



Stanford Center for Biomedical Informatics
Research

THE NON-EXISTING LOD CLOUD

AND HOW IT COULD FINALLY BE (RE-)CREATED

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

Website: <http://polleres.net>

Twitter: @AxelPolleres

About myself: proud member of the Pedantic Web group (ranting unsuccessfully about Linked Data Quality) since 2009 ...

Welcome to the Pedantic Web Group



[pe·dan·tic /pe'dæntɪk/:](#)


overly concerned with formal rules and trivial points of learning

Your data is broken.

You publish RDF data on the web, and thereby contribute to our shared passion: the emerging global information space that we call the web of data. Thank you for that! Thank you for sharing your data!

But your data is broken. Syntax errors, unescaped characters, encoding problems, broken links, ambiguous identifiers, undefined vocabulary terms, mismatched semantics, unintended inferences: If you publish anything on the web of data, chances are that there is some problem.

Digital Enterprise Research Institute www.deri.ie




Linked Broken Data?


Dr Axel Polleres
Digital Enterprise Research Institute, National University of Ireland, Galway

Based on joint work with Aidan Hogan, Andreas Harth, Renaud Delbru, Giovanni Tummarello, Stefan Decker

© Copyright 2009 Digital Enterprise Research Institute. All rights reserved.



OÉ Gaillimh
NUI Galway



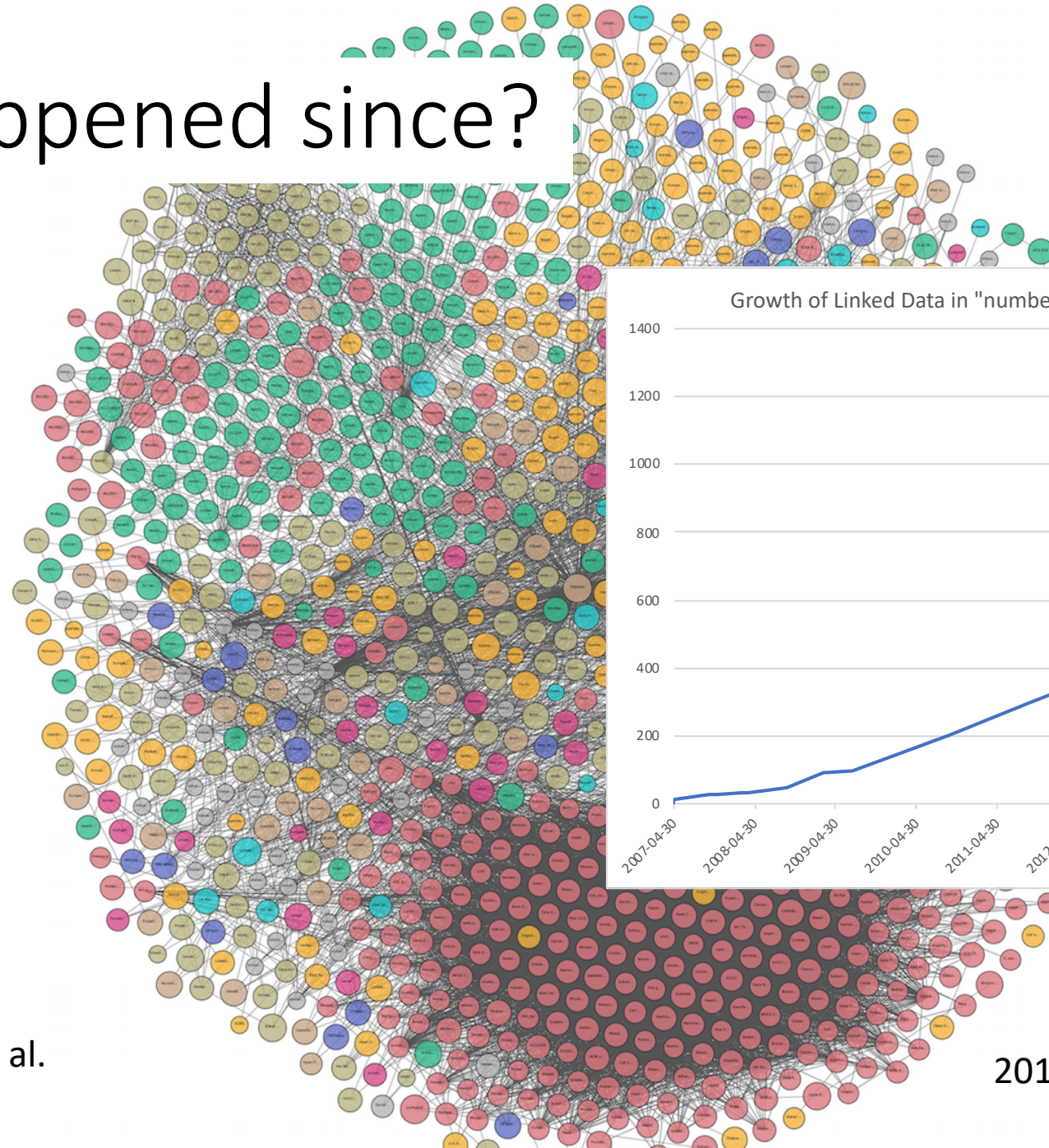
sfi
science foundation ireland
fondúireacht eolaíochta éireann

Linked Data - The four holy commandments:

Linked Data Principles

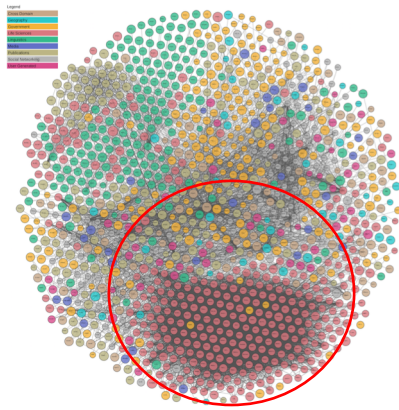
- **LDP1:** use URIs as names for things
- **LDP2:** use HTTP URIs so those names can be dereferenced
- **LDP3:** return useful – RDF? – information upon dereferencing those URIs
- **LDP4:** include links using externally dereferenceable URIs.

What happened since?



