On the Overlooked Challenges of Link Discovery

Peru Bhardwaj, Christophe Debruyne and Declan O'Sullivan,

ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
{peru.bhardwaj, christophe.debruyne, declan.osullivan}@adaptcentre.ie

Abstract. Challenges in interlinking two datasets have been studied extensively in the state-of-art in terms of the complexity of the matching process used for interlinking. However, the challenges in gathering the input datasets to be interlinked and finalizing a link specification, which constitute the preprocessing phase of the link discovery (LD) workflow, are mostly overlooked. In this paper, we highlight these challenges through a case study of interlinking the Ordnance Survey Ireland (OSi) datasets with the geospatial data in the Linked Open Data (LOD) cloud. Our study shows that designing a query and using an interface to retrieve the instances to be interlinked from SPARQL endpoint is difficult. In finalizing a link specification, additional properties can be critical when labels are ambiguous. Also, the selection of similarity measures to compare these properties is unintuitive. These challenges show that interlinking datasets is not very straightforward, even with the availability of link discovery tools. Since the challenges in the preprocessing phase are not obvious, the analysis documented here can provide guidance in undertaking a project in interlinking two datasets.

Keywords: Interlinking, Link Discovery, Geospatial Data, DBpedia, OSi

1 Introduction

Interlinks between Linked Data datasets are the key in realizing the vision of an interconnected Web of Data [1]. Several frameworks have been proposed to support the discovery of links between two datasets. A generic workflow for such link discovery (LD) frameworks is explained by Nentwig *et al.* in [1]. The LD workflow has three phases namely preprocessing, matching (instance matching or ontology matching) and postprocessing. The challenging nature of link creation due to the complexity of the matching phase has been discussed extensively in the state-of-art [2]. However, the importance of preprocessing phase in the LD workflow is mostly overlooked. This phase includes the preparation of input data and finalization of the linking configuration. When interlinking two heterogeneous datasets that have no existing links between them, the preprocessing phase can become as crucial as the effectiveness of the matching technique used for interlinking them.

In this paper, we elucidate the challenges faced in the preprocessing phase of the LD workflow. For this, we use a case study of interlinking authoritative geospatial data of the Republic of Ireland (ROI) to geospatial data from DBpedia in the Linked Open Data (LOD) cloud. The authoritative geospatial data is made available by the ROI's national mapping agency, Ordnance Survey Ireland (OSi)[1] and served as LOD through their

[1] http://www.osi.ie

GeoHive platform [3]. We document the case study as an analysis of challenges faced and lessons learnt during the preprocessing phase. We believe that this documentation will provide guidance for researchers interested in undertaking any project in interlinking datasets using LD frameworks, and experts in geospatial information that may require interlinking their data but do not necessarily have the expertise in Semantic Web technologies. Our experience of the preprocessing phase is divided into two sections: identifying and accessing geospatial data from DBpedia (Section 3); finalizing a link specification to match instances (Section 4).

In summary, the main contribution of this paper is twofold – to highlight the challenges faced during the preprocessing phase in LD workflow; and to provide practical guidance in undertaking an interlinking project using LD frameworks.

2 OSi to DBpedia Case Study Preliminaries

We chose to interlink the counties and townlands from OSi to DBpedia to explore the issue of interlinking geospatial datasets that have not been linked previously. The choice of datasets was motivated by the following three factors: (i) The datasets are semantically heterogeneous and had no existing links between them. (ii) Authoritative geospatial data is used as a reference data by Linked Data applications that require accurate and complete geospatial data available [3]. These applications derive added value from interlinking the authoritative data with crowd-sourced data on the LOD cloud as they can utilize additional information about the geospatial entities referenced. (iii) Since the geospatial data only makes 2% of the LOD cloud [4], adding authoritative OSi data will expand the geospatial section and improve the data quality of the LOD cloud in terms of accuracy and completeness.

The case study was conducted in two parts – (i) interlink the **counties** from OSi to DBpedia; (ii) interlink the **townlands** from OSi to DBpedia. For source datasets, we used the 100 meters boundary generalizations of county and townland datasets available online[2] from OSi. For target datasets, subsets of DBpedia were selected as described in Section 3. The interlinking problem was tackled as an instance matching and not an ontology matching problem because the ontology for OSi data is not readily available. We used the LIMES link discovery framework as it is shown to perform better than other state-of-art LD frameworks in [2], [5] and provides a wide range of similarity measures [1]. The input configuration file for LIMES includes the source and target datasets to be interlinked, the instance properties to be used for interlinking, a metric expression or machine learning algorithm to be used for similarity measurement of properties and the acceptance and review threshold[3].

3 Discovering the Dataset

This section describes our experience in discovering the subsets of DBpedia (version 2016-10) as target datasets to be used as input to the LD workflow.

[2] http://data.geohive.ie/downloadAndQuery.html
[3] http://dice-group.github.io/LIMES/user_manual/

3.1 Identifying the Dataset

Identifying the instances of counties and townlands in DBpedia was not straightforward as we did not know the semantic equivalent for a county or townland in DBpedia. We started by selecting arbitrary instances from the OSi dataset, looking for the corresponding instances in DBpedia (by using the nomenclature for resource URIs in DBpedia) and analyzing the properties of these instances to recognize common properties.

