CITYSPIN

Cyber-Physical Social Systems for City-wide Infrastructures

Deliverable 4.2: Data Integration (v.1)

Authors	:	Armin Haller, Javier D. Fernández, Axel Polleres, Maulik R. Kamdar
Dissemination Level	:	Public
Due date of deliverable	:	31.03.2019 (v1)
Actual submission date	:	07.04.2019 (v1)
Work Package	:	4. Scalable Linked Data Integration
Type	:	Report
Version	:	1.0

Abstract

The overall goal of WP4 is to analyse and provide methods for scalable data integration in CPSS systems. In this deliverable we focus on the *Integration and Quality Improvement* component that leverages Linked Data technologies. In particular, we address current findability and accessability issues of Linked Open Datasets, as they constitute one of the main challenges of data integration. This deliverable provides a rigorous characterization of links between Knowledge Graphs on the Web, an extensive empirical analysis to provide systematic data quality assessment and a discussion towards a more sustainable publication of Linked Open Datasets that can help data integration processes, including those in CPSS systems.

FFG

Project Funded by FFG – IKT der Zukunft Programme

Project Number: 861213 **Start date: 01.10.2017** **Duration: 30 months**

History

Version	Date	Reason	Revisited by
0.1	27.02.2019	First skeleton of the work	AH
0.2	1.03.2019	Ideas dump and initial tests	AH, AP, JF
0.3	8.03.2019	Text elaboration, initial setup and tests	AH, JF
0.4	14.03.2019	Experiments	AH, JF
0.5	16.03.2019	Result analysis	AH
0.6	18.03.2019	Text elaboration, state of the art	AH, MK
0.7	19.03.2019	Algorithms	JF
0.8	21.03.2019	New definitions, state of the art, analysis	AP, AH, MK
0.85	24.03.2019	Text polishing and experiments	AH, MK, JF
0.9	26.03.2019	Update definitions, experiments and analysis	AP, AH, JF
0.95	31.03.2019	Text polishing	AP, AH, JF, MK
1.0	07.04.2019	Polish and submit final version	AH, JF

Author List

Project Partner	Name(Initial)	Contact Information
WU Wien	Javier D. Fernández (JF)	javier.fernandez@wu.ac.at
WU Wien	Axel Polleres (AP)	axel.polleres@wu.ac.at
(*) Australian National University	Armin Haller (AH)	armin.haller@anu.edu.au
(*) Stanford University	Maulik R. Kamdar (MK)	maulik@maulik-kamdar.com

(*) Collaborators

Executive Summary

Cyber-physical social systems typically involve data integration at two levels. First, large amounts of internal data must be integrated from heterogeneous, poly-structured data from a variety of sources, ranging from legacy databases to highly dynamic sensor data. Then, external data, such as, weather data or social media stream can enrich existing information in CPSSs. WP4 regards data integration based on Linked Data technologies.

Linked Open Data promises to provide guiding principles to publish interlinked Knowledge Graphs on the Web in the form of findable, accessible, interoperable and reusable datasets, facilitating integration processes. Thus, although Linked Data may be viewed as a basis for instantiating the so-called FAIR principles, there are still a number of open issues that cause significant data quality issues even when Knowledge Graphs are published as Linked Data. Firstly, in order to define boundaries of single coherent Knowledge Graphs within Linked Data, a principled notion of what a dataset is, or, respectively, what links within and between datasets are, has been missing. Secondly, in order to enable FAIR knowledge graphs, Linked Data misses standardised findability and accessability mechanism, via a single entry link.

In order to address the first issue, in this deliverable we (i) propose a rigorous definition of a naming authority for a Linked Data dataset (ii) define different link types for data in Linked datasets, and (iii) provide an empirical analysis of linkage among the datasets of the Linked Open Data cloud. We base our analyses and link computations on a scalable mechanism implemented on top of the HDT format, which allows us to analyse quantity and quality of different link types at scale; additionally, we argue that HDT itself, as a uniform publication format for interchangeable knowledge graphs, would also potentially address findability and accessability issues, among other features such as enabling scalable querying in the form of Linked Data fragments out-of-the-box.

The results of this deliverable will guide our future steps in WP4 towards a scalable Linked Data platform and the proof of concepts in WP7, as they have to integrate internal and external data.

Table of Content

List of Figures

List of Tables

1 Introduction

Efficient and timely integration of high volume, highly dynamic and heterogeneous data sources is a key process in *Cyber-Physical Social Systems* (CPSSs) [37]. One of the goals of **WP4** of the CitySPIN project is to provide a data integration architecture for CPSS based on Linked Data technologies. The final result of such semantic integration is recently known as a *Knowledge Graph*.

While the term *Knowledge Graph* has only surfaced in 2012 in Google's blog, it has since then significantly changed the game not only for Web search engines, but also for data integration within other enterprises and services. Yet, most of the prominent examples termed "knowledge graphs" are *closed* knowledge bases described, for instance, as "intelligent model[s] [...] that understand real-world entities and their relationships to one another" [1]. They have been developed and curated within single enterprises, and are not available to the public, but represent the relevant entities for a particular domain (e.g., common categories and entities relevant for Web search) very well.

In parallel to the rise of the term *Knowledge Graph*, the *FAIR principles* were published in 2016 [39]. In contrast to the above-mentioned current trend to keep valuable knowledge graphs closed, the FAIR principles have been imposed by the scientific research community to claim the importance of improving the Findability, Accessibility, Interoperability, and Reusability of digital assets, with an emphasis on machine-actionability (i.e., the capacity to automate the task to find, access, interoperate, and reuse data). Thus, FAIR Knowledge Graphs should provide all the necessary building blocks for efficient data integration in CPSSs.

Interestingly – since already long before the term *Knowledge Graph* or *FAIR principles* have become popular – these two trends have both been pre-dated by another initiative to set up to publish graph-shaped data assets in an openly accessible manner using standard Web protocols, extended by four simple publishing principles for data, commonly termed under the name *"Linked Data"*, imposed by Tim Berners-Lee in 2006 (with some refinements in 2009) [5]:

- (**LDP1**) use URIs as identifiers for things;
- (**LDP2**) use HTTP URIs so those identifiers can be dereferenced;
- (**LDP3**) return useful information upon dereferencing of those URIs using a standard format (typically, RDF [31]) ; and
- (**LDP4**) include links using externally dereferenceable URIs.

Data publishers from different domains have published numerous datasets following these principles over the past 10 years. Each of these datasets represents a domain-specific knowledge asset, which can be crawled and collected from the Web. The four Linked Data principles, if followed correctly, provide: *i) accessability* through relying on the commonly implemented HTTP protocol, and *ii) interoperability* through relying on a common data format (RDF), out of the box.

While *re-usability* and *findability* are not directly addressed by the Linked Data principles alone, significant efforts have been made to catalog and curate Linked Data assets in the so-called *"Linked Open Data (LOD) Cloud"* [1]. The LOD cloud provides meta-data (e.g. concerning license information, basic descriptive statistics of datasets, and entry

[1] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

links for crawling domain-specific datasets). Thus, we may argue that Linked Data and its principles both capture and combine the idea of both knowledge graphs and FAIR principles: indeed, the Linked Data principles and the Linked Open Data "cloud" have enabled the growth of a network of interlinked graph-structured knowledge bases publicly accessible on the Web.

As such, the LOD cloud can be viewed as a network of open, interconnected knowledge graphs published on the Web, indeed including, for example, DBpedia [3] and Wikidata[2] [12] as the two most widely known and used open knowledge graphs. So, one may ask in how far Linked Data has been successful in establishing a network of FAIR knowledge graphs or why — so far — has it not? In our context, can we really trust on Linked Data for efficient data integration processes in CPSSs?

1.1 Deliverable Goal

The goal of Deliverable D4.2 is to critically and systematically assess the network of knowledge graphs available and accessible as Linked Data, as this constitutes the basis of efficient semantic data integration. In particular, we focus on current findability and accessability issues of Linked Open Datasets (the main handicaps towards FAIR knowledge graphs), and we analyse the most critical quality aspect of a true "network" of open interconnected knowledge graphs: **links**. While links may be considered the greatest strength of Linked Data, they are also it's greatest vulnerability. The following are a few exemplary reasons:

- references to a large number of inaccessible URIs (i.e., broken links may render a dataset largely useless). In some cases, the information (triples) from the "external" dataset can be copied into the local dataset, which in turn creates redundancies as another downside.
- changes in the external dataset to which one links are out of the control of the data publisher.
- publishing datasets as Linked Data does not necessarily keep the dataset in one place. Thus, when crawling Linked Data it is typically hard to determine which links are actually "internal" (i.e., links between parts of one coherent dataset or "knowledge graph"), and which ones are "external" (i.e., links between different datasets).

These issues are aggravated as the sheer notions of "dataset" and "link" are not even clearly defined in RDF or in the Linked Data principles.

This deliverable first sets the boundaries of datasets with a definition of naming authorities, and defines different types of ontology and instance links in Linked datasets, Then, we provide an extensive analysis of links and their quality in the current Linked Open Data cloud. Finally, we propose HDT [13] as a uniform publication format for interchangeable knowledge graphs, which would also potentially address findability and accessability issues. HDT also provides other features such as enabling scalable querying in the form of Linked Data fragments out-of-the-box.

1.1.1 What is *not* in this deliverable

This deliverable does not analyse which concrete Linked Open datasets can be integrated in particular CPSS scenarios. Data integration processes are highly use case dependent,

[2] Although Wikidata is not part of the LOD cloud "diagram" [1] itself as of yet!

being out of scope of this deliverable. Concrete integration examples can be found in the proof of concepts of WP7 (*PoC Technology Stack Implementation and Evaluation*).

1.2 Relation to other Work Packages

WP4 cooperates with most of the other work packages, as it provides the main scalable Linked Data platform for CPSSs. In particular, the data integration processes involved in this deliverable should strictly adhere to and adopt privacy-aware policies developed in WP6. In this context, our privacy-aware Linked Widget platform already considers data integration processes, which must comply with users' policies. In turn, data integration is a key aspect of the different CPSS scenarios considered in the proof of concepts of WP7 (*PoC Technology Stack Implementation and Evaluation*).

