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

May 10, 2018

Outline

- Motivation
- Approach
- Use cases
- Experimental setup
- Statistics about Datasets
- Conclusions and Future work

Motivation

Data from the URI

- Problems to obtain data from a URI.

Motivation

Data from the URI

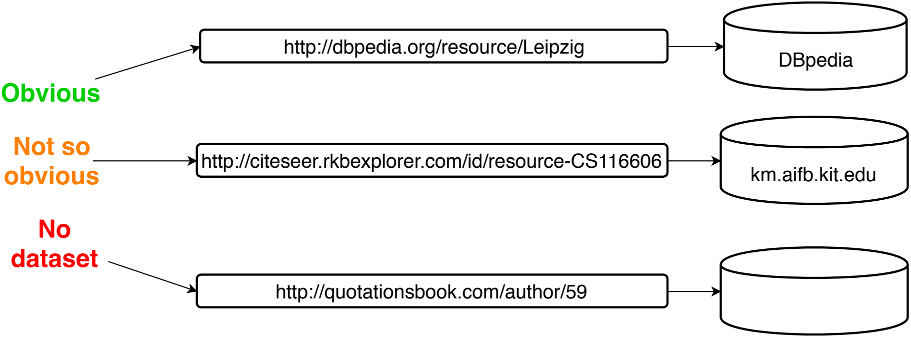
- Problems to obtain data from a URI.

URI dataset

- Which dataset the URI was defined?

Motivation

Example:



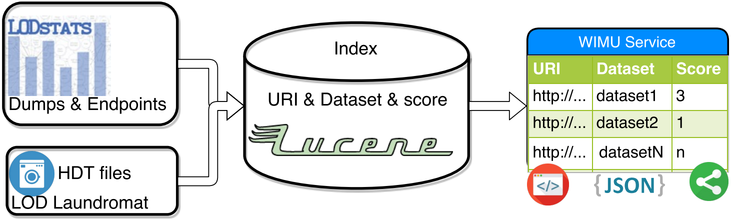
Obvious
http://dbpedia.org/resource/Leipzig
DBpedia
Not so obvious
http://citeseer.rkbexplorer.com/id/resource-CS116606
km.aifb.kit.edu
No dataset
http://quotationsbook.com/author/59

Approach

Goal

Indexing URIs and their use in order to let Linked Data consumers find the respective RDF data source.

- Rank the datasets proportionally to the number of literals.



URI	Dataset	Score
http://...	dataset1	3
http://...	dataset2	1
http://...	datasetN	n

Keeping the provenance of the URI.

Approach

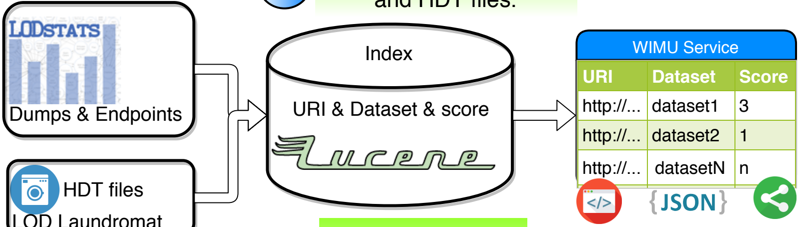
Steps to create the index

1 Retrieve list of datasets from sources.

2 Retrieve data from datasets.

```
SELECT ?s (count (?o) as ?c)
WHERE {
?s [] ?o . FILTER(isliteral (?o))
} GROUP BY ?s
```

3 Build three indexes from dump files, endpoints, and HDT files.



URI	Dataset	Score
http://...	dataset1	3
http://...	dataset2	1
http://...	datasetN	n

4 Merge the indexes into one.

5 Make the index available and browsable via a web application and an API service.

Approach

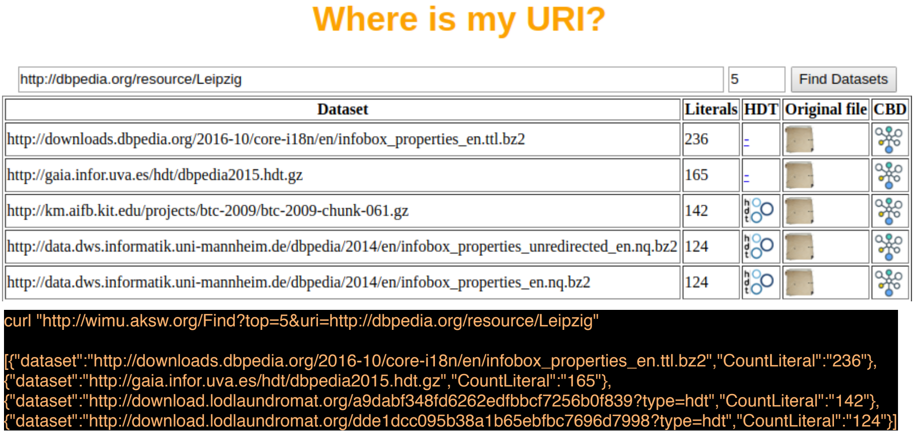
The power of Literals

The heuristic

- The raw data that can disambiguate a URI node in the most direct way.
- The semantic web architecture expects that datasets reusing a URI only refer to it without defining more literal values.