Use case: Drug repurposing

- Drug discovery costs have increased exponentially
 - \$2.85 billion for a bio-pharma company to research and sell a new drug
- Alternative solution: looking for novel uses of existing drugs
 - all information about a drug entity
 - drug–protein target interactions
 - publications mentioning the drug
 - adverse reactions
 - downstream drug targets located in biological pathways,
 - essays that test the binding activity of the “drug active ingredient”

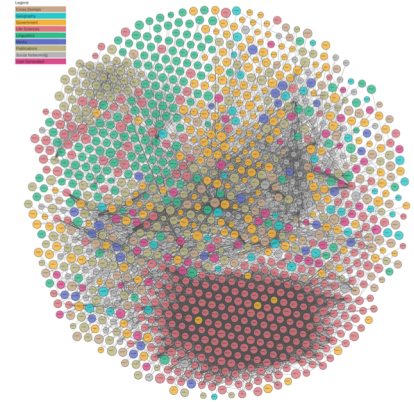


= 0 drug repurposing (AFAIK)

Why it is still (too?) hard... what about Web-scale queries

- E.g. retrieve all entities in LOD with the label “WBGene00000001 (aap-1)”

```
select distinct ?x {  
  ?x rdfs:label "WBGene00000001 (aap-1)"  
}
```

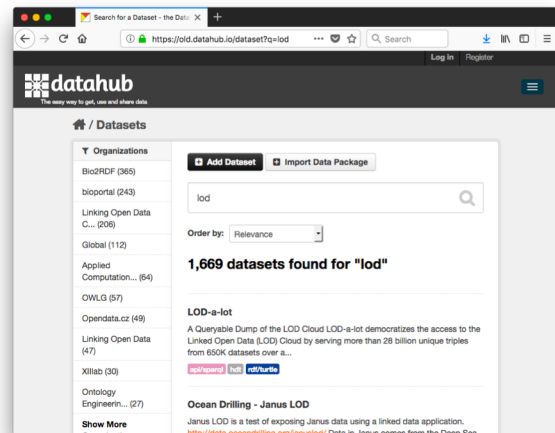
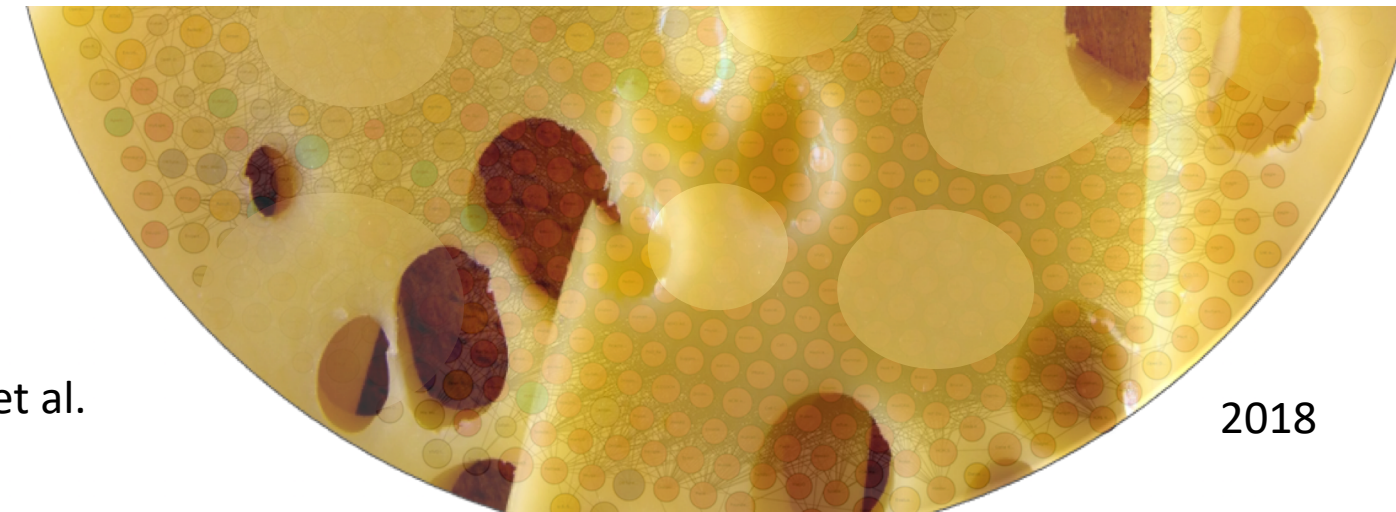


- Options:
 - Crawl and index LOD locally (-no for a single query-)
 - Follow-your-nose (where should I start?)
 - Federated querying (as good as the endpoints you query)
 - Use a seed (e.g. datahub.io or **LOD Laundromat**) as a “good approximation” (still querying potentially many datasets, 650K in LOD Laundromat)

What really happened since?



Among the mentioned **5435** resource URLs in the 1281 "LOD"-tagged datasets on **old.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 😞



Not only our datasets, but also our services and tools disappear....

ソフトバンク光の転用について

sparql.sindice.com/sparql/

ソフトバンク光の転用

ソフトバンク光の転用について

フレッツ光からソフトバンク光に転用を検討中の方は必見、実際の体験談を交えてわかりやすくそのしくみや手続きを紹介します。何かと賛否が分かれるソフトバンク光への転用の現実を知ってスカッとしちゃいましょう

[トップページ](#) > ソフトバンク光の転用について

ソフトバンクのスマホを手に入れたのでついには自宅のネットもソフトバンク光に、なんて人は多いでしょう。スマホとのセット割でオトクに使えることはわかるんだけど、手続きのことや料金・回線速度や安定感、またサービス面やキャンペーンなども気になりますよね。