1.3 Deliverable Structure

This deliverable is structured as follows. We first motivate the need of formally defining datasets and links in Section 2. We present preliminaries, including definitions of what we mean by datasets and links in Sec. 3. Previous work conducted to analyze the availability, quality, and "linked-ness" of the Linked Open Data (LOD) cloud is discussed in Sec. 4. We then present our methodology for analysing links, including the establishment of the dataset corpus, the ontology corpus, and definitions on a dataset authority and namespace, in Sec. 5. We conclude this section by defining link types. We present results of our computation of links in practice on a corpus of datasets registered in the LOD cloud in Sec. 6. We discuss our observations of this analysis in Sec. 7, before we present our HDT-based vision in Section 8. Finally, Section 9 concludes and discusses future work.

2 The Need to Define the Notion of a 'Dataset' and a 'Link' on the LOD Cloud

What is a dataset? When RDF data is published according to Linked Data principles, there is no notion about the sets of triples which form a dataset, or – in other words – a coherent knowledge graph that taken on its own provides a useful asset of information. In fact, Linked Data datasets published on the Web are often partitioned in several files and made available through Linked Data APIs or are in separate named graphs behind SPARQL endpoints [26], where common practices suggest that single datasets and the URIs "belonging" to these datasets can be referred to by sharing a common *namespace*.

However, this notion of a namespace is typically not tied to a notion of authority, as opposed to the original intention of URIs in the Web architecture, cf. Section 3.2 of RFC3986 [32], which defines authority as an integral part of URIs as follows:

```
URI = scheme ":" [//authority] path ["?"query] ["#"fragment]
```

RFC3986 further states that typically "URI schemes include a hierarchical element for a naming authority so that governance of the namespace defined by the remainder of the URI is delegated to that authority". This notion of a namespace and thereby authority, however, is blurry in RDF: it depends on the RDF serialization, whether the prefix of an identifier determining the namespace is clearly recognizable as such or not,

opposed to XML, for instance, which rather considers identifiers as clearly separated pairs of namespace URIs and qualified names [6]. Authority in HTTP URIs (which are prevalently used for IDs in Linked Data and RDF), is typically determined by the pay-level domain, though there might be argued finer-grained notions, or subdivisions of namespaces including parts of the path or specific sub-domains necessary to determine the authoritative namespace part of a URI.

In this sense, the lack of an explicit notion of namespace and the authority of a namespace for a particular URI makes the question to which dataset a certain URI "belongs" difficult, if not impossible to answer by automated means. A dataset may contain several namespaces and a namespace may be authoritative for several datasets.

While not being one of the Linked Data core principles, best practices have been suggested to solve this issue, by declaring certain namespace prefixes to be authoritatively owned by the dataset within metadata. [25] However, Linked Datasets do not consistently publish these authoritatively owned namespace(s) contained within the dataset. For example, according to Polleres et al. [25] and again validated in our analysis, 53.8% of all datasets in the LOD cloud did not explicitly declare their namespace(s).

The lack of notion for namespace and dataset boundaries leads to several problems. First and foremost, it means that users do not know which data and URIs are authoritatively owned by which dataset, while also not knowing what data is reused and potentially extended from other authoritative sources. We argue that without the notion of authoritative namespaces per dataset, it is impossible to determine clear boundaries between datasets and to analyze links between datasets.

What is a link? In contrast to hyperlinks on the traditional document Web which have a clear direction (from one document to another), links in Linked Data, and as such the LOD cloud, do not have a clear definition. Rather, the notion of a triple in RDF (or an edge in the graph) is often taken synonymously with the notion of a typed link that can be used. For example:

t1: [`dbpedia:Wolfgang_Amadeus_Mozart`, owl:sameAs, `wikidata:Q254`] establishes equality between individuals published under different URIs belonging to different datasets (i.e., Wolfgang Amadeus Mozart entities belonging in DBpedia and WikiData are the same)

t2: [`dbpedia:Wolfgang_Amadeus_Mozart`, rdf:type, `dbpedia_ontology:Person`] denotes that an individual is of a certain type (i.e., Wolfgang Amadeus Mozart was a "Person" as defined by the DBpedia ontology [3])

t3: [`dbpedia:Wolfgang_Amadeus_Mozart`, `foaf:name`, "Wolfgang A. Mozart"@en] denotes the name of an individual (i.e., `dbpedia:Wolfgang_Amadeus_Mozart` has the name Mozart, as defined in the FOAF ontology [7])

Note, however that the direction of the link (i.e., whether the first triple **t1** may be considered a link from DBpedia to Wikidata or vice versa) does not depend on whether the respective triple has a DBpedia or a Wikidata URI in its subject, but rather on the fact in which dataset the triple appears. Also, if we assume that the respective triples were all published within the DBpedia dataset, that we can distinguish different kinds of outlinks, **t1** denotes a link to an individual in another dataset, whereas **t3** actually links to an externally defined ontology.

Although previous works (e.g., Schmachtenberg et al. [30]) have analyzed the number of links between a sample of documents in the LOD cloud and discussed their relative

lack, a formal definition of interlinking and distinction between different types of links has been missing from the literature. Also, links have been considered to be directional in previous work, i.e. a link is between the entity identified by the subject and the entity identified by the object [40]. However, a dataset publisher may reuse an external resource in the subject position of a triple in their dataset. Our definitions and analysis of links herein shall capture and clarify these cases.

To address these issues, we first propose a rigorous definition of a naming authority for a Linked Dataset in this deliverable. We aim to distinguish internal references within the dataset from links to data defined in external datasets. Consequently, we can provide concrete definitions of links between datasets and then define different link types in Linked Datasets. We present automated methods to analyze different link types at scale, and provide an empirical analysis of linkage among the datasets of the Linked Open Data cloud.

3 Preliminaries and Definitions

The lack of a clear definition of what a "dataset" (i.e., a coherent knowledge graph) in Linked Data comprises has already been emphasised as early as 2008. Cyganiak et al. [10] propose a metadata-mechanism, in the form of Semantic Sitemaps to scope and describe the set of actually published files that form a dataset. However, as claimed in Polleres et al. [25], this schema is hardly used consistently across datasets. Therefore, we propagate a definition based rather on intuition, which we will thereafter empirically test in our evaluation below.

Definition 3.1. A *dataset* is a collection of one or more associated RDF graphs, published by a single controlling entity either as single or separate files, or accessible via a common SPARQL endpoint. Given a dataset ds, we denote by G_{ds} the merge of all of its graphs.

Here, when we say "published by a single controlling entity", we mean that a single controlling entity has the right or possibility to take the whole dataset offline and/or change RDF triples in the respective graphs composing the dataset. We further assume that datasets authoritatively control a subset of the mentioned URIs in the dataset, by prefixes.

Definition 3.2. We assume each dataset uses a finite set of *namespaces*,[3] (i.e., URI-prefixes), some of which it controls authoritatively. Given a dataset ds, we denote by NS_{ds} the set of its authoritative namespaces for ds. Moreover, we assume each namespace is authoritatively controlled by at most a single dataset. That is, we assume that $ds_1 \neq ds_2$ implies that $NS_{ds_1} \cap NS_{ds_2} = \emptyset$.

Typical mechanisms for namespace authority is the ownership of a certain pay-level-domain. However, disjoint datasets hosted under the same pay-level domain are possible.[4] Next, with reference to common, established notions of standard use of the OWL and RDF vocabularies and under the assumption that triples with non-standard use of these vocabularies are ignored, we distinguish between different types of URIs, depending on their positions in triples.

[3]While, datasets themselves can be infinite in principle, for instance, the dynamically generated Linked Open Numbers dataset [36].

[4]For example, different Linked Data datasets hosted on Github using `https://github.com/USERNAME/`-prefixed URIs, where the username determined the authority instead of the pay-level-domain.

Definition 3.3 (Non-Standard-use, extending Definition 5.5 of Hogan [16])**.** Let RDF, RDFS, OWL and XSD, with URIs `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, `http://www.w3.org/2000/01/rdf-schema#`, and `http://www.w3.org/2002/07/owl#`, respectively, denote the *reserved* namespaces. Let G_{RDF}, G_{RDFS}, and G_{OWL}, resp., denote the RDF graphs accessible at these URIs, where we write $G_{res} = G_{RDF} \cup G_{RDFS} \cup G_{OWL}$. A *non-standard triple* in any RDF graph other than G_{res} is a triple where:

- a class in G_{res} appears in a position other than as the value of `rdf:type`, or
- a property in G_{res} appears outside of the predicate position.

Assuming a triple with standard vocabulary use, we distinguish class positions, property positions, datatype positions, and instance positions of URIs outside of one of the reserved namespaces as follows:

Definition 3.4. A URI u outside of one of the reserved namespaces in an RDF triple $t = (s, p, o)$ is in a *class position* if

- $s = u \ \wedge \ p \in \{p | (p, \texttt{rdfs:domain}, \texttt{owl:Class}) \in G_{res} \vee (p, \texttt{rdfs:domain}, \texttt{rdfs:Class}) \in G_{res}\}$
- $o = u \ \wedge \ p \in \{p | (p, \texttt{rdfs:range}, \texttt{owl:Class}) \in G_{res} \vee (p, \texttt{rdfs:range}, \texttt{rdfs:Class}) \in G_{res}\}$
- $o = u \ \wedge \ p = \texttt{rdfs:type}$

Definition 3.5. A URI u outside of the reserved namespaces in an RDF triple $t = (s, p, o)$ is in a *property position* if

- $s = u \wedge$
  $p \in \{p | (p, \texttt{rdfs:domain}, \texttt{owl:ObjectProperty}) \in G_{res}\} \cup \{p | (p, \texttt{rdfs:domain}, \texttt{rdf:Property}) \in G_{res}\}$
- $p = u$
- $o = u \wedge$
  $p \in \{p | (p, \texttt{rdfs:range}, \texttt{owl:ObjectProperty}) \in G_{res}\} \cup \{p | (p, \texttt{rdfs:range}, \texttt{rdf:Property}) \in G_{res}\}$

Definition 3.6. A URI u outside of the reserved namespaces in an RDF triple $t = (s, p, o)$ is in a *datatype position* if

- $s = u \ \wedge \ p \in \{p | (p, \texttt{rdfs:domain}, \texttt{rdfs:Datatype}) \in G_{res}\}$
- u occurs as the datatype of a typed literal $o = \text{"}l\text{"}\hat{}\hat{}u$
- $o = u \ \wedge \ p \in \{p | (p, \texttt{rdfs:range}, \texttt{rdfs:Datatype}) \in G_{res}\}$

Definition 3.7. A URI u outside of the reserved namespaces in an RDF triple $t = (s, p, o)$ that is neither in a class, nor property, nor datatype position, is in an *instance* position.

