Where is my URI?

Andre Valdestilhas[1] Tommaso Soru[1] Markus Nentwig[2] Edgard Marx[3]
Muhammad Saleem[1] Axel-Cyrille Ngonga Ngomo[1,4]

[1]AKSW Group, University of Leipzig, Germany

[2]Database Group, University of Leipzig, Germany

[3]Leipzig University of Applied Sciences, Germany

[4]Data Science Group, Paderborn University, Germany

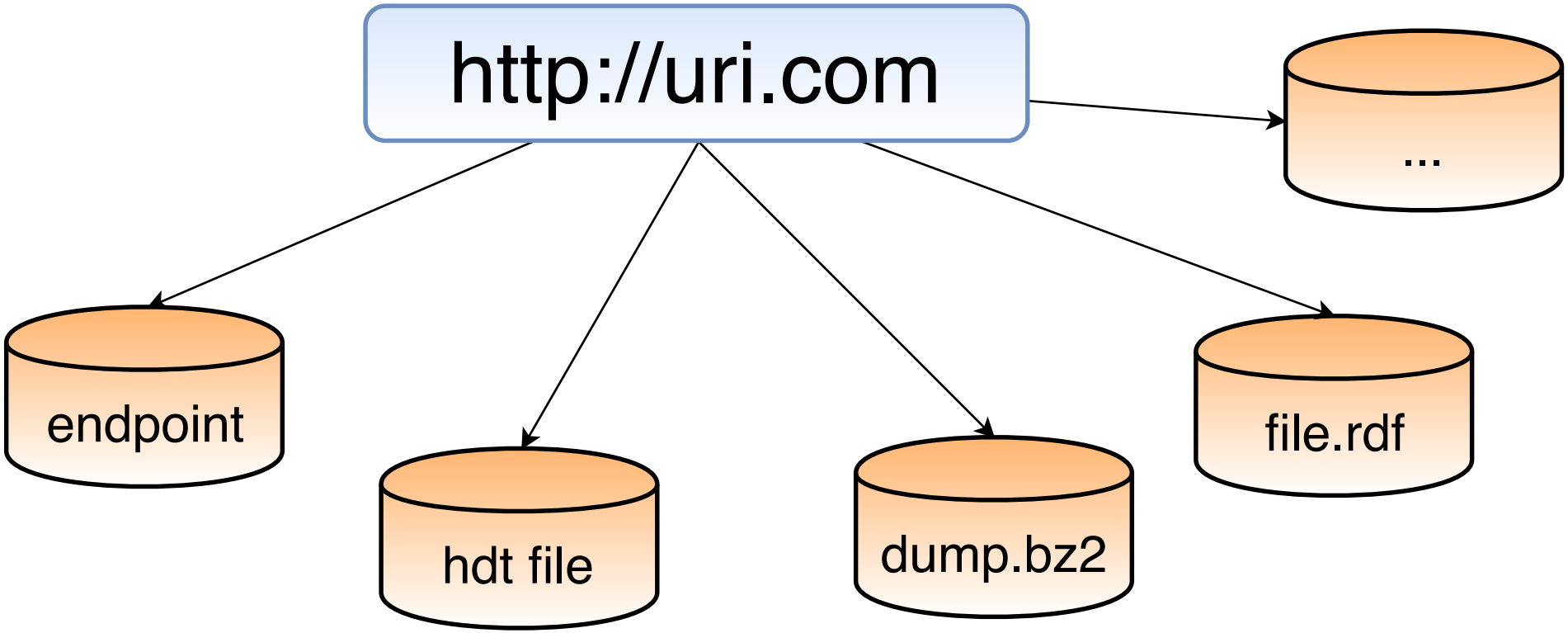
June 29, 2018

Outline

- Motivation
- Approach
- Experiments and Results
- Conclusions and Future work

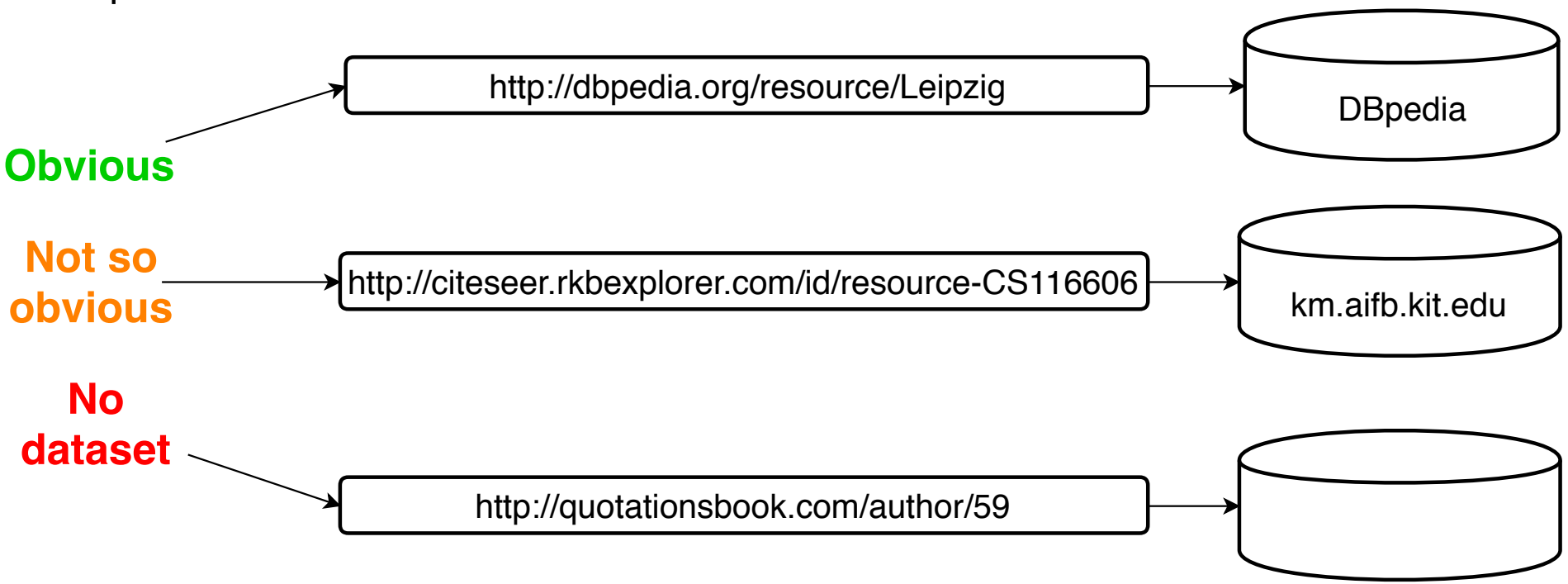
Motivation

Where is my URI?



Motivation

Example:



Motivation

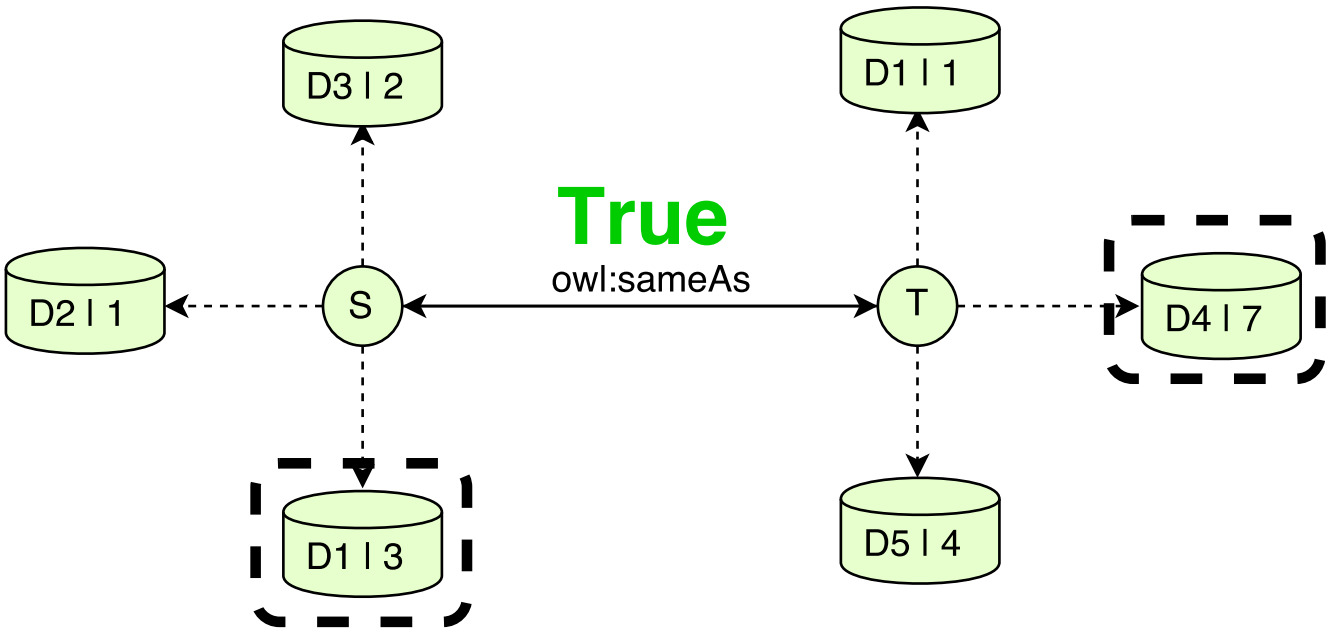
Use cases

Why do we need to know the URI Dataset?

