






Reproducibility Crisis in the LOD Cloud? Studying the Impact of Ontology Accessibility and Archiving as a Counter Measure

Johannes Frey¹, Denis Streitmatter¹, Natanael Arndt²,
and Sebastian Hellmann¹

¹ Knowledge Integration and Linked Data Technologies (KILT/AKSW) DBpedia
Association/InfAI, Leipzig University, Leipzig, Germany
{frey,streitmatter,hellmann}@informatik.uni-leipzig.de
² eccenca GmbH, Leipzig, Germany
natanael.arndt@eccenca.com

Abstract. The reproducibility crisis is an ongoing problem that affects data-driven science to a big extent. The highly connected decentral Web of Ontologies represents the backbone for semantic data and the Linked Open Data Cloud and provides terminological context information crucial for the usage and interpretation of the data, which in turn is key for the reproducibility of research results making use of it.

In this paper, we identify, analyze, and quantify reproducibility issues related to capturing terminological context (e.g. caused by unavailable ontologies) and delineate the impact on the reproducibility crisis in the Linked Open Data Cloud. Our examinations are backed by a frequent and ongoing monitoring of online available vocabularies and ontologies that results in the DBpedia Archivio dataset. We also show the extent to which the reproducibility crisis can be countered with the aid of ontology archiving in DBpedia Archivio and the Linked Open Vocabularies platforms.

1 Introduction

The reproducibility crisis is an ongoing problem in science [2] that has a big impact on data centric disciplines as well [11,12,17]. Cockburn et al. and Miyakawa emphasize the importance of the availability of data and materials for research to be reproducible [5,15]. The Linked Open Data (LOD) cloud provides a huge amount of data relevant for data science. The semantic web architecture, as technological foundation for the LOD cloud and major driver for collecting and publishing globally interlinked knowledge, consists of instance data and terminological data. The terminological data is captured by vocabularies and ontologies that make up a common point of reference for the instance data. Reuse of terms across different ontologies and their formalization are crucial patterns for data engineering on the Web of Data and a major aspect to foster interoperability

and data exchange. Accessing that ontological and terminological context information is crucial for the interpretation and use of the instance data. Often this context also formalizes implicit knowledge (e.g. subclass relationships) that is not explicitly materialized in the data itself.

Moreover, accessibility is one key aspect of the FAIR data principles [21] which also explicitly require the use of FAIR ontologies for FAIR (meta)data. Given the the best practice to reuse and derive from existing terms in ontology development, this typically leads to a recursive problem. If an ontology A , that is (re)used by an ontology B , becomes unavailable and therefore loses its FAIRness, then as a result B also loses its FAIRness. Subsequently, accessibility and reliability of vocabularies and ontologies are fundamental requirements for such a decentralized (FAIR) data architecture. Thus we argue that the reproducibility of research based on or utilizing LOD is influenced to a significant extent by the accessibility of the referenced vocabularies.

However, the accessibility of vocabularies and ontologies is subject to constant evolution and unavailability (link rot, “HTTP error 404”). Stakeholders, like *Ontology Users*, *Ontology Engineers*, and *Ontology Researchers* are affected by the unavailability of ontologies in their work to varying degrees. *Ontology Users* apply the terminology in their knowledge graphs and applications and are interested in having a consistently and permanently working application. *Ontology Engineers* create new ontologies by reusing existing terminology and are interested in the reliability of the ontologies they are reusing, as well as in the reliability of their own ontologies. *Ontology Researchers* retrieve data from the LOD cloud (typically according to schematic criteria, perform analyses or benchmarks using the data and ontologies; they are interested in the reproducibility and reliability of their results over a long period of time. Common to all of these stakeholders is, the demand for the availability of pre-existing ontologies and their own contribution in the future.

Based on these abstract requirements, we pursue four main research questions in this paper to further understand the reproducibility crisis on the LOD cloud with a focus on the ontological context.

- RQ1** How does the reproducibility crisis look like in the Linked Open Data cloud in terms of accessing the ontological context?
- RQ2** How big is **(a)** the problem of vocabulary and ontology accessibility issues and **(b)** the impact on the reproducibility crisis in the Linked Open Data cloud?
- RQ3** How much of the terminology used in the Linked Open Data cloud is and is not **(a)** accessible in a formal way (i.e. RDFS/OWL ontologies or SKOS concept schemes) such that it can be automatically preserved, and **(b)** how much is preserved already.
- RQ4** Can archiving contribute as a countermeasure to the accessibility issues of ontologies.

The contribution of this paper is subdivided into the following steps. We provide an analysis of the aspects contributing to the reproducibility crisis on the

Linked Open Data cloud. These aspects are then quantified with the aid of DBpedia Archivio (a unified online ontology interface and open augmented ontology archive). In this way we can depict the impact of the reproducibility crisis on the Linked Open Data cloud. Finally, based on the quantification, a categorization of the impacted vocabularies can be performed to indicate counter-measures, such as the automatic preservation, which leads to an evaluation of two archiving approaches to tackle the impact of the reproducibility crisis.

The remainder of the paper is structured as follows: Sect. 2 gives an overview what material and methods were used while Sect. 3 presents the results. In Sect. 4 we describe related work and Sect. 5 concludes the results and gives an overview over possible future work.

2 Material and Methods

In the following section we describe the tools and their methods which we selected for the analysis setup to answer the research questions. To perform the analysis, we were in need of unified access to, on the one hand a vast amount of ontologies published in the Web of Ontologies, and on the other hand datasets of the LOD cloud. We used the DBpedia Archivio Ontology archive and Linked Open Vocabularies for the former and LOD-a-lot for the latter, which are described in more detail in the next subsections.