Approach

The interface (Web and Json)



Where is my URI?

http://dbpedia.org/resource/Leipzig | 5 | Find Datasets

Dataset	Literals	HDT	Original file	CBD
http://downloads.dbpedia.org/2016-10/core-i18n/en/infobox_properties_en.ttl.bz2	236	-	📦	✳
http://gaia.infor.uva.es/hdt/dbpedia2015.hdt.gz	165	-	📦	✳
http://km.aifb.kit.edu/projects/btc-2009/btc-2009-chunk-061.gz	142	hdt	📦	✳
http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/infobox_properties_unredirected_en.nq.bz2	124	hdt	📦	✳
http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/infobox_properties_en.nq.bz2	124	hdt	📦	✳

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

[{"dataset":"http://downloads.dbpedia.org/2016-10/core-i18n/en/infobox_properties_en.ttl.bz2","CountLiteral":"236"},
{"dataset":"http://gaia.infor.uva.es/hdt/dbpedia2015.hdt.gz","CountLiteral":"165"},
{"dataset":"http://download.lodlaundromat.org/a9dabf348fd6262edfbbcf7256b0f839?type=hdt","CountLiteral":"142"},
{"dataset":"http://download.lodlaundromat.org/dde1dcc095b38a1b65ebfbc7696d7998?type=hdt","CountLiteral":"124"}]
```

Approach

Usage

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

Parameter	Default	Description
top	0	Top ocurrences of the datasets.
uri	-	URI expected to search.
link	-	URL from a linkset.
cbd	-	The URI that will 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

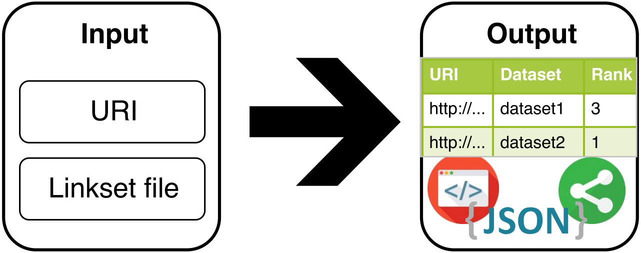
CBD

https://wimu.aksw.org/Find?cbd=http://dbpedia.org/resource/Leipzig&ds=http://download.lodlaundromat.org/a9dabf348fd6262edfbbcf7256b0f839?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).



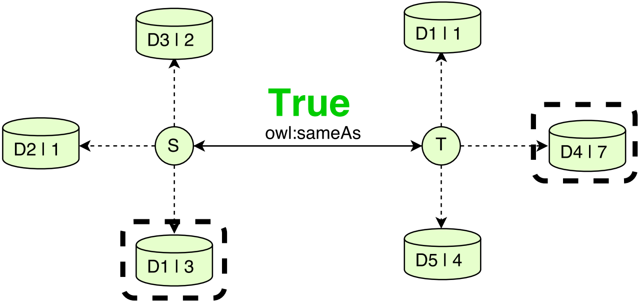
URI	Dataset	Rank
http://...	dataset1	3
http://...	dataset2	1

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

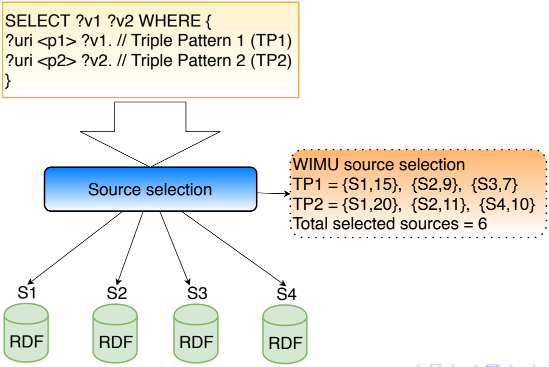


Use cases

Why do we need to know the URI Dataset?

Federated Query Processing

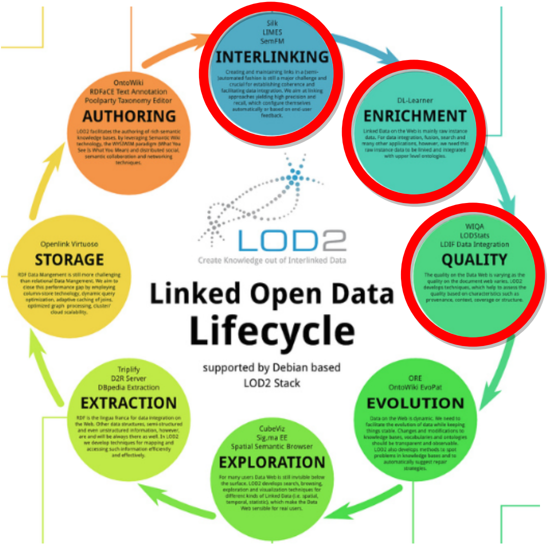
- Query planning and Source (dataset) selection.
- WIMU will find relevant sources against the individual triple patterns of a given SPARQL query.