Based on its position we can now distinguish link types for URIs:

Definition 3.8. Let ds_1, ds_2 be datasets. Then, we call triple $t \in G_{ds_1}$ a *link* from ds_1 to ds_2, if t contains a URI u from a namespace in NS_{ds_2}. Depending on the position of u we further distinguish:

- t is called an *instance link*, if u is in an instance position in t.

- t is called an *ontology link*, otherwise, where we further distinguish TBox-Links as follows:
  - t is called a *class link*, if u is in a class position other than the o position of an `rdf:type` triple, i.e., a link to a class from an external dataset in a TBox statement.
  - t is called an *instance typing link*, if u in a the class position $o = p$ of an `rdf:type` triple, i.e., a link from an individual to a class from an external dataset in an ABox statement.
  - t is called a *property link*, if u in in a property position other than p, i.e., a link to a property of an external namespace in a TBox statement.
  - t is called an *instance role link*, if u is in the property position $u = p$, i.e., a link between individuals, referring to a property from an external dataset in an ABox statement.

Finally, if u does not appear in G_{ds_2}, we call t a *broken* link.

Before we analyze how these notions apply to knowledge graphs published as Linked Data datasets "in the wild", let us review related works on "linked-ness" and link quality in the context of Linked Open Data cloud.

4 Related Work

Starting from 2007 onwards, publishers have used Semantic Web technologies, such as RDF, OWL, and SPARQL querying language, to publish and link their datasets on the Web. These datasets may be available as RDF/OWL data dumps and may also be exposed through an interface that enables the users to formulate SPARQL queries (i.e., a SPARQL endpoint).

To keep track of all the sources whose datasets have been published and linked on the Web, the Semantic Web community proposed a starting point of entry for any new user who wishes to use these linked datasets in their research. The `LOD-cloud.net` [1] is this starting point, and different snapshots of the Linked Open Data (LOD) cloud show the growth and evolution of the cloud from 12 linked sources in 2007 (as the first prototype) to more than 1,200 linked sources, as of June 2018, with datasets being published from several different domains, such as the life sciences, geography, economics, politics, and media. Until recently, the LOD cloud diagram at `LOD-cloud.net` has been generated by looking at the linked dataset descriptions and metadata catalogued at the DataHub repository[5]. Several efforts have been undertaken to evaluate the availability, quality, and the "linked" nature of the LOD cloud using a myriad of approaches.

4.1 Availability and Discoverability of Linked Open Data sources

There have been numerous studies that investigate and evaluate the availability and discoverability of the LOD cloud using the list of SPARQL endpoints and RDF data dumps access URIs that are listed on the (now discontinued) DataHub repository (which has been the basis of the creation of the LOD cloud diagram on `LOD-cloud.net`). Vandenbussche et al. [34] found that many of the SPARQL endpoints in the LOD cloud had

[5] `http://old.datahub.io`

issues with availability and only 32.2% were available for more than 95% of the time over a 27 month period between 2013 and 2015. Debattista et al. [11] evaluated the 2014 version of the LOD cloud, and found that out of 569 linked data sources, only around 42% (i.e., 239 sources) had an available linked data access point (i.e., a data dump URI or a SPARQL endpoint). On conducting a preliminary analysis in 2017, Polleres et al. [25] found that while the 2017 version of the LOD cloud had 1,281 sources, only 50% (i.e., 646 sources) had a possible linked data access point. In this deliverable, we demonstrate that the availability of SPARQL endpoints in the LOD Cloud has dropped even further in 2019.

It has to be emphasized again that the LOD cloud diagram is created from the source metadata descriptions from the DataHub repository – thus, not all the metadata entries may have been updated to reflect the current resources and access points, and sources that provide a linked data access point may not even be listed on the DataHub repository, and hence not included in the LOD cloud diagram. To the best of our knowledge, there are no approaches that can evaluate, at scale and without seed URIs, all possible linked data access points available currently on the Web.

4.2 Metadata Representation and Quality

Representation of metadata of a linked dataset (i.e., class and property characteristics, number of instances and assertions, and also the incoming and outgoing links from a dataset) has been a widely-discussed issue within the Semantic Web community. Alexander et al. [2] proposed the Vocabulary of Interlinked Datasets (VoID) specification to achieve this goal. VoID statistics and metrics can be used for SPARQL query federation (i.e., the methodology to process and execute SPARQL queries across multiple sources on the LOD cloud), and some query federation engines, such as SPLENDID [14], support the processing of VoID-annotated metadata. However, Debattista et al. [11] found that most SPARQL endpoints and RDF data dumps, in the current state of the LOD cloud, do not provide the VoID statistics along with the linked dataset. While, Debattista et al. [11] extensively analyzed a small subset of LOD datasets using 27 Linked Data quality metrics (e.g., licensing, provenance, availability, metadata) that are proposed by Zaveri et al. [40], this study did not perform any analysis to detect authoritatively-owned namespaces.

Hogan et al. [19] proposed a set of fourteen guidelines (e.g., dereferenceable and short HTTP URIs, licensing, metadata) to publish good quality linked data on the Web. They evaluate ≈4 million RDF/XML documents constituting of over 1 billion quadruples. Certain guidelines are widely adhered to by data publishers (e.g., HTTP URIs, stable URIs) whereas certain guidelines pertaining to data licensing and human-readable metadata representation are almost always ignored.

Rietveld et al. [27] presented an automated approach to compute metadata statistics of the different datasets in the LOD Laundromat [4], a catalogue of (re)published and cleaned LOD datasets. The LOD Laundromat Meta-Dataset contains provenance annotations and uses de-facto Semantic Web vocabularies (e.g., VoID) for publishing the metadata. However, no analysis has yet been performed to detect authoritatively-owned namespaces across the datasets.

4.3 Authoritative Namespaces and Links Between Linked Datasets

Schmachtenberg et al. [30] crawled the LOD cloud in 2014 with a seed set of URIs and retrieved more than 900 thousand documents describing more than 8 million resources.

They found that only 56% of all datasets in their corpus link to other datasets. The analysis did not determine an authoritative namespace for a dataset to determine the link statistics, but they considered two datasets to be linked if there exists at least one RDF link between resources belonging to both datasets. As such, the number and type of links between datasets were only captured if both resources in the link existed in the corpus. It was observed that owl:sameAs is the most important linking predicate within most linked dataset categories, followed by rdfs:seeAlso. As shown in our analysis (**Section 5.4**), owl:sameAs and rdfs:seeAlso predicates play an insignificant role in the number of links between datasets. We consider all links, even if the resource that is linked to does not exist in our corpus, but is outside the authoritative namespace. Although their analysis did record the predicate type that was used to link, they did not distinguish between ontology links and instance links, whereas our analysis shows that the majority of links are ontology links.

Harth et al. [15] introduces the notion of a naming authority (i.e., a data source with the power to define identifiers of a certain structure). The authors use the PageRank algorithm to assign authority values to data sources based on a naming authority graph, and then propagate the authority values to identifiers referenced in the sources. In this deliverable, we are also interested in a naming authority, more specifically the authoritative namespace of data (i.e., classes, properties, and individuals).

Hogan et al. [18] crawled the LOD cloud in 2010 and analyzed the crawled corpus with ≈ 150 million URIs. The analysis discovered several issues pertaining to the accessibility and dereferenceability of the URIs, lack of structured data retrieved on lookup (**LDP3**), misreported content types, syntax errors, reasoning errors due to ontology hijacking (i.e., new ontologies published on the Web re-defining the semantics of existing concepts resident in other ontologies), misplaced classes or properties, misuse of established OWL and RDFS built-ins, and errors due to use of deprecated URIs. We will showcase that some of these issues are still prevalent in the LOD cloud a decade later.

Hogan et al. [17] later define authoritative sources for ontologies and discuss the problem of ontology hijacking in greater detail. Although we also consider this as bad practice, all links from ontologies to other ontologies are considered in our analysis, that is, we are also interested in links from an ontology that redefines the semantics of classes or properties defined in the authoritative source URI for these corresponding classes or properties.

Butt et al. [8] published a collection of ontologies that was retrieved by crawling a seed set of ontology URIs derived from `prefix.cc`. Several ranking algorithms were used to compute the centrality of concepts within the ontology they were defined in and within the ontology corpus. In this deliverable, we also use a crawl of `prefix.cc` to establish a set of classes and properties and their authoritative namespace.

4.4 Domain-specific Analyses of Life Sciences Linked Open Data

There have been several domain-specific efforts to evaluate the availability, quality, and reuse across linked data sources. Several data and knowledge publishers in biomedical domains have published and linked their sources on the Web [38, 9, 28, 29, 23]. Indeed, several linked biomedical data and knowledge sources (i.e., biomedical ontologies) are present in the current LOD cloud diagram (available at `LOD-cloud.net`), listed under the 'Life Sciences' region. Hu et al. [20] conducted a link analysis on the datasets published by the Bio2RDF project [9] in the LOD cloud. Specifically, they evaluated the links between different Bio2RDF datasets, estimated symmetry and transitivity of links

		% of total	Available	Available as % of total
Total # of Datasets	1359	100%	—	—
SPARQL Endpoint	459	33.5%	125	9.1%
Available Download	890	65.4%	226	16.6%

Table 1: Availability of a Linked Dataset as SPARQL Endpoint or as a Donwloadable RDF Dumps.

between Bio2RDF domain-specific entities (e.g., drugs and genes), and exhaustiveness of different predicates (e.g., owl:sameAs, bio2rdf:x-ref) to link similar entities. While the study offered promising results, with room for improvement, it was only focused on a small set of linked datasets published under the same Bio2RDF project.