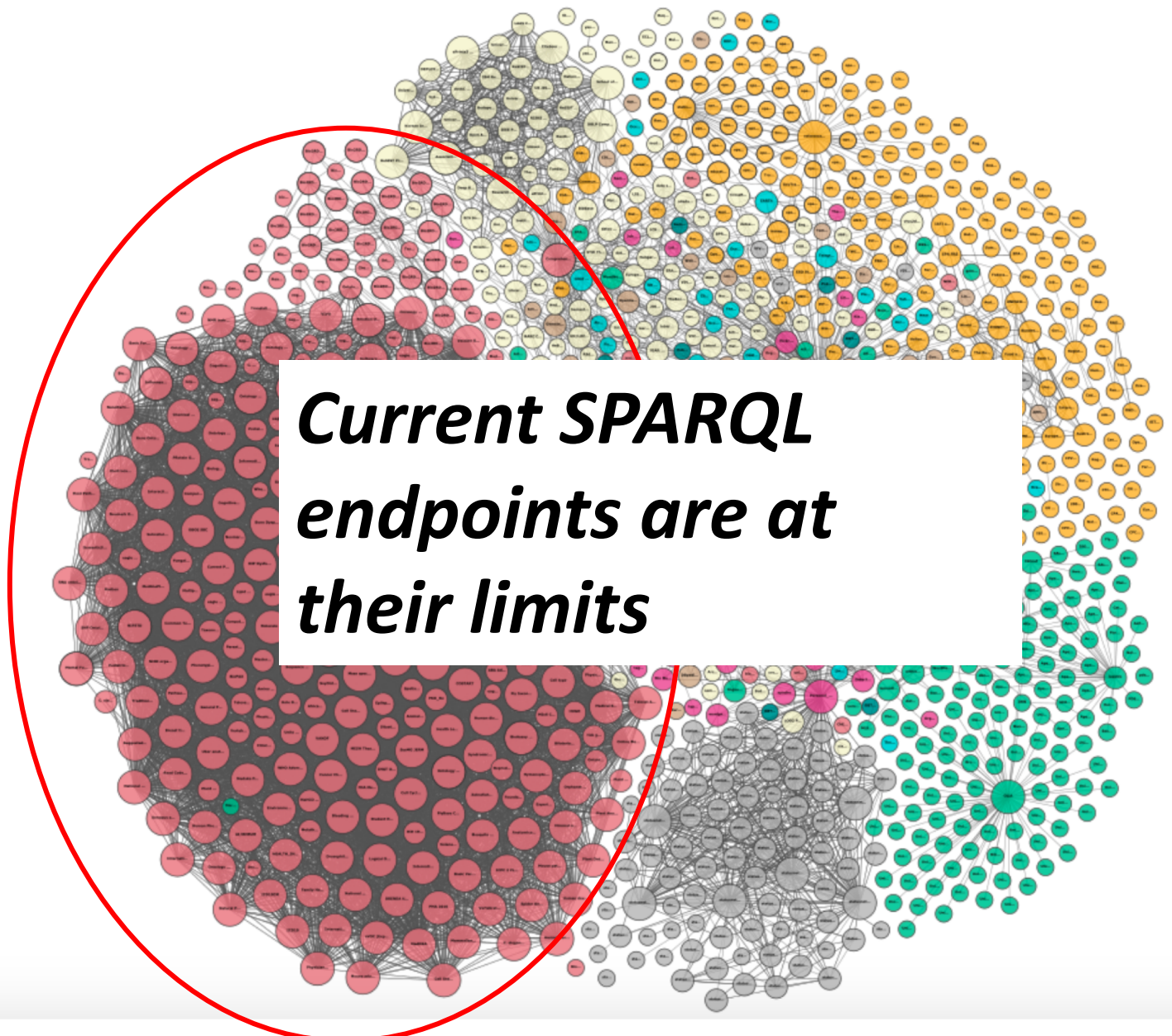
そんな訳で今回は、実際フレッツ光からソフトバンク光に転用した私（筆者：以下省略）自身の体験談を交えて、そのしくみやメリットデメリットを盛り込みながら転用へのモヤモヤをイッキに解決しちゃいます。

[そもそもソフトバンク光に転用とは？](#)

Some concrete challenges...

Current LOD Cloud:

- Challenge 1: Size
 - single datasets (e.g. Pubmed dump has 7b triples, Wikidata [ttl.gz dump](https://www.wikidata.org/wiki/Wikidata:List_of_dump_files) +30GB, 5.7b triples)
 - bigger than a significant rest of the LOD cloud (whole LOD-a-lot experiment 28b triples)
 - Current Triple stores scale probably up to ~20-30b?
 - But:
 - Provide significant bottlenecks in access (e.g. limits for timeouts in wikidata query-services)
 - Bio2RDF endpoint has only ~1b triples
 - Where's the rest??



Current LOD Cloud:

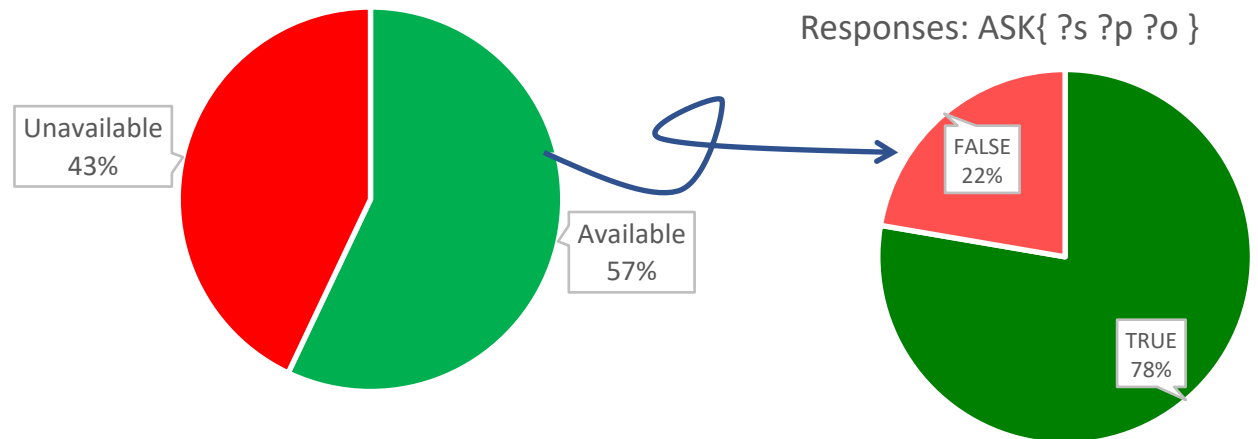
- Challenge 2: SPARQL endpoints availability, findability and limits:
 - <http://pubchem.bio2rdf.org/sparql> is down
 - <http://pubmed.bio2rdf.org/sparql> redirects to
 - <http://download.bio2rdf.org/#/current/pubmed/>
- cf. also: <http://sparqls.ai.wu.ac.at/>

→ Better try dumps?

