A More Decentralized Vision for Linked Data

Axel Polleres [a,b,*], Maulik Rajendra Kamdar [c], Javier David Fernández [a,b], Tania Tudorache [c], Mark Alan Musen [c]

[a] *Institute for Information Business, Vienna University of Economics and Business, Vienna, Austria*
[b] *Complexity Science Hub Vienna, Vienna, Austria*
E-mails: axel.polleres@wu.ac.at, javier.fernandez@wu.ac.at
[c] *Center for Biomedical Informatics Research, Stanford University, CA, USA*
E-mails: maulik@maulik-kamdar.com, tudorache@stanford.edu, musen@stanford.edu

Editors: Pascal Hitzler, Kansas State University, Manhattan, KS, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited reviews: Anna Lisa Gentile, IBM Research, CA, USA; Jens Lehmann, University of Bonn, Germany

Abstract. In this *deliberately provocative* position paper, we claim that more than ten years into Linked Data there are still (too?) many unresolved challenges towards arriving at a truly machine-readable *and* decentralized Web of data. We take a deeper look at key challenges in usage and adoption of Linked Data from the ever-present "LOD cloud" diagram.[1] Herein, we try to highlight and exemplify both key technical and non-technical challenges to the success of LOD, and we outline potential solution strategies. We hope that this paper will serve as a discussion basis for a fresh start towards more actionable, truly decentralized Linked Data, and as a call to the community to join forces.

Keywords: Linked Data, Decentralization, Semantic Web

1. Decentralization Myths on the Semantic Web

Let us start with a rant, arguing that the Semantic Web may well be considered a story of failed promises with regards to decentralization:

- We had hopes (as a community) to revolutionize Social Networks in a way that every data subject owns and controls their social network data in **decentralized FOAF** [13] files published in their personal Web space – we got siloed, centralized social networks (Facebook, LinkedIn). Attempts to re-decentralize the Social Web, for instance, through the work of the W3C Social Web WG[2] appear not to have found major adoption at a level comparable with these siloed sites.[3]
- We envisioned a **decentralized network of ontologies on the Web** that would enable smart agents to seamlessly talk to each other, and that would enable easy integration of data by following the guiding principles of ontology engineering and Gruber's often cited vision of ontologies as shared conceptualizations [25].[4] While there are indeed certain areas in which ontolo-

[*] Corresponding author. E-mail: axel.polleres@wu.ac.at.

[1] The Linking Open Data cloud diagram, available at http://lod-cloud.net/, which has been regularly updated since 2007 by Andreas Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak, with its latest version having been created in March 2019 [1].

[2] https://www.w3.org/wiki/Socialwg

[3] While ActivePub has been picked up by several implementations, cf. https://en.wikipedia.org/wiki/ActivityPub#Implementations, reversing the network effects that have drawn a critical mass of users to these siloed sites seems still far away.

[4] Or, as Dan Brickley, one of the inventors of FOAF stated slightly sarcastically in personal communication: "we took one useful feature of RDF/RDFS (fine grained vocabulary composition) and elevated it to a cult-like holy law, to the extent that anyone who created a useful RDF vocabulary and wanted to keep improving it, found themselves pushed instead into combining it with dozens of other half-finished, poorly documented efforts that weren't really designed to fit together nicely."

1570-0844/19/$35.00 © 2019 – IOS Press and the authors. All rights reserved

gies are used to share conceptualizations of a domain, mostly these are insular efforts that do the job well for a particular community. However, on Web scale, ontology and vocabulary reuse is still limited or tied to inherent challenges, see also [26, 27]. Rather, we see a rise of main *central* schema (schema.org), and fast-growing community projects like Wikidata [50] refusing to buy into the need for re-using terms from other ontologies.[5]

- We put a lot of effort into **formal semantics and clean axiomatization** of those ontologies – we got logical inconsistency.[6] Whereas, serious attempts to apply such reasoning about Web Data in the wild have either had to restrict themselves to lightweight ontologies or have not been further developed in the past five years, with *(a)* the semantics of OWL [45] and even parts of RDF(S) [12] turning out to be too hard to grasp for normal Web users and developers to survive in the World Wild Web [24, 36]; and *(b)* the DL community mostly having turned their back to seriously taking the challenge of decentralized applications at Web scale.
- Berners-Lee et al. in their original Semantic Web article [7] promised **Web-scale automation**: automated calendar synchronisation, personalised health care assistance, home automation – some of these applications are a reality now (Amazon Alexa's home control, or Google's and Apple's widely used services), but rather than relying on a decentralized Semantic Web, use single companies' curated knowledge bases – also now called "Knowledge Graphs" – that enhance these companies' services' backend systems.
- More specifically, we see **knowledge graphs** evolve and embrace them as a success story of the Semantic Web. Yet a good definition of what a Knowledge Graph is and what differentiates it from an "ontology" is still to be provided – apart from the single distinguishing feature that all known examples of knowledge graphs (Google's, Bing's, and Yahoo's knowledge graphs as well as their open pendants DBpedia and Wikidata) are NOT decentralized.

So, here we are after more than ten years. However, there is one lighthouse project that clearly has implemented the vision of a decentralized Semantic Web. This single project that we, as a community, hinge upon and tend to accept as a clear success to wipe away all the failed promises mentioned above is: Linked Data [9]. The promise to be able to publish structured data in a truly decentralized fashion, with a couple of simple principles to enable the automatic retrieval and integration by just "following your nose" (i.e., dereferencing HTTP links). This principle is the most powerful promise that filled the community with new enthusiasm through the so-called "LOD cloud", cf. Fig. 1. If we measure the number of datasets published according to the *four Linked Data principles* [6] and that link to each other, we find evidence of growth and prosperity (cf. Fig. 2), and hope to finally make the vision of a decentralized Web of data come true. Meanwhile, indeed this "cloud" contains over 1,184 datasets, which should be considered good news.

However, as we will discuss in the present paper, there are still serious barriers to consume and use Linked Data from the "cloud", wherein we have to question the usefulness of the current LOD cloud. Thus, we would like to take a step back and assess the situation. We call for a more principled and, in our opinion, more useful restart and for more collaboration and decentralization in the community itself.

Along these lines, in the remainder of this paper, we start with some background on the genesis of the current LOD cloud in Section 2. We will then highlight five perceived main challenges we deem important to be addressed to make Linked Data more usable and, therefore, useful. These challenges will be presented – by examples and discussing their implications – in the course of Section 3. Finally, we conclude with a call to collaboratively and openly address these challenges as a community in order to (re-)decentralize the Semantic Web again in Section 4.

