LINKED DATA MEETS ONTOLOGY MATCHING
Enhancing Data Linking through Ontology Alignments

François Scharffe[1] and Jérôme Euzenat[2]

[1]*LIRMM, University of Montpellier, 161 Rue Ada, Montpellier, France*

[2]*INRIA, 655 Avenue de l'Europe, Montbonnot, France*

Keywords: Semantic web, Linked data, Data linking, Ontology alignment, Ontology matching, Entity reconciliation, Object consolidation.

Abstract: The Web of data consists of publishing data on the Web in such a way that they can be connected together and interpreted. It is thus critical to establish links between these data, both for the Web of data and for the Semantic Web that it contributes to feed. We consider here the various techniques which have been developed for that purpose and analyze their commonalities and differences. This provides a general framework that the diverse data linking systems instantiate. From this framework we consider the relation between data linking and ontology matching activities. Although, they can be considered similar at a certain level (they both relate formal entities), they serve different purposes: one acts at the schema level and the other at the instance level. However, they would find a mutual benefit at collaborating. We thus present a scheme under which it is possible for data linking tools to take advantage of ontology alignments. We present the features of expressive alignment languages that allows linking specifications to reuse ontology alignments in a natural way.

1 INTRODUCTION

The Web of data is the network resulting from publishing structured data sources in RDF and interlinking these data sources with explicit links. Web datasets are expressed according to one or more vocabularies or ontologies, which range from simple database schema exposure to full-fledged ontologies.

The Web of data requires to interlink the various published data sources. Given the large amount of published data, it is necessary to provide means to automatically link those data. Many tools were recently proposed in order to solve this problem, each having its own characteristics.

In many cases, datasets containing similar resources are published using different ontologies. Hence, data interlinking tools need to reconcile these ontologies before finding the links between entities. This could be done automatically, but more often this is done manually and built in the link specifications. This has two drawbacks: (a) this prevents to reuse the work made in ontology matching for reconciling ontologies and (b) the information about reconciling the ontologies is mixed with the information about how to identify entities.

Hence, the goal of this work is to consider data interlinking and ontology matching and to determine how these two activities are related and how they could better cooperate.

For that purpose, after briefly introducing the challenges of data interlinking and ontology matching (Section 2), we provide a general framework for data interlinking (Section 3). This framework clearly separates the data interlinking and ontology matching activities and we show how these can collaborate through three different languages for links, data linking specification and ontology alignment. We provide examples of an expressive alignment language and a modified linking specification language that can implement this cooperation (Section 4).

An extended version of this paper is available as a technical report (Scharffe and Euzenat, 2011).

2 WEB OF DATA, DATA INTERLINKING, AND ONTOLOGY ALIGNMENT

We briefly introduce the data interlinking problem. We provide examples of this problem and why it would require specific linking tools. We then present

Scharffe F. and Euzenat J..
LINKED DATA MEETS ONTOLOGY MATCHING - Enhancing Data Linking through Ontology Alignments.
DOI: 10.5220/0003666002790284
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 279-284
ISBN: 978-989-8425-80-5

why these tools could take advantage of ontology matching and alignments.

The main problem on the Web of data is to create links between entities of different datasets. Most often, this consists of identifying the same entity across different datasets and publishing a link between them as a `sameAs` statement. We call this task data interlinking and summarize it in Figure 1.

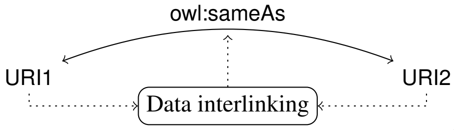

Figure 1: The data interlinking problem.

Once identified, the links discovered between two datasets must also be published in order to be reused. The VoiD vocabulary (Alexander et al., 2009) allows for describing linksets as special datasets containing sets of links between resources of two given datasets. Once linksets are constructed, two approaches are proposed to retrieve equivalences between resources: it is possible to assign to each real world entity a global identifier that will then be related to every URIs describing this entity. This is the approach taken in the OKKAM project (Bouquet et al., 2008) that proposes the usage of Entity Name Servers taking the role of resource name repositories. The other approach uses equivalence lists maintained with interlinked resources across datasets. There is thus no global identifier in this approach but equivalence links can be followed using a third-party Web service, e.g., http://sameas.org, or a bilatteral protocol (Volz et al., 2009).

