Engaging Content
Engaging People

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

On the Overlooked Challenges of Link Discovery

Peru Bhardwaj, Christophe Debruyne, Declan O'Sullivan

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

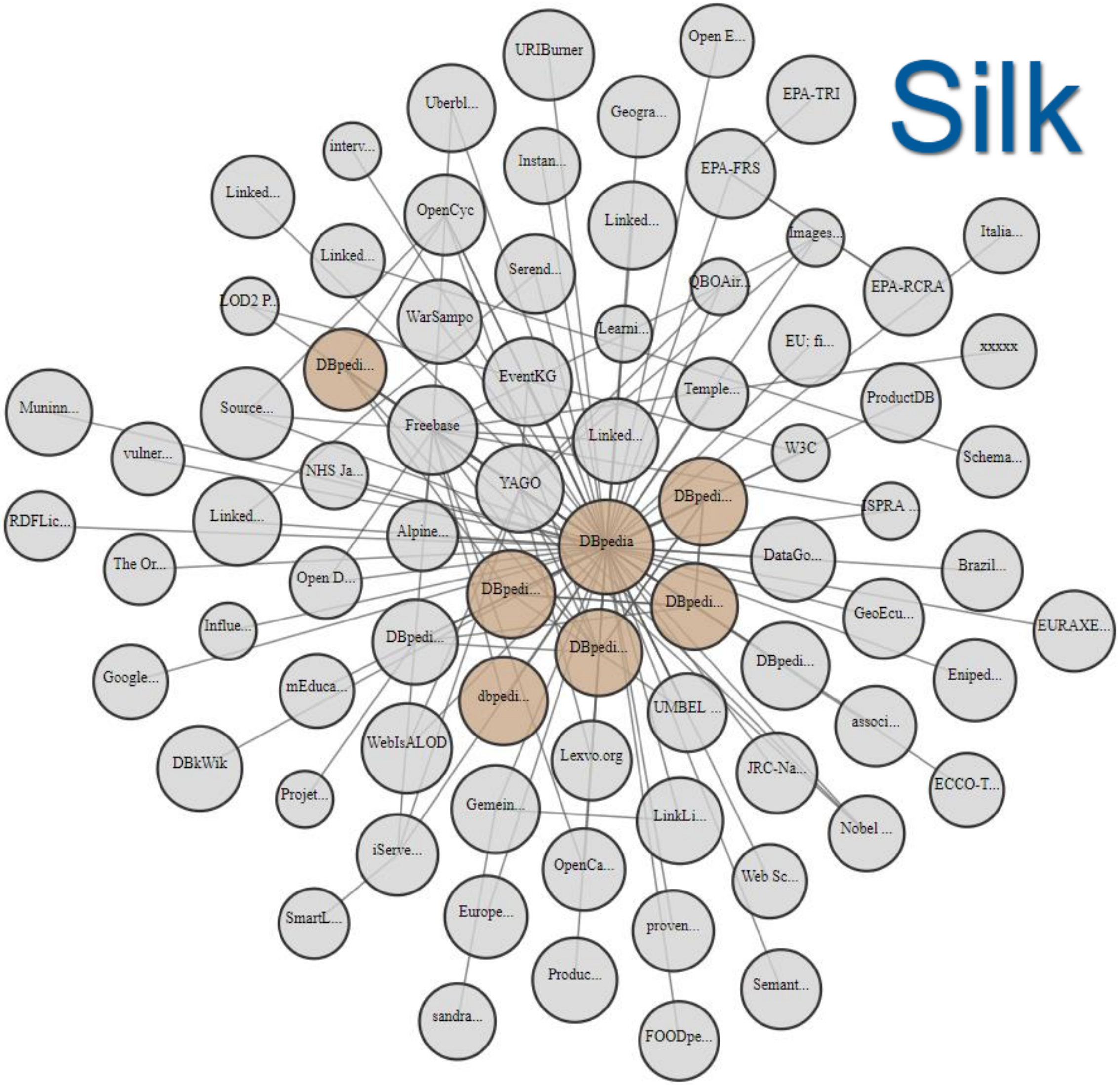
Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

