A More Decentralized Vision for Linked Data

Axel Polleres, Maulik R. Kamdar, Javier D. Fernandez, Tania Tudorache, and Mark A. Musen

Arbeitspapiere zum Tätigkeitsfeld
Informationsverarbeitung, Informationswirtschaft und Prozessmanagement
Working Papers on Information Systems, Information Business and Operations

Nr./No. 02/2018
ISSN: 2518-6809
URL: http://epub.wu.ac.at/view/p_series/S1/

Herausgeber / Editor:
Department für Informationsverarbeitung und Prozessmanagement
Wirtschaftsuniversität Wien · Welthandelsplatz 1 · 1020 Wien
Department of Information Systems and Operations · Vienna University of Economics and Business · Welthandelsplatz 1 · 1020 Vienna

A More Decentralized Vision for Linked Data

Axel Polleres[1,2], Maulik R. Kamdar[1], Javier D. Fernandez[2], Tania Tudorache[1], and Mark A. Musen[1]

[1] Stanford University, CA, USA
[2] Vienna Univ. of Economics & Business / Complexity Science Hub Vienna, Austria

Abstract. In this *deliberately provocative* position paper, we claim that 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 the biomedical domain—currently, one of the most promising "adopters" of Linked Data—if we believe the ever-present "LOD cloud" diagram.[3] Herein, we try to highlight and exemplify 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.

1 Decentralization Myths on the Semantic Web

Let us start with a rant. The Semantic Web is 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[4] appear not to have found major adoption at a level comparable with these siloed sites.[5]
- 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 [30].[6] While

[3] The Linking Open Data cloud diagram, available at http://lod-cloud.net/, which has been regularly updated since 2007 by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak, with its latest version having been created in April 2018 [2].

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

[5] While there is some hope left, in ActivePub being picked up by several implementations, cf. https://en.wikipedia.org/w/index.php?title=ActivityPub&oldid=841568831#Implementations, reversing the network effects that have drawn a critical mass of users to these siloed sites seems still far away.

[6] or, as Dan Brickley, one of he 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

there are indeed certain areas in which ontologies 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 extremely limited. Instead, we got a main *central* schema (schema.org), and fast-growing community projects like Wikidata [64] refusing to buy into the need for re-using terms from other ontologies.[7]

- We put a lot of effort into **formal semantics and clean axiomatization** of those ontologies – we got logical inconsistency.[8] 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 [59] 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 [28,45]; 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... 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

---

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."

[7] 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 to choose a stronger semantics - like OWL - or a weaker semantic - like SKOS[47] or not" (personal communication Denny Vrandečić).

[8] Even within DBpedia [49,8], the central crystalization point of the LOD cloud [10].

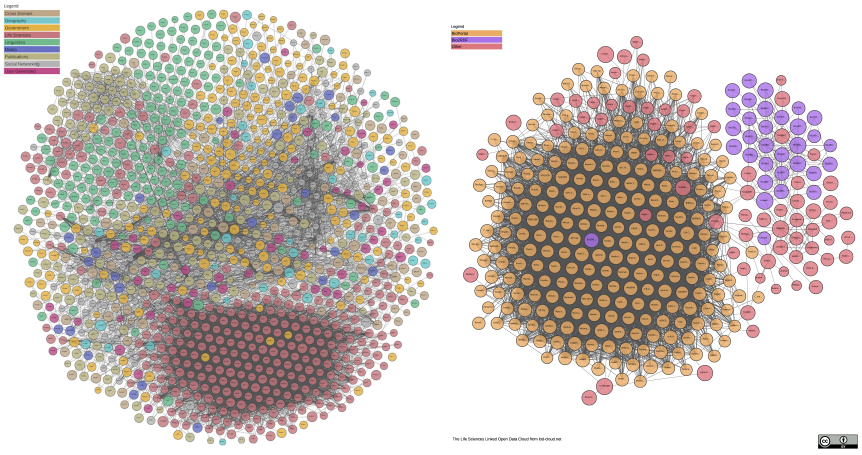

Fig. 1. The latest "LOD cloud" diagram [2], from April 2018, counting 1,184 datasets (left), and the significant portion of life-sciences and biomedical data amongst it, with 339 datasets (right).

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 this data. Thus, we would like to take a step back and assess the situation. We will identify some serious challenges in consuming and using Linked Data from the "cloud", with a deeper look at the biomedical domain. Then, we will question the usefulness of the current LOD cloud, and, finally, we call for a more principled 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 argue why the biomedical domain has had such a strong interest in ontologies and Linked Data, and is, therefore, well-suited as a show case to highlight the most urging challenges in Section 3. Based on Linked Data from the biomedical domain, 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 4. 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 5.