2. Background: The genesis of the LOD Cloud

The creation of a complete Web index is a never-ending story. Since the early days of the Linked Data Web, several attempts have been created and failed to sustain exhaustive Linked Data Search en-

[5] The main reason for Wikidata not to prescribe existing vocabularies was to leave the community freedom to link and use what they deem useful within one consistent scheme/namespace: one of the reasons was to avoid the needed buy-in to existing ontologies the popularity of which or agreement about could shift over time. Therefore, they "left it to the community whether to choose stronger semantics (e.g., OWL) or weaker semantics (e.g., SKOS [37]) or not" (personal communication Denny Vrandečić).

[6] Even within DBpedia [8, 40], the central crystallization point of the LOD cloud [10].

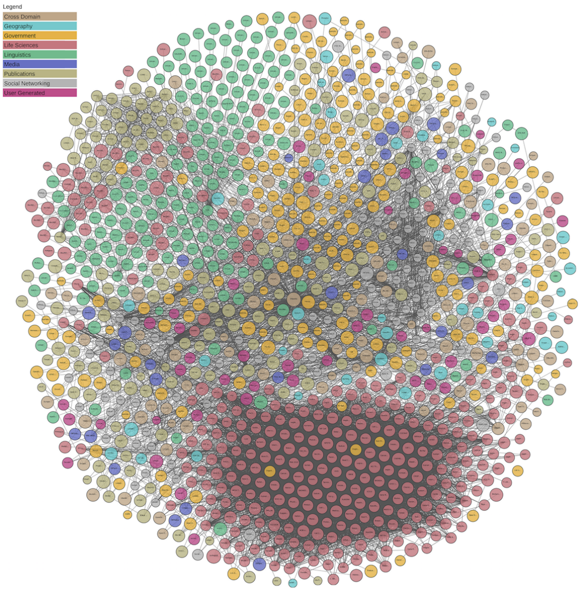

Figure 1. The “LOD cloud” diagram [1], from April 2018, counting 1,184 datasets.

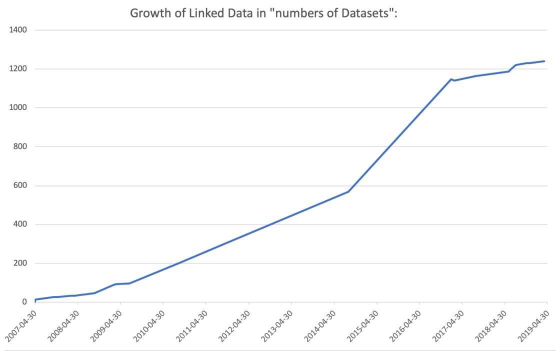

Figure 2. The growth of the “LOD cloud” in number of datasets seems to indicate steady, while not rapid or even overwhelming adoption; we still have to view this as opposed to the probably much more rapid growth of other parts of the Web [41]

.

gines, such as Sindice [39], SWSE [30], Watson [17], Swoogle [20], just to name a few. Typically based on bespoke, crawler-based architectures, these search engines relied on either *(i)* collecting data published under the Linked Data principles and particularly applying the “follow-your-nose” approach enabled through these principles (i.e., find more Linked Data by dereferencing links appearing in Linked Data), and sometimes *(ii)* relying on “registry” or “pingback” services to collect and advertise linked data assets, such as Semantic Pingback [46]. In the meantime, unfortunately all of these search engines have been discontinued, and we are not aware of any active, public Semantic Pingback services. Recently, the LOD-Laundromat project [5] offers an URL lookup service[7] generated from/for the (accessible parts of) the LOD cloud. Moreover, the LOD Cache by OpenLinkSW[8] remains available for LOD entity lookups and SPARQL queries, although it does not provide a detailed specification of which datasets it indexes.

Both of these more recent efforts though claim to refer to datasets in the LOD cloud. The LOD cloud diagram [1], which took a different approach, has been generated from metadata provided by the community at a (CKAN-driven) Open Data portal, namely http://datahub.io. Interestingly, this is only confirmed for its prior version in August 2017, as references to datahub.io have been removed from current, later LOD diagram versions. Also note that datahub.io has moved in the meantime and the "old" LOD-cloud dataset metadata descriptions are only available via the suggestively “deprecated” URL https://old.datahub.io/. While the LOD-cloud initiative itself seems to have been suffering from starvation as well, the current noble effort is depending on a few individuals, such as Abele et al. [1] (the creators and maintainers of the LOD cloud diagram), which seems to put the initiative at risk. At least, there is recent active development, with regular bimonthly updates on the lod-cloud.net between April 2018 and April 2019.

Still, the LOD-cloud at lod-cloud.net and the metadata at datahub.io seem to remain the single most popular entry points to Semantic Web data (with the exception of domain-specific portals such as BioPortal [44]), and therefore a bottleneck.

The metadata the LOD cloud relies on and comprises of metadata fields such as:[9]

- **tags**, where as a pre-filter, only those datasets are included in the cloud that have the tag “lod”,
- **link descriptions**, i.e. declarations of numbers of links to other datasets,
- **resources**, that is, URLs to access the dataset in the form of e.g. dumps, as SPARQL endpoints,

[7] http://lotus.lodlaundromat.org/

[8] lod.openlinksw.com/

[9] Disclaimer: Note that our observations base on metadata from datahub.io collected in April 2018; since then, lod-cloud.net has discontinued on datahub.io and now provides an own form-based submission system for metadata on its Web page.

or semantic descriptions (e.g. in the form of a Void [2] descriptions) or an XML sitemap.

Apart from the LOD cloud, a similar effort exists to collect and register Linked Data *vocabularies* and document their interconnections in the Linked Open Vocabularies (LOV) project by Vandenbussche et al. [47]. As opposed to the purely metadata based approach of the LOD cloud collection, LOV relies on curation and quality checks, verification of parsable vocabulary descriptions, etc. We note that the distinction between Linked "vocabularies" and "data" is not always straightforward, with for instance the entries of the BioPortal [44], a registry of ontologies (which could by definition be considered as vocabularies), being (and, in fact, a significant) part of the LOD cloud, but not being present in LOV.

So, where does this leave us? We have seen a lot of resources being put into publishing Linked Data, but yet a publicly widely visible "killer app" is still missing. The reason for this, in the opinion and experiences of the authors, lies all to often in the frustrating experiences when trying to actually *use* Linked Data for building actual applications. Many attempts and projects end up still using a centralized warehousing approach, integrating a handful of data sets directly from their raw data sources, rather than being able to leverage their "lifted" Linked Data versions: the use and benefits of RDF and Linked Data over conventional databases and warehouses technologies, where more trained people are available, remain questionable. In the following, we will elaborate on the main reasons for this current state, as we perceive them, however, with a hopeful perspective, that is, sketching solution paths to overcome these challenges.