Introduction



Silk The Linked Data Integration Framework

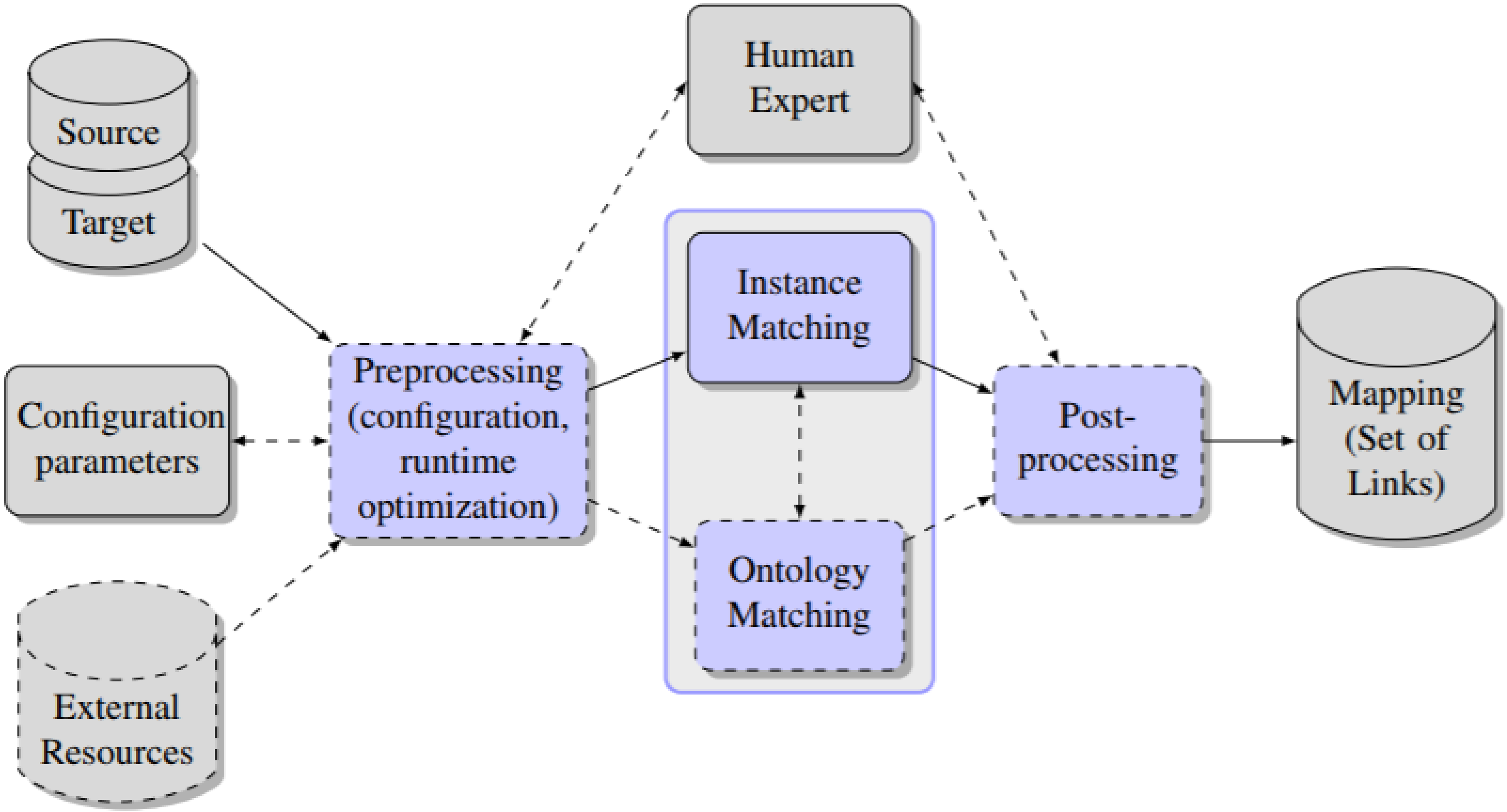
LIMES

Link Discovery Framework for Metric Spaces

KNOFUSS

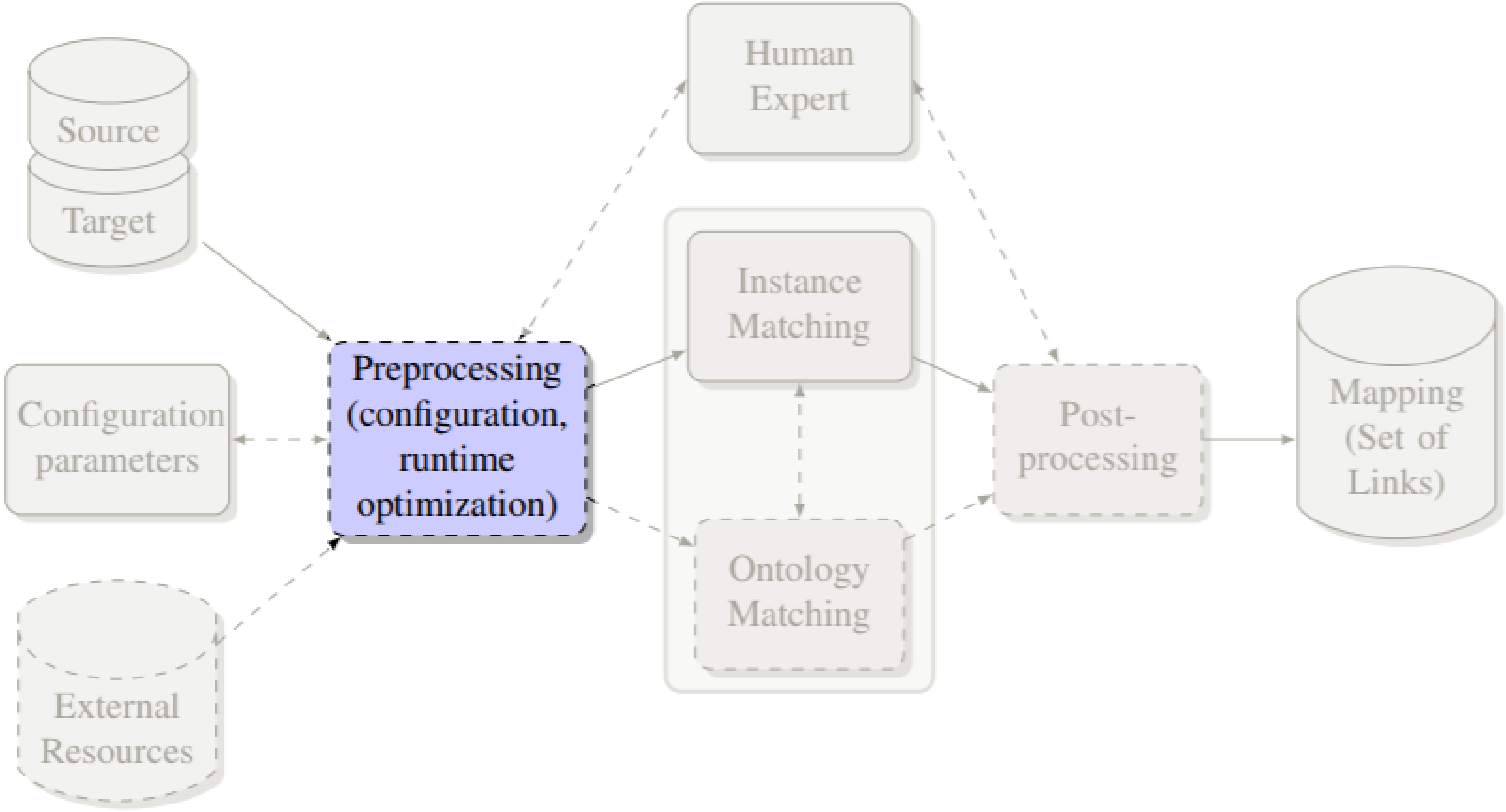
Linking Open Data cloud diagram 2018, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Introduction