The high level perspective on the analysis method is, that we analyze terminology reproducibility aspects on instance data using LOD-a-lot, and accessibility issues of ontologies in general using Archivio’s accessibility statistics to get an impression of the dimension of the reproducibility crisis. We create an index on the terms contained in Archivio & LOV and another index on the terms in LOD-a-lot, that could in general be subject to accessibility issues. By joining the index information, it is possible to determine the minimal number of terms where accessibility issues can be countered by archiving (reproducibility support). In the term index for LOD-a-lot we incorporate frequency (triple) count information, to study the effects also weighted by term adoption. In contrast, we integrate information about the accessibility rate for every term in the Archivio index based the ontology that defines it. In a final step, we measure the effectiveness or impact of this theoretical reproducibility support of DBpedia Archivio by calculating the amount of LOD data (number of triples) that fall into different reliability classes. To complete the picture, we use Archivio’s crawling engine in a sandboxed experiment to preserve terms that are not covered by Archivio and report on issues preventing an inclusion but also the potential of ontologies that could be included in the future.

2.1 DBpedia Archivio - Augmented Ontology Archive

DBpedia Archivio’s initial vision was to create a fully automated, persistent ontology archive that can serve as a backbone for the Semantic Web [8] and to serve as a convenient and stable interface for ontology consumers [9].

Launched in May 2020, Archivio has meanwhile become one of the most exhaustive and recent ontology archives, providing alternative, persistent, and unified access to over 1,600 ontologies¹ in more than 5,000 versions. The daily checks for new ontology versions and automated tests monitor the evolution and accessibility of a huge portion of the ontologies used in the LOD cloud and allow to get a picture of the state of affairs on a global scale. As of September 2021 growth has not reached a plateau, yet and it is steadily growing at a pace of around 12.6 ontologies per week (6 month average, see Fig. 1) [10]. While more than 1440 ontologies were archived automatically via web-scale discovery mechanisms, Archivio also performed over 160 successful ontology inclusions suggested by the community (i.e. submitting the ontology URL manually at <https://archivo.dbpedia.org/add>). This fact and around 90 ontology downloads on an average day (plus 640 daily downloads from major bots) show that Archivio is already being adopted by the community.

2.2 Archivio Ontology Discovery and Monitoring

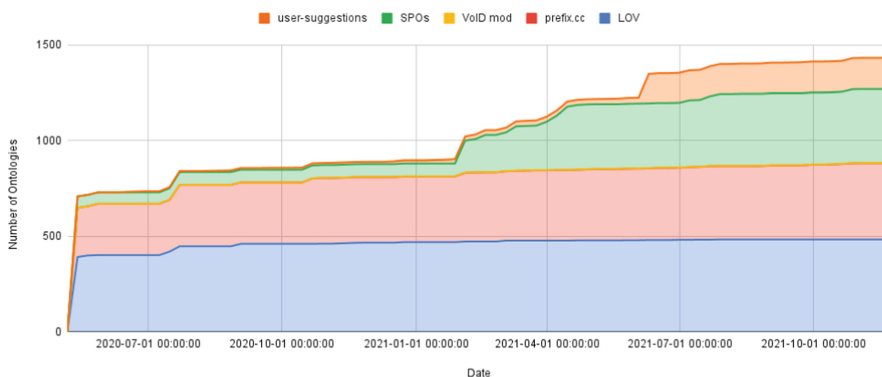


Fig. 1. Development of ontology archive growth, divided by discovery source (State of November 22nd 2021).

Archivio implements four generic approaches to discover RDFS & OWL² ontologies and SKOS³ schemes to be archived. First, it queries already existing ontology repositories and catalogs (currently Linked Open Vocabularies [19] and prefix.cc). Second, it performs a vocabulary usage analysis of all RDF assets on the DBpedia Databus⁴ utilizing VoID Mods that analyse the usage of classes and properties in the datasets. Moreover, it discovers (transitive) dependencies and imports ontologies from previous iterations of Archivio crawls. Finally, users can

¹ <https://archivo.dbpedia.org/list>.

² <https://www.w3.org/TR/owl-overview/>.

³ <https://www.w3.org/TR/skos-reference/>.

⁴ <https://databus.dbpedia.org/>.

request missing ontologies to be included in the automated runs via a Web interface. These approaches allow Archivo to have a good coverage of meaningful and relevant ontologies of the Semantic Web, while preventing the upload of incorrect ontologies (ontology hijacking or spamming) by users. However, in order to ensure this, Archivo uses a strict technical definition of an ontology: it requires an RDF file that types the resolvable ontology (document) identifier with either `owl:Ontology` or `skos:ConceptScheme`. Note, that this requirement does not exclude RDFS ontologies, since these can be declared as an `owl:Ontology` but use plain RDFS semantics (a prominent example is the RDFS vocabulary itself).

2.3 Linked Open Vocabularies (LOV)

Linked Open Vocabularies [19] is a very prominent semi-automatically curated catalog of vocabularies that hosts snapshots of ontologies and provides an index to search for terms and vocabularies. New vocabularies are discovered by analyzing (re)use of terms from archived ontologies and can be suggested by users, but are subject to manual review and approval procedures. LOV provides an API⁵ for easy access to the archived ontologies. Note, that the definition of an ontology slightly differs and that while Archivo uses the list of ontology identifiers in the LOV catalog, it performs its own automated crawling, access, versioning, monitoring, and approval strategies. As a consequence, there is no full overlap in terms of archived ontologies between the two approaches.