3. Key Challenges for Linked Data Adoption

Reasons for LOD not yet having reached its full potential are manifold and not simple, and we do not claim to be exhaustive herein; rather, we provide a list from the experiences of the authors to help explain some major challenges in the current state of affairs around LOD. We have chosen to divide reasons into technical and non-technical underlying challenges.

3.1. Technical challenges

The current mode of collection of LOD by metadata published once-off by dataset creators has lead to mainly a nice drawing, rather than making Linked Data accessible and usable. We see the following major challenges when attempting to use Linked Data, parts of which we underpin by some preliminary analyses on the metadata from old.datahub.io. We are obviously not the first ones to recognize these as such, and will accompany them with similar analyses and references where available. Yet, we focus on challenges which we believe need a solution first, before we can dream about federated queries or optimizing query answering over Linked Data (which is what we do mostly in our research now — before addressing practical applications over *several datasets* in *real existing Linked Data*).

3.1.1. Availability and resource limits.

As a result of a recent analysis we did over the metadata on datahub.io, we unfortunately confirmed a very low level of availability of resources, which was already identified as one of the main challenges in the biomedical domain [34]. Among the mentioned 5435 resources in the 1281 "LOD"-tagged datasets on datahub.io, there are only 1917 resources URLs that could be dereferenced. Among all the datasets only 646 dataset descriptions contain such dereferenceable (not counting links to PDF, CSV, TSV files) resource URLs. That is, almost half, 635 dataset descriptions contain no dereferenceable resource URLs that would point to data at all. We applied a best effort here, that is dereferencing both HTTP and FTP URLs with a timeout of 10 seconds awaiting a potential response, counting all 2xx return codes for a HEAD request for HTTP (and following redirects), or, resp. LIST requests for the containing directory for FTP as positives. This confirms the similar experiments by Debattista in his thesis [18, Section 9] and in a more recent article [19]; many LOD cloud datasets are indeed not even being mentioned in his quality assessment framework[10], which only covers 130 accessible datasets.

We note that even a best effort of availability could be viewed as optimistic, if we look in a finer grained analysis of the different formats in these URLs (cf. Figure 3). For example, concerning SPARQL endpoints, our small experiment reconfirms that, among the mentioned 444 potential SPARQL endpoint URLs in metadata, only 252 responded at all, and only 195 responded "true" to a simple `ASK {?S ?P ?O}` query.[11] Table 1 shows the numbers for responding

[10] http://jerdeb.github.io/lodqa

[11] Also, some endpoint implementations returned non-SPARQL-protocol-conformant results such as http://identifiers.org/services/

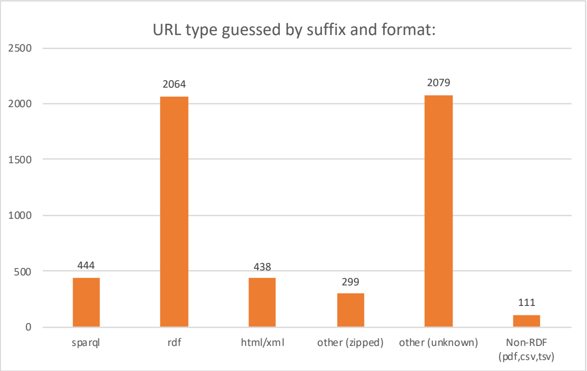

Figure 3. Types of URLs in the “LOD cloud” guessed by declared metadata format and suffixes.

query	conformant responses
ASK {?S ?P ?O }	195 (true) + 7 (false)
ASK { }	150 (true) + 7 (false)
ASK {GRAPH ?G {?S ?P ?O }}	192 (true) + 9 (false)
ASK {GRAPH ?G { }}	146 (true) + 11 (false)
SELECT (count(*) AS ?C) WHERE {?S ?P ?O }	143 (137 non-zero)
SELECT (count(*) AS ?C) WHERE { GRAPH ?G { ?S ?P ?O }}	134 (132 non-zero)

Table 1

SPARQL protocol conformant responses out of the 251 of overall 440 endpoints that responded at all.

endpoint (without timeouts) to a set of test queries, indicating a considerable number of non-responding or non-SPARQL-protocol-conformant endpoints.

Towards a solution path: In order to avoid stale or outdated datasets and SPARQL endpoints, a “live” LOD cloud would need to be producible in a regular and automated fashion. As a part of a solution path, we view regular monitoring frameworks, such as SPARQLES [48],[12] or the Dynamic Linked Data Observatory[32],[13], as essential, which both *(i)* assess which parts of the LOD cloud are still “alive”, and also *(ii)* could notify the providers and publishers about potential problems. Similarly, Debattista’s fine-grained quality framework mentioned above [18, Section 9], aimed originally at re-assessing and testing LOD data regularly could be a valid starting point, but seems to not have been updated since 2016. The same applies for the LOD Laundromat crawl, which is not updated on a regularly basis. LODVader [4] and IDOL [3] are two further projects that have attempted to monitor and generate “live” LOD clouds and metadata in an automated manner directly from the data sources which could serve as starting points.

Outdated, as well as non-available, data is worthless and the frustrating experiences of not finding half the resources when trying to retrieve Linked Data, rather jeopardizes the LOD initiative than inviting externals to our own close community to buy in to the ideas of Linked Data. That is, the LOD cloud itself needs to be “live” and providers that do not comply with minimal availability over a certain duration should be notified and removed. Also, notoriously outdated, stale data should not be listed.

3.1.2. *Size and Scalability*

The situation in terms of dataset sizes have changed dramatically since the early days of semantic search engines, where relatively small amounts of triples could be feasibly managed in a single triple store: few datasets generated from big databases reach dramatic sizes. For instance, the latest edition of DBpedia (2016-10), consists of more than 13 billion triples, Wikidata comprises +5B triples and the whole LOD-Laundromat project, which attempts to process and cleanse the accessible part of the LOD cloud, reports at the moment 38.8B indexed triples.

We also note that, to the best of our knowledge, current triple stores on commodity servers do not scale up to more than 50B triples, apart from lab experiments on hardware probably not yet available to most research labs in our community. AllegroGraph and Oracle triple stores have reported dealing with up to 1 trillion triples.[14]

We already see the number of triples reported on the LOD cloud diverging from what a simple `SELECT (COUNT (*) AS ?C) WHERE {?S ?P ?O}` to their respective endpoints reports in various examples, just to name some: the Pubmed-Bio2RDF endpoint [15], reports 1.37B triples on the query above,[16] whereas the

sparql which returns “false” on above `ASK` query, although clearly its default graph is not empty.