The data interlinking task can be achieved manually or with the help of data interlinking tools. These tools take as input two datasets and ultimately provide a linkset. In addition, they use what we call a linking specification, i.e., a “script” specifying how and/or what to link. Indeed, given dataset sizes, the search space for resources interlinking can reach many billion resources, e.g., DBPedia. It is thus necessary to use heuristics giving hints to the interlinking system where to look for the corresponding resources in the two datasets. These linking specifications can be specific to a pair of datasets and can be reused for regenerating linksets (we provide an example of such a specification in the Silk language in Section 4).

Mining for similar resources in two Web datasets raises many problems. Each datasets having its own namespace, resources in different datasets are given different URIs. Also, although naming conventions exist, there is no formal nor standard way of naming resources. For example, if we take the URI for the famous musician Johann Sebastian Bach in various Web datasets we obtain very different results even though they all represent the same real world object.

Fortunately, dereferencing URIs can be used for retrieving more information about entities: property values and related resources can be observed. But for a same real-world entity, the same property can take different values, making the interlinking process more difficult. This can be because of varying value approximations across datasets, because of different units of measure, because of mistakes in the datasets, or because of loose ontological specifications. For instance, the property `foaf:name` does not specify in what format should the name be given. “J.S. Bach”, “Bach, J.S.” or “Johann Sebastian Bach” are possible values for this property. Hence, data interlinking tools have to compare property values in order to decide if two entities are the same, and must be linked, or not. For that purpose, tools use similarity measures based on the type of values (e.g., string, numbers, dates) and aggregate the results of these measures. This activity is reminiscent of record linkage which has been given considerable attention in database (Fellegi and Sunter, 1969; Winkler, 2006; Elmagarmid et al., 2007).

Another problem is caused by the usage of heterogeneous ontologies for describing datasets. In this case, a same resource is typed according to different classes and described with different RDF predicates belonging to different ontologies. For example, a name in a dataset can be attributed using the `foaf:name` data property from the FOAF ontology while it is attributed using the `vcard:N` object property from the VCard ontology in another dataset.

Hence, for the interlinking techniques to work, it is necessary that the datasets use the same ontology or that data interlinking tools are aware of the correspondences between ontologies.

The goal of this paper is to investigate the relationships between data interlinking and ontology matching (Euzenat and Shvaiko, 2007). In particular, we want to understand if these two activities would benefit to be merged into a single activity and sharing the same formats.

3 A FRAMEWORK FOR DATA INTERLINKING

We provide in this section a general framework encompassing the various approaches used to interlink resources on the Web of data. We first consider each case that may happen when interlinking data and describe them abstractly and through an example. In the end, we unify all this cases in a common framework.

In the first case resources are manually interlinked. Manually linking resources can be performed using collaborative tools in the case of large datasets. In some cases, illustrated below, resources can be trivially linked using a simple transformation of their URIs.

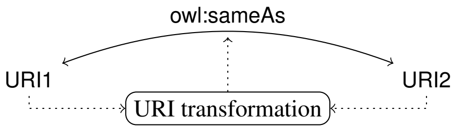

Figure 2: URI transformation.

A set of rules can be defined to identify equivalent resources from their identifier. For example, in the dataset LastFM[1], the URI representing an artist is built on the pattern "First_name+Last_name". Person URIs in DBPedia[2] are built around the pattern "FirstName_LastName". A trivial algorithm can be developed to find equivalent artists based on their URIs.

Further that the URIs, it may be necessary to consider the ontologies in order to identify entities. In a first case, the two datasets to interlink are described by the same ontology. The role of the interlinking system is to analyze resources of the same type in order to detect the equivalent ones. To do this, the system compares resource properties with a similarity measure. Systems in this category take as input the properties to compare, the type of comparison algorithm to use for each property, and the method to aggregate the similarity measures of the various properties in order to construct a measure between two resources.