```
SELECT ?v1 ?v2 WHERE {
?uri <p1> ?v1. // Triple Pattern 1 (TP1)
?uri <p2> ?v2. // Triple Pattern 2 (TP2)
}
```

WIMU source selection
TP1 = {S1,15}, {S2,9}, {S3,7}
TP2 = {S1,20}, {S2,11}, {S4,10}
Total selected sources = 6

Use cases



Silk
LIMES
SemFM

INTERLINKING

Creating and maintaining links in a (semi-)automated fashion is still a major challenge and crucial for establishing coherence and facilitating data integration. We aim at linking approaches yielding high precision and recall, which configure themselves automatically or based on end-user feedback.

DL-Learner

ENRICHMENT

Linked Data on the Web is mainly raw instance data. For data integration, fusion, search and many other applications, however, we need this raw instance data to be linked and integrated with upper level ontologies.

WIQA
LODStats
LDIF Data Integration

QUALITY

The quality on the Data Web is varying as the quality on the document web varies. LOD2 develops techniques, which help to assess the quality based on characteristics such as provenance, context, coverage or structure.

ORE
OntoWiki EvoPat

EVOLUTION

Data on the Web is dynamic. We need to facilitate the evolution of data while keeping things stable. Changes and modifications to knowledge bases, vocabularies and ontologies should be transparent and observable. LOD2 also develops methods to spot problems in knowledge bases and to automatically suggest repair strategies.

CubeViz
Sig.ma EE
Spatial Semantic Browser

EXPLORATION

For many users Data Web is still invisible below the surface. LOD2 develops search, browsing, exploration and visualization techniques for different kinds of Linked Data (i.e. spatial, temporal, statistic), which make the Data Web sensible for real users.

Triplify
D2R Server
DBpedia Extraction

EXTRACTION

RDF is the lingua franca for data integration on the Web. Other data structures, semi-structured and even unstructured information, however, are and will be always there as well. In LOD2 we develop techniques for mapping and accessing such information efficiently and effectively.

Openlink Virtuoso

STORAGE

RDF Data Mangement is still more challenging than relational Data Mangement. We aim to close this performance gap by employing column-store technology, dynamic query optimization, adaptive caching of joins, optimized graph processing, cluster/ cloud scalability.

OntoWiki
RDFaCE Text Annotation
Poolparty Taxonomy Editor

AUTHORING

LOD2 facilitates the authoring of rich semantic knowledge bases, by leveraging Semantic Wiki technology, the WYSIWIM paradigm (What You See Is What You Mean) and distributed social, semantic collaboration and networking techniques.

LOD2

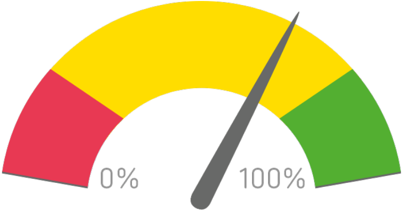
Create Knowledge out of Interlinked Data

Linked Open Data Lifecycle

supported by Debian based
LOD2 Stack

Experimental Setup

- Creation of the index: 3 days and 7 hours.
- Hardware: 64 CPU cores, 126 GB RAM, 2 TB hard disk.



Statistics about Datasets

Not all datasets are ready to use

	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

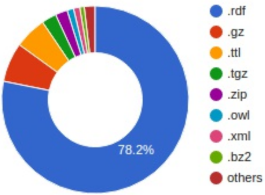
- 60% offline, 14% empty.
- 8% triples with literals as objects are blank nodes.
- 35% online datasets present some error using Jena.
- 69.8% datasets with parser errors.

LODLaundromat

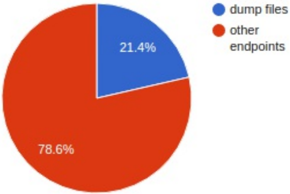
- 2.3% parsing errors. 99% indexed by WIMU.

Statistics about Datasets

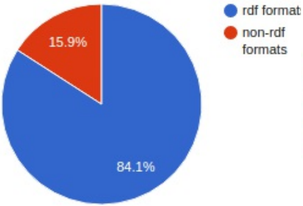
dumps by file extension



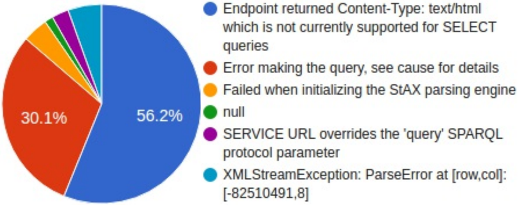
endpoints by types



dumps by format



errors JENA parsing



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 inform which dataset most likely defines 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?

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

Contact e-mail: valdestilhas@informatik.uni-leipzig.de