[12] http://sparqles.ai.wu.ac.at/

[13] http://km.aifb.kit.edu/projects/dyldo/

[14] cf. https://www.w3.org/wiki/LargeTripleStores, last retrieved 2018-05-16, where we note that these experiments have been conducted on synthetic LUBM data, which does not necessarily reflect the characteristics of Linked Data “in the wild”.

[15] http://pubmed.bio2rdf.org/sparql

[16] The same number is returned on a query for quads, i.e. `SELECT (COUNT (*) AS ?C) WHERE {GRAPH ?G {?S ?P ?O}}`, which is of course not necessarily the case for all SPARQL endpoints.

dump[17] reports 1.8B triples. Yet again, on a side note, different to both of that, the metadata at datahub.io reports 5B triples for the same dataset,[18], where however it cannot be easily determined in how far these numbers refer to different versions or subsets of the dataset. Likewise, Wikidata's query service responds to the same query a number of 5.2B triples, which is significantly lower than the 5.7B triples we retrieved from the dump mentioned above.

In addition to that, it is mostly impossible to indeed retrieve all triples from a SPARQL endpoint, due to result size restrictions that many endpoints apply, either in the form of timeouts or only returning a certain maximum number of results/triples. For details, Buil et al. [14] discusses some of these restrictions, and also explains, why in general they cannot be trivially circumvented (e.g., by "paging" results with `LIMIT` and `OFFSET`). As another example of related problems, UniProt [15], reported to have +39B triples served on its public endpoint,[14] times out on the simple query to count its triples mentioned above.

Another potential challenge in terms of size and scalability is the amount of duplicates in current dumps. As an example, the PubMed RDF dump from Bio2RDF we mentioned above (cf. Footnote 17) consists of +7.22B nquads spread over 1151 dump files. A lot of triples are actually duplicated across these dump files from the same dataset. Downloading all of these and de-duplicating them locally both wastes bandwidth and makes processing such dumps unnecessarily cumbersome.

Towards a solution path: It seems that in order to avoid both such discrepancies and bottlenecks for downloads and query processing, a combination of *(i)* dumps provided in HDT [21], a compressed and queryable RDF format, as well as *(ii)* Triple Pattern Fragments (TPF) endpoints [49] as the standard access method for Linked Datasets could alleviate some of these problems The Triple Pattern Fragments interface – essentially limits queries to an endpoint to simple triple matching queries which offloads processing of complex joins and other operations to the client-side, while still not having to download complete dumps. HDT,[19] on the other hand is an already compressed dump-format that allows such triple pattern queries without decompression and also guarantees duplicate-freeness. Notably, there are already several TPF endpoints available,[20] most of them powered by HDT in the backend, thus creating a small server-footprint and -load, for either answering triple pattern queries or downloading the whole dump. HDT has also been recently extended to handle also nquads besides RDF triple dumps, thus also being usable for datasets consisting of different (sub)graphs [22]; an analogous extension of the TPF interface to nquads would be straightforward. We also note that metadata such as the number of triples encoded are stored/computed during dump generation in the header in HDT files, thus providing a single, reliable entry to metadata.

As an additional starting point, work on load balancing under concurrency for SPARQL endpoints [38] also promises better resource management on server-side query processing. Eventually, we expect lots of room for research and potential solutions to the challenges for truly decentralized, scalable linked data querying, by combining client-side and server-side query processing in a demand-driven manner.

3.1.3. Findability and (Meta-)Data Formats

While the current metadata available on the LOD cloud does not tell us a lot about how to access single datasets, over time, various dataset description formats and mechanisms have been proposed, namely *(i)* VoID descriptions, *(ii)* (Semantic) Sitemaps, and *(iii)* SPARQL service endpoint descriptions. In the following, we analyze their state of use in the LOD cloud.

The Vocabulary of Interlinked Datasets (VoID) had been designed as a minimalistic entry point for describing datasets and how to access them, containing properties for locating dumps (`void:dataDump`), finding SPARQL endpoints (`void:sparqlEndpoint`) or describing the size of the dataset, i.e., numbers of triples (`void:triples`), and other structural statistics. In order to find the VoID description, it is suggested to place the dataset description in the root directory of a Web-server under `/.well-known/void`.

There are various problems with this approach: firstly, different datasets hosted under one common domain/server cannot provide different dataset descriptions; as illustration, obviously for Github hosted data, `https://github.com/.well-known/void` would not return a valid VoID description, although Github is gaining popularity for hosting Linked Data sets. Secondly, even the "epicenter" of the LOD-

[17] http://download.bio2rdf.org/#/release/4/pubmed/
[18] https://old.datahub.io/dataset/bio2rdf-pubmed
[19] http://rdfhdt.org
[20] http://linkeddatafragments.org/

cloud, dbpedia.org does not follow the rules and provides a VoiD description at the non-obviously findable URL `http://dbpedia.org/void/page/Dataset` instead. Lastly, indeed, among all 881 hostnames mentioned in URLs in datahub.io's metadata, 159 respond to an HTTP GET with this recipe, at least 75 of which though seem to be HTML responses, and only 56 valid RDF;[21] without going into further detail, even if the HTML contained RDFa (which in the cases we inspected it did not), it seems that easy-to-parse RDF results with valid VoId descriptions are the exception.

(Semantic) Sitemaps: XML Sitemaps[22] seem to be a more commonly implemented pattern to discover data and pages accessible via an HTTP server, not least because of their recommendation by search engines. It is a simple XML format that should guide crawlers across sites, where Tummarello et al. had even proposed an extension of the Sitemaps protocol to link to RDF datasets specifically [16], that has been implemented in Sindice [39]. Sitemaps are expected to be found under the root of a dataset's directory on a host in a file called 'sitemap.xml', that is, not necessarily directly underneath root directory of the host address. datahub.io's metadata contains hints (by filename) to such sitemaps for 57 datasets, 56 indeed returning valid sitemaps, and 55 of which indeed use the semantic sitemap extension [16] (52 containing a `sc:dataDump` attribute and 53 containing a sc:sparqlEndpoint field). Overall, while semantic sitemaps are only used for a marginal $\sim 5\%$ of datahup.io datasets, they seem to be fairly consistent.