Data quality in Link Repositories

Regenerate mappings using the CBDs to reapply link discovery algorithms in order to validate the mappings.

Part of LinkLion 2.0



Motivation

Motivation

Use cases

Federated Query Processing

Query planning and Source (dataset) selection

WIMU will nd relevant sources against the individual triple patterns of a given SPARQL query

Approach

Goal

Index URIs and their use to enable Linked Data consumers to nd relevant RDF data sources

Rank the datasets proportionally to the number of literals

Keep the provenance of the URI

Approach

Goal

Index URIs and their use to enable Linked Data consumers to nd relevant RDF data sources

Rank the datasets proportionally to the number of literals

Keep the provenance of the URI

Approach

Steps to create the index

Approach

The interface (Web and Json)

Approach

Usage

Service: https://wimu.aksw.org/Find

Parameter	Default	Description
top	0	Top occurrences of the datasets
uri		URI expected to search
link		URL from a linkset
cbd		The URI that will be the origin the CBD
ds		URL to download the dataset

Approach

Examples

Single URI, top 5 datasets

https://wimu.aksw.org/Find?top=5&uri=http://dbpedia.org/resource/Leipzig

Linkset

https://wimu.aksw.org/Find?link=http://www.linklion.org/download/mapping/sws.geonames.org---purl.org.nt

CBD

http://wimu.aksw.org/Find?cbd=http%3A%2F%2Fciteseer.rkbexplorer.com%2Fid%2Fresource-CS116606&ds=http://download.lodlaundromat.org/b7081efa178bc4ab3ff3a6ef5abac9b2?type=hdt

Approach

Relevant points

Rank the datasets from LODStats and LODLaundromat using a score function

is able to process linksets (more than one URI per request)

Experiments and results

Experimental Setup

Hardware: 64 CPU cores, 126 GB RAM, 2 TB hard disk.

Creation of the index: 3 days and 7 hours.

Experiments and results

The heuristic behind the Literals

A sample of 100 DBpedia URIs and occurrences of the top-1 datasets according to the number of Literals

Manually checked: 100% precision

Experiments and results

The heuristic behind the Literals

Top 100 datasets from http://dbpedia.org/resource/Leipzig

Experiments and results

The heuristic behind the Literals

Rank top 5 datasets

Experiments and results

Datasets

	LOD Laundromat	LODStats	Total
URIs indexed	4,185,133,445	31,121,342	4,216,254,787
Datasets checked	658,206	9,960	668,166
Triples processed	19,891,702,202	38,606,408,854	58,498,111,056