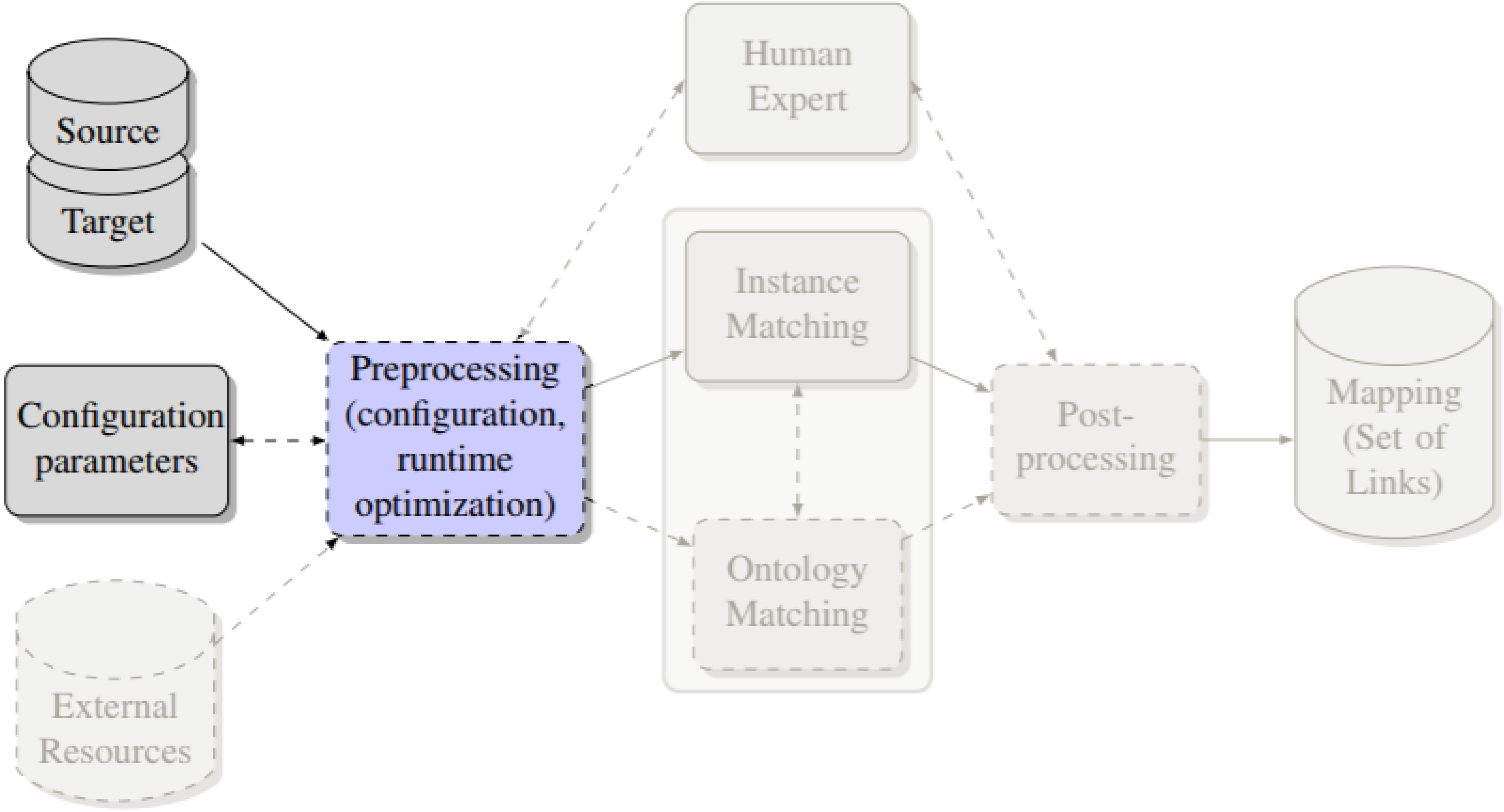
Nentwig, M., Hartung, M., Ngonga Ngomo, A.C. and Rahm, E.: A survey of current link discovery frameworks. Semantic Web, 8(3), 419-436 (2017)