Current SPARQL endpoints are non-available



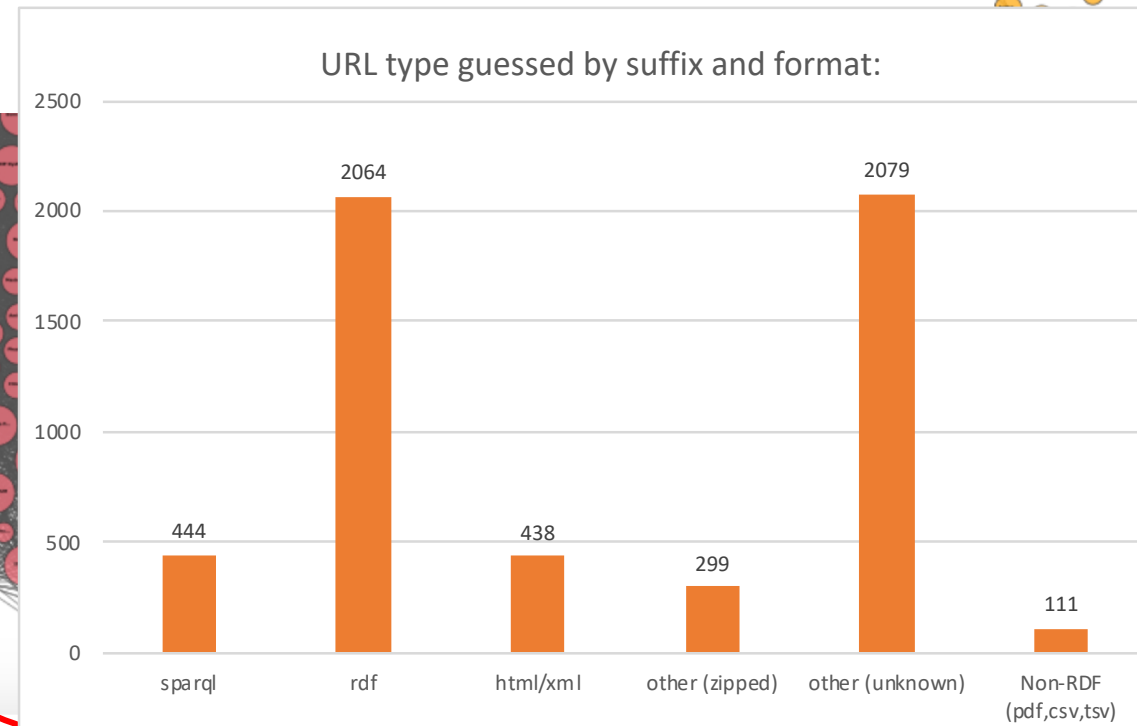
Availability - 440 SPARQL Endpoints



Current LOD Cloud:

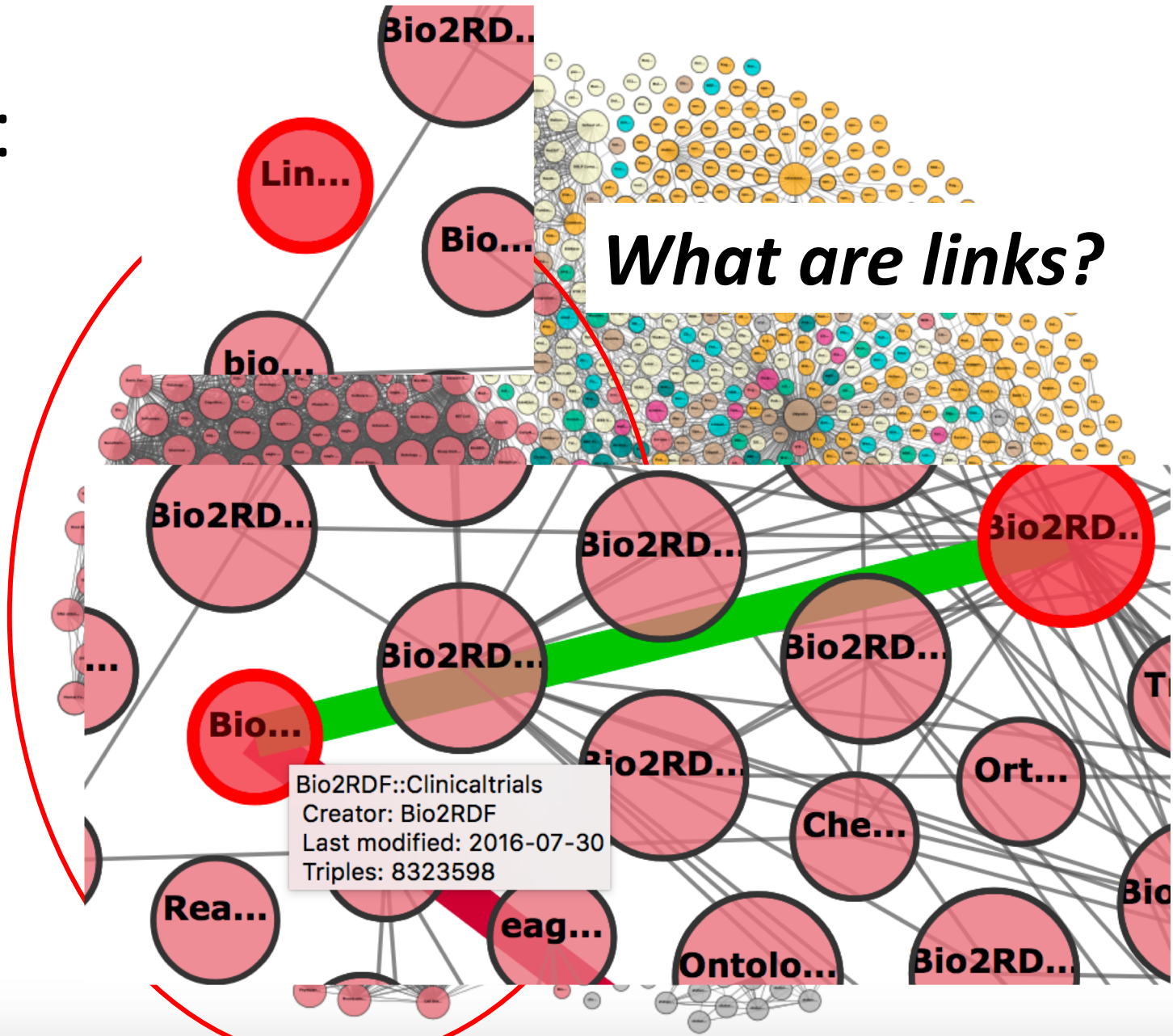
- Challenge 3: APIs not uniformly findable
 - Dumps not easily accessible
 - <http://download.bio2rdf.org/#/current/pubmed/>
 - Javascript page which can only be crawled with a headless browser
 - <https://www.ebi.ac.uk/rdf/datasets/#BulkDownloads> →
 - Bulk download links as result of [SPARQL queries](#) against VOID descriptions, e.g.
 - [<ftp://ftp.ebi.ac.uk/pub/databases/RDF/biomodels/r31/biomodels-rdf.tar.bz2>](ftp://ftp.ebi.ac.uk/pub/databases/RDF/biomodels/r31/biomodels-rdf.tar.bz2)
 - Various compressions used, etc.
 - SPARQL service description vocabulary does NOT have an attribute for pointing to alternative dumps or proper description of limitations imposed