SPARQL service endpoint descriptions: according to the SPARQL1.1 specification, "*SPARQL services made available via the SPARQL Protocol SHOULD return a service description document at the service endpoint when dereferenced using the HTTP GET operation without any query parameter strings provided. This service description MUST be made available in an RDF serialization, MAY be embedded in (X)HTML by way of RDFa [RDFA], and SHOULD use content negotiation [CONNEG] if available in other RDF representations.*" Yet, out of the 251 potential respondent endpoint addresses mentioned above only 136 respond to this recipe, out of which in fact 63 return HTML (mostly query forms), even if attempting CONNEG.[23]

We note that while some of these mentioned HTML responses *might* contain RDFa, it is still an extra step to extract and parse and each such extra step will bloat a potential consuming client unnecessarily. Similarly, when attempting to find data dumps, without a semantic sitemap or a VoID file in place, our best guess would be to guess and try parsers from "format" descriptors in the metadata or from filename suffixes. An additional complication here are compressed formats, where attempting different decompression formats (gzip, bzip, tar, zip, just to name a few), sometimes even used in combination, further complicate accessibility. Some of the the guessed formats we found in all URLs are listed again in Fig. 3 above.

We note that by manual inspection, some endpoint addresses or accessibility of datasets could be recovered, but since we emphasize on machine accessibility, manual "recovery" seems an undesirable option.

Towards a solution path: We feel that as for automatic findability, Semantic Sitemaps with pointers to a VoID description, with concrete pointers to primarily a dump, preferably in HDT as well as (optionally) a pointer to a SPARQL endpoint (or TPF endpoint) should be the commonly to be agreed upon practice. We note here, that the use of HDT makes this task even simpler, as indeed the Header part of an HDT dump file holds a place for metadata descriptions about the dataset readily.[24] Also, SPARQL endpoints should provide service descriptions in easily accessible RDF (not RDFa) available via CONNEG, where again these SPARQL service descriptions should describe service limitations (such as e.g. result size limits or connection limits and timeouts). Also, the service description should declare potential differences between the data in the dump and in the endpoint, if any. We emphasize here, that to the best of our knowledge there is no agreed upon vocabulary for SPARQL endpoint restrictions and capabilities.

[21] We tested all hosts from URLs that provided non-error results.

[22] https://www.sitemaps.org/protocol.html

[23] with sending an `'Accept: text/turtle, application/n-triples, application/trig, application/n-quads, application/rdf+xml, *'` header.

[24] In fact, some automatically computable VoID properties are already computed and included in HDT's header per default, and it is possible to add additional properties such as pointers to (SPARQL or Linked Data fragments) endpoints, or used namespaces within this header, as a single point of access through an HDT dump file.

3.1.4. Linked Data Quality & Semantics of Links.

The Linked Data principles define rough guidelines on derefenceability and linkage of datasets, yet in order for RDF datasets, once downloaded, to be truly machine-processable and being able to traverse and interpret those links fruitfully, more detailed guidelines seem to be indispensable: in an early approach, Hogan et al. proposed the "Pedantic Web"[29] alongside with the discontinued tool, RDFAlerts, to check and assess the quality, dereferenceability, and finally syntactical (e.g. use of ill-defined literals) logical consistency (in terms of RDFS/OWL inferences, use of literals in place of object properties, availability of definitions for used properties and classes, etc.) of RDF datasets. A lot of these checks though, were not necessarily designed to scale to datasets of billions of triples, or, resp., should be reassessed in terms of feasibility: HDT could serve as a basis for scalable, out-of -the-box implementations of such checks on a dataset level [26].

For example, link counts in the LOD cloud diagram, which shall indicate in how far one dataset links to another dataset, could be checked and computed automatically. To the best of our knowledge, these links and their strength, have been created so far from datahub.io's metadata field `links:<Dataset-acronym>` (i.e., been typically manually specified by the contributors of said metadata). The definition for how such links should be declared on lod-cloud.net provides the following inclusion/exclusion criterion for datasets in the LOD cloud: "*The dataset must be connected via RDF links to a dataset that is already in the diagram. This means, either your dataset must use URIs from the other dataset, or vice versa. We arbitrarily require at least 50 links.*" An older version of the page also provided a slightly more concrete definition of what is meant by a link here: "*A link, for our purposes, is an RDF triple where subject and object URIs are in the namespaces of different datasets.*" We however find this definition hard to assess. Since no concrete guideline with regards to "ownership" of namespaces is provided here, any attempt to compute such links automatically is doomed to fail. From our observations on different datasets, it is by no means always clear

1. to which namespace a URI belongs, or
2. to which dataset a namespace belongs

As for 1, we note that in many cases it is not even clear entirely purely from the RDF data which part of the URIs in a dataset denote namespaces: namespaces and qnames in RDF have no special status as in XML, they simply denote prefixes; while certain "recipes" for such prefixes exist, such as most commonly used '/' and '#' prefixes, some ontologies use completely different recipes to separate identifiers from prefixes. In fact, various datasets "mint" URIs with differing recipes, for instance, we find the prefix scheme `http://bioonto.de/sbml.owl#Uniprot:` within the BIOMDELS ontology from Bioportal, with 562 identifiers using this scheme (e.g., `http://bioonto.de/sbml.owl#Uniprot:Q9UJX6`). In this case, what is the namespace prefix? It seems intuitive that this URI minting scheme refers to UNIPROT which indeed means the dereferenceable URL `https://www.uniprot.org/uniprot/Q9UJX6`. Now, at a closer look this example[25] illustrates several problems at once: it is unclear which prefix denotes the "namespace": `http://bioonto.de/sbml.owl#` or rather `http://bioonto.de/sbml.owl#Uniprot:`? Likewise, the same entity appears in the LOD cloud under a different, disconnected namespace prefix: `http://purl.uniprot.org/uniprot/Q9UJX6`

In fact, the "#-namespace" in our example, `http://bioonto.de/sbml.owl#`, does not refer to a dereferenceable URI. Here data itself comes from a dataset dump in an old version of BioPortal, that has been fixed in the meantime, but nonetheless serves for illustration. Whole datasets, such as the BIOMODELS ontology now exists on different places in the LOD cloud, within Bio2RDF, within BioPortal, but also as an RDF dataset directly published by EBI[26] in three different "RDF exports" of the same database.

While – depending on the serialisation – namespaces could be filtered out based on being explicitly represented (e.g., marked with XML namespaces in RDF/XML or by @prefix declaration in Turtle, respectively, this seems not to be a reliable way of recognizing all used namespaces within an RDF datadump in a declarative machine-readable manner. Plus, as the example illustrates, even if we had all namespaces occurring within a dataset, various URL schemes used refer to either non-dereferenceable or non-RDF publishing third-party namespaces, that cannot be simple assigned to "belonging" to a single dataset. More issues about URI schemes and namespaces and term (non-)re-use have been described in [26, 33, 35].