2.4 LOD Vocabulary Usage

In order to gain insight into the vocabulary usage of the Linked Open Data cloud, we utilized the LOD-a-lot HDT dump [4]. It contains more than 28.36 billion triples, 3.17 billion distinct objects, 3.21 billion distinct subjects, and 1,168,932 properties. Over 650,000 datasets are integrated summing up to 524 GB of compressed HDT [7] data. This dump data was crawled and cleaned by LOD-Laundromat [3]. A list of properties was retrieved by filtering the triples for predicates; a list of classes was retrieved by collecting all IRIs that occur in the object position of an `rdf:type` assertion.

3 Analysis

3.1 Ontology Accessibility Study

While it may be quite inconvenient if vocabularies are temporarily unavailable due to server failures, this unavailability leads to anomalies when using datasets built on top of them (e.g. varying or incomplete query results due to temporarily missing subclass axioms). Moreover, completely unreachable ontologies (e.g. due to publishers losing control over the domain) that are likely to be never accessible again, impede the reproducibility of existing work based on it significantly. In

⁵ <https://lov.linkeddata.es/dataset/lov/api>.

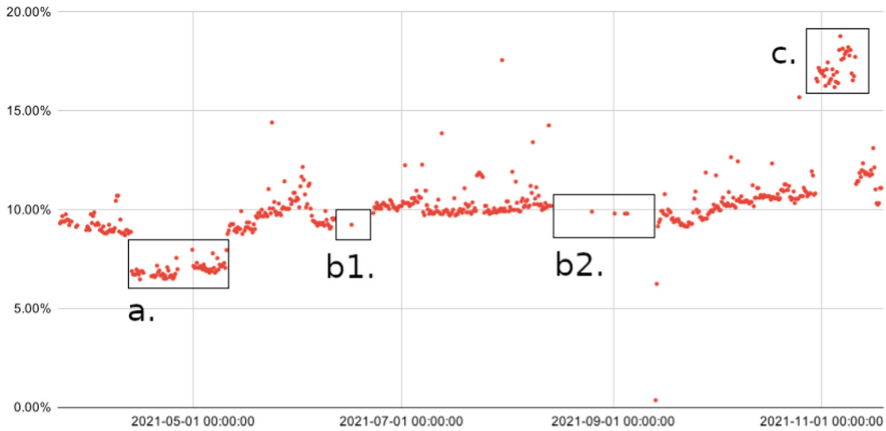


Fig. 2. Fraction of inaccessible ontologies per crawl from March to November 2021.

this first study we want to quantify how many ontologies are affected and how severely they are affected by unavailabilities.

Since Archivio runs multiple checks on every included ontology to potentially fetch new updates, three times a day, the Archivio logs⁶ can be used to measure downtimes and outages of these ontologies. An outage occurs if a `HTTP-HEAD` request or the subsequent `HTTP-GET` request returns a status code ≥ 400 or reaches a timeout (Archivio waits 30s for a response), the host name can not be resolved via the DNS, or if the RDF document is available but does not conform to the respective RDF syntax (i.e. if any error occurs when parsing the document⁷).

Figure 2 shows the outages in relation to the total number of included ontologies for the period of roughly eight months (240 days, from March 23rd to November 18th, 2021). While in average the total outage ratio is around 10%, four areas stand out, as denoted in the diagram:

- a. April 12th - May 10th: the vocabularies hosted on the domain `vocab.deri.ie` were temporarily brought online but since then were unavailable again due to Linked Data configuration failures
- b. June 11th - June 22nd and August 13th - September 14th: The Archivio crawling monitor had issues
- c. October 29th - November 11th: A lot of vocabularies from `purl.org` were not available, but the problem was fixed eventually

Table 1 lists statistics of the downtimes of ontologies, measured over the same time period as Fig. 2. But unlike Fig. 2 it is aggregated per day and not per Archivio-Crawl, i.e. an ontology is considered as “down” for a particular day if it was inaccessible at least for one of the three crawling attempts that day. We

⁶ See <https://github.com/dbpedia/archivio/tree/master/paper-supplement/iswc2022>.

⁷ For this purpose Archivio uses the RaptorRDF library: <https://librdf.org/raptor/>.

excluded the days with crawling gaps from areas b1 and b2 as there is no reliable accessibility data (in total 29 days were excluded). The rows represent statistical values, i.e. *minimum*, *first quartile*, *median*, *third quartile*, *maximum*, *average*, and *total ontology count*. The columns stand for certain subsets of Archivo ontologies: *All onts* stand for the complete set of evaluated ontologies, *all failing* stands for all ontologies that fail at least once and *temp. failing* is the group of ontologies failing at least once, but excluding the vocabularies that fail over the whole monitoring period. The other four columns group the temporarily failing ontologies by downtime fractions, i.e. $[0.01, 5)\%$ is the set of all ontologies being inaccessible 0.01% (included) up to 5% (excluded) of the time since their addition to Archivo.

Table 1. The distribution of downtimes of Archivo ontologies. Columns 1 to 3 group ontologies into failure classes. Columns 4 to 7 break down the temporarily failing ontologies into downtime intervals.