Kamdar et al. [22] performed a systematic analysis over heterogeneous biomedical ontologies in BioPortal repository to detect and estimate class reuse (i.e., when a class URI from one ontology is reused in another ontology) and class overlap (i.e., when similar classes are present in different ontologies). The study observed minimal reuse of classes (with the correct URI representation) but high levels of overlap across these biomedical ontologies (e.g., multiple ontologies use different URIs for the class Cardiac Muscle). Kamdar [21] conducted a similar analysis on vocabulary reuse and label mismatch (i.e., when different class or property URIs are used in different linked datasets to model similar information, such as drug–protein target interaction). Moreover, both studies document 'intent for reuse' in data and knowledge publishers. That is, publishers wish to link and reuse to classes, properties, and instances in existing sources, but end up using different and often incorrect URI representations, with faulty namespaces and deprecated versions. While these studies do not rely on the list of endpoints from the DataHub repository, and exhaustively analyze the quality, reuse, and "linked" characteristics of the linked datasets (or ontologies) in the corpus, they are limited in focus (i.e., only life sciences LOD) and require domain-specific knowledge.

Since the LOD cloud diagram is often represented to be the face of the Semantic Web movement, the lack of availability of resources on the Web as well as quality issues (i.e., lack of reuse, intent for reuse, semantic mismatch) have negative implications. If the LOD sources do not have available linked data access points with high availability and quality, then the research and development of Semantic Web-based methods (e.g., query federation) and tools is severely impacted.

5 Methodology

In the following sections, we describe a generic methodology to define and analyze link types in a corpus of Linked Datasets using a set of automated SPARQL queries.

5.1 Establish Dataset Corpus

It has been shown that although the LOD cloud is still growing, albeit at a slow pace, many datasets and SPARQL endpoints that service a dataset registered in the LOD cloud are not available anymore [34]. To establish our corpus we, therefore, first checked for all datasets registered in the LOD cloud[6] if they are still available. That is, we checked if there is either a functioning SPARQL endpoint or at least one usable download file available.

[6] https://lod-cloud.net/lod-data.json

Table 1 shows the statistics of our analysis. As evident from the analysis, only about a quarter (i.e., 25.6%) of all datasets in the LOD cloud still have a functioning SPARQL endpoint or provide a downloadable file. The status of SPARQL endpoints was tested with the same queries proposed in Vandenbussche et al. [34] as shown in **Listing 1**.

Listing 1: Queries used to test the status of SPARQL endpoints on the LOD cloud

```
ASK WHERE {?S ?P ?O .}
SELECT ?S WHERE {?S ?P ?O .} LIMIT 1
```

As we are using several computationally expensive SPARQL queries (i.e., queries that operate on all triples in the graph), we can not use those SPARQL endpoints directly but need to perform the queries locally. We therefore focused our attention on the downloadable datasets and checked the availability of downloaded RDF dumps. 64.38% of all datasets offer some form of downloadable file (i.e., one or many "full_download" and/or "other_download" locations) while the remaining 1.1% of datasets do not provide any data. However, of those that do, only 226 are still available, representing 16.6% of all download URLs. Although this is still more than the 9.1% availability of SPARQL endpoints, it is a first indication of the relatively poor health of the LOD cloud.

Therefore, to increase the size of our corpus we also included historical datasets from the LOD cloud that were cached in the LODLaundromat [4] and provided as a downloadable corpus in HDT by Debattista et al. [11]. The resulting corpus consists of 430 Linked Datasets (i.e. 214 more than currently available in the LOD cloud), each encoded in HDT for a total size of 51 GB (uncompressed 204 GB), with a total number of 3,262,929,887 triples (i.e., $\approx$ 3.3 billion triples).

5.2 Establish ontology corpus

For our link analysis, we distinguish between *Ontology Links* and *Instance Links* as defined in **Section 3**. To distinguish between the two, we first need to establish a corpus of ontologies available and used in the LOD cloud.

Although ontologies typically only consist of terminological axioms T (TBox), they may also include a set of assertional axioms A (ABox). In the latter case, codelists or thesaurological terms can be defined as assertional axioms in an ontology. Contrarily, datasets registered in the LOD cloud typically consist of only assertional axioms. This is confirmed by our analysis, where only three datasets registered in the current LOD cloud are, in fact, ontologies: *i)* `opencyc.org` dataset (an upper level ontology), *ii)* `umbel.org` dataset (an upper ontology mapping and binding exchange layer that defines a large set of supertypes used to map individuals), and *iii)* `onto.beef.org.pl` (an ontology that forms the core of the OntoBeef Domain Thesaurus that is registered as a separate dataset). We excluded these three ontologies from our analysis of Linked Datasets (cf. **Section 6**, but included their axioms in our corpus of classes and properties.

Linked Datasets themselves, however, may also include terminological axioms, either, because an ontology is contained within the dataset, but using a different namespace, or because some new terminological axioms are defined within the same namespace as the assertional axioms in the dataset. Although the latter can be considered bad practise, it is possible, and as our analysis shows, also common (cf. **Section 6.2**).

To distinguish ontologies and their namespace from and within datasets we need to establish a corpus of ontology namespaces and the classes and properties contained within. While registration of an ontology on `prefix.cc` is often regarded as a common best

practice in the Linked Open Data community, this is voluntary. Consequently, it is difficult to establish such a corpus by just looking at those ontologies that are registered on prefix.cc, since many ontologies in the LOD cloud (and on the Web for that matter) may not be registered on prefix.cc. Hence, we use a two-step process to mitigate this situation and establish such a corpus:

Step 1: We crawl all ontology namespaces of prefix.cc and stored each unique class and property contained within those ontologies. This crawl is performed four times over the span of two months and yields a combined unique number of classes and properties as shown in **Table 2**.

# of unique Classes	204,616
# of unique Properties	1,821

Table 2: Ontology Corpus Statistics

Step 2: We also crawl each dataset in our corpus for declared classes and properties. To check for all classes that are declared within a dataset we perform the query shown in **Listing 2**[7]. We then record if all of the declared classes are contained within the prefix.cc corpus.

Listing 2: SPARQL query used to retrieve all classes that are declared within a dataset.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?C WHERE {
    {?C a owl:Class. } UNION
    {?C a rdfs:Class. }
}
```

To retrieve all properties that are declared within a dataset, we follow a similar process and use the query shown in **Listing 3**. We compare the retrieved set of all unique properties with the properties contained within the prefix.cc corpus.

Listing 3: SPARQL query used to retrieve all properties that are declared within a dataset.

```
SELECT DISTINCT ?P WHERE {
    {?P a rdf:Property. } UNION
    {?P a owl:ObjectProperty. } UNION
    {?P a owl:DataTypeProperty. }
}
```

5.3 Establish dataset authority and authoritative namespace

At this point, there is an absence of a central authority and the presence of incomplete metadata. Note that, although the metadata in the old DataHub repository allows for manually defining the namespace of a given dataset, this information is rarely completed and is subject to manual error as described earlier. For instance, 53% of the 1,359 datasets registered in the LOD Cloud have an empty namespace. Thus, we are interested in identifying all used namespaces in the dataset and ontology corpus, and establishing the dataset authority (i.e., responsible) of each namespace. This information allows us to define, in an automatic way, ontology and instance links between datasets (see **Section 5.4**).

[7]Please note that we will define prefixes for SPARQL queries only once in the deliverable

Towards this aim, we first convert all datasets to HDT [13], a well-known RDF compressed format. HDT splits the RDF graph into three main components: *(i)* the Header, providing general metadata of the RDF datasets (publisher and other provenance information, number of triples, etc.), *(ii)* the Dictionary, that assigns and provides a mapping between each term in the RDF graph (URIs, literals and blank nodes) and a numeric identifier, and *(iii)* the Triples, that makes use of the HDT Dictionary to replace and index the original graph of terms with a graph of ids. HDT provides built-in indexes for the Dictionary and Triples components [24] that allow for efficient term and id retrieval in the dictionary, and triple pattern resolution at triple level. In particular, the HDT Dictionary splits terms by roles and lexico-graphically indexes four different subdictionaries:

SO: Shared subject-objects (i.e., all subject terms that also appear in the graph as objects).

S: Unique subjects (i.e., all terms occurring in the subject position that are not objects).

O: Unique objects (i.e., all terms occurring in the object position that are not subjects).

P: Predicates (i.e., all predicates, irrespective whether they also appear as subjects or objects).

Thus, we make use of the HDT Dictionary functionality to efficiently iterate through all different roles (subject, object and predicate) in each RDF dataset and extract all different namespaces in each RDF dataset. This method is shown in **Algorithm 1**. Given that subdictionaries are sorted lexico-graphically, the process is just limited to a series of simple steps such as namespace finding (line 4) and counting (line 5). We then compute the 'relative occurrence' of each namespace in the dataset as the percentage of each namespace over the total terms in the subdictionary (line 9), discounting blank nodes (line 3) if present.

Note that we also disregard those namespaces with a small number of occurrences. In our experiments this threshold was practically set to 50 occurrences.

In order to assign the authoritativeness of each namespace, we then compare all relative occurrence in the corpus. From this, we assign the dataset authority of each namespace, as the dataset with the maximum relative occurrence.

Finally, a namespace that is extensively used in a dataset may be classified as its authoritative namespace. However, we need to consider special cases, where the namespace is in fact an external link to a dataset that might not be present or available in the LOD corpus. For example, an automatic inspection on DBpedia can incorrectly determine that it is the authoritative dataset of the `wikimedia.org` namespace. To minimize this effect in our analysis, we restrict to defining only one authoritative namespace for each Linked Dataset. That is, the namespace that (i) has been assigned as an authoritative namespace of the dataset and (ii) it has the maximum relative occurrence of all authoritative namespaces in the dataset. In order to consider a wider range of URIs, for our further analysis, we only consider the Pay Level Domains (PLD) of the authoritative namespace.

Table 3 shows statistics of the process. In general, our process finds an authoritative namespace for 92% of the datasets (395 out of 430 datasets in our corpus). The missing 8% corresponds to datasets with few triples (less than our minimum threshold) and/or namespaces that are further represented in a different dataset. Note that only

Algorithm 1 Computing the namespaces and their relative percentage of a dataset

Input: `HDT(G)`, the HDT version of an RDF dataset G, and MINOCCS, the minimum number of terms in a namespace

Output: `namespaces`, a map with the relative percentage of each namespace occurring in G, namespaces:$string \rightarrow \{0..1\}$