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 engines, such as Sindice [48], SWSE [35], Watson [18],

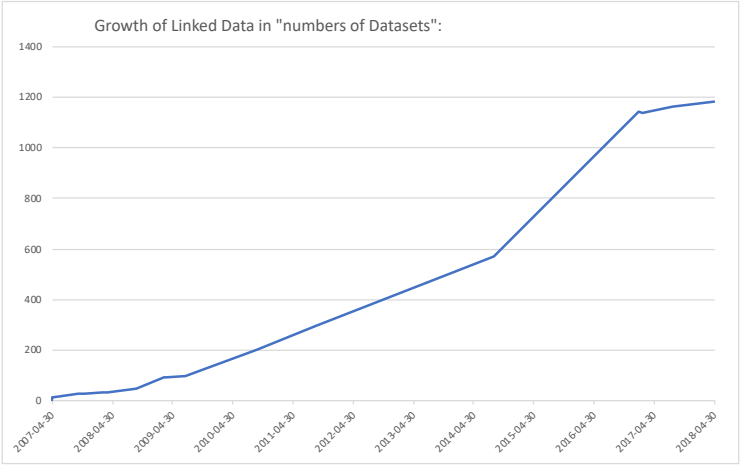

Fig. 2. The growth of the "LOD cloud" in number of datasets seems to indicate steady, while not rapid or even overwhelming adoption; we still have to view this as opposed to the probably much more rapid growth of other parts of the Web in the same time period [50]

Swoogle [23], 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 [60]. In the meantime, unfortunately all of these search engines have been discontinued, and we are not aware of any active, public Semantic Pingback services. As more recent efforts, the LOD-Laundromat project [5] offers an URL lookup service[9] generated from/for the (accessible parts of) the LOD cloud, and also the LOD Cache by OpenLinkSW[10] remain 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 [2] took a different approach, that is, it 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 the current, latest LOD diagram version from April 2018; 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/. We thus have to conclude that the LOD-cloud is suffering from starvation as well, and the current noble effort by John McCrae et al. (the creators and maintainers of the LOD cloud diagram) seems to be likewise at risk.

[9] http://lotus.lodlaundromat.org/
[10] lod.openlinksw.com/

Still, the LOD-cloud at lod-cloud.net and its metadata at datahub.io seem to remain the single most popular entry point to Semantic Web data (with the exception of domain specific portals such as Bioportal [56], which we will discuss further in the next section). The metadata the LOD cloud relies on, comprises metadata fields such as:

- **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, or semantic descriptions (e.g. in the form of a Void [3] 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. [61]. 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 [56], a registry of ontologies (which could by definition be considered as well as vocabularies), being (an in fact significant) part of the LOD cloud, but not being present in LOV.

3 Linked Data in the Biomedical Domain

The biomedical domain is one of the earliest adopters of Semantic Web technologies and Linked Data principles for representing, publishing, linking and querying data on the Web. This adoption is starkly obvious in the well-known LOD cloud diagram, in which the biomedical datasets make up the largest portion of the cloud (cf. Fig. 1). We would expect to see a plethora of applications of LOD in biomedicine, however, they are conspicuously missing. In this section, we briefly describe the LOD adoption in major biomedical projects, and we discuss the challenges that make it very difficult to use LOD effectively in biomedicine with the help of a use case. In the next section, we will generalize these challenges and articulate our thoughts about potential steps to alleviate these issues.

3.1 Linked Data Adoption in Biomedicine

Several key biomedical initiatives use Semantic Web technologies for the integration of diverse datasets in fields, such as, neurosciences (e.g., NeuroCommons [52]), cancer research (e.g., Granatum and Linked TCGA [31,55]), and drug discovery (e.g., Linking Open Drug Data [36]). One of the most notable open-source projects, Bio2RDF, uses Semantic Web technologies to build and provide the largest network of Linked Open Data for the Life Sciences (LSLOD) from a diverse set of heterogeneously formatted sources obtained from multiple data providers [16]. The Bio2RDF Release 3 consists of around $\approx$ 11 billion triples generated from 35 important biomedical data sources, such as, DrugBank, PharmGKB and KEGG [16].