	Failure classes			Temp. failing classes			
	All onts	All failing	Temp. failing	[0.01,5)%	[5,25)%	[25,75)%	[75,100)%
Min	0.00%	0.50%	0.50%	0.50%	5.15%	26.87%	75.12%
Q1	0.00%	1.00%	1.00%	0.50%	6.47%	32.84%	88.56%
Med	0.50%	4.98%	3.72%	1.00%	7.46%	36.32%	88.56%
Q3	5.97%	12.19%	7.96%	1.99%	10.45%	69.40%	89.90%
Max	100.00%	100.00%	99.00%	4.98%	24.88%	74.62%	99.00%
Avg	10.64%	19.67%	12.20%	1.59%	9.17%	47.27%	88.90%
#	1439	775	709	394	224	51	40
% all	100.00%	53.86%	49.27%	27.38%	15.57%	3.54%	2.78%
% tmp	–	–	100.00%	55.57%	31.59%	7.19%	5.64%

Of all ontologies included during the evaluation (1439), Archivo detected no outages for 664 ($\sim 46\%$) ontologies, showing that at least roughly a half of the ontologies are quite well maintained. On the other hand, 66 ($\sim 5\%$) were inaccessible at every day Archivo crawled, which renders a huge problem for datasets depending on them. At least some of them are completely unmaintained and will likely continue to be inaccessible in the future. The rest (709) was inaccessible at least once (but not the whole time) in the time interval. Column 4 to 7 in Table 1 break down these temporarily inaccessible ontologies into smaller bits: more than half of them ($\sim 56\%$) fall into the lowest category of outages (max. $\sim 5\%$ downtime), with an average of 1.59% unavailability. Only 40 ($\sim 6\%$) of the temporarily failing ontologies are in the worst category (inaccessible for more than 75% of measurement).

Overall, as it can be seen in Fig. 2 and Table 1, there is a total average of 10% downtime for all ontologies. This shows the clear need for a backup in form of an archive for ontologies, keeping track of older versions, and making backups of inaccessible ontologies easily accessible for reproducibility.

Table 2. LOD vocabulary term/namespace share.

Filter step	Properties			Classes		
	Terms	t. fract.	Triple fract.	Terms	t. fract.	Triple fract.
NONE	1,168,933	100.00%	100.00%	833,232	100.00%	100.00%
http(s) based	1,163,128	99.50%	100.00%	831,955	99.85%	99.99%
w/o dbr	1,090,550	93.29%	99.36%	785,351	94.25%	99.86%
w/o freebase	1,077,753	92.20%	99.08%	774,755	92.98%	99.57%
w/o dbp	145,820	12.47%	95.50%	–	–	–
w/o DBpYago	–	–	–	291,818	35.02%	98.19%
w/o Wikidata	142,424	12.18%	95.05%	291,555	34.99%	94.28%
w/o RDF-Seq	109,945	9.41%	94.35%	–	–	–
min 10 triples	52,721	4.51%	94.35%	145,870	17.51%	94.27%

3.2 LOD Term Usage Analysis

In a first step, we analyzed the used terminology of the LOD cloud based on LOD-a-lot. We retrieved in total 1,168,933 terms that were used as predicate identifier and 833,232 class identifiers used within instance type assertions (see Table 2).

Although the LOD-a-lot data was subject to LOD-Laundromat cleaning procedures [3], we discovered more than 5,000 irretrievable identifiers that were using a namespace that was not http(s) based. Typical representatives were unexpanded namespace prefixes, file URI schemes, or URN schemes. We consider these types of identifiers as a burden for reproducibility since it is not possible to automatically retrieve the semantics via Linked Data principles. Fortunately, these identifiers make up only less than half a percent of all terms and are neglectable when it comes to the amount of filtered LOD triples affected.

During further investigation of the LOD term lists, we identified more terms and namespaces that affect a meaningful outcome of the coverage study and which we subsequently excluded in cascaded filtering steps and comment potential implication of these properties on the reproducibility. Table 2 reports how many terms remain after each filtering step, as well as the remaining fraction compared to the distinct number of terms and triples respectively.

A well-known error is to use DBpedia entity resource identifiers (namespace prefix *dbr*) as a class reference, but surprisingly also as property identifier. These triples are semantically incorrect and are therefore excluded. In the next step, we additionally exclude Freebase identifiers, because these can be considered as unreproducible, since Freebase did not publish an ontology. Furthermore the project is deprecated and does not serve Linked Data anymore. We discovered more prominent terms that are not captured systematically in an ontology. A huge fraction (almost 80%) for property terms originates from the DBpedia property (*dbp*) namespaces that are produced by the DBpedia Generic extraction [13] for each language version. These properties represent the raw value

of Wikipedia infobox parameters and therefore have no RDF or OWL semantics. The meaning can change over time and depend on the entity type, which significantly affects reproducibility. So-called DBpedia-YAGO class identifiers proxy the YAGO ontology but are neither captured in the DBpedia ontology nor resolvable via Linked Data. This leads to reproducibility problems for more than 57% of the class terms but less than 1.5% for the type statements. We also pruned almost 4% of the Wikidata class assertions since Wikidata’s class hierarchy is not expressed using the common OWL/RDFS axioms and multiple namespaces do not resolve via Linked Data (as of December 2021). As a consequence, in total, at least 87% of property and at least 65% of class terms have issues in capturing the terminology context and semantics in an automatically reproducible way. Fortunately, this only affects less than 6% of the data.