Last, but not least, as an open problem, links in one dataset always refer to a particular *version* of the linked dataset, which cannot be guaranteed to persist or being dereferenceable in the future. For more sustain-

[25] which is one of many, we emphasize it is not our intention to point fingers to anyone!

[26] at ftp://ftp.ebi.ac.uk/pub/databases/RDF/biomodels/

able Linked Open Data, we therefore deem *versioned* Linked Data as well as archives a necessity.

Towards a solution path: We feel that in order to avoid such issues, established best practices for Linked Data publishing would need to provide more guidelines for URL minting and reuse. Namespace and ID minting should probably be restricted to machine-recognizable patterns (such as strict adherence to '/' and '#'-namespaces), with dereferenaceable namespace URLs). Ownership of a namespace could – for instance – be restricted to pay-level-domain, that is, definition of namespaces being restricted to the own pay-level domain, and URL and namespace schemas given a clear machine-readable ownership relation. We leave a concrete definition of such a machine-readable and assessable ownership open for now, but refer to similar concepts and thoughts about URI "authority" having been discussed before in the context of ontological inference by Hogan in his thesis [28, Section 5] as a potential starting point. Hogan's thesis also contains some details on scalable implementations of the above-mentioned checks that have been described in RDFAlerts [29] earlier, which we believe could be implemented directly and efficiently on top of indexed compressed formats HDT, which we leave to future work on our agenda for now.

As for archiving and versions, we refer to [23] and references therein in terms of starting points. While no standard exists at this point for how to publish versioned RDF archives, let us again refer to possible HDT-based solutions, particularly enabled through the recent extension of HDT to handle nquads [22].

3.2. Non-Technical Challenges

Even if we will be able to solve all the above technical challenges, there are several more pertinent issues in the critical path to the success of LOD. Many non-technical challenges should be fixed in order to stimulate adoption of linked data, a non-exhaustive list of which we briefly describe hereafter.

3.2.1. Completeness/Consistency

Several well-known and important RDF datasets are missing in the LOD cloud. For example, EBI RDF [31] is not there (plus various other well-known data bases from the biomedical and life sciences domain [34]), which have gone through the effort of publishing RDF, but not taken the additional hurdle of manually adding and updating their metadata in yet another centralized catalog such as datahub.io. For similar reasons, Wikidata is not a dataset in the LOD cloud, although it is clearly linked well with several datasets present.

Overall, the burden of manually and pro-actively needing to provide and maintain LOD cloud metadata on the publisher-side has proven unsustainable.

3.2.2. Trust

Besides the pervasive issues of availability and reliability, developers are rightfully worried that published data in the cloud is not kept up to date, and as such these technical issues might overall give rise to (or have already given rise to, possibly) doubts on the technologies and principles of Linked Data. Stale datasets, while still available, but with outdated, once-off RDF exports of in the meantime evolved databases, likewise raise trustworthiness issues in Linked Data.

While it seems to have been a sufficient incentive to "appear" in the LOD cloud to publish datasets adhering to Linked Data principles, a similarly strong incentive to sustain and maintain quality of published datasets seems to be missing. It is therefore important for us as a community to keep the LOD project up and alive *and* maintained, by creating sustainable publishing and monitoring processes.[27]

3.2.3. Governance

We note that not only trust in the LOD cloud itself, but also mutual trust between LOD providers may be a problem that is difficult to circumvent. For instance the presence of various different unlinked "RDF dumps" or LOD datasets arising from exports of the same legacy database (BIOMODELS given as *one* illustrative example of many above) could be potentially related to many of our exports and datasets being created in isolation, by closed groups, without incentives for collaboration, sharing infrastructures, and evolving those exports jointly. This issue can only be solved by a more collaborative, and truly open governance.

3.2.4. Documentation and Usability

Besides the technical issues discussed above, usability issues and documentation standards have been long overlooked in many Linked Data projects. Industry-strength tools to consume and use Linked Data with sufficient documentation are still under-developed.

We believe this issue can be ameliorated by: *(1)* better metadata for describing the datasets; *(2)* better documentation for using the datasets, including sample queries; *(3)* better tool support for enabling reuse of

[27] with the questionable alternative to re-brand under another name after an "LOD winter" from unfulfilled expectations.

existing vocabularies; and *(4)* Supporting and promoting developer-friendly formats, such as JSON-LD.

In addition, in terms of positive examples, apart from the aforementioned HDT and TPF projects, useful SPARQL query editing tools such as YASGUI [43] or Wikidata's query interface, have appeared in the last two years; we need more tools like those.

3.2.5. Funding & Competition.

Last, but not least, while the EU and other funding agencies have supported the creation of a Web of data greatly, we also feel that there are problematic side effects which need discussion and counter-strategies:

- cross-continental research initiatives are not being funded
- EU project consortia are typically being judged by complementary partner expertise

Both these factors prevent research groups working on overlapping topics collaboratively, and rather stimulate an environment of isolated closed research than open collaboration to jointly build sustainable solutions and address the issues mentioned so far.

Lack of collaboration may in other cases also just be caused by the disconnect of research communities. This is for instance exemplified by the Semantic Web in Life Sciences community, for instance, seemingly having recently started efforts very similar to SPARQLES [48] in building up a completely independent SPARQL endpoint monitoring framework [52],[28], not even citing SPARQLES (sic!), which seems unnecessarily duplicating efforts instead of collaboratively developing and maintaining such services.

4. Conclusions and Next Steps

So, is Linked Data doomed to fail? In this paper we did not present a lot of new insights, but our deliberatively provocative articulation of rethinking Linked Open Data and its principles. It is not too late to counteract and join forces. We hope that our summary of problems and challenges, reminders of valuable past attempts to address them, and outline of potential solution strategies can serve as a discussion basis for a fresh starts ahead towards more actionable Linked Data. Yet, we should not conceal there is still a lot of research to be done and open problems to work on, for instance in demonstrating how the sketched solution paths can be realized at Web scale.

On the bright side, specific communities, such as the biomedical community have been very successful in using OWL and Semantic Web technologies for the management of large biomedical vocabularies and ontologies, for a detailed overview of successes in this area we refer to [33, 34]. Main factors for success projects are: *(1)* Having a dedicated and very active development team behind it with continuous funding over several years; *(2)* Actively building a strong community of domain users from different areas, and using their needs as the driver for the ontology development; *(3)* Having an exemplary documentation, about both ontology, but also about how to use Linked Data in applications targeted to domain users, as well as documentation about the processes for building and maintaining collaboratively generated Linked Data sources; *(4)* Using a principled approach for developing the underlying ontology and maintaining the vocabularies used; *(5)* Using automated pipelines to check and ensure data and vocabulary quality.