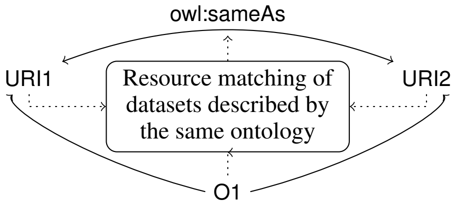
owl:sameAs
URI1
URI2
Resource matching of datasets described by the same ontology
O1

Figure 3: Matching two datasets described according to the same ontology.

For example, Jamendo and MusicBrainz, two datasets containing musicological data, are both described according to a common music ontology (Raimond et al., 2007). The artist J.S. Bach can be identified in both datasets by observing the first name and last name properties of the class *MusicArtist*. It is not possible in this case to identify the equivalence of resources based on their URIs. This example is illustrated in Figure 4.

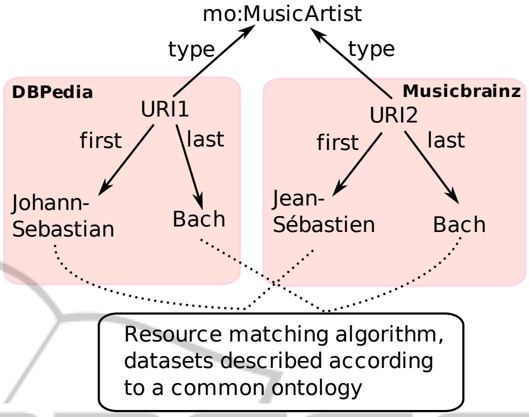

Figure 4: Example of matching two datasets described according to the same ontology.

Datasets can be described by different ontologies. This case is illustrated in Figure 5. In order to know which types of entities have to be linked together, the system needs to know the correspondences between these types of entities. Then it can work similarly as if there were a single ontology.

We represent this case in Figure 5 by introducing the correspondences between ontology classes as an alignment. This alignment is presented as implicit because it does not exist as such, but it is mixed with the data interlinking specification or system.

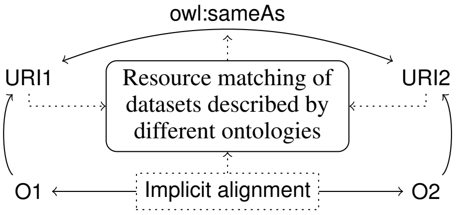
owl:sameAs
URI1
URI2
Resource matching of datasets described by different ontologies
O1
Implicit alignment
O2

Figure 5: Two datasets matched using an implicit alignment.

For example, OpenCyc[3] represents the artist J.S. Bach using a different ontology than the one used to describe MusicBrainz. The properties "firstname" and "lastname" correspond to a property "EnglishID" in which both names are concatenated. The class *MusicArtist* in the Music Ontology corresponds to a class *Classical Music Composer* in OpenCyc. An alignment between classes and properties needs to be specified in order to find an equivalence between the two resources. This example is illustrated in Figure 6.

[1] http://last.fm

[2] http://dbpedia.org

[3] sw.opencyc.org

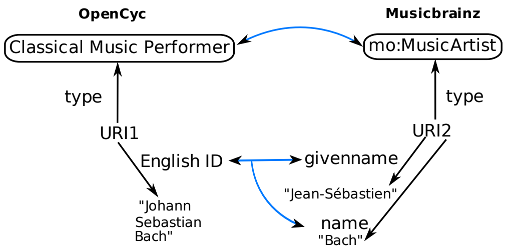

Figure 6: Example of two datasets described with heterogeneous ontologies.

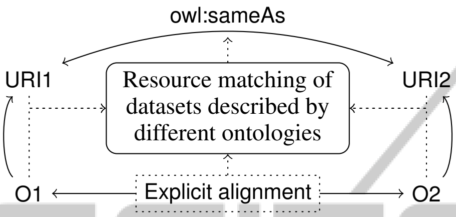
owl:sameAs
URI1
URI2
Resource matching of datasets described by different ontologies
O1
Explicit alignment
O2

Figure 7: Two datasets matched using an explicit alignment.

Another approach, illustrated in Figure 7, takes advantage of an already existing explicit alignment between the two ontologies used by the datasets.

An additional possibility, not found in existing systems, would be for the data linking system to first match the two ontologies before using the resulting alignment for supporting data interlinking. In such a system, ontology matching and data interlinking would be merged.