Introduction



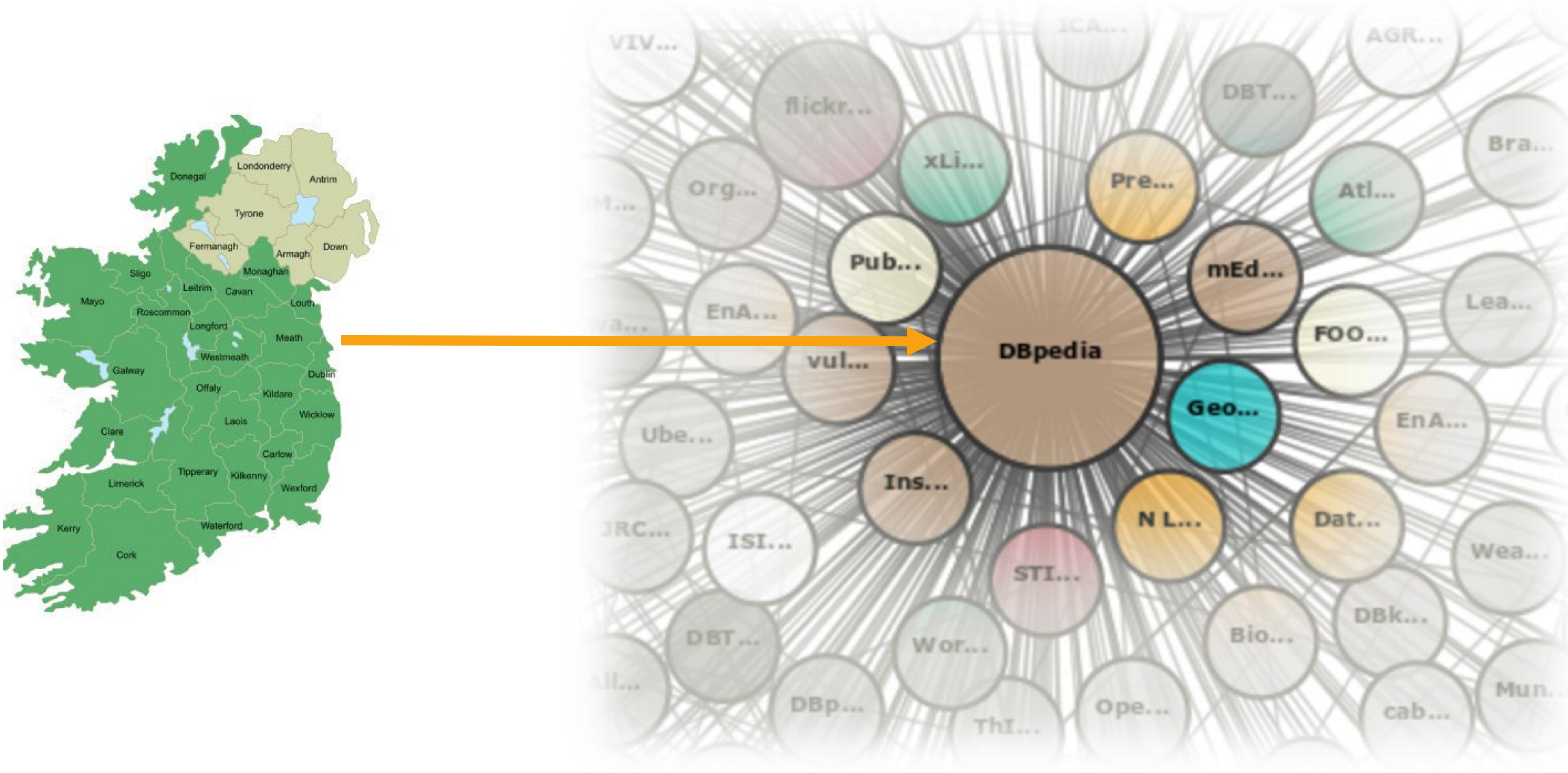
Nentwig, M., Hartung, M., Ngonga Ngomo, A.C. and Rahm, E.: A survey of current link discovery frameworks. Semantic Web, 8(3), 419-436 (2017)