Data providers in the domain themselves, are now embracing Semantic Web technologies and started providing data dumps in RDF as alternative downloads. Some even incorporate SPARQL functionality or standard enpoints in their web portals. For example, the European Bioinformatics Institute (EBI) provides SPARQL access to their proprietary databases (e.g., UniProt, ChEMBL, and Reactome) in the EBI-RDF platform [37]. The National Center for Biotechnology Information (NCBI) publishes the entire PubChem data repository [11] of biological assays and activities of compounds, as RDF data dumps [27]. In February 2014, the National Library of Medicine's (NLM) Linked Data Infrastructure Working Group released an RDF version of the Medical Subject Headings (MeSH) taxonomy [15]. All these initiatives are highly promising and illustrative for the LOD adoption.

3.2 Challenges in Using Linked Data for Biomedical Applications

So far, significant resources have been invested in publishing biomedical data on the LOD Cloud, however, yet to the best of our knowledge we did not find any major applications that use *multiple* Linked Data sources to generate new insights, or to discover novel implicit associations *serendipitously*. In most cases, the publishers mention the "potential" use cases achieved by publishing and querying biomedical data on the LOD cloud in a controlled environment [55,39]. However, for most biomedical researchers (and autonomous agents) querying against the LOD sources in the wild does not bear fruitful results (or any useful results in most cases). We have identified the most important show-stoppers for using LD in biomedicine as:

1. The **availability** of datasets (either as RDF dumps, or as functioning SPARQL endpoints with significant uptime and latency);
2. The **semantic heterogeneity** among datasets (correct reuse of existing vocabularies and ontologies, as well as existing entity URIs); and
3. The **steep learning curve** for understanding and using LD and Semantic Web technologies for biomedical researchers on the one hand as opposed to the steep learning curve for Semantic Web researchers to understand published biomedical data on the other hand.

We will further describe these challenges with the help of a use case for which Linked Data would be a natural fit, however impossible to use with the current state of the LOD.

Use Case: Drug Repurposing. According to a recent study, drug discovery costs have exploded. It now costs $2.87 billion (in 2013 dollars) for a bio-pharma company to research and sell a new drug, and these costs are bound to increase exponentially [22]. To mitigate the costs of drug discovery, researchers have started looking for novel uses of existing drugs, often called drug repurposing [58]. Once a drug is released in the market after clinical trials, federal regulators monitor the prevalence of adverse drug reactions (ADR) in patients who are administered the particular drug, often called pharmacovigilance or drug safety. Adverse drug reactions may not always be detected during the

clinical trials, and ADRs often manifest in patients who are prescribed multiple concomitant drugs that interact with each other [40].

For biomedical research pertaining to drug discovery, drug repurposing, and drug safety, it is often necessary to retrieve and integrate all data and knowledge pertaining to a given DRUG entity, and provide an aggregated summary to the biomedical researcher (e.g., drug–protein target interactions, publications that mention drugs and their adverse reactions, downstream drug targets located in biological pathways, assays that test the binding activity of the "drug active ingredient").

To address this problem, researchers have used conventional methods of data integration: download and process the data in varied formats (CSV, XML, MySQL databases), reconcile entities, and then publish the systems pharmacology network [32,44]. For example, Himmelstein, et al. [32] created a systems pharmacology network composed of different biological entities (drugs, proteins, pathways, diseases) using a common data model by integrating 29 different data and knowledge sources, manually. However, in an ideal world and LSLOD Cloud, such a systems pharmacology network should just be easily created by using a federated SPARQL CONSTRUCT query. Indeed, most of these 29 sources are already available on the LSLOD Cloud, and should be query-able according to the LOD Cloud Diagram.

Challenges. Unfortunately, the current state of the LSLOD Cloud (and by extrapolating, the LOD Cloud) is not suitable for achieving seamless integration of data and knowledge from multiple sources. We detail further the three challenges we mentioned before.

The accessibility and availability of LSLOD sources as RDF dumps or SPARQL endpoints are one of the major reasons why the LSLOD cloud cannot be queried and consumed by biomedical researchers. We will discuss these issues in the Technical Challenges section below in more detail. To give some examples from biomedicine: even though the SPARQL endpoint of BioPortal (http://sparql.bioontology.org/) is mostly available, it has not been updated in several years. Bio2RDF dumps are available only via a Javascript-based page (http://download.bio2rdf.org/#/) that cannot be easily crawled by machines (without, for example, using a headless browser interpreting the Javascript). The liveliness of SPARQL endpoints depends heavily on the continuing support and interest of their maintainers. Once the main maintainer moves on from the project, often, the SPARQL endpoints are not updated anymore, and ultimately, they stop working. Just to give a few example of such defunct biomedical datasets: the DERI HCLS workbench (http://hcls.deri.org:8080/openrdf-sesame/repositories/), the SIDER–Side Effect Resource (http://wifo5-03.informatik.uni-mannheim.de/sider/), the STITCH–Chemical-Protein Interactions dataset (http://wifo5-03.informatik.uni-mannheim.de/stitch/), and the list can go on much longer.