Additional 30 thousand `rdfs:ContainerMembershipProperty`s (e.g. used in RDF sequences) can be excluded, since the semantics is specified in the RDF standard, and this infinite set of properties is not materialized in the RDF(S) ontology. From these over 109 thousand property terms and 291 thousand class terms, we further filtered out all terms that had less than 10 occurrences in LOD. We consider these terms as noise/errors and removing them has an impact of less than 0.01% of ignored triples but cuts more than half the amount of terms from the previous filter step.

The remaining 4.51% resp. 17.51% of terms occur following Zipf’s Law in around 94% of the LOD statements, which ensures that the reduced list of terms still accurately represents a huge and relevant portion of LOD data.

3.3 Reproducibility Support and Archiving Impact Study

Based on the filtered term list we can evaluate how many terms are captured in Archivo and LOV and the amount of LOD data that can be supported in terms of a more robust reproducibility. We loaded the latest ontology snapshot of every ontology contained in Archivo as of April 19th 2022 into a SPARQL endpoint to verify if a term is defined in one of the archived ontologies. The same was done with all archived ontologies of the LOV repository of that time by using its API to fetch the latest version of each vocabulary.

We define a class as any subject that is typed as `rdfs:Class` or as a class that is `rdfs:subclassOf` of `rdfs:Class`. Note that `owl:Class` is a subclass of `rdfs:Class` and therefore OWL classes are included as well. The properties were retrieved in a similar manner, only with the type being either `rdf:Property` or any subclass of it⁸. These terms were then mapped to the frequency counts per term measured in Sect. 3.2.

The results can be seen in Table 3 for properties and Table 4 for classes. Out of the 52,721 property terms, 8.25% (4,350) were archived by Archivo and 9.23% by LOV, which in turn increases the reproducibility robustness for over 44% (almost 12 billion triples) respectively 52% for LOV out of the 26.76 billion triples. In contrast, more than 80% (2.52 billion out of 3.13 billion triples) and

⁸ See the Supplemental Material section at the end of the paper for further details.

74% of the type statements can be supported by Archivo resp. LOV. However, the support boost for individual class terms is on a similar level compared to property terms with approximately 10.82% (15,786 terms) in the case of Archivo but significantly lower with 2.41% in the case of LOV.

Although these numbers indicate that with LOV and Archivo the reproducibility of at least half of the LOD data is given, the effectiveness or impact of archiving as countermeasure is still unclear. All of these covered triples could have an ontological context defined in ontologies that are very reliable, such that the effect of archiving would be negligible at the current stage. In order to study RQ4, we therefore join the term frequency with the ontology accessibility monitoring information (as described in Sect. 2.2) of the ontology that defines the term. Figure 3 shows the impact of archiving ontologies by breaking down the fraction of triples that are covered by Archivo into the different accessibility categories of the ontology where the term is defined. The categories correspond with the ones in Table 1, Note that no data exists about the accessibility over time for ontologies only contained in LOV since the monitoring is a feature of Archivo. As a result this breakdown is only possible for terms that are covered by Archivo. We found that over 54% of these triples have their context in ontologies that did not show any problems in the monitoring time span (cf. Fig 2). However, from the remaining 46%, 15% would lack reproducibility without archiving, since the ontological context is permanently failing. The remaining 31% have temporary failures. These break down into 17% failing very often (75%–99.99% failure downtime), 2% often (25%–74.99%), 9% that fail sometimes (5%–24.99%), and 3% that fail rarely (0.01%–4.99%).

Table 3. LOD Property term coverage and reproducibility support of Archivo and LOV.

	Archivo		LOV	
	Count	Rep. factor	Count	Rep. factor
Terms covered	4,350	8.25%	4,865	9.23%
Studied terms	52,721	–	52,721	–
Triples covered	11,950,908,409	44.66%	14,025,673,856	52.41%
Studied triples	26,760,669,318	–	26,760,669,318	–

3.4 Archiving Potential and Barriers

Although Table 3 and Table 4 show that the fully automated ontology discovery, archiving, and evaluation of Archivo achieves all in all a similar performance for covering LOD terms compared to LOV, we wanted to study what major failure categories prevent an automatic retrieval and archiving of the corresponding ontologies (by Archivo) and whether there is a potential of ontologies that were not discovered yet but could be included. Therefore, we used the term list of the coverage study as input for the discovery and crawling mechanism in an

Table 4. LOD Class term coverage and reproducibility support of Archivio and LOV.

	Archivio		LOV	
	Count	Rep. factor	Count	Rep. factor
Terms covered	17,362	11.90%	3,516	2.41%
Studied terms	145,870	–	145,870	–
Triples covered	2,516,568,507	80.38%	2,322,889,414	74.19%
Studied triples	3,130,912,310	–	3,130,912,310	–

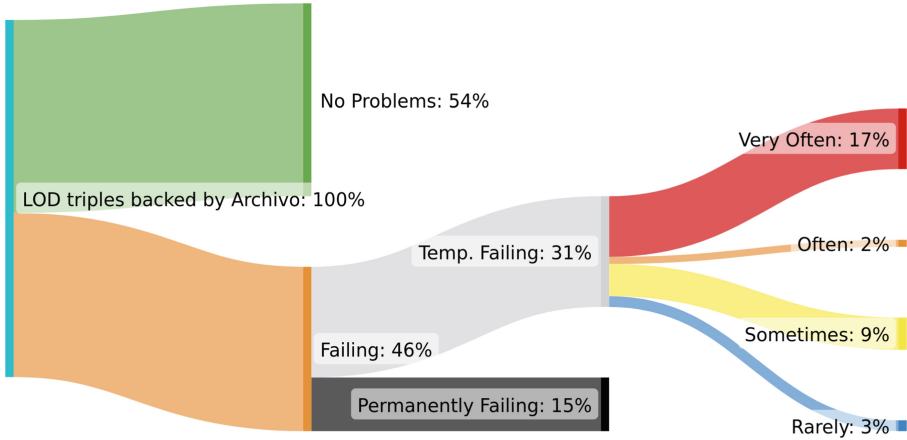


Fig. 3. Archivio Archiving Impact: Breakdown of LOD triples covered by Archivio, into the failure rate of the ontology defining the property/class term.