Introduction

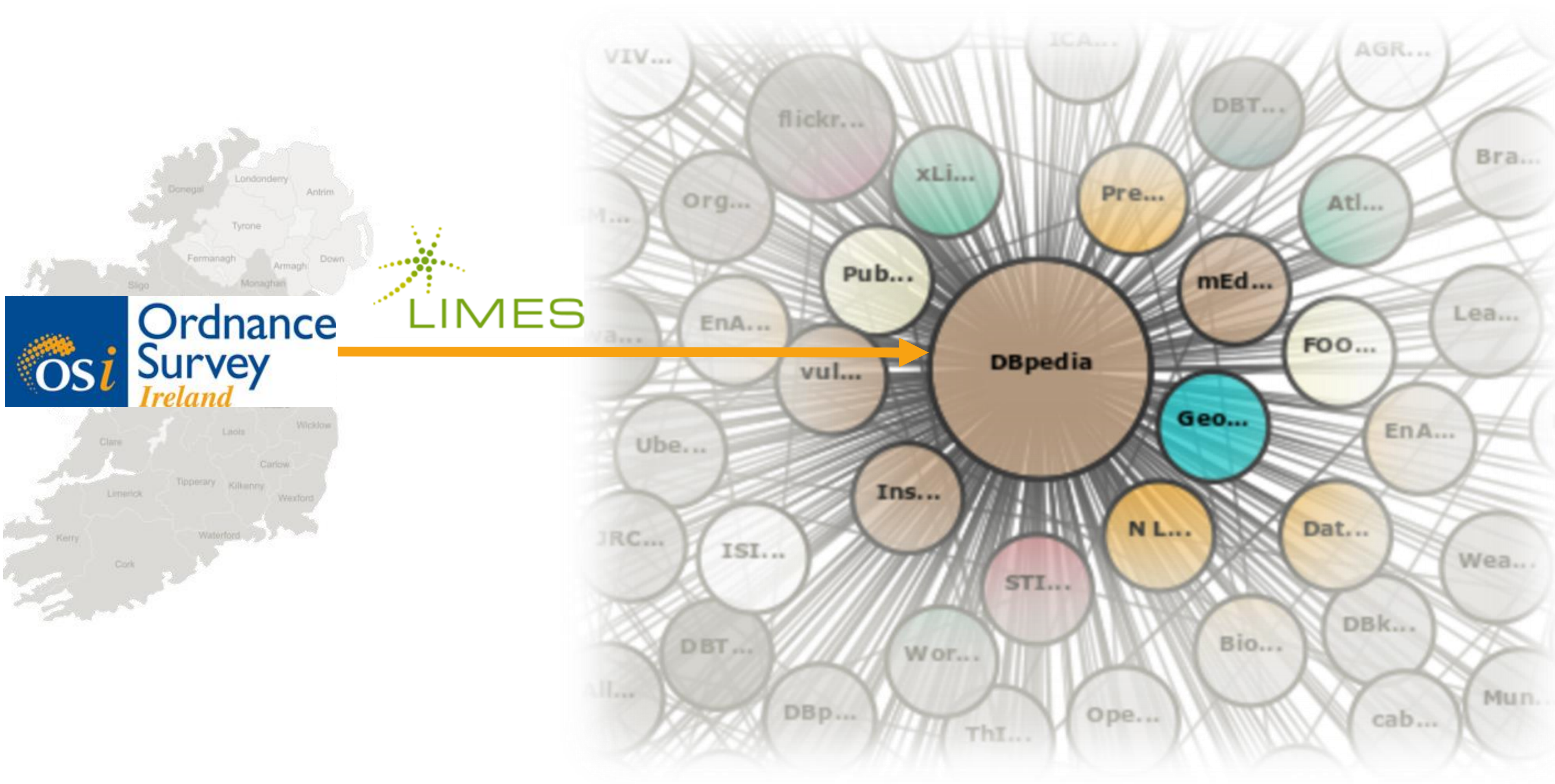


Nentwig, M., Hartung, M., Ngonga Ngomo, A.C. and Rahm, E.: A survey of current link discovery frameworks. Semantic Web, 8(3), 419-436 (2017)

Introduction



Introduction



Contribution

Highlight the challenges faced during the Preprocessing phase in Link Discovery workflow

Provide practical guidance in undertaking an interlinking project using Link Discovery frameworks

Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

Counties and Townlands

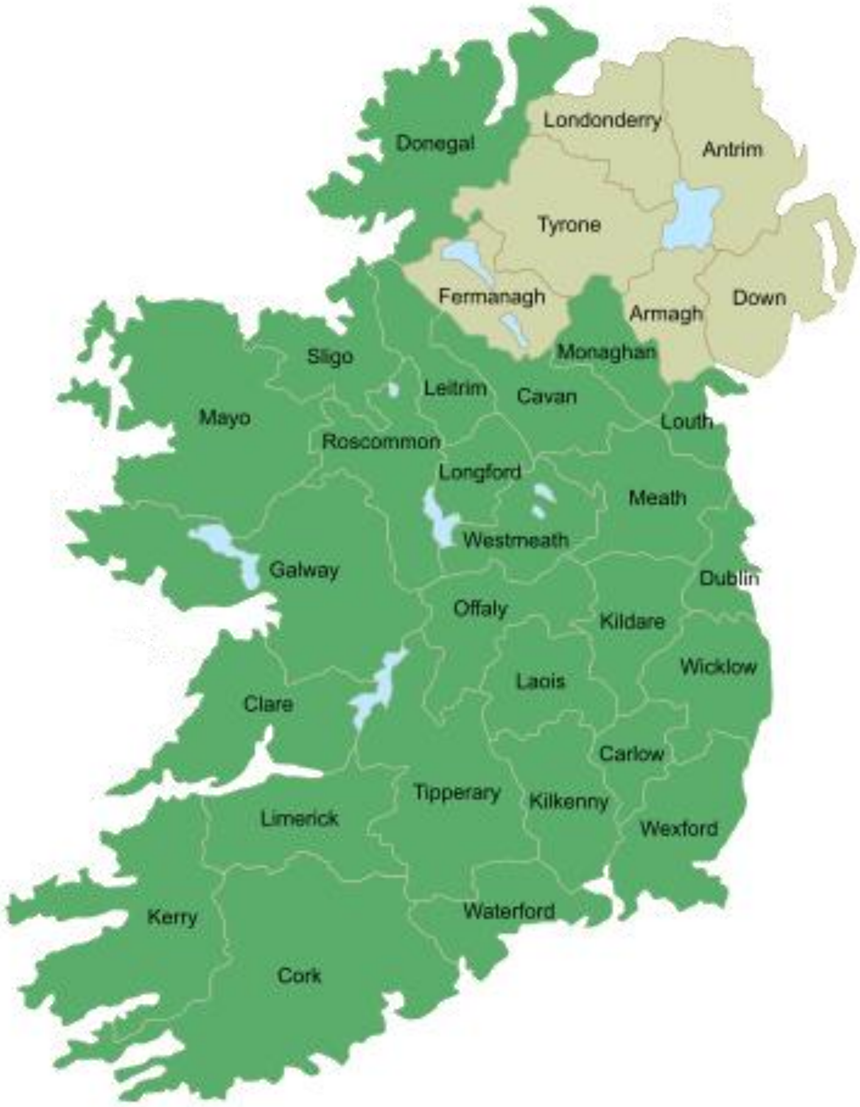
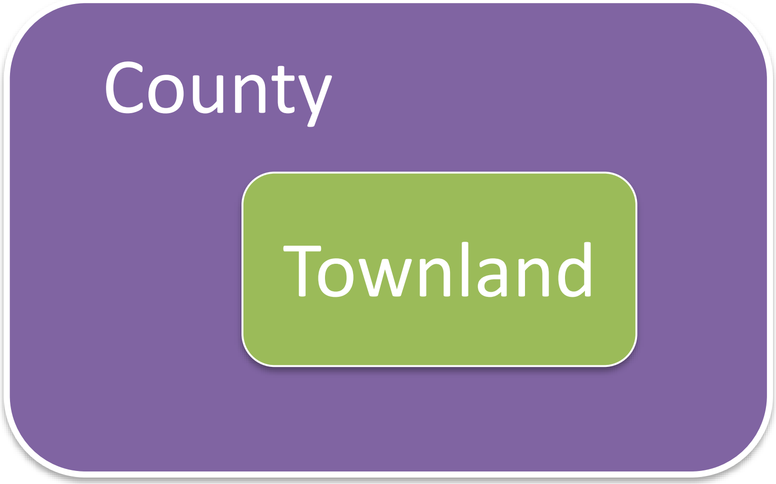