The semantic heterogeneity of LSLOD is another major issue for its use within biomedical applications. Automated traversal across datasets or integration work only if the datasets are linked using the same identifiers for the same terms consistently. While this might seem like a trivial requirement to satisfy, it is not what we find in practice. For example, rather than using common vocabularies (e.g., from the LOV or biomedical ontologies), data publishers use their own vocabularies to generate RDF

data (e.g., they use different URIs for the DRUG class). Even worse, publishers reuse inconsistently (and often, incorrectly) the representations of URIs and IRIs for namespaces. For example, we found the following UniProt URI representations in different datasets:

- `http://purl.uniprot.org/uniprot/`
- `http://bio2rdf.org/uniprot:`
- `http://purl.obolibrary.org/obo/UniProt:`
- `http://identifiers.org/uniprot/`
- `http://bioonto.de/sbml.owl#Uniprot:`

All these URI prefixes are meant to refer to the same UniProt identifiers, e.g. `Q9UJX6`. An application developer unaware of all these representations within different datasets would have a hard time using data from different sources. This issue creates significant burden on the side of application developers who query linked data (i.e., link traversal/conventional query federation methods will just not work across these sources). This "intent for reuse" is also present across biomedical ontologies – we have documented all these cases in Kamdar, et al. [41,43].[11].

Many biomedical projects had to address the semantic heterogeneity problem in order to build applications. The most common solution is to use warehousing in which all data is transformed under a common schema using a uniform set of notations. Projects, such as, OpenPhacts [65], ReDrugs [46], DisQover[12] De Witte et al. [19] use such a warehousing approach, and they have been quite popular for developing biomedical applications using Semantic Web technologies. However, all the respective projects require a lot of centralization and maintenance, and they also need to be updated when the underlying content changes. Such a high effort requires significant resources and therefore this solution can only be implemented as part of a consortium of companies.

Essentially, these approaches enforce our view that decentralization is not possible in the current state of the LOD Cloud (where for instance, as we will later discuss, neither versioning nor accessability can be monitored easily, nor the ownership or re-use of certain namespace prefixes or identifier schemes is in any way regulated or findable, which would be prerequisites).

Conventional query federation methods (e.g., FedX [57], SPLENDID [29]) are often evaluated as closed-box systems (i.e., the datasets are refined and deployed locally and known), and are never evaluated in the wild [54,53]. To give our reader an idea, to retrieve and integrate all drug–protein interactions from 4 sources on the LSLOD Cloud—which should be a straightforward task—will require a complex federated query with > 20 triple patterns (not considering the inconsistent URI representations of entities) [40]. Compiling such a complex query is not only impossible for most biomedical researchers, but also difficult for most computer scientists, leaving alone the executability on SPARQL endpoints that —if available— typically impose strict limits and easily run into timeouts on complex queries. Trying to generate a systems pharmacology network as presented by Himmelstein et al. [32] using query federation methods is almost impossible with the current LSLOD.

[11] We also present these discrepancies visually at: http://onto-apps.stanford.edu/lslodminer

[12] DisQover uses federation in the sense that all data sources are first transformed under the common schema and notations, and then they are divided into different endpoints.

The steep learning curve in understanding and using LD and Semantic Web technologies is another major impediment to the effective use of LD in biomedicine. Data publishers lack tools and guidelines to help them discover and reuse existing content in a correct way, hence reducing the risk of semantic heterogeneity. Our recent study on the analysis of 4 years of BioPortal usage logs found that most users have not explored or queried a large proportion of most ontologies (especially lower levels) either through the BioPortal web interface, or through the API [42]. BioPortal ontologies comprise of a major portion of the LOD Cloud[13]. If biomedical users never explore or query ontological content in BioPortal, how will they reuse this ontological content? It is probably naive to expect that biomedical researchers will formulate sophisticated SPARQL queries, without minimal automated support, over heterogeneous LOD sources that are rarely updated (and in many cases, do not have available RDF dumps or functioning SPARQL endpoints) for their data and knowledge integration needs.

So, where does this leave us? We have seen a lot of resources being put into publishing and using Linked Data in biomedicine, but a "killer app" is still missing. We tried to explain in this section our (often frustrating) experiences with using LD for building biomedical applications. Similarly to the projects that chose in the end to use a centralized warehousing approach, we found the current state of the LSLOD to be largely insufficient to sustain any real application, plus if a central warehouse is used, the use and benefits of RDF and Linked Data over conventional databases and warehouses technologies, where more trained people are available, remain questionable.