isolated, temporary Archivio instance. Table 6 and Table 5 show the results. In these tables, multiple reasons are given for ontologies not being accessible to Archivio. The percentage refers to the total number of terms resp. triples noted in Table 4/3. A minor reason for the outage for both classes and properties is that the crawling robot was not allowed to fetch the ontology. While this says nothing about the actual availability of the ontology, it completely prevents the ontology to be archived by Archivio and therefore no stable backup is provided. The by far most prominent reason for retrieval failure was the inaccessibility of any valid RDF at the term IRI. This could be due to link rot, server issues, losing control over the domain of the ontology, or providing unparseable RDF. This is the case for roughly 84% of uncovered properties and 59% of uncovered classes. If any RDF was discovered, the most common error was the missing ontology declaration statement, meaning the retrieved RDF document was not recognizable as an ontology and it does also not link to one. Interestingly, the share for this reason is far higher for classes (38%) than for properties (3%). Rather minor reasons were an error in the linked data deployment (the wrong identifier typed as ontology, or other errors in the RDF) or the ontology is

contained in Archivo, but the term is not defined (usual typos in identifiers or deprecated terms). The last row denotes terms for which an ontology could be found and which could be archived permanently without problems, so these terms may be covered by Archivo in the future.

Table 5. Distribution of reasons for inaccessibility of properties not covered by Archivo. The percentage is based on the total number of terms/triples listed in Table 3.

	# of terms	% terms	# of triples	% triples
Total terms	48,371	91.75%	14,809,760,909	55.34%
Robots disallowed	2,336	4.43%	707,758,083	2.64%
No valid RDF accessible	40,851	77.49%	13,584,634,580	50.76%
Not linked to ontology/not recognizable	1,630	3.09%	112,504,895	0.42%
Ontology LD deployment error	729	1.38%	5,174,428	0.02%
Ontology in Archivo but term not defined	1,208	2.29%	300,318,286	1.12%
Coverable in the future	1,617	3.07%	99,370,637	0.37%

Table 6. Distribution of reasons for inaccessibility of classes not covered by Archivo. The percentage is based on the total number of terms/triples listed in Table 4.

	# of terms	% terms	# of triples	% triples
Total terms	128,508	88.10%	614,343,803	19.62%
Robots disallowed	1,894	1.30%	33,102,294	1.06%
No valid RDF accessible	76,409	52.38%	523,303,990	16.71%
Not linked to ontology/not recognizable	48,914	33.53%	5,093,624	0.16%
Ontology LD deployment error	280	0.19%	318,618	0.01%
Ontology in Archivo but term not defined	304	0.21%	29,854,614	0.95%
Coverable in the future	707	0.48%	22,670,663	0.72%

4 Related Work

Related and previous work can be grouped into three areas: archiving or mirroring of LOD-related data, data availability monitoring, and LOD evolution analysis studies.

Linked Open Vocabularies [19] (as described in Sect. 2.3) is a well-known, extensive cross-domain catalog for ontologies. There are also further efforts to host, archive, version, index, or catalog ontologies and vocabularies like OBO-Foundry [18] and BioPortal [20]. For an in-depth comparison of these approaches we refer the reader to [9].

LOD Laundromat [3] is a tool that crawls and cleans data from the LOD cloud. However, as of December 2021, the service <http://lodlaundromat.org> did not provide any access to the cleaned files anymore for several months and the GitHub page states that it is closed for maintenance since July 2021. Fortunately, a subset of the data is available in **LOD-a-lot** [4], that has been used for this analysis.

OpenLink's **LOD Cloud Cache** data space⁹ is a SPARQL endpoint that gives access to data for a selected subset of the LOD cloud.

The **LOD Cloud**¹⁰ website is a LOD metadata catalog which is also monitoring LOD datasets. The service provides a history for a set of accessibility crawls and the evolution of the catalog. Moreover, there is an effort to preserve LOD data on the IPFS filesystem [16]. The type of data being preserved on IPFS varies from dataset to dataset, ranging from metadata (e.g. VoID summary) to RDF snippets for example entities, but most importantly also ontologies and vocabularies.

The **LODStats** system¹¹ [6] lists 9,960 datasets that are monitored with regard to their accessibility and reports comprehensive statistics about its content. The statistics comprise the access methods to datasets, number of triples, issues when processing the datasets, the usage of classes, properties, datatypes, vocabularies, namespaces and many more. The datasets listed sum up to over 192 million triples, almost 50 thousand properties and 3,480 classes. The service provides insights into the accessibility and structure of the analyzed datasets, and also on the overall linked data cloud and the usage of the ontologies. However, the statistics shown on the website have some inconsistencies (e.g. almost 50 thousand properties overall are reported and the list shows only 32,634) and the project seems not active anymore, since the last update is reported over 6 years ago (as of July 2022). The fact that LOD-a-lot provides more data and access to the triples itself to calculate our own terminology usage statistics, were reasons why we picked LOD-a-lot.