Current SPARQL endpoints don't provide metadata nor point to accessible dumps and ... too many formats:



Current LOD Cloud:

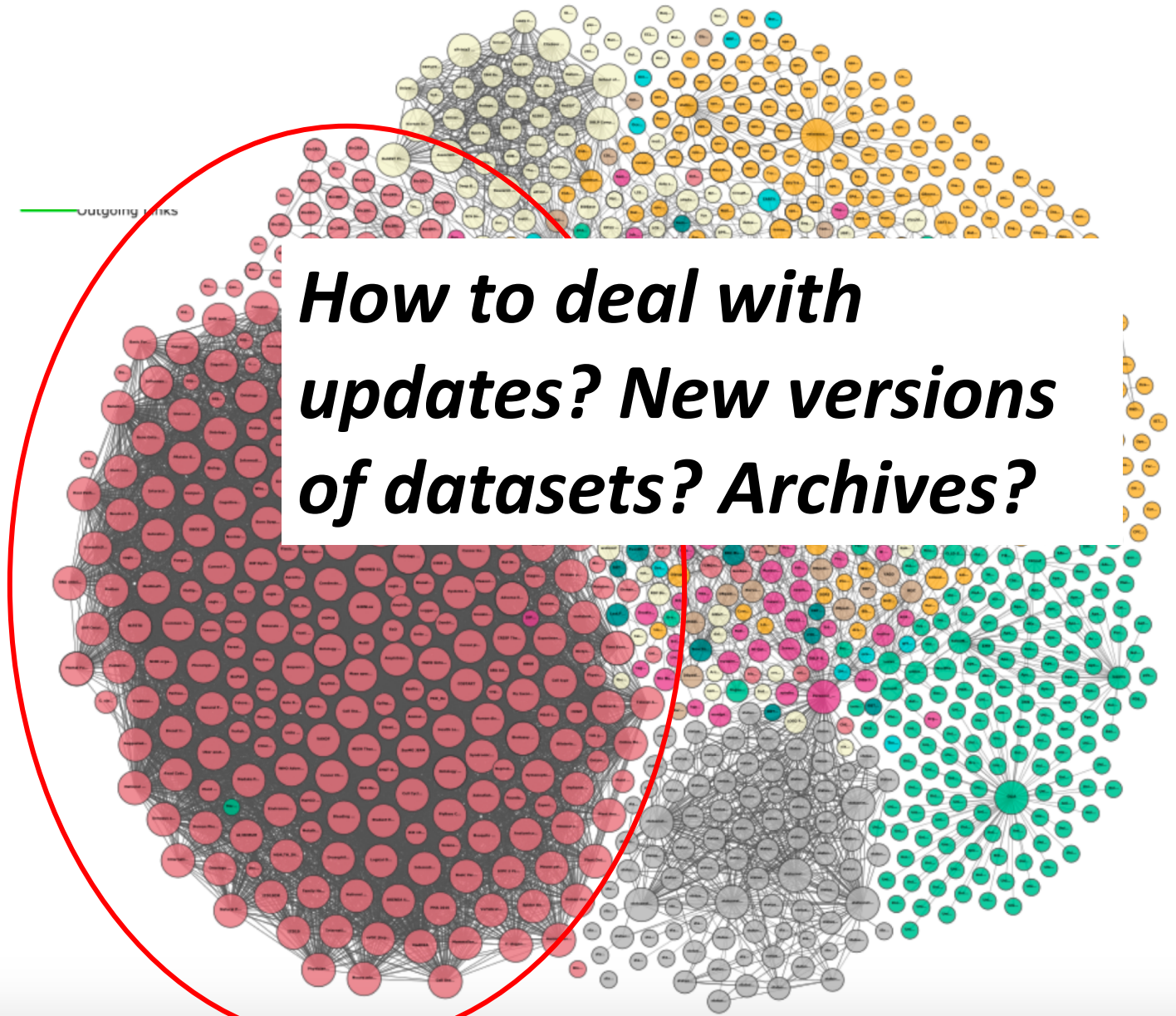
- Challenge 4: What do Links in LOD cloud actually mean?
 - What are in-links/out-links?
 - Computed from meta-data on datahub.io
 - But description is ambiguous:
 - **Definition:** “either your dataset must use URIs from the other dataset, or vice versam”
- What does it actually mean?
 - Ontology reuse?
 - Instance Links?
 - Joint reuse of entities from 3rd dataset?
 - Who does a namespace belong to?



Current LOD Cloud:

- Challenge 5: Completeness/consistency
- Well known RDF datasets missing
 - E.g. EBI RDF not there (plus around 10 other well known Bio data bases), or even [wikidata not there](#) (sic!)
- Datasets no longer available or moved elsewhere... how do I find them?
 - How does the party Linking to a dataset get notified that the dataset has been updated?

→ ***Needs monitoring & ability to link to versions!***



Last, but not least: Non-Technical Challenges



Other non-technical challenges

- The **steep learning curve**

- Often frustrating experiences with using LD
- Data publishers lack tools and guidelines to help them discover and reuse existing content
- How to perform sophisticated (federated) SPARQL queries, without minimal automated support?

- **Trust & Sustainability**

- Outdated, once-off RDF exports
- Doubts of the maintenance: will dataset X be kept up to date?

- **Documentation and Usability**

- Often inexistent or poor documentation, support or maintenance

- **Funding**

- No cross-continental projects, no overlapping topics (rather complementary)
- Duplicated services among communities, e.g. SPARQLES and <http://yummydata.org/endpoints> (life sciences)

Still there is hope! Brave PhD candidates defend and “stitch” the Web of Data!