4 Key Challenges in usage and adoption of Linked Data

Reasons for LOD not yet having reached its full potential are manifold and not simple, and we do not claim to be exhaustive herein; yet, following on from the aforementioned challenges in the biomedical domain, we would like to 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.

4.1 Technical challenges

As already hinted in Section 2, the current model of collection of LOD by meta-data published once-off by the creators of datasets has lead to mainly a nice drawing, rather than making Linked Data accessible and usable. In fact, 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, wherefore we will accompany them with similar analyses and references where available. Yet, we focus on challenges which we believe to 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 papers now — without practical applications over *several datasets* in *real existing Linked Data*).

[13] Obviously these ontologies are also available through other portals that may be used by most biomedical researchers, however it only strengthens our point that the LOD Cloud diagram has resources that are never accessed.

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.

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: 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; i.e., 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 [20, Section 9] and in a more recent article [21]; many LOD cloud datasets are indeed not even being mentioned in his quality assessment framework[14], 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 different formats in these URLs, cf. Figure 3, e.g. concerning SPARQL endpoints: indeed 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.[15] Table 1 shows the numbers for responding endpoint (without timeouts) to a set of test queries, which seem to indicate a considerable number of non-responding and also non-SPARQL-protocol-conformant endpoints.

Towards a solution path: As a part of a solution path, we view regular monitoring frameworks like SPARQLES [62],[16], or the Dynamic Linked Data Observatory[38],[17] 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,

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

[15] Also, some endpoint implementations returned non-SPARQL-protocol-conformant results such as http://identifiers.org/services/sparql which returns "false" on above ask query, although clearly its default graph is not empty.

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

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

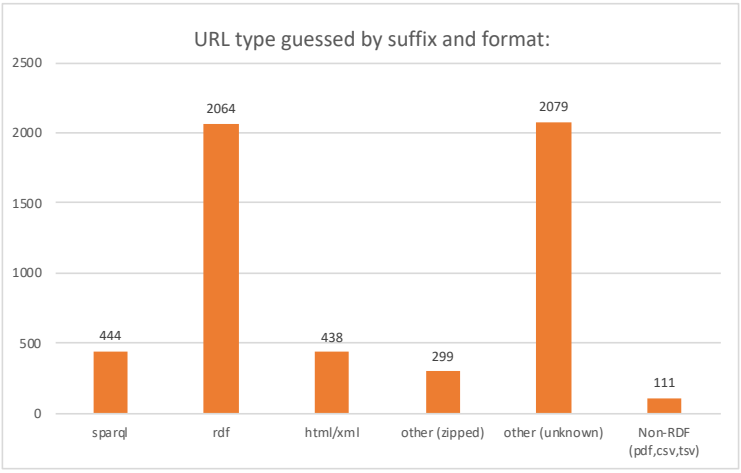

Fig. 3. types of URLs in the "LOD cloud" guessed by declared metadata format and suffixes.

Debattista's fine-grained quality framework mentioned above, 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.

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.

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.[18]

[18] 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".

We already see sizes 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 [19], reports 1.37b triples on the query above,[20] whereas the dump[21] 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,[22], 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, see also [14], which 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, reported to have +39b triples served on its public endpoint, cf. Footnote 18, 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 21, 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 [24], a compressed and queryable RDF format, as well as (ii) Triple Pattern Fragments (TPF) endpoints [63] as the standard access method for Linked Datasets could alleviate some of these problems: the triple-patterns fragment 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,[23] 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,[24] 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 quads besides RDF triple dumps, thus also being usable for datasets consisting of different (sub)graphs [25]; an analogous

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

[20] 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.

[21] http://download.bio2rdf.org/#/release/4/pubmed/

[22] https://old.datahub.io/dataset/bio2rdf-pubmed

[23] http://rdfhdt.org

[24] http://linkeddatafragments.org/

extension of the TPF interface to quads would be straightforward. Lastly, we note that e.g. the number of triples it encoded and stored during dump generation in the metadata header of HDT files, thus providing a single, reliable entry to the dataset size.

Findability and (Meta-)Data Formats. The current metadata available on the LOD cloud does not tell us a lot about how to access the single datasets.

Over time, various dataset description formats and mechanisms have been proposed, typically (i) VoID descriptions, (ii) (Semantic) Sitemaps, and (iii) SPARQL service endpoint descriptions. In the following, we analyze the current state of affairs 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 SPAQL endpoints (`void:sparqlEndpoint`) or describing the size of the dataset in terms of numbers of triples (`void:triples`) and other structural statistics. In order to find the VoID description, it is suggested to place the dataset description under `/.well-known/void` in the root directory of a Web-server.