```
1: namespaces = {}, tempCounter = {}, numSubjectURIs = 0
2: for subject ∈ HDT(G).getSubjects() do
3:   if subject ∉ BlankNodes then
4:     namespace_subj = getNamespace(subject)
5:     tempCounter[namespace_subj]++
6:     numSubjectURIs++
7:   end if
8: end for
9: for (namespace, count) ∈ tempCounter do
10:  if (count >= MINOCCS) then
11:    namespaces[namespace] = (count/numSubjectURIs)
12:  end if
13: end for
14: return namespaces
```

65% of the datasets with authoritative namespace (i.e. 257) had an assigned namespace in the LOD cloud metadata and, from them, only 63% (i.e. 162) exactly correspond with our assigned namespaces[8]. A manual inspection of the remaining 38% reveals different errors in the metadata declaration in the LOD Cloud metadata. For example, the dataset `bbc-music` defines *http://www.bbc.co.uk/music/artist/* as the namespace, while the data actually contains *http://purl.org/ontology/mo/*. In other cases, such as `didactalia`, the dataset includes the VOiD descriptive metadata with a different namespace (e.g. *http://didactalia.net* vs. *http://didactalia.com/*). A similar problem can be found with SPARQL endpoints, which we currently do not crawl, such as in `dbpedia-es`, and can be subject of future work.

# of Datasets in our corpus	430
# of D. with Auth. namespace	395
# of D. with namespace in LOD Cloud metadata	257
# of D. matching Auth. namespace and LOD Cloud metadata	162

Table 3: Authoritative namespace statistics

.

5.4 Link Type Analysis

As of our definitions in **Section 3** we distinguish two general types of links, Ontology (TBox) Links and Instance (ABox) Links. In the following sections, we provide more details on the SPARQL queries that correspond to the different links defined above.[9]

[8]We compare the PLDs of both our authoritative namespace and the LOD cloud metadata.

[9]The detailed statistics per authoritative namespace are published at: https://github.com/arminhaller/LinksInLOD

5.4.1 Ontology (TBox) Links

With the query shown in **Listing 4** that instantiates the definitions from **Section 3**, we retrieve all external classes (i.e., classes using a namespace other than the authoritative namespace) that are not explicitly declared as a class, but are used to *i)* define an instance (i.e., they are used in an assertional axiom), *ii)* define a terminological axiom that either extends a class through a *subclass* or *superclass* relationship, *iii)* define a class' equivalence, disjointedness, unionOf, disjointUnionOf, intersectionOf, complementOf, or "enumeration" kind, *iv)* define the domain or key of a property or range of a property, or *v)* describe a universal or existential object property expression.

Listing 4: SPARQL query used to retrieve all external classes.

```
SELECT DISTINCT ?C WHERE {
    {[] a ?C. } UNION
    {[] rdfs:SubClassOf ?C. } UNION {?C rdfs:SubClassOf []. } UNION
    {?C owl:disjointWith [].} UNION {[] owl:disjointWith ?C.} UNION
    {?C owl:disjointUnionOf [].} UNION
    {?C owl:equivalentClass [].} UNION {[] owl:equivalentClass ?C.} UNION
    {?C owl:intersectionOf [].} UNION
    {?C owl:unionOf [].} UNION
    {[] rdfs:complementOf ?C. } UNION {?C rdfs:complementOf []. }
    {?C owl:oneOf [].} UNION
    {[] rdfs:domain ?C. } UNION
    {[] rdfs:range ?C. } UNION
    {[] rdfs:onClass ?C. } UNION
    {[] owl:allValuesFrom ?C. } UNION
    {[] owl:someValuesFrom ?C. }
    FILTER (!regex(?C, "AUTHORITATIVENAMESPACEURI","i")) .
}
```

For each class URI retrieved through this query, we check its occurrence in either the subject or object position in any triple in the dataset through the query shown in **Listing 5**.

Listing 5: SPARQL query used to determine subject/object position in any triple in a given dataset.

```
SELECT ?C WHERE {
    {[] [] ?C . } UNION
    {?C [] []}
    FILTER (regex(?C, "CLASSURI","i")) .
}
```

The number of resulting triples constitutes the number of *Class Links* in the dataset.

For *Property Links* we follow a similar process. With the query shown in **Listing 6**, we retrieve all external properties (i.e. properties using a namespace other than the authoritative namespace) that are not explicitly declared as a property are but used: *i)* within a subproperty relation, *ii)* within a property chain, *iii)* in a property restriction, or negative property assertion *iv)* to define a properties' equivalence, disjointedness or inverseness with/to another property, or *v)* to define the domain or range of a class.

Listing 6: SPARQL query used to retrieve external properties.

```
SELECT DISTINCT ?P WHERE {
    {?P rdfs:SubPropertyOf []. } UNION {[] rdfs:SubPropertyOf ?P. } UNION
    {?P owl:propertyChainAxiom []. } UNION
    {[] owl:onProperty ?P. } UNION
    {[] owl:assertionProperty ?P. } UNION
```

```
    {?P owl:equivalentProperty []. } UNION {[] owl:equivalentProperty ?P. } UNION
    {?P owl:propertyDisjointWith []. } UNION {[] owl:propertyDisjointWith ?P. }
        UNION
    {?P owl:inverseOf []. } UNION {[] owl:inverseOf ?P. } UNION
    {?P rdfs:domain []. } UNION
    {?P rdfs:range []. }
    FILTER (!regex(?P, "AUTHORITATIVENAMESPACEURI","i")) .
}
```

For each property URI retrieved through this query, we check its occurrence in the predicate position in any triple in the dataset through the query below.

Listing 7: SPARQL query to check position for each property URI in any triple in a given dataset.

```
SELECT ?P
WHERE {
    [] ?P [] .
    FILTER (regex(?P, "PROPERTYURI","i")) .
}
```

The number of resulting triples constitutes the number of *Property Links* in the dataset.

5.4.2 Instance Links (ABox Links)

Before we can compute the number of instance links from an individual in the authoritative namespace to any individual in an external namespace, we first need to find all unique individuals in a dataset.

1. We find all individuals of classes/properties that are declared (i.e., individual that are defined as a type of a class/property).

   Listing 8: SPARQL query to retrieve all individuals defined as a type of a class/property.

   ```
   SELECT DISTINCT ?S WHERE { ?S a ?O. }
   ```

   For each retrieved individual, we check if they are defined in the authoritative namespace. If not, they are counted as an *Instance Typing Link*.

2. We then find all individuals that are reused from a non-authoritative namespace URI in the subject position without being explicitly declared as a type of a class or property. To retrieve those, we first query all triples in the dataset and then check for each unique subject URI that is not in the authoritative namespace, if it is already in the set of declared instances (as of the previous step), or if it is in the set of classes and properties (cf. **Section 5.2**). If it is neither, we count it as an *Instance Link*.

3. We then follow a similar process for each individual reused from a non-authoritative namespace URI in the object position. For each unique object URI, we check the following conditions: *i)* if the subject is not a blank node, *ii)* the subject URI does not contain the authoritative namespace URI, *iii)* the predicate is not an RDF type relation, and *iv)* the object URI is not already contained within the set of declared instances. If none of these conditions are satisfied, we record it as an *Instance Link*.

For each of these *Instance Links*, we also check if they are explicitly using an owl:sameAs, owl:differentFrom, or owl:AllDifferent relation for the link.

Listing 9: SPARQL query to determine the semantics for Instance Links