Our hope is that the Linked Data community can learn from such specific projects, and that it will try to apply some of the same approaches that proved to be so successful. We believe the community needs to work on those by joining forces, rather than by competition. We also argued that HDT, a compressed and queryable dump format for Linked Datasets, could play a central role as a starting point to address some (but not all) of the technical challenges we have outlined, i.e., implicitly suggesting a "fifth" Linked Data principle [6]:

> 5. Publish your dataset as an **HDT dump**, including **VoID metadata** as part of its header, declaring (i) the (authoritatively) **owned namespaces**, (ii) links to previous and most current **versions** of the dataset, (iii) and – whenever you use namespaces owned by other datasets or ontologies – **links to specific versions of** these **other datasets**.

In fact, we would argue that more principled Linked Data publishing could allow to auto-generate LOD clouds from a set of HDT dumps, which to demonstrate is on our agenda for future work. While we admit of course that HDT is not the only possible solution to this problem, what we aim at emphasizing here is the necessity for a principled approach to overcome the obvious synchronization problems arising from separate maintenance of metadata outside of the actual knowledge source in current dataset catalogues: an interesting very recent initiative in this context is the

[28] available at http://yummydata.org/endpoints

DBpedia Databus project[29], currently in public beta, which aims at providing an end-to-end pipeline easing auto-extraction, metadata-generation and publishing of Linked Data Knowledge Graphs at scale.

Apart from technical challenges, other issues arose, that seem equally important, such as the establishment of collaborative and shared research infrastructures to guarantee sustainable funding and persistence of Linked Data assets, as we have seen many promising efforts and initiatives mentioned in this paper having discontinued unfortunately. In the meanwhile, we also emphasize that initiatives like the recently US-founded "Open Knowledge Network"[30] initiative or a Dagstuhl seminar on "New Directions for Knowledge Representation on the Semantic Web" [11] have provided platforms to openly discuss such a fresh start, in the context of new trends and efforts around Knowledge Graphs and the FAIR principles [51], that parallel and complement the Linked Data movement.

Acknowledgements

A preliminary version of this work [42] was presented at the DESEMWEB2018 workshop: we thank the participants and particularly Dan Brickley, Sarven Capadisli, Amrapali Zaveri, and Denny Vrandečić for their comments on this first revision. Axel Polleres' work was supported under Stanford University's Distinguished Visiting Austrian Chair program. Javier Fernández' work was supported by the EC under the H2020 project SPECIAL and by the Austrian Research Promotion Agency (FFG) under the project "CitySpin". Maulik Kamdar, Tania Tudorache, and Mark Musen were supported in part by the grants GM086587, GM103316, and U54-HG004028 from the US National Institutes of Health.

References

[1] A. Abele, J. P. McCrae, P. Buitelaar, A. Jentzsch, and R. Cyganiak. Linking Open Data cloud diagram (2018-04-30), 2018. From http://lod-cloud.net/; retr. 2018/06/01.

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note 03 March 2011, 2011. From https://www.w3.org/TR/void/; retr. 2018/06/01.

[3] C. Baron Neto, D. Kontokostas, A. Kirschenbaum, G. Publio, D. Esteves, and S. Hellmann. IDOL: Comprehensive & complete LOD insights. In *Proceedings of the 13th International Conference on Semantic Systems (SEMANTiCS 2017)*, 2017. DOI: 10.1145/3132218.3132238.

[4] C. Baron Neto, K. Müller, M. Brümmer, D. Kontokostas, and S. Hellmann. LODVader: An interface to LOD Visualization, Analytics and DiscovERy in Real-time. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 163–166. International World Wide Web Conferences Steering Committee, 2016. DOI: 10.1145/2872518.2890545.

[5] W. Beek, L. Rietveld, H. R. Bazoobandi, J. Wielemaker, and S. Schlobach. LOD Laundromat: a uniform way of publishing other people's dirty data. In *13th International Semantic Web Conference (ISWC)*, pages 213–228. Springer, 2014. DOI: 10.1007/978-3-319-11964-9_14.

[6] T. Berners-Lee. Linked Data. W3C Design Issues, July 2006. From http://www.w3.org/DesignIssues/LinkedData.html; retr. 2018/06/01.

[7] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, pages 29–37, May 2001.

[8] S. Bischof, M. Krötzsch, A. Polleres, and S. Rudolph. Schema-agnostic query rewriting in SPARQL 1.1. In *13th International Semantic Web Conference*, LNCS. Springer, Oct. 2014. DOI: 10.1007/978-3-319-11964-9_37.

[9] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009. DOI: 10.4018/jswis.2009081901.

[10] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *J. Web Sem.*, 7(3):154–165, 2009. DOI: 10.1016/j.websem.2009.07.002.

[11] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti. Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web (Dagstuhl Seminar 18371). *Dagstuhl Reports*, 8(9):29–111, 2019. DOI: 10.4230/DagRep.8.9.29.

[12] D. Brickley and R. Guha. RDF Schema 1.1. W3C Recommendation, Feb. 2014. http://www.w3.org/TR/rdf-schema/.

[13] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.99, Jan. 2014. http://xmlns.com/foaf/0.1/.

[14] C. Buil-Aranda, A. Polleres, and J. Umbrich. Strategies for executing federated queries in SPARQL1.1. In *13th International Semantic Web Conference (ISWC)*. Springer, 2014. DOI: 10.1007/978-3-319-11915-1_25.

[15] U. Consortium et al. The universal protein resource (UniProt). *Nucleic acids research*, 36(suppl 1):D190–D195, 2008. DOI: 10.1093/nar/gkm895.

[16] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. 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, 2008.

[17] M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semantic Web*, 2(1):55–63, 2011. DOI: 10.3233/SW-2011-0031.

[18] J. Debattista. *Scalable Quality Assessment of Linked Data*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Oct. 2016.

[19] J. Debattista, C. Lange, S. Auer, and D. Cortis. Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web*, 9(6):859–901, 2018. DOI: 10.3233/SW-180306.

[20] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the*

[29] https://blog.dbpedia.org/2019/10/02/one-billion-derived-knowledge-graphs/

[30] http://ichs.ucsf.edu/open-knowledge-network/

ACM International Conference on Information and Knowledge Management (CIKM), pages 652–659. ACM, 2004. DOI: 10.1145/1031171.1031289.