There are various problems with this approach: firstly, different datasets hosted under one common domain/server cannot provide different dataset descriptions; as an illustration obviously `https://github.com/.well-known/void` does not return a valid VoID description, although github is gaining popularity for hosting Linked Data sets. Secondly, even the "epicenter" of the LOD-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;[25] 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 seem to be the exception.

(Semantic) Sitemaps XML Sitemaps [26] 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 [17], that has been implemented in Sindice [48]. 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 [17] (52 containing a `sc:dataDump` attribute and 53 containing a sc:sparqlEndpoint field). So, overall, while semantic sitemaps are only used for a marginal 5% of the datasets in datahup.io, they seem to be fairly consistent.

[25] We tested all hosts from the URLs that provided non-error results

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

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.[27]

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 also note that by manual inspection, some endpoint addresses or accessibility of datasets could be recovered, but since we herein would like to 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.[28] 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.

"RDF Data Quality" of Datasets and the "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

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

[28] In fact, some automatically computable VoID properties are already computed and included in HDT's header per default, and it is well 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.

be indispensable: in an early approach, Hogan et al. proposed the "Pedantic Web"[34] alongside with an in the meanwhile 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. Again, HDT could serve as a basis for scalable, out-of -the-box implementations of such checks on a dataset level.

Besides the aforementioned "semantic heterogeneity" issue in the biomedical domain, as a particular additional example of checks that should be automatically performed on a dataset level, we mention the links in the LOD cloud diagram, shall indicate in how far one dataset links to another dataset; 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 so concrete guideline with regards to "ownership" of name spaces is provided here, any attempt to compute such links automatically is doomed to fail. As from our observation when investigating 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 ontolgy 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 is referring to UNIPROT which indeed means the dereferenceable URL

`https://www.uniprot.org/uniprot/Q9UJX6`.

Now, at a closer look this example[29] 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:`?

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

- the same entities exist in the LOD cloud under different, disconnected namespace prefixes, such as the Uniprot identifier Q9UJX6, the "official" prefix of which is `http://purl.uniprot.org/uniprot/Q9UJX6` as per the authoritative pay-level-domain uniprot.org.
- the overall "#namespace" `http://bioonto.de/sbml.owl#` does not refer to a dereferenceable URI; the data itself comes in fact from a dataset dump in an old version of bioportal, that has been fixed in the meantime, but nonetheless it serves for illustration; a detailed analysis of present such quality issues in the LOD-cloud is still on our agenda, but we have reason to believe that many such issues still persist also in the current LOD cloud. In fact the example BIOMODELS ontology dataset now exists on different places in the LOD cloud, within BIO2RDF, within BIOPORTAL, but also as an RDF dataset directly published by EBI at `ftp://ftp.ebi.ac.uk/pub/databases/RDF/biomodels/` 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 [43] and [41].

Last, but not least, as an open problem, links in one dataset always refer to a particular *version* of the linked dataset, which if not archived cannot be guaranteed to persist or being dereferenceable in the future. For a more sustainable version of 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, to be 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 [33, 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 [34] 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 [26] and references therein in terms of starting points; although no single agreed proposal exists at this point for how to publish

versioned RDF archives we again refer to possible HDT-based solutions, particularly enabled through the recent extension of HDT to handle quads [25].

4.2 Non-Technical Challenges

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

Completeness/Consistency. Several well-known and important RDF datasets are missing in the LOD cloud, e.g. EBI RDF is not there (plus various other well-known data bases from the biomedical and life sciences domain), 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, e.g., 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.

Trust. Besides the pervasive issues of availability and reliability, developers are rightfully worried that the published data in the cloud is not kept up to date, and as such the technical issues mentioned above might overall give rise to (or have already given rise to, possibly) doubts on the technology 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 this project up and alive, by creating sustainable publishing and monitoring processes.[30]

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 that actually arise 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 having been created in isolation, by closed groups, without inviting collaboration or being based on infrastructures to share and evolve those exports jointly. We feel that this issue can only be solved by a more collaborative, and truly open governance.

[30] Of course, with the alternative to eventually re-brand it under a different name after survival of an "LOD winter" from unfulfilled expectations)

Documentation and Usability. Besides the technical accessibility 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 existing vocabularies; and (4) Supporting and promoting the use of developer-friendly formats, such as JSON-LD.

In addition, in terms of positive examples, we would again like to name the aforementioned HDT and TPF projects, as well as useful SPARQL query editing tools such as YASGUI [51] or Wikidata's query interface, which have appeared in the last two years; we need more tools like those.