Image: www.wikipedia.org



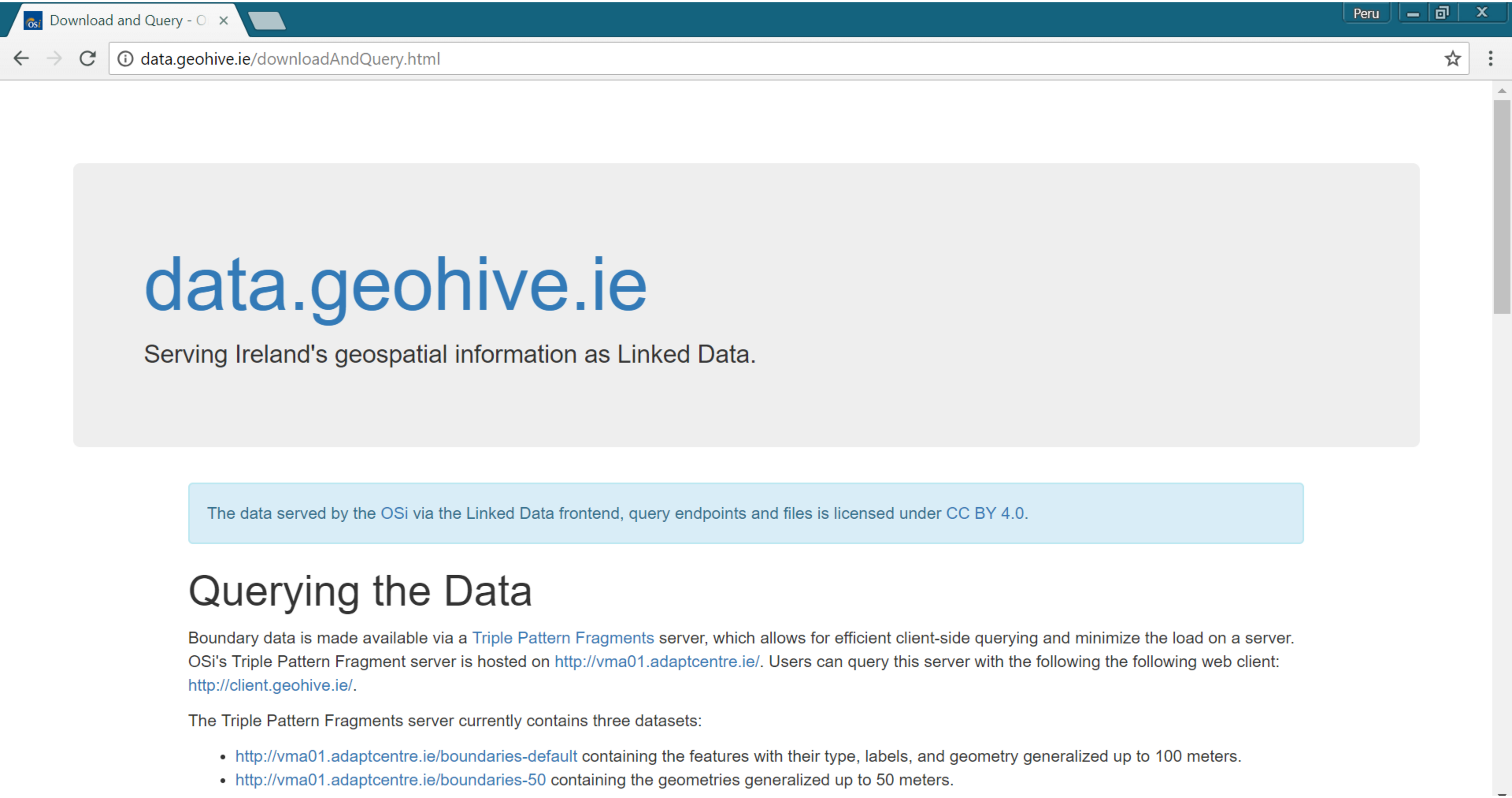
26 counties

~50,000 townlands

Why OSi and DBpedia?

- Semantically heterogeneous datasets
- Added value for Linked Data applications that require authoritative geospatial data
- Expand the geospatial section of the LOD cloud

OSi Dataset



data.geohive.ie

Serving Ireland's geospatial information as Linked Data.

The data served by the OSi via the Linked Data frontend, query endpoints and files is licensed under CC BY 4.0.