[21] J. D. Fernández, M. A. Martınez-Prieto, C. Gutiérrez, A. Polleres, and M. Arias. Binary RDF Representation for Publication and Exchange (HDT). *J. Web Sem.*, 19(2), 2013. DOI: 10.1016/j.websem.2013.01.002.

[22] J. D. Fernandez, M. A. Martínez-Prieto, A. Polleres, and J. Reindorf. HDTQ: Managing RDF datasets in compressed space. In *European Semantic Web Conference (ESWC)*. Springer, 2018. DOI: 10.1007/978-3-319-93417-4_13.

[23] J. D. Fernandez, J. Umbrich, A. Polleres, and M. Knuth. Evaluating query and storage strategies for RDF archives. *Semantic Web*, 10(2):247–291, 2019. DOI: 10.3233/SW-180309.

[24] B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres. OWL: Yet to arrive on the web of data? In *WWW2012 Workshop on Linked Data on the Web (LDOW)*, 2012.

[25] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995. DOI: 10.1006/ijhc.1995.1081.

[26] A. Haller, J. D. Fernández, M. R. Kamdar, and A. Polleres. What are Links in Linked Open Data? A Characterization and Evaluation of Links between Knowledge Graphs on the Web. Technical Report 2/2019, Department für Informationsverarbeitung und Prozessmanagement, WU Vienna University of Economics and Business, 2019.

[27] A. Haller and A. Polleres. Are We Better Off With Just One Ontology on the Web? *Semantic Web*, in this volume, 2019.

[28] A. Hogan. *Exploiting RDFS and OWL for Integrating Heterogeneous, Large-Scale, Linked Data Corpora*. PhD thesis, Digital Enterprise Research Institute, National University of Ireland, Galway, 2011.

[29] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *International Workshop on Linked Data on the Web (LDOW) at WWW*, 2010.

[30] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. *J. Web Sem.*, 9(4):365–401, 2011. DOI: 10.1016/j.Websem.2011.06.004.

[31] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, et al. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014. DOI: 10.1093/bioinformatics/btt765.

[32] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing linked data dynamics. In *Proceedings of Extended Semantic Web Conference (ESWC)*, pages 213–227. Springer, 2013. DOI: 10.1007/978-3-642-38288-8_15.

[33] M. R. Kamdar. *A web-based integration framework over heterogeneous biomedical data and knowledge sources*. PhD thesis, Stanford University, 2019.

[34] M. R. Kamdar, J. D. Fernández, A. Polleres, T. Tudorache, and M. A. Musen. Enabling web-scale data integration in biomedicine through linked open data. *NPJ digital medicine*, 2(1):1–14, 2019. DOI: 10.1038/s41746-019-0162-5.

[35] M. R. Kamdar, T. Tudorache, and M. A. Musen. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic web*, 8(6):853–871, 2017. DOI: 10.3233/SW-160238.

[36] A. Mallea, M. Arenas, A. Hogan, and A. Polleres. On Blank Nodes. In *Proceedings of the International Semantic Web Conference (ISWC)*, volume 7031 of *LNCS*. Springer, 2011. DOI: 10.1007/978-3-642-25073-6_27.

[37] A. Miles and S. Bechhofer. Simple knowledge organization system reference. Recommendation, W3C, August 18 2009.

[38] T. Minier, H. Skaf-Molli, and P. Molli. SaGe: Web preemption for public SPARQL query services. In *The World Wide Web Conference*, pages 1268–1278. ACM, 2019. DOI: 10.1145/3308558.3313652.

[39] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *IJMSO*, 3(1):37–52, 2008. DOI: 10.1504/IJMSO.2008.021204.

[40] H. Paulheim and A. Gangemi. Serving dbpedia with DOLCE - more than just adding a cherry on top. In *14th International Semantic Web Conference (ISWC)*, pages 180–196, 2015. DOI: 10.1007/978-3-319-25007-6_11.

[41] A. Polleres, A. Hogan, A. Harth, and S. Decker. Can we ever catch up with the Web? *Semantic Web*, 1(1-2):45–52, 2010. DOI: 10.3233/SW-2010-0016.

[42] A. Polleres, M. R. Kamdar, J. D. Fernández, T. Tudorache, and M. A. Musen. A more decentralized vision for linked data. In *Decentralizing the Semantic Web (Workshop of ISWC2018)*, volume 2165 of *CEUR Workshop Proceedings*. CEUR-WS.org, Oct. 2018.

[43] L. Rietveld and R. Hoekstra. The YASGUI family of SPARQL clients. *Semantic Web*, 8(3):373–383, 2017. DOI: 10.3233/SW-150197.

[44] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semantic Web*, 4(3):277–284, 2013. DOI: 10.3233/SW-2012-0086.

[45] M. K. Smith, C. Welty, and D. L. McGuinness. OWL Web Ontology Language Guide. W3C Recommendation, Feb. 2004. http://www.w3.org/TR/owl-guide/.

[46] S. Tramp, P. Frischmuth, T. Ermilov, and S. Auer. Weaving a Social Data Web with Semantic Pingback. In *Knowledge Engineering and Knowledge Management by the Masses (EKAW)*, volume 6317 of LNAI, pages 135–149. Springer, 2010. DOI: 10.1007/978-3-642-16438-5_10.

[47] P. Vandenbussche, G. Atemezing, M. Poveda-Villalón, and B. 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.

[48] P.-Y. Vandenbussche, J. Umbrich, L. Matteis, A. Hogan, and C. Buil-Aranda. SPARQLES: Monitoring public SPARQL endpoints. *Semantic Web*, 8(6):1049–1065, 2017. DOI: 10.3233/SW-170254.

[49] R. Verborgh, M. V. Sande, O. Hartig, J. V. Herwegen, L. D. Vocht, B. D. Meester, G. Haesendonck, and P. Colpaert. Triple pattern fragments: A low-cost knowledge graph interface for the web. *J. Web Sem.*, 37-38:184–206, 2016. DOI: 10.1016/j.websem.2016.03.003.

[50] D. Vrandecic and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. DOI: 10.1145/2629489.

[51] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The FAIR guiding princi-

ples for scientific data management and stewardship. *Scientific data*, 3, 2016. DOI: 10.1038/sdata.2016.18.

[52] Y. Yamamoto, A. Yamaguchi, and A. Splendiani. Umaka-Yummy Data: A Place to Facilitate Communication between Data Providers and Consumers. In *International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, volume 1795 of *CEUR*, 2016.