Funding & Competition. Last, but not least, while the EU and other funding agencies have supported our endeavor to create 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, which prevent research groups working on overlapping topics from collaboration, and rather stimulate an environment of isolated closed research than open collaboration to jointly 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 [62] in building up a completely independent SPARQL endpoint monitoring framework [66],[31], not even citing SPARQLES (sic!), which seems unnecessarily duplicating efforts instead of collaboratively developing and maintaining such services.

5 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.

On the bright side, the biomedical community has been very successful in using OWL and Semantic Web technologies for the management of large biomedical vocabularies and ontologies. The poster child is the development of the Gene Ontology (GO) [1,4], arguably the most important biomedical ontology in existence and with the highest impact in the community. In our opinion, the main factors contributing to the big success of the GO are: (1) Having a dedicated and very active development team

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

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, not only about the ontology itself, but also about how to use it in applications targeted to domain users, as well as documentation about the processes for building and maintaining the GO; (4) Using a principled approach for developing the ontology; (5) Using automated pipelines to check and ensure the quality of the ontology (and also document the whole process).

Our hope is that the Linked Data community can learn from the development of GO, 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 these challenges, 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 and 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 – the **links to specific versions of** these **other datasets**.

In fact, the original goal we had with this paper was to demonstrate how you can auto-generate LOD clouds from a set of HDT dumps, but as we got stuck already with so many of the other issues mentioned throughout what you just read, we got – let's say – dragged away a bit; this original goal is still on our agenda for future work: 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 hope that the upcoming DeSEMWEB workshop at ISWC2018, but also initiatives like the recently US-founded "Open Knowledge Network"[32] initiative or the upcoming Dagstuhl seminar on "New Directions for Knowledge Representation on the Semantic Web"[33] will provide platforms to openly discuss such a fresh start.

Finally, we admit the present paper is a bit long for a vision paper: we decided not to shorten it and rather would like it to be understood as a vision paper with a strong survey character, as we considered it necessary to give an – even if not exhaustive – account to prior work that should be potentially re-considered for jointly addressing the challenges mentioned herein.

Acknowledgements We thank Dan Brickley, Sarven Capadisli, and Denny Vrandečić for comments on the first revision of this paper, which led to some additional footnotes making it even longer. Axel Polleres' work was supported under the Distinguished Visiting Austrian Chair program hosted by The Europe Center of Stanford University. Javier Fernandez' work was supported by the EC under the H2020 project SPECIAL and by the Austrian Research Promotion Agency (FFG) under the project "CitySpin".

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

[33] https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=18371

References

1. Gene ontology consortium: going forward.
2. 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.
3. 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.
4. M. Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. DOI:10.1038/75556.
5. W. Beek, L. Rietveld, H. R. Bazoobandi, J. Wielemaker, and S. Schlobach. Lod laundromat: a uniform way of publishing other peoples dirty data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 213–228. Springer, 2014.
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 *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)*, LNCS. Springer, Oct. 2014.
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.
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.
11. E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant. Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, volume 4, pages 217–241. Elsevier, 2008.
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 *Proceedings of the International Semantic Web Conference (ISWC)*. Springer, 2014.
15. B. Bushman, D. Anderson, and G. Fu. Transforming the medical subject headings into linked data: creating the authorized version of MeSH in RDF. *Journal of library metadata*, 15(3-4):157–176, 2015.
16. A. Callahan et al. Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In *The Semantic Web: Semantics and Big Data*, pages 200–212. Springer, 2013.
17. 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.
18. M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semant. web*, 2(1):55–63, 2011.
19. D. De Witte, L. De Vocht, F. Pattyn, H. Constandt, E. Mannens, and R. Verborgh. Scaling out federated queries for life sciences data in production. In *SWAT4LS*, pages 1–10, 2016.
20. J. Debattista. *Scalable Quality Assessment of Linked Data*. PhD thesis, Rheinische Friedrich-Wilhelms-Universitat Bonn, Oct. 2016.