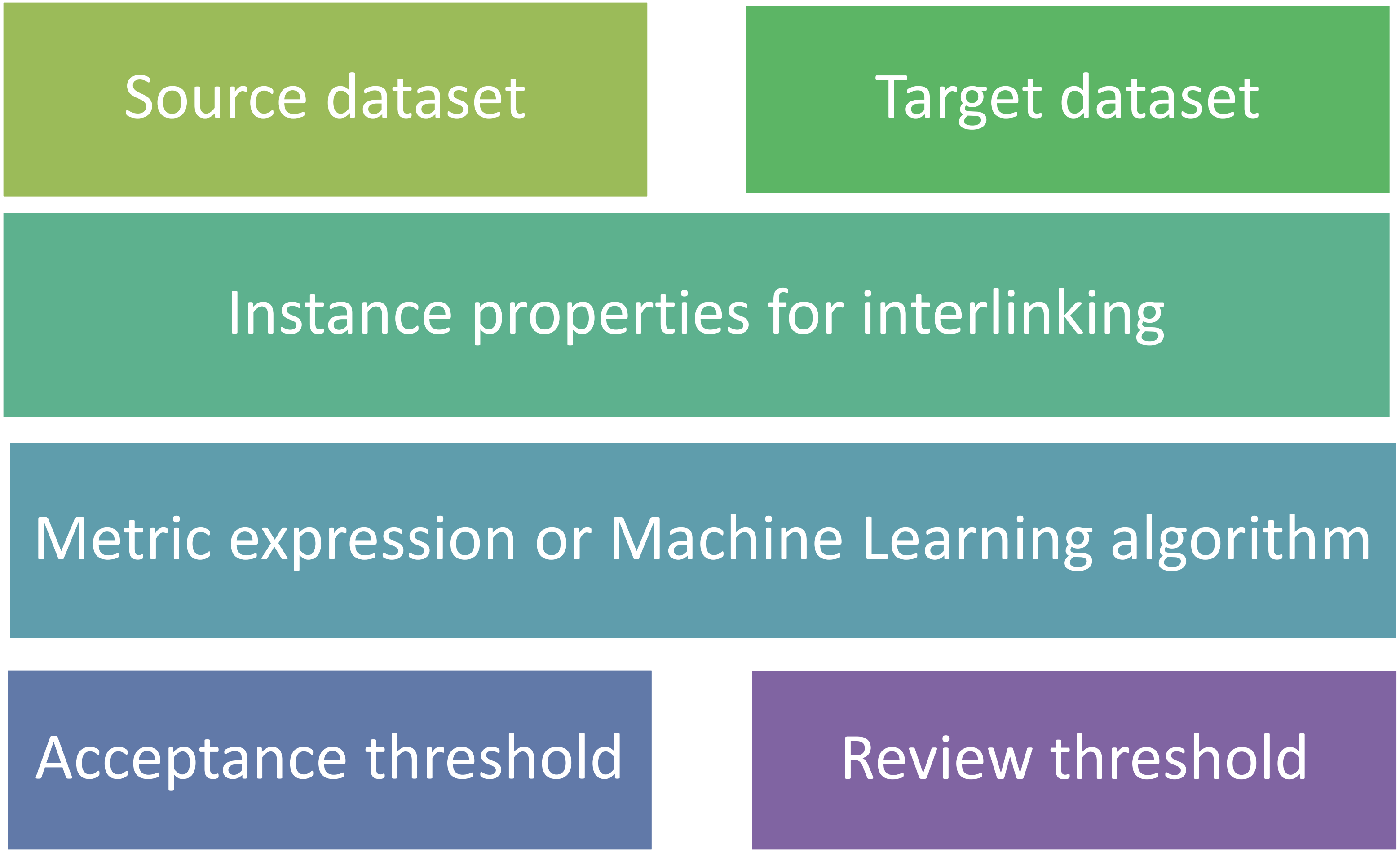
Querying the Data

Boundary data is made available via a Triple Pattern Fragments server, which allows for efficient client-side querying and minimize the load on a server. OSi's Triple Pattern Fragment server is hosted on http://vma01.adaptcentre.ie/. Users can query this server with the following the following web client: http://client.geohive.ie/.

The Triple Pattern Fragments server currently contains three datasets:

- http://vma01.adaptcentre.ie/boundaries-default containing the features with their type, labels, and geometry generalized up to 100 meters.
- http://vma01.adaptcentre.ie/boundaries-50 containing the geometries generalized up to 50 meters.

LIMES Configuration File



Source dataset

Target dataset

Instance properties for interlinking

Metric expression or Machine Learning algorithm

Acceptance threshold

Review threshold

Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

DBpedia

Version 2016-10

Identifying the Dataset

Query 1

```
?townland dct:subject ?subject .
FILTER(REGEX(?subject, "townland", 'i'))
```

Query 2

```
?townland dbo:type dbr:Townland .
```

Query 3

```
?townland dbo:abstract ?abstract.
?abstract bif:contains '"is a townland"'
```

Query 4

```
?townland <http://purl.org/linguistics/gold/hypernym> dbr:Townland .
```

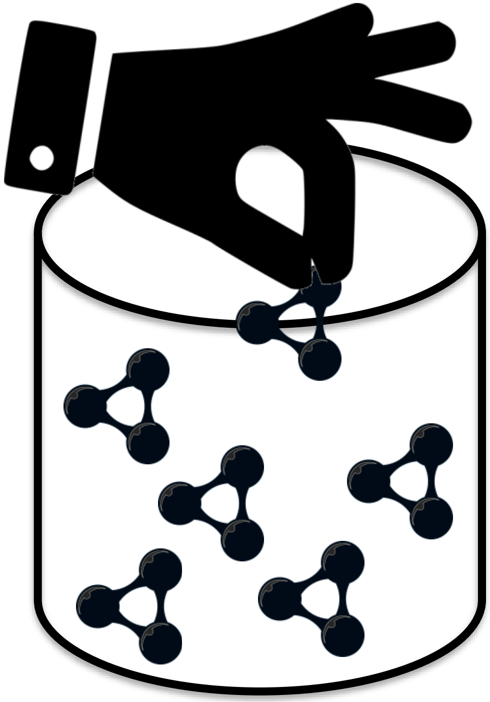
Lesson 1

Identifying the most suitable query to isolate the instances to be interlinked is a trial and error based iterative process.