A very simple LOD monitoring service is **LODservatory**¹², which reports the availability and service status of SPARQL interfaces of a list of endpoints (including ones from the LOD cloud) every hour. The **Dynamic Linked Data Observatory** (Dyldo) project [14] performs weekly crawls on Linked Open Data. Based on an IRI seed list it crawls and archives RDF data, subsequently all discovered IRIs are used to perform another crawl, finally the retrieved RDF data, HTTP headers, and redirections are persisted. This process captures also terms from ontologies or could even persist entire ontologies. However, there are no guarantees on completeness for terms and ontologies. Nevertheless, the availability and functioning of the Linked Data mechanisms for particular namespaces can be analyzed over time.

⁹ <http://vos.openlinksw.com/owiki/wiki/VOS/VirtuosLODSampleTutorial>.

¹⁰ <https://lod-cloud.net>.

¹¹ <https://lodstats.aksw.org/>.

¹² <https://github.com/SmartDataAnalytics/lodservatory>.

An analysis on the evolution of vocabulary terms and their impact on the LOD Cloud has been carried out in [1]. The authors investigated to which extent changes in vocabularies were adopted in the evolution of three datasets (the Billion Triples Challenge datasets, the Dynamic Linked Data Observatory dataset, and Wikidata). The results show that the frequency of term changes was rather low, but a huge portion of deprecated terms was still used in the datasets.

To the best of our knowledge, this work is the first effort that specifically studies the accessibility of a huge corpus of ontologies for a longer period of time while also trying to analyze the potential impact of preserving this vocabularies for the LOD cloud to get a better picture of the state of affairs in terms of reproducibility of ontological context.

5 Discussion, Conclusion and Future Work

To conclude, we would like to summarize the results in terms of our research questions. Initially, we gathered reproducibility problems (RQ1) by looking at the namespaces, that are rooted in data or terminology representation itself: term identifiers were not using the HTTP protocol or not formalized with the standards RDFS, OWL or SKOS, formalization was not accessible as dump, or the dump file was not delivered or announced in a way to be accessed via Linked Data when resolving the term and ontology identifiers. Moreover, we discovered a huge portion of proxy identifiers. While it sounds alarming that these issues affected around 88% of the property and 65% of class terms used in LOD-a-lot, it fortunately affected less than 5% of the LOD-a-lot data. We excluded this portion of data from further being used in the studies, since the data or the ontological context modeling needs to be fixed in the first place, in order to be considered a meaningful amount of Linked Open Data.

In RQ2 we measured the problem from two angles. In RQ2a we were looking at the ontologies and in RQ2b at the data affected by the problem through their use of the ontologies. In terms of RQ2a we found that, while 46% of the Archivo-backed ontologies were fully reliable, 5% were permanently inaccessible. 3% of the ontologies were effectively inaccessible (more than 75% downtime) and around 4% were very unreliable (25–75% downtime). For the portion of LOD data, for which the Archivo-backed ontologies provide ontological context, we measured w.r.t. RQ2b that 46% of the statements are affected by accessibility failures of ontologies. 15% of that data is affected by permanently failing ontologies, and 17% by the basically inaccessible ontologies. As a result 32% of data is impacted by ontologies with very severe accessibility issues that make up a fraction of 8% of the backed ontologies. Surprisingly in contrast to that, the ontologies that are failing rarely (56%) only affected 3% of the data.

Based on the reduced and filtered LOD terms list, that excluded terms where we spotted general issues that affect the accessibility and reproducibility beforehand, we found with regard to RQ3b that only 8 to 9% of the property terms are covered, whereas for class terms around 12% are covered by Archivo and 2% by LOV. With the help of the Archivo crawling engine, we measured for over

77% of the property terms and over 52% of class terms that no RDF file could be retrieved (RQ3a). For around 3% of the property terms that are currently not covered by Archivo, we are optimistic that their ontologies can be preserved in future work by feeding them into the discovery mechanism. Additionally, 34% of the class terms are currently inaccessible to Archivo due to its strict protocol requirements. In the future, heuristics and more sophisticated crawling approaches could help here to also include these.

Fortunately in terms of RQ4, having these single digit fractions of terms preserved, covers a significant large amount of LOD triples. Around 50% of the statements are currently having a backup in Archivo or LOV. In the case of Archivo w.r.t. ontology properties for at least 44% of the LOD data and even 80% w.r.t. type assertions. Even more than half of the statements have reproducibility support by LOV for the property. For this portion of backed triples, we have shown that 46% were affected by accessibility issues. When the percentages as shown in Fig. 3 are set into relation to the entire amount of LOD triples in the experiment (i.e. are divided by 2, since roughly half of the triples are covered), this translates into a rough estimate that Archivo could have provided failover for up to 23% of the statements, if data would have been requested at the time of inaccessibility. Subsequently, for roughly $\frac{15}{2}\% + \frac{17}{2}\% = 16\%$ of the LOD triples we effectively consider archiving as an important countermeasure since the ontological context would be not accessible for at least 75% of the time.