```
SELECT ?S ?O WHERE {
?S ?P ?O .
FILTER ((?P = owl:sameAs || ?P = owl:differentFrom || ?P = owl:AllDifferent)
    && (!regex(?S, "AUTHORITATIVENAMESPACEURI","i") || (!regex(?O,
    "AUTHORITATIVENAMESPACEURI","i"))
       }
```

6 Computation of Links in Practice

In the following sections, we discuss the results of the analysis of the LOD cloud corpus.

6.1 General characteristics of the LOD corpus

In the first step, we computed general statistics of the datasets in the LOD cloud (cf. **Table 4**). The first observation we can make is that the majority of the datasets are rather small in size, that is, 50% of all datasets have less than 4,478 triples. Although the mean (17,860,436 triples) is much larger, it is skewed by some few much larger datasets (e.g., DBpedia, the Zeitschriftendatenbank database, the WebIsA database, and the catalogue of the German National Library).

On average, the number of subjects is about an order of magnitude smaller than the number of triples, implying that there are on average 10 statements made about each subject. The mean number of unique predicates is interestingly very small — only 31 unique predicates (including RDF(S) and OWL predicates) are used in each dataset. Again, the mean is larger, but is, in fact, largely skewed by just one dataset, namely DBpedia with 68,687 unique predicates. The dataset defining the second most predicates, the B3Kat dataset, has only 3,259 unique predicates. The large and unusual number of predicates in DBpedia can be explained by the automated generation of its triples and a lack of reconciliation of similar properties with slightly different names (labels). Not surprisingly, the average number of unique objects in linked datasets is larger than the number of unique subjects. This is an indication of the existence of links between datasets (i.e., the reuse of objects). However, again the mean is much larger — more than three orders of magnitude larger than the median. This is again due to some few large datasets, in particular, the Zeitschriftendatenbank, DBpedia, the WebIsA database and the catalogue of the German National Library.

	Median	**Mean**
Number of Triples	4,478	17,860,436
Number of Unique Subjects	613	1,774,578
Number of Unique Predicates	31	455
Number of Unique Objects	2,245	5,296,390

Table 4: General statistics of the corpus

6.2 Ontology Links

Our analysis of *Ontology Links* in the corpus revealed some interesting usage patterns of ontologies. However, before we discuss the number of *Ontology Links* we present some general statistics on the use of classes and properties in the LOD cloud, which are shown in **Table 5**:

	Median	**Mean**
Number of Declared Classes	0	52
Number of Undeclared Classes:	7	54
Number of Declared Properties:	0	550
Number of Undeclared Properties:	24	226

Table 5: General statistics on the use of classes and properties in the LOD cloud

Not surprisingly, the median number of declared classes and properties for Linked Datasets is 0. In fact, 67% of all datasets do not declare any classes or properties. In terms of undeclared classes, we can see that 50% of all datasets reuse at least 7 classes, while the average number of reused classes is 54. All, but three datasets, include at least one reused class (which for some datasets is just an owl:Class or rdfs:Class).

We also compared the resulting class URIs for each dataset to the class URIs retrieved from prefix.cc to check how many classes in our corpus are not registered on prefix.cc. Out of 36,970 unique classes used, in total, in our corpus, 20,217 classes are not registered. The low number of registered classes on prefix.cc is quite surprising, given that its corpus includes 204,616 classes. Although we can not determine the individual ontology namespace from a class URI, if we group those class URIs by their Pay Level Domains (PLDs), we end up with only 135 "ontology" PLDs that are not registered with prefix.cc. These PLDs account for the total number of unregistered classes. The top ten of these PLDs are listed in **Table 6**.

Table 6: PLDs with the highest number of unregistered Class URIs

PLD	**# of Class URIs**
dbpedia.org	10,579
sli.uvigo.gal	1,818
semanticscience.org	1,427
purl.org	1020
www.productontology.org	990
purl.obolibrary.org	855
semanticweb.org	809
minsky.gsi.dit.upm.es	714
wikidata.dbpedia.org	455
www.wikidata.org	219

Table 7: PLDs with the highest number of unregistered Property URIs

PLD	**# of Property URIs**
sw.opencyc.org	54,916
umbel.org	27,919
dbpedia.org	10,591
www.orpha.net	6,198
purl.obolibrary.org	5,609
www.ebi.ac.uk	4,712
onto.beef.org.pl	2,369
sli.uvigo.gal	1,818
semanticscience.org	1,429
purl.org	1,239

Some of these point to the use of deprecated or wrong URIs in Linked Datasets. For example, the complete set of class URIs for dbpedia.org is registered with prefix.cc and therefore the class URIs used in linked datasets not registered are either wrong or deprecated. On the other hand, there are no ontologies registered in prefix.cc for the sli.uvigo.gal, the semanticscience.org, and surprisingly, for the www.productontology.org namespace.

Repeating the process for property URIs shows that the ratio of unregistered unique

properties in prefix.cc is much bigger even than for class URIs (cf. Table 7). Namely, out of 142,694 unique properties used, 122,199 are not registered with prefix.cc. In total, these belong to 160 PLDs, few examples of which are listed in **Table 7**. The largest number are from sw.opencyc.org and umbel.org, both of which are not registered in their entirety with prefix.cc.

Table 8 and **Table 9** show the most commonly used class and property URIs (other than RDFS/OWL URIs) in datasets in our corpus, respectively.

Class URI	# datasets
http://rdfs.org/ns/void#Dataset	118
http://rdfs.org/ns/void#Linkset	90
http://xmlns.com/foaf/0.1/Person	74
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Word	65
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Sentence	64
http://www.w3.org/2004/02/skos/core#Concept	56
http://xmlns.com/foaf/0.1/Organization	51
http://vivoweb.org/ontology/core#CoreLaboratory	30
http://vivoweb.org/ontology/core#Center	28
http://xmlns.com/foaf/0.1/Agent	24

Table 8: Number of datasets that use a specific Class URI

Property URI	# datasets
http://purl.org/dc/terms/title	163
http://purl.org/dc/terms/creator	140
http://purl.org/dc/terms/description	134
http://xmlns.com/foaf/0.1/homepage	125
http://purl.org/dc/terms/publisher	112
http://purl.org/dc/terms/subject	105
http://rdfs.org/ns/void#vocabulary	103
http://purl.org/dc/terms/modified	98
http://rdfs.org/ns/void#exampleResource	96
http://rdfs.org/ns/void#subset	88

Table 9: Number of datasets that use a specific Property URI

6.2.1 Class Links

Only a few datasets include *Class Links*, which is not particularly surprising, considering the low number of declared classes in datasets in the corpus. However, 44% of all datasets link to classes outside of the authoritative namespace. This is ≈ 10% points more than datasets declaring classes, which points to the reuse of external classes in terminological axioms. The mean number of *Class Links* with 1,299 triples is largely influenced by the top ranked authoritative namespaces http://vivo.iu.edu with 119,358 links and http://vivo.scripps.edu with 63,128, both more than two orders of magnitude larger than the 10^{th} ranked namespace, http://vivoweb.org with 847 *Class Links*. The distribution of *Class Links* per the size of the dataset is shown in **Figure 1**. The distribution although slightly linearly correlated, shows that a large proportion of authoritative namespaces include between 10 and 1,000 Class Links.

Median:	0
Mean:	1,299
Proportion above 0:	44%

Table 10: Class Links Statistics

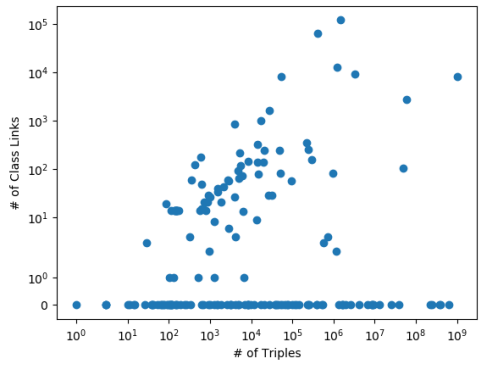

Figure 1: Class Links per # of Triples

http://vivo.iu.edu	119,538
http://vivo.scripps.edu	63,128
http://www.imagesnippets.com	12,874
http://core.kmi.open.ac.uk	9,143
http://commons.wikimedia.org	8,258
http://vivo.psm.edu	8,036
http://datos.bne.es	2,778
http://dbpedia.org	1,614
http://www.productontology.org	1,000
http://vivoweb.org	847

Figure 2: Authoritative namespaces with most Class Links

6.2.2 Property Links

Very few authoritative namespaces use *Property Links* (only 18%). The maximum with 4,995 such links in http://commons.wikimedia.org is more than two orders of magnitudes larger than the 10^{th} ranked namespace, http://tkm.kiom.re.kr with 60. Those authoritative namespaces that include *Property Links* largely use between 10 and 1,000 of those (**Figure 3**). As with *Class Links*, one would expect *Property Links* mostly in ontologies, and therefore the low number of such links in our corpus is, in fact, a positive sign of the reuse of ontologies, rather than the redefinition/extension of properties in the local dataset namespace.

Median:	0
Mean:	47
Proportion above 0:	18%

Table 11: Property Links Statistics

6.2.3 Instance Typing Links

Instance Typing Links are actually the most common link type in the datasets in our corpus with a median of 206 and a mean of 1,967,570 such links per authoritative namespace. All, except eight datasets, use external classes to type individuals in the authoritative namespace. We can also observe a strong linear correlation between the number of triples and the number of instances using external class types (cf. **Figure 5**). Unsurprisingly, the authoritative namespace with the most such links is http://webisa.webdatacommons.org (cf. **Table 6**), as its purpose is to define IsA relations for hypernymy relations extracted from the Common Crawl.

We also analysed the distinct external classes used in such links. The median is a relatively high 11 external class types used and the mean is 108. http://commons.wikimedia.org is the authoritative namespace with the most distinct external classes

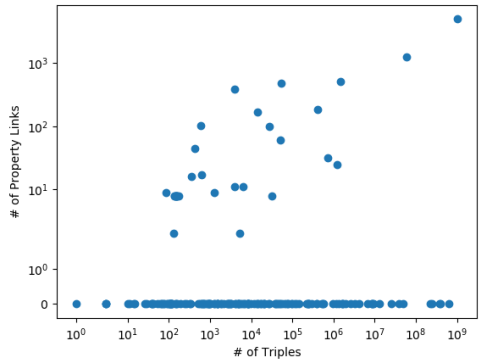

Figure 3: Property Links per # of Triples

http://commons.wikimedia.org	4,995
http://datos.bne.es	1,255
http://vivo.iu.edu	510
http://vivo.psm.edu	481
http://vivoweb.org	386
http://vivo.scripps.edu	187
http://semanticscience.org	168
http://www.iupac.org	102
http://dbpedia.org	101
http://tkm.kiom.re.kr	60

Figure 4: Authoritative Namespaces with most Property Links

Median:	206
Mean:	1,967,570
Proportion above 0:	97%

Table 12: Instance Typing Links Count

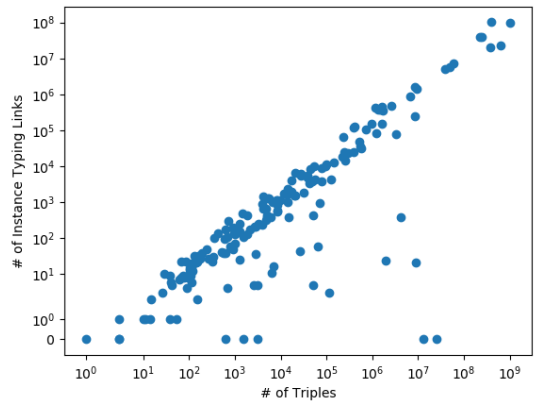

Figure 5: Instance Typing Links per # of Triples

http://webisa.webdatacommons.org	101,491,507
http://commons.wikimedia.org	100,022,186
http://lod.b3kat.de	40,674,519
http://lod.hebis.de	39,160,423
http://d-nb.info	20,096,228
http://datos.bne.es	7,419,630
http://data.ordnancesurvey.co.uk	5,653,997
http://data.europeana.eu	4,987,332
http://id.loc.gov	1,570,877
http://data.bibsys.no	1,440,011

Figure 6: Authoritative namespaces with most `Instance Typing Links`

(mostly from the DBpedia ontology namespace) used for typing individuals with 3,197 in total.