21. J. Debattista, C. Lange, S. Auer, and D. Cortis. Evaluating the quality of the lod cloud: An empirical investigation. *Semantic Web*, (Preprint):1–42, 2017.
22. J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics*, 47:20–33, 2016.
23. 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 ACM International Conference on Information and Knowledge Management (CIKM)*, pages 652–659. ACM, 2004.
24. 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.
25. J. D. Fernandez, M. A. Martínez-Prieto, A. Polleres, and J. Reindorf. HDTQ: Managing RDF datasets in compressed space. In *Proceedings of the European Semantic Web Conference (ESWC)*. Springer, 2018.
26. J. D. Fernandez, J. Umbrich, A. Polleres, and M. Knuth. Evaluating query and storage strategies for RDF archives. *Semantic Web*, 2018. to appear (accepted for publication).
27. G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton. Pubchemrdf: towards the semantic annotation of pubchem compound and substance databases. *Journal of cheminformatics*, 7(1):34, 2015.
28. 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.
29. O. Görlitz and S. Staab. Splendid: Sparql endpoint federation exploiting void descriptions. In *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, pages 13–24. CEUR-WS. org, 2011.
30. 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.
31. A. Hasnain, M. R. Kamdar, P. Hasapis, D. Zeginis, C. N. Warren, H. F. Deus, D. Ntalaperas, K. Tarabanis, M. Mehdi, and S. Decker. Linked biomedical dataspace: lessons learned integrating data for drug discovery. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 114–130. Springer, 2014.
32. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6, 2017.
33. 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. From http://aidanhogan.com/docs/thesis/; retr. 2010/10/27.
34. 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.
35. 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.
36. A. Jentzsch, J. Zhao, O. Hassanzadeh, K.-H. Cheung, M. Samwald, and B. Andersson. Linking Open Drug Data. In *Proceedings of I-SEMANTICS*, 2009.
37. 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.
38. T. Käfer, A. Abdelrahman, J. Umbrich, P. OByrne, and A. Hogan. Observing linked data dynamics. In *Proceedings of Extended Semantic Web Conference (ESWC)*, pages 213–227. Springer, 2013.
39. M. R. Kamdar et al. An Ebola virus-centered knowledge base. *Database*, 2015, 2015.
40. M. 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)*, 2017.

41. 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.
42. M. R. Kamdar, S. Walk, T. Tudorache, and M. A. Musen. Analyzing user interactions with biomedical ontologies: A visual perspective. *Journal of Web Semantics*, 49:16 – 30, 2018.
43. M. R. e. a. Kamdar. An empirical meta-analysis of the life sciences (linked?) open data cloud, 2018. Unpublished Manuscript, available at http://onto-apps.stanford.edu/lslodminer.
44. J. Li and Z. Lu. Pathway-based drug repositioning using causal inference. *BMC bioinformatics*, 14(16):S3, 2013.
45. 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.
46. J. P. McCusker, M. Dumontier, R. Yan, S. He, J. S. Dordick, and D. L. McGuinness. Finding melanoma drugs through a probabilistic knowledge graph. *PeerJ Computer Science*, 3:e106, 2017.
47. A. Miles and S. Bechhofer. Simple knowledge organization system reference. Recommendation, W3C, August 18 2009.
48. 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.
49. H. Paulheim and A. Gangemi. Serving dbpedia with DOLCE - more than just adding a cherry on top. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, pages 180–196, 2015.
50. 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.
51. L. Rietveld and R. Hoekstra. The YASGUI family of SPARQL clients. *Semantic Web*, 8(3):373–383, 2017.
52. A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall. Life sciences on the semantic web: the neurocommons and beyond. *Briefings in bioinformatics*, 10(2):193–204, 2009.
53. M. Saleem et al. A fine-grained evaluation of SPARQL endpoint federation systems. *Semantic Web*, (Preprint):1–26, 2015.
54. M. Saleem, A. Hasnain, and A.-C. N. Ngomo. Largerdfbench: a billion triples benchmark for sparql endpoint federation. *Journal of Web Semantics*, 2018.
55. M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. N. Ngomo. Big linked cancer data: Integrating linked tcga and pubmed. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:34–41, 2014.
56. 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.
57. A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. Fedx: Optimization techniques for federated query processing on linked data. In *International Semantic Web Conference*, pages 601–616. Springer, 2011.
58. M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96), 2011.
59. 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/.
60. S. Tramp, P. Frischmuth, T. Ermilov, and S. Auer. Weaving a Social Data Web with Semantic Pingback. In *Proceedings of the Knowledge Engineering and Knowledge Management by the Masses (EKAW)*, volume 6317 of LNAI, pages 135–149. Springer, 2010.
61. 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.

62. 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.
63. 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.
64. D. Vrandecic and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.
65. A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, et al. Open PHACTS: semantic interoperability for drug discovery. *Drug discovery today*, 17(21-22):1188–1198, 2012.
66. A. S. Yasunori Yamamoto, Atsuko Yamaguchi. Umaka-Yummy Data: A Place to Facilitate Communication between Data Providers and Consumers. In *Proceedings of the International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, volume 1795 of *CEUR*. 2016.