A More Decentralized Vision for Linked Data

Axel Polleres[1,2], Maulik R. Kamdar[1], Javier D. Fernández[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. We claim that ten years into Linked Data there are still many unresolved challenges towards arriving at a truly machine-readable *and* decentralized Web of data. With a focus on the the biomedical domain—currently, one of the most promising "adopters" of Linked Data, we highlight and exemplify key technical and non-technical challenges to the success of Linked Data, and we outline potential solution strategies.

1 Introduction

Linked Data promises 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 enthusiasm through the so-called "LOD cloud". If we measure the number of datasets published according to the *four linked data principles* [2] and that link to each other, we find evidence of growth and prosperity, with currently over 1,184 datasets, which sounds like good news.

However, there are still serious barriers to consume and use this data: we identify some serious challenges in consuming and using Linked Data from the "cloud". In this paper, which is accompanied by an extensive technical report [9], (i) we argue why the biomedical domain has had such a strong interest in ontologies and Linked Data, (ii) we highlight perceived main challenges important to be addressed to make Linked Data more usable; finally, (iii) we conclude with a call to collaboratively and openly address these challenges as a community in order to (re-)decentralize the Semantic Web.

2 Background: the LOD Cloud

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. The metadata this LOD cloud relies on, comprises 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 [1] descriptions) or an XML sitemap.

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. We would expect to see a plethora of applications of LOD in biomedicine, however, they are conspicuously missing.

Several key biomedical initiatives use Semantic Web technologies for the integration of diverse datasets in fields, such as, neurosciences, cancer research, and drug discovery. 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 .

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. The National Center for Biotechnology Information (NCBI) publishes the entire PubChem data repository of biological assays and activities of compounds as RDF data dumps. The National Library of Medicine's (NLM) Linked Data Infrastructure Working Group released an RDF version of the Medical Subject Headings (MeSH) taxonomy. All these initiatives are highly promising and illustrative for the LOD adoption.

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. However, For most biomedical researchers (and autonomous agents) querying against the LOD sources in the wild does not bear fruitful results. We have documented some main difficulties in this domain from our own experiences [6,7].

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, 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 see the following major challenges when attempting to use Linked Data, where 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*).

Availability and resource limits. 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 resource URLs; i.e., almost half, 635 dataset descriptions contain no dereferenceable resource URLs that would point to data at all.

Concerning SPARQL endpoints, among the mentioned 444 potential SPARQL endpoint URLs in metadata, only 252 responded at all. Only 195 responded ”true” to a simple `ASK {?S ?P ?O}` query, 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,[3], or the Dynamic Linked Data Observatory,[4] 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.

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.

We also see sizes of triples reported on the LOD cloud diverging from what a simple `SELECT (COUNT (*) AS ?C) WHERE {?S ?P ?O}` to their endpoint reports in various examples.

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.

Another potential challenge in terms of size and scalability is the amount of duplicates in current dumps: a lot of triples are actually duplicated across 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.

[3] http://sparqles.ai.wu.ac.at/
[4] http://km.aifb.kit.edu/projects/dyldo/

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 [4], a compressed and queryable RDF format, as well as (ii) Triple Pattern Fragments (TPF) endpoints [10] 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,[5] on the other hand is an already compressed dump-format that allows such triple pattern queries without decompression and also guarantees duplicate-freeness. Notably, 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.

As for VoID, it is suggested to place the dataset description under `/.well-known/void` in the root directory of a Web-server; among all 881 hostnames mentioned in URLs in datahub.io's metadata, 159 respond to an HTTP Get with the recipe, at least 75 of which though seem to be HTML responses, and only 56 valid RDF. 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.

Tummarello et al. had proposed an extension of the Sitemaps protocol to link to RDF datasets specifically [3], that has been implemented in Sindice [8]. 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. So, overall, while semantic sitemaps are only used for a marginal 5% of the datasets in datahup.io, they seem to be fairly consistent.

The SPARQL1.1 specification, suggests that endpoints return a service description document at the service endpoint when dereferenced using the HTTP GET operation without any query parameter string. 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).

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.

[5] http://rdfhdt.org

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.[6] We emphasize, 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 be indispensable. Again, HDT could serve as a basis for scalable, out-of -the-box implementations of quality 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

[6] 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.

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.

As an additional 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.

As for archiving and versions, we refer to [5] and references therein in terms of starting points.

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:

- Completeness/Consistency. Several well-known and important RDF datasets are missing in the LOD cloud; the burden of manually and pro-actively needing to provide and maintain LOD cloud metadata on the publisher-side has proven unsustainable.
- Trust. 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. 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.
- 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 could be potentially related to many of our exports and datasets having been created in isolation. We feel that this issue can only be solved by a more collaborative, and truly open governance.
- 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.

– Funding & Competition. (i) Cross-continental research initiatives are not being funded by the EU or the US easily, (ii) EU project consortia are typically being judged by complementary partner expertise: both these factors 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.

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), 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 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” [2]:

> 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"[7] initiative or the upcoming Dagstuhl seminar on "New Directions for Knowledge Representation on the Semantic Web"[8] 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 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".

References

1. 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.
2. T. Berners-Lee. Linked Data. W3C Design Issues, July 2006. From http://www.w3.org/DesignIssues/LinkedData.html; retr. 2018/06/01.
3. 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.
4. 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.
5. 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).
6. M. R. Kamdar, T. Tudorache, and M. A. Musen. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic Web*, (Preprint):1–19, 2016.
7. 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.
8. 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.
9. A. Polleres, M. R. Kamdar, J. D. Fernandez, T. Tudorache, and M. A. Musen. A more decentralized vision for linked data. Technical report, Working Papers on Information Systems, Information Business and Operations. Available at http://epub.wu.ac.at/6371/.
10. 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.

[7] http://ichs.ucsf.edu/open-knowledge-network/
[8] https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=18371