Table 13: Distinct Classes used in Instance Typing Links

Median:	11
Mean:	108
Proportion above 0:	97%

Table 14: Authoritative Namespaces with most distinct classes used in Instance Typing Links

http://commons.wikimedia.org	3,197
http://sli.uvigo.gal	1,830
http://semanticscience.org	1,595

6.2.4 Instance Links

Our analysis of the LOD cloud shows that there are relatively few instance links defined in Linked Datasets. In fact, 28% of all datasets do not include any link from any individual in the authoritative namespace to any other individual in an external namespace, either in the subject or object position. The mean number of links with 1,984,955 is highly skewed by the top two ranked authoritative namespaces which are listed in Table 8, with the median being a mere 24 *Instance Links*, while the 90^{th} percentile is still only 3,863.

Median:	206
Mean:	4,240,890
Proportion above 0:	72%
90^{th} percentile:	3,863%

Table 15: Instance Links Count

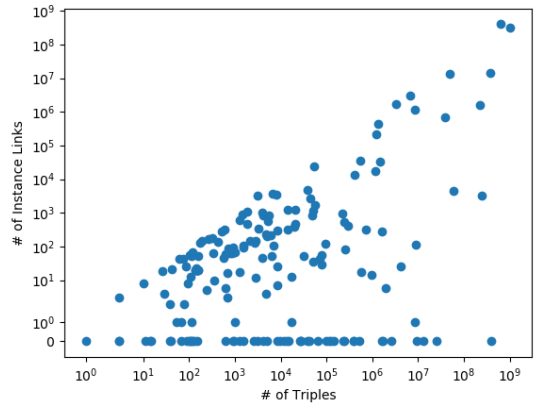

Figure 7: `Instance Links` per # of Triples

http://ld.zdb-services.de	398,381,851
http://commons.wikimedia.org	319,988,690
http://d-nb.info	14,160,649
http://data.ordnancesurvey.co.uk	13,277,718
https://data.gov.cz	3,081,559
http://core.kmi.open.ac.uk	1,696,618
http://lod.hebis.de	1,624,579
http://id.loc.gov	1,143,545
http://data.europeana.eu	687,735
http://spraakbanken.gu.se	451,081
http://www.imagesnippets.com	214,362
http://data.coi.cz	34,277

Figure 8: Authoritative namespaces with most `Instance Links`

As shown, the authoritative namespaces with the most `Instance Links` are http://ld.zdb-services.de and http://commons.wikimedia.org, while the 10^{th} ranked `http://data.coi.cz` already uses four orders of magnitude fewer links. Although there is some slight linear correlation between the size of the dataset and the number of *Instance Links* (cf. **Figure 7**), there is a large cluster of authoritative namespaces that only uses between 10 and 10,000 *Instance Links*.

Looking at some specific predicate types that are used in those links we can see that the often considered popular owl:sameAs link is not particularly widely used. In fact, it is only used in 53% of all datasets, while some few authoritative namespaces, in particular, http://commons.wikimedia.org (linking mostly to http://dbpedia.org/resource)

and some of the authoritative namespaces of the German library community (i.e., http://ld.zdb-services.de, http://d-nb.info, http://lod.b3kat.de and http://lod.hebis.de) account for a large part of the mean number of owl:sameAs links of 503,859. The owl:differentFrom predicate is only used by one authoritative namespace, once again http://commons.wikimedia.org, while owl:allDifferent is not used in any dataset to link an individual in the authoritative namespace to an external individual. The rdfs:seeAlso relation is used slightly more, but it is again http://commons.wikimedia.org that uses it extensively (to link to http://dbpedia.org/resource), whereas the third ranked http://data.nobelprize.org includes only 5,827 *Instance Links* using the rdfs:seeAlso predicate.

	owl:sameAs	owl:DifferentFrom	rdfs:seeAlso	owl:AllDifferent
Median	0	0	0	0
Mean	503,859	581	2,735	0
Proportion > 0	53%	¡1%	14%	0
90th Percentile	1,460	0	1	0
1st	http://commons.wikimedia.org			N/A
1st #	40,636,493	103,439	324,659	
2nd	http://ld.zdb-services.de	N/A	http://stitch.cs.vu.nl	N/A
2nd #	18,049,155	N/A	153,699	N/A
3rd	http://d-nb.info	N/A	http://data.nobelprize.org	N/A
3rd #	17,410,586	N/A	5,827	N/A

Table 16: Selected usage of predicates for linking

6.2.5 Total Number of Links

In **Table 17** some statistics on the total number of links per authoritative namespace are presented. There is a strong linear correlation between the number of triples and the total number of links in the authoritative namespace. However, since the number of *Instance Typing Links* per authoritative namespace is by far the largest, while also showing a strong linear correlation, this result is not surprising. Surprisingly, though, 4% of all authoritative namespaces do not use any link type to an external namespace. The namespaces with the most number of total links are http://ld.zdb-services.de and http://commons.wikimedia.org.

Median:	416
Mean:	6,209,808
Proportion above 0:	96%

Table 17: Total Links Count

7 Discussion

There are several observations from our analysis of the Linked Open Data cloud corpus that are worth discussing.

Ontologies are reused widely: With 36,970 classes and 142,694 properties reused in authoritative namespaces in our corpus, the popularity of ontologies can not be denied. Also, while there is a relative lack of Instance (ABox) links, external classes are used extensively to type individuals in the authoritative namespace of datasets in our corpus. Only a few datasets define their own ontology or extend/narrow the semantics of classes and properties of external ontologies. This is a sign that: *1)* dataset publishers follow best practices and separate the ontology namespace from the authoritative namespace of

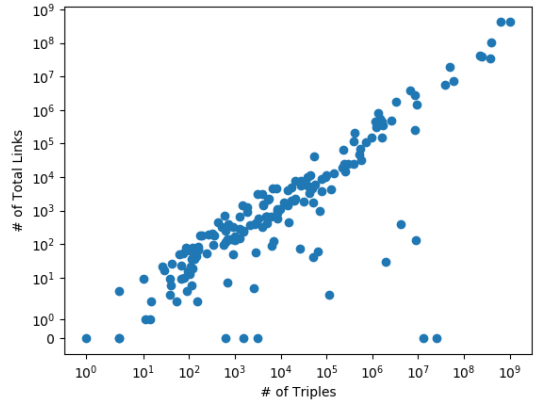

Figure 9: `Total Links` per # of Triples

http://ld.zdb-services.de	421,206,061
http://commons.wikimedia.org	420,024,129
http://webisa.webdatacommons.org	101,491,507
http://lod.hebis.de	40,785,002
http://lod.b3kat.de	40,677,795
http://d-nb.info	34,256,877
http://data.ordnancesurvey.co.uk	18,931,817
http://datos.bne.es	7,428,111
http://data.europeana.eu	5,675,067
https://data.gov.cz	3,958,043

Figure 10: Authoritative namespaces with highest number of links

the dataset, and *2)* it is also a sign that there exists a large number of ontologies that cover already many domains that can be readily reused.

Need for Ontology publishing best practices: As our analysis showed, many ontology namespaces, and as such, their classes and properties are not registered on `prefix.cc`. Even if they are registered, their historical namespace and/or deprecated class and property URIs are often not available anymore. Although there are attempts to establish domain-specific ontology repositories (e.g. BioPortal [38]) and general domain ontology repositories (i.e. the LOV portal [33]), an authoritative ontology register and a persistence mechanism beyond `prefix.cc` is missing. Such a mechanism should assign a DOI to an ontology and persist the document itself in perpetuity (attributes offered by portals such as `zenodo.org`), but also register its authoritative namespace(s), preferred authoritative prefix and resolve its class URIs and property URIs in perpetuity. While the latter are partly covered by using `prefix.cc` in combination with `https://w3id.org` or `http://purl.org`, a repository and mechanism offering all these features is lacking.

Lack of ABox Links: Many (28% of all) datasets do not use any *Instance Links*. Although this number is significantly higher the results reported in earlier work on samples of the Linked Open Data Cloud (i.e., 56%) [30], together with the median number of Instance Links (i.e.) this is still disappointingly low. Furthermore, the authoritative namespaces that actually do use *Instance Links* use mostly other predicates than owl:sameAs relations, that were thought of as the most popular relations for linking [30], while also being the relation that is most useful to reconcile similar individuals in different datasets. The lack of Instance Links can be explained by several factors: 1) these links are expensive to establish manually 2) expensive to maintain and 3) even if they exist, there is no incentive to publish them openly. Evidence that these factors play a large part in explaining the relative lack of Instance Links is the fact that datasets (other than the community-built DBpedia) that do include *Instance Links* are largely from the Humanities discipline where there is a strong community that follows standardised publishing principles and where data is largely historic and static, i.e. once a link is established, it does not have to be updated.

Lack of and incorrect namespace declarations: Only 59% of all datasets in our corpus publish their namespace in the LOD cloud metadata, and of those 257 that do, only 162 match the namespace that we obtain through an analysis of the triples in the graph (cf. Sec. 5.3). Although based on a a rigorous analysis of the triples in a dataset, our algorithm may not always choose the correct authoritative namespace. However, as

discussed in **Section 5.3**, if available, the namespace in the metadata is incorrect in many cases, as there are no guidelines or best-practices what actually constitutes a namespace in a Linked Dataset. With the definitions in this deliverable, we hope to provide both the necessary rigour but also a tool for future data publishers to be able to publish the authoritative namespace of a Linked Dataset.

8 Towards a 5th Linked Data Principle

Running the analyses we did in our deliverable might seem tedious, but we argue that this is mainly due to the heterogeneous publication formats used in linked data: as we have shown, link computations can be done at scale on even large datasets in HDT, and due to the extensible header format of HDT, the respective metadata about links and authoritative namespaces per dataset can be easily published and computed in place at HDT generation time. The respective source code of scripts we used in this deliverable are available at `https://github.com/arminhaller/LinksInLOD`.

We argue that having a this way generated HDT dump with up-to-date link statistics and namespace metadata in place dereferenceable at the namespace URL would solve issues with other publication methods: as shown in prior analysis [34] and again confirmed in this deliverable, for instance SPARQL endpoints are an unreliable access point for Linked Data. Also, for most larger datasets, many typical queries (such as the exploratory queries used in our analysis) time out and as such do not provide a result. HDT [13] as a scalable mechanism to reuse and analyze Linked Datasets published on the Web, can circumvent many of these issues: (1) HDT requires far less ressources than running a SPARQL endpoint for maintainances on the publisher side, with SPARQL clients being readily available; moreover, Triple Pattern Fragments [35] servers are readily available as an interface for HDT, supporting lightweight querying that balances query processing between clients and servers.

We therefore recommend to make a published dataset available as one file in the HDT format, along with the respective meta-data, directly in the HDT header, as a 5th Linked Data principle, extending LDP:

- (**LDP5**) Make your dataset available in HDT and publish dataset namespace authority and link statistics in dataset metadata