An analysis of county instances showed that the counties of the ROI have the article category of *dbc:Counties_of_the_Republic_of_Ireland*. A query for the places (*rdf:type* is *dbo:Place*) with this article category gave us the county dataset from DBpedia.

After analyzing multiple instances of townlands, we observed the following four different ways of querying the townlands from DBpedia.

QRY1. As the Irish townlands fall under the administrative unit of county, the article category for a townland is assigned by the county to which it belongs. For example, *dct:subject* for townlands of county Laois is *dbc:Townlands_of_County_Laois*. To get townlands of multiple counties, the resources which have the string *"townland"* in the subject can be queried. However, the query is faulty for townlands that do not have this pattern for their article category; for example, *dbr:Ballynoe,_Great_Island*.

QRY2. The DBpedia ontology type (*dbo:Type*) for some townlands is *dbr:Townland*. Hence, a query can be executed to retrieve all resources for which the ontology type is townland. However, there are some resources like *dbr:Kilnaboy* which are not a townland but which have the ontology type townland; these would be included in the query results. Also, there are multiple townlands in the OSi dataset which do not have an ontology type in DBpedia; for example, *dbr:Saggart* and *dbr:Arywee*.

QRY3. Most townlands have the word *"townland"* in the abstract. However, the word is also present in the definition of townland as well as the resource for the list of townlands in a county. To remove these resources from the results, a query for the phrase *"is a townland"* in the abstract can be executed. But there are some abstracts with phrases like *"is a small townland"* (for example, *dbr:Kildaree*) or *"is a residential townland"* (for example, *dbr:Drumardagh*). These will be missed by the query.

QRY4. Townlands in DBpedia have a hypernym value of *dbr:Townland*. However, all resources with this hypernym are not townlands; for example, *dbr:Clooniffe* and *dbr:Tooban*. Also, there are some townlands in the OSi dataset that are assigned a hypernym for village in DBpedia; for example, *dbr:Saggart*[4].

For each type of query, the number of results were different and none of the results contained the instances from all the other queries. In addition, the inconsistencies in assignment and distribution of properties of resources in DBpedia made it unintuitive in the design of a query to find all townlands in DBpedia. For example, many townlands do not have the property *dbo:country* to enable discovery of townlands specific to ROI.

Lesson 1 – In the interlinking of two semantically heterogeneous datasets, the existence of semantically equivalent concepts in them is unlikely and one cannot count on the existence of consistent, more complex "patterns" to discover instances to be interlinked. In the case of DBpedia townlands, instances could be found using multiple queries. In such a case, identifying the most suitable query to isolate the instances of a concept is a trial and error based iterative process.

[4] http://data.geohive.ie/resource/townland/2AE1962A048C13A3E055000000000001

3.2 Accessing the Dataset

Querying the DBpedia SPARQL endpoint is possible through Virtuoso and Snorql interfaces. We noticed that the Snorql interface ensured reliable results, while the Virtuoso interface was, for some reason, giving different results. It turns out that DBpedia's Virtuoso infrastructure is configured to return incomplete results when a certain query timeout is exceeded. This is based on the "Anytime Queries" functionality which has been implemented as part of their fair use policy.[5] As we were unaware of this behavior, we decided to resort to the latest version of DBpedia dumps (2016-10). It was however tricky and laborious to figure out which dumps to use as they are partitioned. We could reproduce results of only QRY3 (Section 3.1) using the data dump of short abstracts.

Lesson 2 – Interfaces for SPARQL endpoints may be "unreliable", as demonstrated by incomplete results from DBpedia endpoint. If complete results are needed for a query, one should use the data dump(s) provided by the data publisher for validation. An incomplete view via the interface might lead to errors and the ingestion of whole dumps requires additional skills and resources.

4 Finalizing the Link Specification

This section details our experience in the selection of properties to be compared and the selection of similarity measures to compare these properties in the preprocessing phase.

4.1 Selecting Properties

Using labels is popular for matching instances between two datasets to be interlinked, as demonstrated in the experiments by Ngomo and Auer [2]. For us, the use of labels was insufficient for interlinking townlands because there is significant similarity in the labels of different townlands. For example, 4451 English labels contain the string "*bally"*, 878 English labels contain the string "*derry"*. Also, there are multiple townlands with the same name; for example, 21 townlands are named "*Ballina"*. Hence, the use of geometry was essential for matching the townlands.

Lesson 3 – Though it is common practice to use labels for matching instances in datasets, the use of geometries can be essential when the labels are ambiguous. Thus, there is added value in the geospatial information of entities in a LD workflow.

4.2 Selecting Similarity Measures

Selection of a suitable similarity measure is difficult but a deciding factor on whether links can be discovered [2]. Supervised techniques to learn a link specification proved to be unusable in our case study due to lack of training data. The use of unsupervised version of the WOMBAT algorithm in LIMES for comparing geometries of townlands did not generate any links. Hence, we had to manually select the similarity metric. Instead of guessing the similarity measure to use, we examined the number of links

[5] We became aware of this only recently and thank Kingsley Idehen for pointing this feature out: https://lists.w3.org/Archives/Public/public-lod/2018Apr/0030.html

generated by all the similarity measures.