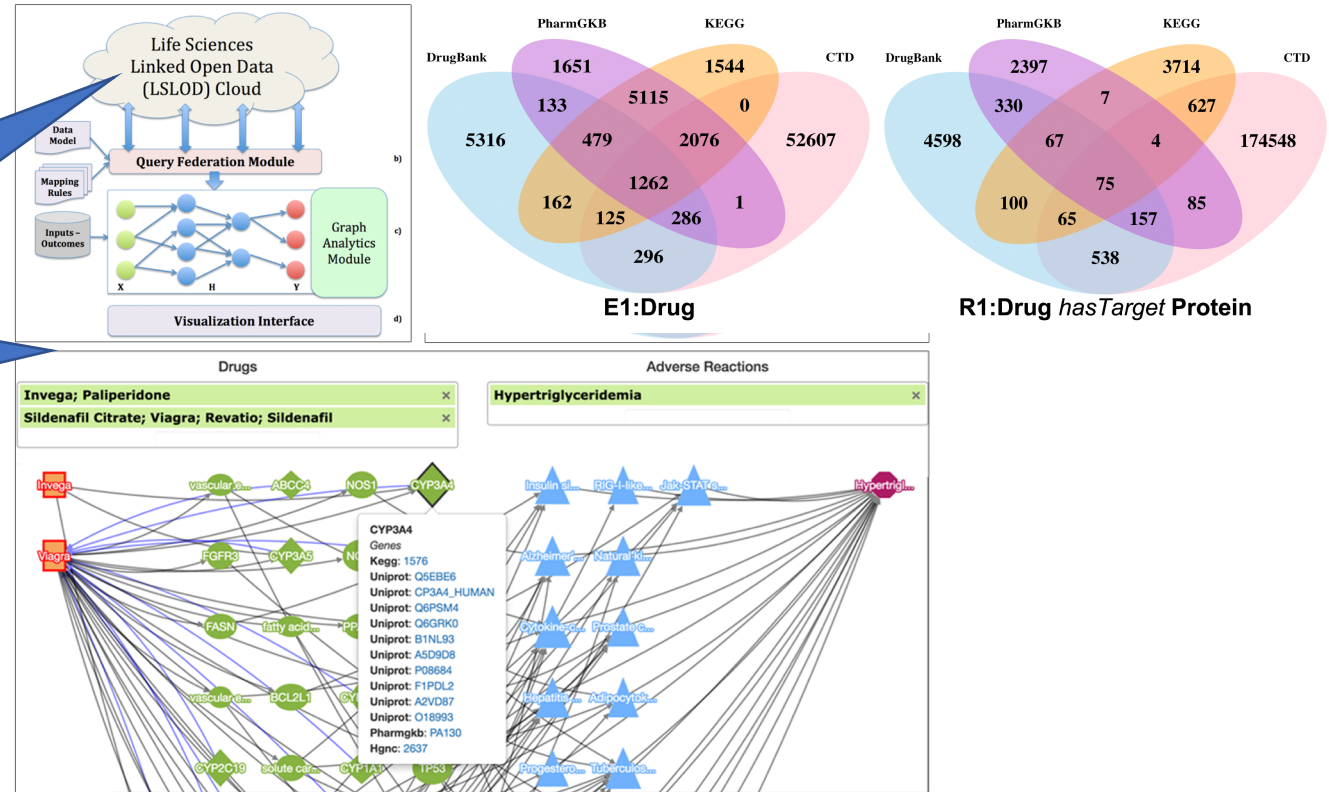


***Phlegra
Spiderman***



What did Maulik solve in his PhD?

Advances in Drug Repurposing through the integration of different LOD sources!



- **Good news:** Great use cases out there for Linked Data applications!
- **Bad news:** needs a superhero with a PHD to “stitch” & integrate the Web of data

Conclusion:

The LOD cloud is as messy as my slides 😞

...

It is **NOT** a machine-readable entry-point to the Web of Linked Data

Some good starting point (but not yet a solution)

- LDF+HDT: “a swiss-army-knife for large RDF datasets”
 - Emerged from the PhD thesis of another Linked Data superher
- Provides a **uniform compressed exchange format for dumps**
- Enables **Linked Data fragments endpoints**
(= of-the-box lightweight (“SPARQL light”) endpoints)
- Keeps **data and metadata together** and in sync (in header)
 - MetaData such as dataset size, number of links could be computed at publishing time.
 - Implicit notion of “dataset”: 1 dataset = 1 HDT file
- Active developer & user community (some are in the room!)



#LDFman



Mr. LOD
Laundromat



Captian HDT

Active Research and Development in HDT

- HDTQ: Enable quads & and versioning:

Semantic Web 0 (0) 1
IOS Press

accepted at SWJ

Evaluating Query and Storage Strategies for RDF Archives

Javier D. D. Fernández^{a,*} Jürgen Umbrich^a Axel Polleres^a Magnus Knuth^b
^a *Vienna University of Economics and Business, Vienna, Austria*
Email: {javier.fernandez,juergen.umbrich,axel.polleres}@wu.ac.at
^b *Hasso Plattner Institute, University of Potsdam, Potsdam, Germany*
Email: magnus.knuth@hpi.de

Abstract.
There is an emerging demand on efficiently archiving and (temporal) querying different versions of evolving semantic Web data. As novel archiving systems are starting to address this challenge, foundations/standards for benchmarking RDF archives are needed to evaluate its storage space efficiency and the performance of different retrieval operations. To this end, we provide theoretical foundations on the design of data and queries to evaluate emerging RDF archiving systems. Then, we instantiate

... allow to deal with versioned dataset dumps and integration of different datasets in one HDT.

ESWC2018

HDTQ: Managing RDF Datasets in Compressed Space

Javier D. Fernández^{1,2}, Miguel A. Martínez-Prieto³,
Axel Polleres^{1,2,4}, and Julian Reindorf¹

¹ Vienna University of Economics and Business, Austria
² Complexity Science Hub Vienna, Vienna, Austria
³ Dept. of Computer Science, Universidad de Valladolid, Spain
⁴ Stanford University, CA, USA

{javier.fernandez, axel.polleres}@wu.ac.at,
migumar2@infor.uva.es, julian.reindorf@gmail.com

Abstract. HDT (Header-Dictionary-Triples) is a compressed representation of RDF data that supports retrieval features without prior decompression. Yet, RDF datasets often contain additional graph information, such as the origin, version or validity time of a triple. Traditional HDT is not capable of handling this additional parameter(s). This work introduces HDTQ (HDT Quads), an extension of HDT that is able to represent quadruples (or quads) while still being highly compact and queryable. Two HDTQ-based approaches are introduced: Annotated Triples

Active Research and Development in HDT

- k-shortest path queries:

SEMANTiCS 2016

Counting to k or how SPARQL1.1 Property Paths Can Be Extended to Top- k Path Queries*

Vadim Savenkov
Vienna University of Economics and Business
Welthandelsplatz 1
Vienna, Austria 1020
vadim.savenkov@wu.ac.at

Jürgen Umbrich
Vienna University of Economics and Business
Welthandelsplatz 1
Vienna, Austria 1020
juergen.umbrich@wu.ac.at