Further, for datasets, as is common best practice for ontologies, the authoritative namespace of the data contained within should be published in its metadata. The `void:uriSpace` property offered by the VoID vocabulary is a suitable property to do so. Link statistics that so far could have been provided manually using `void:Linksets` can now be computed directly using the HDT link analysis script developed here and published at `https://github.com/arminhaller/LinksInLOD`.

Together, HDT and the namespace authority and link analysis annotations published/linked from the namespace URI provide a simple and effective publishing principle that enables a more findable/accessible and interoperable way of publishing interlinked knowledge graphs.

9 Summary and Future Work

Efficient data integration (from internal and external heterogeneous data sources) is a fundamental but challenging requirement in the context of Cyber-Physical Social Systems (CPSSs). This deliverable focuses on critically and systematically assessing the network of knowledge graphs available and accessible as Linked Data, in terms of analysing the most critical quality aspect of a true "network" of open interconnected knowledge graphs: **links**. We first proposed a rigorous definition of a naming authority for a Linked Dataset. This definition of an authoritative namespace allows us to distinguish internal references within a dataset from links to data defined in an external namespace. Consequently, we provided concrete definitions of links between datasets, distinguishing between *Ontology (TBox) Links* and *Instance (ABox) Links*.

We presented automated methods to analyse different link types at scale, and provided an empirical analysis of linkage among the datasets of the Linked Open Data (LOD) cloud. For this analysis we established a corpus of classes and properties defined within the corpus and within ontologies registered on `prefix.cc`. This ontology corpus allowed us to distinguish TBox links from ABox links.

The analysis of a corpus of 430 datasets from the LOD cloud showed that almost all datasets use external ontologies for the typing of individuals, i.e. *Instance Typing Links*, while links on the data level, i.e. *Instance Links*, are relatively sparse, with a median number of such links of 206 per authoritative namespace. Also, only 72% of all authoritative namespaces include links to other individuals at all, either in the subject or object position of a triple. The previously thought to be popular of owl:sameAs relations are, in fact, only used in 53% of all datasets. Although this low number and quality of links between datasets on the ABox level is concerning and somehow undermines the idea of Linked Data, the number and quality of links on the TBox level is promising. It shows a strong propensity of reuse of classes and properties defined in ontologies on the Web.

To better enable the reuse of data and to be able to link to existing resources, we propose an addition to the four Linked Data principles, i.e. "Make your dataset available in HDT and publish dataset namespace authority and link statistics in dataset metadata". Only if users of a dataset are 1) able to download and manipulate the file in an efficient matter and 2) know which data is authoritatively owned and managed in the dataset, can they make informed decisions and potentially reuse the data. Our ongoing work focuses on applying our results to the Linked Widget platform data and the integration processes of the proof of concepts of WP7 (*PoC Technology Stack Implementation and Evaluation*).

References

[1] Andrejs Abele, John P McCrae, Paul Buitelaar, Anja Jentzsch, and Richard Cyganiak. Linking open data cloud diagram 2017. Url: http://lod-cloud.net (accessed: 31.12.2018), Insight-Centre, 2017.

[2] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3c interest group note 03 march 2011, W3C, 2011.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: a nucleus for a web of open data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 722–735. LNCS, Busan, South Korea, 2007.

[4] 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 *Proceedings of the International Semantic Web Conference (ISWC)*, pages 213–228, Riva del Garda, Italy, 2014. LNCS.

[5] Tim Berners-Lee. Linked Data. W3C Design Issues, July 2006. From http://www.w3.org/DesignIssues/LinkedData.html; (Accessed: 27.10.2010.

[6] Tim Bray, Dave Hollander, Andrew Layman, Richard Tobin, and Henry S. Thompson. Namespaces in xml 1.0 (third edition), December 2009. Available at https://www.w3.org/TR/xml-names/.

[7] Dan Brickley and Libby Miller. Foaf vocabulary specification 0.91, 2007.

[8] Anila Sahar Butt, Armin Haller, and Lexing Xie. Ontology search: An empirical evaluation. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 130–147, Riva del Garda, Italy, 2014. LNCS.

[9] Alison Callahan et al. Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, pages 200–212. LNCS, Montpellier, France, 2013.

[10] Richard Cyganiak, Holger Stenzhorn, Renaud Delbru, Stefan Decker, and Giovanni Tummarello. Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Canary Islands, Spain, June 2008. LNCS. doi: 10.1007/978-3-540-68234-9_50.

[11] Jeremy Debattista, Christoph Lange, Sören Auer, and Dominic Cortis. Evaluating the quality of the lod cloud: An empirical investigation. *Semantic Web*, (Preprint): 1–43, 2017.

[12] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing wikidata to the linked data web. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 50–65, Riva del Garda, Italy, 2014. LNCS.

[13] Javier D Fernández, Miguel A Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary rdf representation for publication and exchange (hdt). *Web Semantics: Science, Services and Agents on the World Wide Web*, 19:22–41, 2013.

[14] Olaf Görlitz and Steffen Staab. Splendid: Sparql endpoint federation exploiting void descriptions. In *Proceedings of the International Workshop on Consuming Linked Data, in conjunction with International Semantic Web Conference (ISWC)*, pages 13–24, Bonn, Germany, 2011. CEUR-WS.org.

[15] Andreas Harth, Sheila Kinsella, and Stefan Decker. Using naming authority to rank data and ontologies for web search. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *Proceedings of the International Semantic Web Conference (ISWC)*, pages 277–292, Washington, DC., USA, 2009. LNCS. ISBN 978-3-642-04930-9.

[16] Aidan Hogan. *Exploiting RDFS and OWL for Integrating Heterogeneous, Large-Scale, Linked Data Corpora*. PhD thesis, 2011.

[17] Aidan Hogan, Andreas Harth, and Axel Polleres. Saor: Authoritative reasoning for the web. In John Domingue and Chutiporn Anutariya, editors, *Proceedings of the International Semantic Web Conference (ISWC)*, pages 76–90, Karlsruhe, Germany, 2008. LNCS.

[18] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW*, volume 628 of *CEUR Workshop Proceedings*, Raleigh, USA, 2010. CEUR-WS.org. URL http://ceur-ws.org/Vol-628/ldow2010_paper04.pdf.

[19] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *Journal of Web Semantics*, 14:14 – 44, 2012. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2012.02.001. URL http://www.sciencedirect.com/science/article/pii/S1570826812000352. Special Issue on Dealing with the Messiness of the Web of Data.

[20] Wei Hu, Honglei Qiu, and Michel Dumontier. Link analysis of life science linked data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 446–462, Bethlehem, PA, USA, 2015. LNCS.

[21] Maulik R. Kamdar. *A web-based integration framework over heterogeneous biomedical data and knowledge sources*. PhD thesis, Stanford University, 2019. URL https://purl.stanford.edu/jr863br2478.

[22] Maulik R Kamdar, Tania Tudorache, and Mark A Musen. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic web*, 8(6): 853–871, 2017.

[23] Maulik R. Kamdar et al. PhLeGrA: Graph analytics in pharmacology over the web of life sciences linked open data. In *Proceedings of the World Wide Web Conference (WWW)*, Perth, Australia, 2017.

[24] Miguel A Martínez-Prieto, Mario Arias Gallego, and Javier D Fernández. Exchange and consumption of huge rdf data. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, pages 437–452, Heraklion, Crete, Greece, 2012. LNCS.

[25] Axel Polleres, Maulik R. Kamdar, Javier D. Fernández, Tania Tudorache, and Mark A. Musen. A more decentralized vision for linked data. In *Proceedings of the 2nd Workshop on Decentralizing the Semantic Web, co-located with the International Semantic Web Conference (ISWC), DeSemWebISWC*, Monterey, CA, USA, 2018. CEUR-WS.org.

[26] Bastian Quilitz and Ulf Leser. Querying distributed rdf data sources with sparql. In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 524–538, Tenerife, Canary Islands, Spain, 2008. LNCS.

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

[28] Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, et al. Advancing translational research with the semantic web. *BMC bioinformatics*, 8(3): S2, 2007.

[29] Matthias Samwald, Anja Jentzsch, Christopher Bouton, Claus Stie Kallesøe, Egon Willighagen, Janos Hajagos, M Scott Marshall, Eric Prud'hommeaux, Oktie Hassanzadeh, Elgar Pichler, et al. Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics*, 3(1):19, 2011.

[30] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 245–260, Riva del Garda, Italy, 2014. LNCS. ISBN 978-3-319-11963-2. doi: 10.1007/978-3-319-11964-9_16.

[31] Guus Schreiber and Yves Raimond. RDF 1.1 primer. W3C Note, June 2014.

[32] Larry Masinter Tim Berners-Lee, Roy Fielding. Uniform resource identifier (URI): Generic syntax. IETF Network Working Group Request for Comments: 3986 (RFC3986), 2005. Available at https://tools.ietf.org/html/rfc3986.

[33] Pierre-Yves Vandenbussche, Ghislain Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452, 2017. doi: 10.3233/SW-160213. URL https://doi.org/10.3233/SW-160213.

[34] Pierre-Yves Vandenbussche, Jürgen Umbrich, Luca Matteis, Aidan Hogan, and Carlos Buil Aranda. SPARQLES: monitoring public SPARQL endpoints. *Semantic Web*, 8(6):1049–1065, 2017. doi: 10.3233/SW-170254. URL https://doi.org/10.3233/SW-170254.

[35] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. Triple pattern fragments: A low-cost knowledge graph interface for the web. *Journal of Web Semantics*, 37-38:184–206, 2016. doi: 10.1016/j.websem.2016.03.003. URL https://doi.org/10.1016/j.websem.2016.03.003.

[36] Denny Vrandecíc, Markus Krötzsch, Sebastian Rudolph, and Uta Lösch. Leveraging non-lexical knowledge for the linked open data web. *Review of April Fool's day Transactions (RAFT'2010)*, 5, 2010.

[37] Fei-Yue Wang. The emergence of intelligent enterprises: From cps to cpss. *IEEE Intelligent Systems*, 25(4):85–88, 2010.

[38] Patricia L Whetzel et al. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545, 2011. DOI:10.1093/nar/gkr469.

[39] 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, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[40] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016. doi: 10.3233/SW-150175. URL https://doi.org/10.3233/SW-150175.