Accessing the Dataset

Virtuoso

Snorql



Lesson 2

- Interfaces for SPARQL endpoints can be **unreliable**
- An incomplete view via the interface might lead to errors
- The ingestion of whole dump requires additional skills and resources

Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

Selecting Properties

Similarity between labels of different townlands

“bally”
4451

“derry”
878

“Ballina”
21

Lesson 3

There is added value in the geospatial information of entities in a Link Discovery workflow

Selecting Similarity Measures

Machine Learning

- No training data
- Geometry comparison not supported by LIMES

Dublin

Dublin

String Similarity Measures

- Excessive links by some measures
- Several measures with same number of links

Topological Similarity Measures

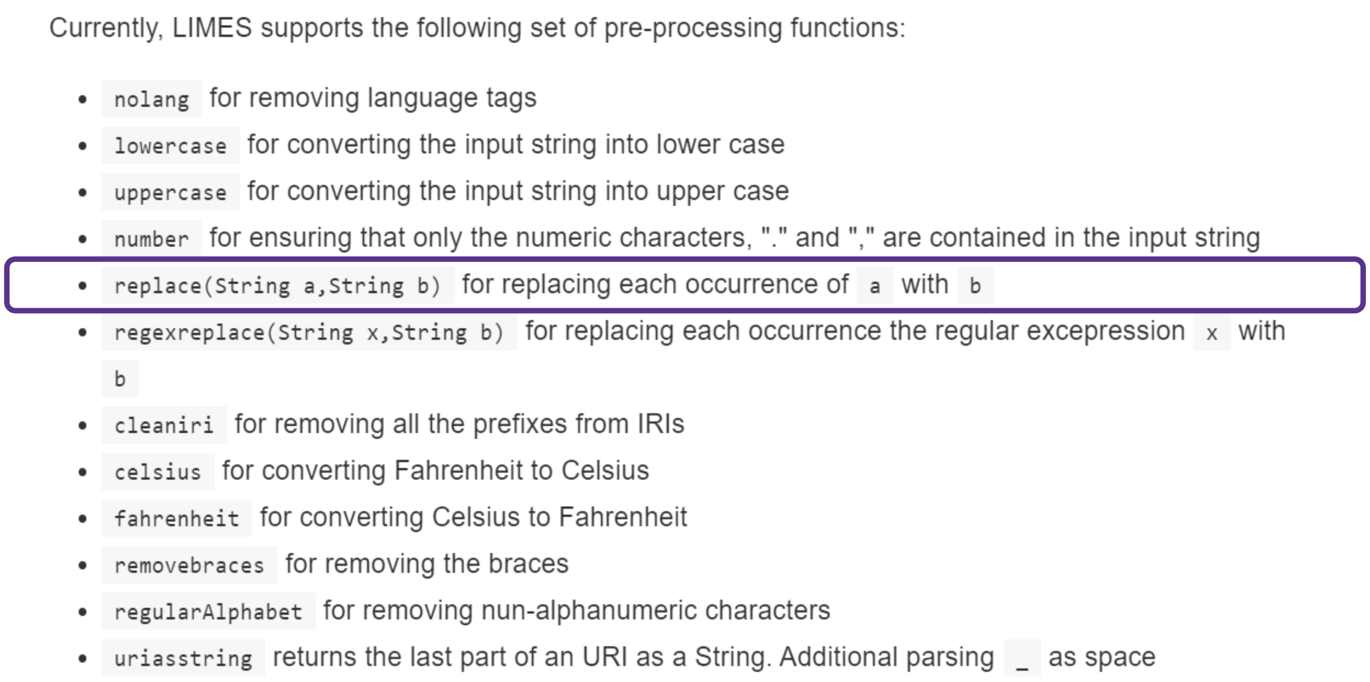
- Dissimilar representations of geometry
- Relative comparison

Lesson 4

The selection of a suitable distance measure is **unintuitive** even though it is crucial in ensuring the effectiveness of the matching phase in LD workflow

Adding Functions and Metric Operations

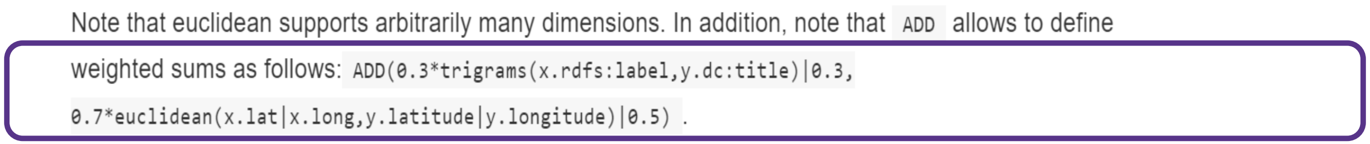
Pre-processing Functions



Currently, LIMES supports the following set of pre-processing functions:

- `nolang` for removing language tags
- `lowercase` for converting the input string into lower case
- `uppercase` for converting the input string into upper case
- `number` for ensuring that only the numeric characters, "." and "," are contained in the input string
- `replace(String a,String b)` for replacing each occurrence of `a` with `b`
- `regexreplace(String x,String b)` for replacing each occurrence the regular excepression `x` with `b`
- `cleaniri` for removing all the prefixes from IRIs
- `celsius` for converting Fahrenheit to Celsius
- `fahrenheit` for converting Celsius to Fahrenheit
- `removebraces` for removing the braces
- `regularAlphabet` for removing nun-alphanumeric characters
- `uriasstring` returns the last part of an URI as a String. Additional parsing `_` as space

Metric Operations



Note that euclidean supports arbitrarily many dimensions. In addition, note that `ADD` allows to define weighted sums as follows: `ADD(0.3*trigrams(x.rdfs:label,y.dc:title)|0.3, 0.7*euclidean(x.lat|x.long,y.latitude|y.longitude)|0.5)` .

Lesson 5

Availability of comprehensive documentation and elaborate examples is critical to avoid significant effort being expended in trial and error

Outline

1. Introduction
2. OSi to DBpedia Case Study Preliminaries
3. Discovering the Dataset
4. Finalizing the Link Specification
5. Conclusion

Conclusion

Discovering the Dataset

- Identifying the dataset
- Accessing the dataset

Finalizing the Link Specification

- Selecting Properties
- Selecting Similarity Measures
- Adding Functions and Metric Operations in LIMES

Questions?

Contact:
Peru Bhardwaj
peru.bhardwaj@adaptcentre.ie