LODStats

60% o ine, 14% empty

8% triples with literals as objects are blank nodes

35% online datasets present some errors using Jena

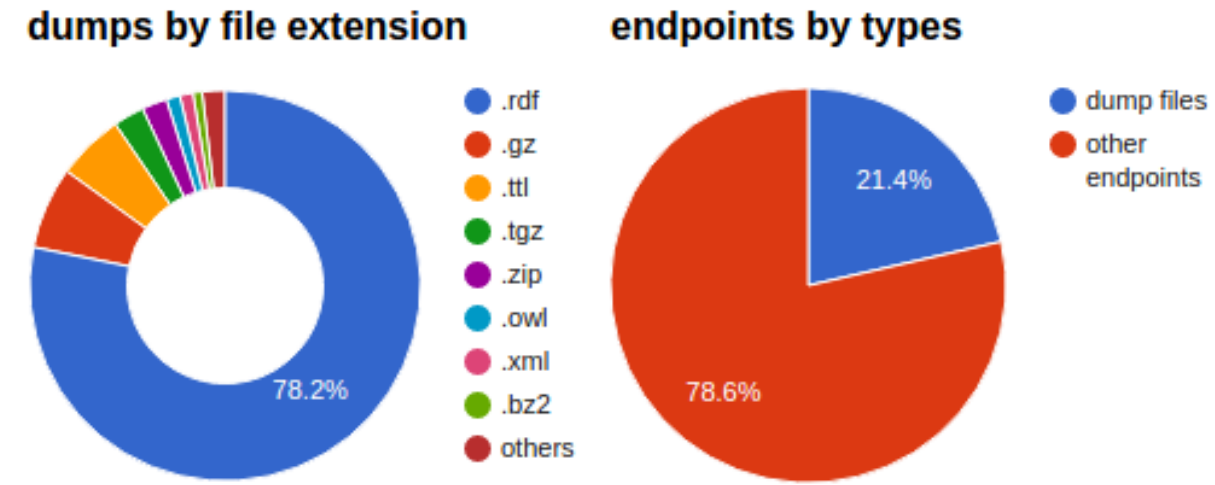
69.8% datasets with parser errors

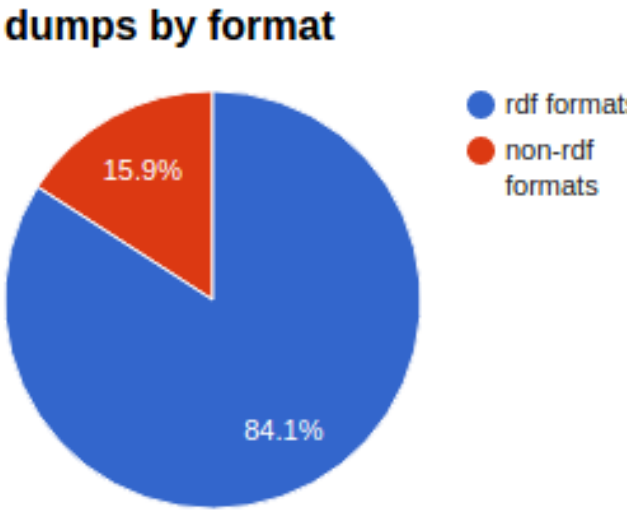
LODLaundromat

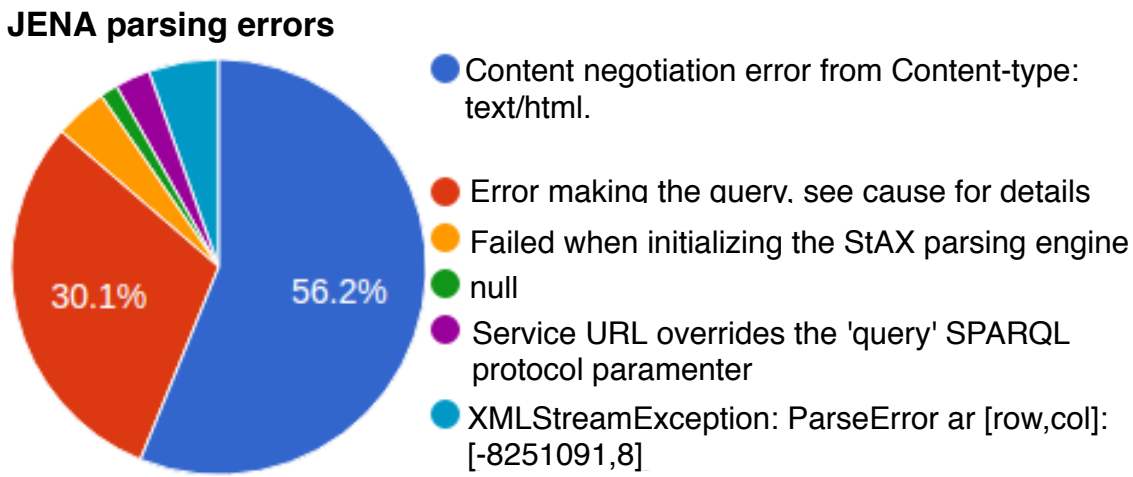
2.3% parsing errors. 99% indexed by WIMU

Experiments and results

Datasets







Not all datasets are ready to use

Conclusions and Future work

- A regularly updated database index of more than 660K datasets from LODStats and LODLaundromat
- An efficient service on the web that can determine which dataset will most likely define a URI
- Various statistics of datasets indexed from LODStats and LODLaundromat
- **Future work**: Integrate the second version of LinkLion
- `http://www.linklion.org`

That's all Folks!

Thanks!

Questions?

Website: `http://wimu.aksw.org/`

Contact: valdestilhas@informatik.uni-leipzig.de

HOBBIT

Holistic Benchmarking of Big Linked Data

This work was supported by grants from the EU H2020 Framework Programme provided for the project HOBBIT (GA no. 688227).