Table 1 shows the number of links accepted and the number of links to be reviewed using the **string similarity measures** for comparing labels of townlands in OSi and DBpedia. As is evident from the table, *Soundex, JaroWinkler* and *MongeElkan* generated excessive links and could not be used. Similarity measures like *ExactMatch, Jaccard* and *Levenshtein* with 545 accepted links and 0 links for review looked the most reliable. Of these, the Levenshtein distance was arbitrarily picked to compare labels.

Combining similar spatial objects is difficult if the dimensions in which they are represented differ [6]. The geometric representation of counties and townlands in OSi is a WKT polygon while their representation in DBpedia is a WKT point. Because of their dissimilar representations, geometries could not be compared directly. Hence, we used topological relations to do a relative comparison. The number of links accepted and the number of links to be reviewed using the **topological similarity measures** for comparing geometries of townlands in OSi and DBpedia are shown in Table 1. As the number of links generated by *top_contains* and *top_intersects* is same, *top_contains* was picked arbitrarily to compare geometries.

Lesson 4 – The selection of a suitable distance measure is tricky and unintuitive even though it is crucial in ensuring the effectiveness of the matching phase in LD workflow.

Table 1. Number of links generated using string and topological similarity measures in LIMES. Acceptance threshold was 0.95 and review threshold was 0.8.

	Similarity Measure	Links accepted	Links for review
String	**ExactMatch, Jaccard, Levenshtein**	**545**	**0**
	Cosine, Overlap, Trigram	545	16
	Qgrams	545	32
	RatcliffObershelp	708	21852
	Jaro	924	178396
	MongeElkan	1268	7843
	JaroWinkler	2144	441384
	Soundex	18493	238701
Topological	**Top_contains, Top_intersects**	**286**	**0**
	Top_touches	0	0

4.3 Adding Functions and Metric Operations in LIMES

Once the input parameters are selected, using the LIMES framework should be straightforward. However, we discovered that the terse documentation available and the lack of variety in examples provided with the framework, resulted in several unexpected mistakes being made on our part. Two examples are – (i) In matching the counties from OSi to DBpedia, we used the *replace* function to remove the word *"county"* in labels. However, we did not realize that the string to be replaced should be passed without the standard double quotes. (ii) While combining two similarity measures to compare both label and geometry, we used the "ADD" operation. However, due to difficulty in understanding the documentation, similarity score of the generated links turned out to be above 1.0, which is erroneous. Hence, we had to use "AND". Clearer documentation or more complex examples could have been helpful in avoiding these mistakes.

Lesson 5 – Even though LD frameworks have the appearance of being easy to use, there can be pitfalls in configuring them that can lead to unexpected results. Availability

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

5 Conclusion

We illustrated the challenges faced in the preprocessing phase of the LD workflow, through a case study of interlinking OSi's authoritative geospatial datasets with the geospatial data in DBpedia. We found that it was difficult to isolate the datasets to be interlinked from DBpedia because of the trial and error in identifying a suitable query and the unreliability of interfaces to the SPARQL endpoint. In finalizing the properties for matching instances, geometric representations had added value as the labels alone were ambiguous. But it was tricky to select a similarity measure and configure the LD framework. We believe that comprehensive documentation and more complex examples could be helpful in reducing the effort in configuring LD frameworks.

A useful insight from our case study is that the process of interlinking two heterogeneous datasets is not as straightforward as documented in the literature about the link discovery mechanisms. A possible reason is that the choices made in the preprocessing phase can impact the number of links generated by the matching phase in the LD workflow. Hence, a project in interlinking heterogeneous datasets should allocate sufficient time and resources for the preprocessing phase.

It is perceptible that the interlinking community is keen on enhancing the efficiency and effectiveness of the LD workflow. Our case study shows that the semantic heterogeneity between datasets can be the bottleneck in realizing this vision.

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

References

1. 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)
2. Ngomo, A.C.N. and Auer, S.: LIMES - a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI, 2312-2317 (2011)
3. Debruyne, C., Meehan, A., Clinton, É., McNerney, L., Nautiyal, A., Lavin, P. and O'Sullivan, D.: Ireland's Authoritative Geospatial Linked Data. In: Proc. of ISWC 2017 (II), pp. 66-74. Springer, Cham (2017)
4. Schmachtenberg, M., Bizer, C. and Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Proc. of ISWC 2014 (I), pp. 245-260, Springer, Cham (2014)
5. Sherif, M.A., Dreßler, K., Smeros, P. and Ngomo, A.C.N.: Radon-Rapid Discovery of Topological Relations. In: *AAAI*, pp. 175-181 (2017)
6. van den Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., Fathy, Y., Castro, R.G., Haller, A., Harth, A., Janowicz, K., Kolozali, S., van Leeuwen, B., Lefrançois, M., Lieberman, J., Perego, A., Le-Phuoc, D., Roberts, B., Taylor, K., Troncy, R.: Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. Semantic Web (in press)