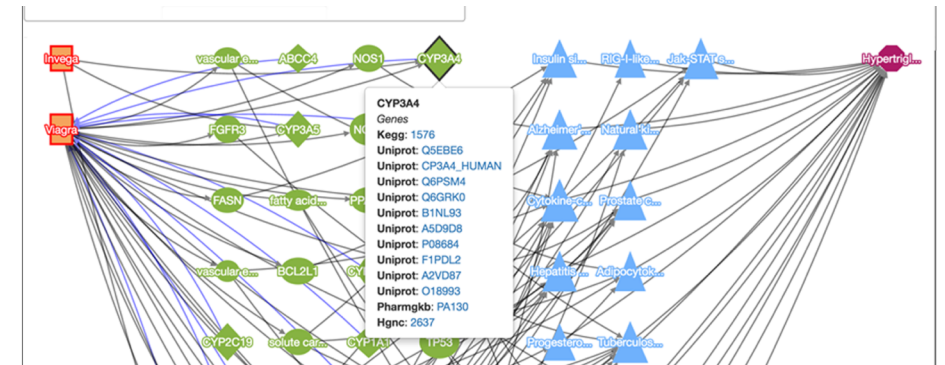
Qaiser Mehmood
Insight Centre for Data Analytics, NUI Galway
P.O. Box 1212
Galway, Ireland 43017-6221
qaiser.mehmood@insight-centre.org

Axel Polleres
Vienna University of Economics and Business and
Complexity Hub Vienna
Vienna, Austria
axel.polleres@wu.ac.at

ACM Reference format:
Vadim Savenkov, Qaiser Mehmood, Jürgen Umbrich, and Axel Polleres. 2017. Counting to k or how SPARQL1.1 Property Paths Can Be Extended to Top- k Path Queries. In *Proceedings of Semantics2017, Amsterdam, Netherlands*, September 11–14, 2017, 7 pages. DOI: 10.1145/3132218.3132239

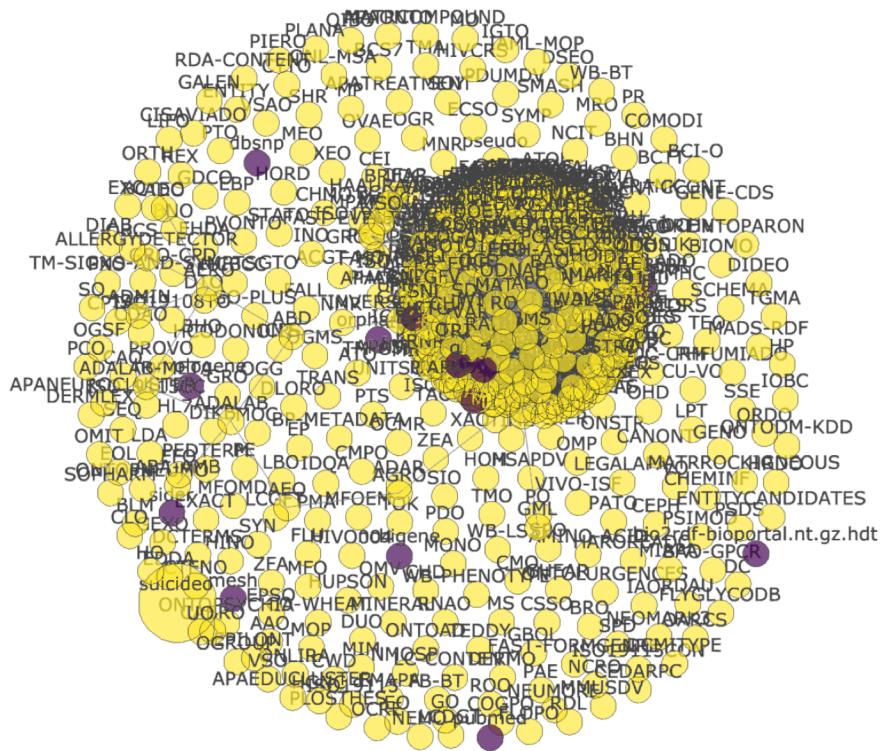
ABSTRACT
While the volume of graph data available on the Web in RDF is steadily growing, SPARQL, as the standard query language for RDF still remains effectively unusable for the basic task of finding paths through the graph between selected nodes. Property Paths, as introduced in SPARQL 1.1 are unfit for this purpose, as they

... could eventually enable use cases like this:



Active Research and Development in HDT

- Work in progress: re-compute a LOD Cloud from a set of HDTs



- Completely auto-created from “HDTized” Bioportal and Bio2RDF
- Idea:
 - Treat each dump-file as a dataset
 - Heuristically assign namespace authority of to datasets heuristically
 - Compute links numbers based on dataset dictionaries using HDT

- (at the moment, uses heuristics to “guess” ownership of namespaces)

The 5th element
and two routes ahead:

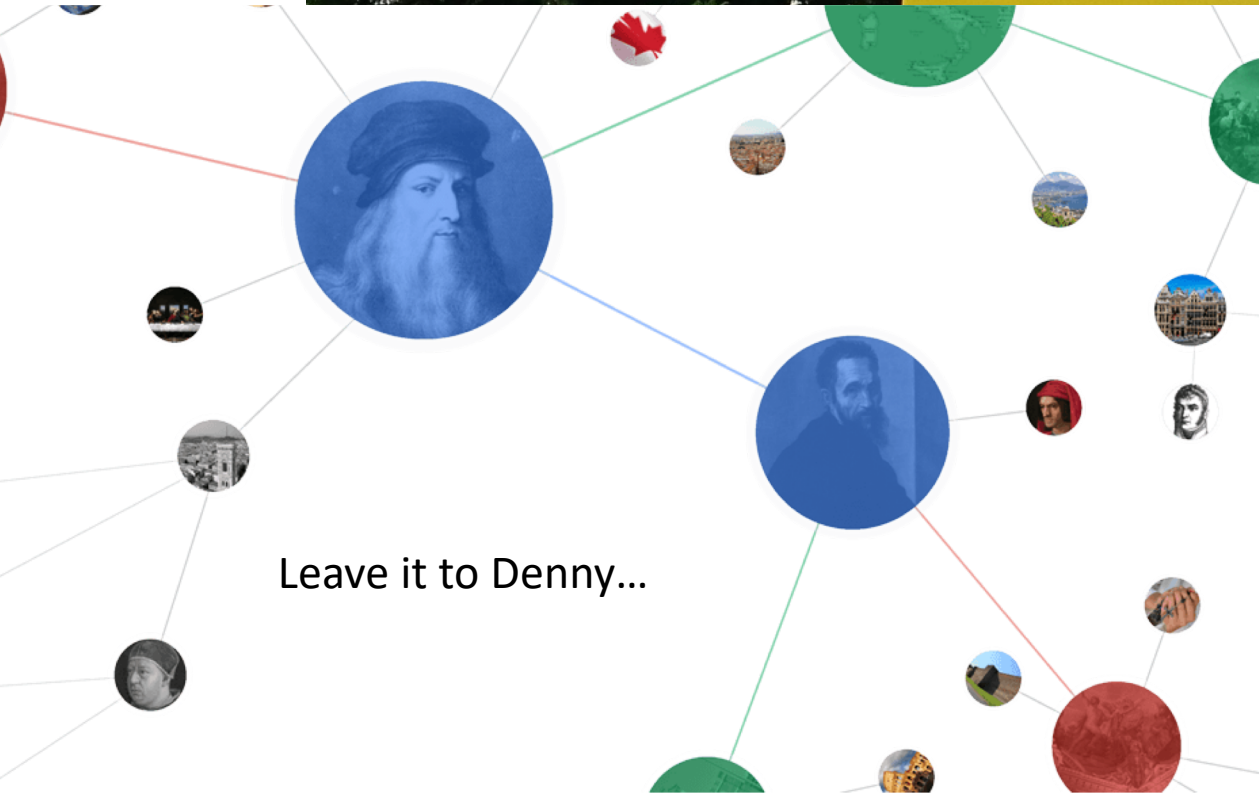


- **LDP5:** Publish your dataset as an **HDT dump**, including **VOID metadata** as part of its header and declaring (i) the **owned namespaces**, (ii) **links to previous versions** of the dataset, (iii) whenever you use namespaces owned by other datasets or ontologies – the **link to specific versions** of these other datasets.

?



?



... or work together?

Can only PhD superheroes integrate Linked Data?
Let's collaborate to make it easier for sheer mortals!



More rants, starting points and a call for collaboration:

Thank you!

Didn't talk about:

- Linking with ML,
- Provenance,
- Quality monitoring, ...

A More Decentralized Vision for Linked Data

Axel Polleres^{1,2}, Maulik R. Kamdar¹, Javier D. Fernandez², Tania Tudorache¹, and Mark A. Musen¹

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.

Backup Slides:



Giovanni Tummarello Actually read it, thanks for the citations 😊 . I was excited to read about GO as possible example of success but disappointed in visiting the site, its pretty abandonware too.

Axel you guys cite problems, but IMO you dont mention the only one: *why* why should people do that.

Without a business reason (broad definition: fsomething that pays you back directly so that you feel compelled and justified - in your non grant non academic work - to do it today as opposed to do other things) nothing can move past the toying around - by people receiving grants to toy around.

(had posted too early previously 😊) now for the second part

Like · Reply · 2w · Edited



Axel Polleres more input, great... thanks! yes, as long as incentives are only acadmic fame, competition among research groups is one of the show-stoppers... this is there in the paper, in section 4.2 - implicitly, but we could maybe make it more explicit.

Like · Reply · 1m · Edited



Dan Brickley "We envisioned a decentralized network of ontologies on the Web that would enable smart agents to seamlessly talk to each other"

I think you mean "we took one useful feature of RDF/RDFS (fine grained vocabulary composition) and elevated it to a cult-... See more

Haha · Reply · 2w



Axel Polleres that would've been even too provocative for me to dare to write, while i like the wording :)))

Like · Reply · 2w



Axel Polleres may i quote you on that?

Like · Reply · 2w



Dan Brickley Axel Polleres sure. it's in fair part my fault, [https://www.w3.org/2001/sw/RDFCore/Schema/200203/...](https://www.w3.org/2001/sw/RDFCore/Schema/200203/) advertised the feature.



W3.ORG
RDF Vocabulary Description Language 1.0: RDF Schema

Like · Reply · Remove Preview · 2w



Axel Polleres FWIW, added your quote in a revised submission, hope that's ok!

Like · Reply · 2w



Tobias Käfer Hi, a very nice overview 😊 although a bit biased towards the LOD cloud. How about all the Linked Data off the LOD cloud? For instance, the Linked Data Platform or the Web of Things?

Like · Reply · 1w

^ Hide 13 Replies



Jürgen Umbrich There exists something else than the lod cloud? And it is called Linked Data? 😊

Like · Reply · 1w



Tobias Käfer Impertinent, how dare people not to register their dataset 😊
Of course Linked Data not registered in the cloud is obvious, but at least the LDP deserves a name-drop I think.

Like · Reply · 1w



Jürgen Umbrich I am seriously confused with all the terminology.
Semantic Web was not about the web but about RDF and OWL, next Linked Data (principles) which is RDF on the Web, next decentralising the SW (again?) and now Linked Data needs decentralization..... See more

Like · Reply · 1w · Edited



Axel Polleres @Tobias, true, that is an aspect... in the paper, we focused on the4 LOD cloud, since it seems to be the single most cited entry point these days.... LDP would have been worth a mention, true, do you have any link that says something about adoption? I wonder how many of the LOD datasets (if any) adhere to the LDP interface. Good one!

Like · Reply · 1w



Axel Polleres p.s.: @Tobias Käfer, we emphasize that we do not mean to be exhaustive, but please, by all means, can you add this as a comment to Openreview? Then, I will try to include it later one!

Like · Reply · 1m · Edited



Dan Brickley "We promised (as a community) to revolu Social Networks in a way that every data subject owns their social network data in decentralized FOAF"

Actually the FOAF project never promised that. We promised to make a machine-readable ve... See more

Like · Reply · 2w

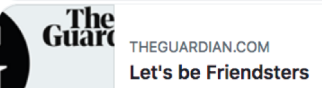


Axel Polleres i would say many understood it like that, at least as having the potential, which you seem to confirm? anyway, the paper is meant to raise discussion, and happy to reword this if it gets accepted to the workshop in the final version... comment appreciated!

Like · Reply · 2w



Dan Brickley Axel Polleres I think the early press - <https://www.theguardian.com/.../2004/feb/19/newmedia.media> - was reasonable, that it was more about better search over public data. The thing that killed that was that none of us had the tools to even deal with L... See more



Like · Reply · Remove Preview · 2w



Axel Polleres thanks for the pointer!!! ...

Like · Reply · 2w



Axel Polleres thanks for the pointer!!!

Like · Reply · 2w



Axel Polleres " it was more about better search over public data" ... pretty much what we're doing now 😊

Like · Reply · 2w



Dan Brickley Axel Polleres the voices of practical RDF were drowned out by a decade of over ontologizing

Like · Reply · 2w

