analysis on Complexity and Evolution Based on Conceptual Model. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Ulf Leser, Felix Naumann, and Barbara Eckman, editors, *Data Integration in the Life Sciences*, volume 4075, pages 216–223. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. Series Title: Lecture Notes in Computer Science.

[54] Haining Yao, Anthony Mark Orme, and Letha Etzkorn. Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science*, 1(1):107–113, January 2005.