We conclude that the archiving approaches presented in this paper provide a foundation to work against the reproducibility crisis. As an approach that builds on top of Archivo and to counter the reproducibility crisis in the future, we plan to implement a transparent proxy tool for reasoners and other semantic tools, that allows reliable and deterministic repeatability and reproducibility of experiments referencing or accessing ontologies (ontology terms), by retrieving the correct, persistent ontology snapshot via Archivo. This approach would allow to fetch data via the original URL, but independent of the data that is actually returned when dereferencing it. Instead, the proxy could serve ontology versions that existed at a specific time span (like a time machine or wayback machine) or could serve as fail-over system if the current deployment of the ontology suffers from availability issues.

Supplemental Material Availability: Source code, scripts, queries and tables are available online. Please refer to <https://purl.org/paper/iswc2022/archivo/material> for further information and guidance.

Acknowledgments. This work was partially supported by grants from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) to the projects LOD-GEOSS (03EI1005E), PLASS (01MD19003D), and CoyPu (01MK21007C).

References

1. Abdel-Qader, M., Scherp, A., Vagliano, I.: Analyzing the evolution of vocabulary terms and their impact on the LOD cloud. In: Gangemi, A., et al. (eds.) *ESWC 2018*. LNCS, vol. 10843, pp. 1–16. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_1
2. Baker, M.: 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016). <https://doi.org/10.1038/533452a>
3. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: a uniform way of publishing other people’s dirty data. In: Mika, P., et al. (eds.) *ISWC 2014*. LNCS, vol. 8796, pp. 213–228. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_14
4. Beek, W., Fernández, J.D., Verborgh, R.: Lod-a-lot: a single-file enabler for data science. In: Hoekstra, R., Faron-Zucker, C., Pellegrini, T., de Boer, V. (eds.) *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017*, Amsterdam, The Netherlands, 11–14 September 2017, pp. 181–184. ACM (2017). <https://doi.org/10.1145/3132218.3132241>
5. Cockburn, A., Dragicevic, P., Besançon, L., Gutwin, C.: Threats of a replication crisis in empirical computer science. *Commun. ACM* **63**, 70–79 (2020). <https://doi.org/10.1145/3360311>
6. 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
7. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). *J. Web Semant.* **19**, 22–41 (2013). <https://doi.org/10.1016/j.websem.2013.01.002>
8. Frey, J., Hellmann, S.: Fair linked data - towards a linked data backbone for users and machines. In: *WWW Companion* (2021). <https://doi.org/10.1145/3442442.3451364>
9. Frey, J., Streitmatter, D., Götz, F., Hellmann, S., Arndt, N.: DBpedia Archivo: a web-scale interface for ontology archiving under consumer-oriented aspects. In: Blomqvist, E., et al. (eds.) *SEMANTICS 2020*. LNCS, vol. 12378, pp. 19–35. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59833-4_2
10. Frey, J., Streitmatter, D., Hellmann, S.: DACOC3 - dbpedia archivo challenging ontology consistency check collection. In: Singh, G., Mutharaju, R., Kapanipathi, P. (eds.) *Proceedings of the Semantic Reasoning Evaluation Challenge (SemREC 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021)*, Virtual Event, 27 October 2021, CEUR Workshop Proceedings, vol. 3123, pp. 32–36. CEUR-WS.org (2021). <https://ceur-ws.org/Vol-3123/paper4.pdf>
11. Gundersen, O.E., Shamsaliei, S., Isdahl, R.: Do machine learning platforms provide out-of-the-box reproducibility? *Future Gener. Comput. Syst.* **126**, 34–47 (2022). <https://doi.org/10.1016/j.future.2021.06.014>
12. Haibe-Kains, B., et al.: Transparency and reproducibility in artificial intelligence. *Nature* **586**(7829), E14–E16 (2020). <https://doi.org/10.1038/s41586-020-2766-y>
13. Hofer, M., Hellmann, S., Dojchinovski, M., Frey, J.: The new dbpedia release cycle: increasing agility and efficiency in knowledge extraction workflows. In: *Semantic Systems* (2020). https://doi.org/10.1007/978-3-030-59833-4_1
14. Käfer, T., Abdelrahman, A., Umbrich, J., O’Byrne, P., Hogan, A.: Observing linked data dynamics. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) *ESWC 2013*. LNCS, vol. 7882, pp. 213–227. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_15

15. Miyakawa, T.: No raw data, no science: another possible source of the reproducibility crisis. *Molec. Brain* **13** (2020). <https://doi.org/10.1186/s13041-020-0552-2>
16. Nasir, J.A., McCrae, J.P.: ilod: interplanetary file system based linked open data cloud. In: Orlandi, F., Graux, D., Vidal, M., Fernández, J.D., Debattista, J. (eds.) *Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with the 19th International Semantic Web Conference (ISWC 2020)*, Virtual event (instead of Athens, Greece), 1 November 2020, *CEUR Workshop Proceedings*, vol. 2821, pp. 27–32. CEUR-WS.org (2020). <https://ceur-ws.org/Vol-2821/paper4.pdf>
17. Pineau, J., et al.: Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *J. Mach. Learn. Res.* **22**(164), 1–20 (2021). <https://jmlr.org/papers/v22/20-303.html>
18. Smith, B., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**(11), 1251–1255 (2007)
19. Vandenbussche, P., Ateazing, G., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. *Semant. Web* **8**(3), 437–452 (2017). <https://doi.org/10.3233/SW-160213>
20. Whetzel, P.L., et al.: Biportal: enhanced functionality via new web services from the NCBO to access and use ontologies in software applications. *Nucl. Acids Res.* **39**, 541–545 (2011). <https://doi.org/10.1093/nar/gkr469>
21. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**(1), 1–9 (2016)