Figure 8 unifies all these processes in a single description. This framework leads to clarify interactions between data interlinking and ontology matching. It would be useful to use it in order to make tools interoperate. This would present many advantages, in particular the possibility to share, distribute and improve link specifications, as well as reuse them or extend them instead of computing them again whenever a dataset is modified. This would also allow to compose linking specifications such that it would be possible to go from one dataset to another without going through an intermediary. We consider below many possible ways to realize this integration.

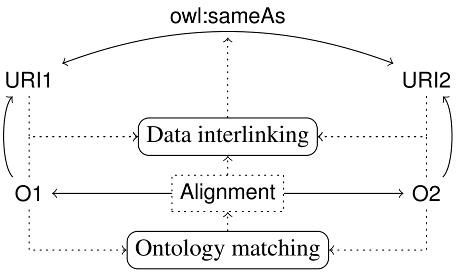

Figure 8: General framework for data interlinking involving ontology matching.

The next section discusses how using ontology alignments could lead to more automation for the interlinking task, as well as how linked data could provide evidence for obtaining better ontology alignments.

4 DATA INTERLINKING USING ONTOLOGY ALIGNMENTS

Although ontology matching and data interlinking can be similar at a certain level (they both relate formal entities), there are important differences as displayed by the the previous framework. Indeed, one acts at the schema level and the other at the instance level. These differences are reflected in the types of specification involved in these processes - a`sameAs` statement tells which `City` in wikipedia correspond to which `P` (place) in geonames, - a linking specification tells how to find the former, an ontology alignment tells which components from one ontology corresponds to which components in the other. This results in two process specifications – interlinking and matching – and their results – linksets between data and alignments between ontologies. By clearly establishing these differences, we obtain a natural partitioning between data links, linking specifications and ontology alignments:

The assertion expression language allows for representing equivalence between resources in datasets, e.g., RDF and VoiD;

The linking specification language allows for defining how to search for equivalence between resources, e.g., Silk;

The alignment representation language allows for specifying equivalence rules between ontological entities, e.g., the Alignment format or EDOAL (see below).

However, data interlinking and ontology matching could find a mutual benefit at collaborating. We propose a scheme under which it is possible for data linking tools to take ontology alignments as a way to constrain their solution space. The of Section 3, as displayed in Figure 8, provides a natural way to implement this collaboration.

EDOAL (Expressive Declarative Ontology Alignment Language) is the new name of the OMWG mapping language for expressing ontology alignment (Euzenat et al., 2007) that has been available through the Alignment API since version 3.1. This language is an extension of the Alignment format (David et al., 2011) that can be generated by most matchers. Its main purpose is to offer more expressiveness in the way alignments are expressed so it can represent complex alignment patterns (Scharffe and Fensel, 2008; Scharffe, 2009). It presents the advantage to be declarative and also to specify transformations like those needed in order to construct links between resources.

In particular, EDOAL allows for expressing contextual relations between elements. For instance, the typical example in Silk documentation is the linking of DBpedia cities and geoname P(laces) through comparing their names and populations. Expressing this with a simple alignment does not express the expected meaning because, of course, `rdfs:label` is not equivalent to `gn:name`. One could consider expressing that `gn:name` is more specific than `rdfs:label`. This is correct but still not precise enough. The intended meaning is that, in the context of `dbpdia:City` and `gn:P`, these two properties are equivalent. This is what EDOAL can express through the following alignment:

```
:dbp-geo a align:Alignment;
    align:onto1 <http://dbpedia.org/ontology/>;
    align:onto2 <http://www.geonames.org/ontology#>;
    align:map [ :map1 a align:Cell;
      align:entity1 dbpedia:City;
      align:entity2 gn:P;
      align:relation align:subsumedBy.
    ];
    align:map [ :map2 a align:Cell;
      align:entity1 [ a align:Property;
        edoal:and dbpedia:populationTotal.
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain dbpedia:City.
];
      align:entity2 [ a align:Property;
        edoal:and gn:population;
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain gn:P. ];
      align:relation align:equivalent.
    ];
    align:map [ :map2 a align:Cell;
      align:entity1 [ a align:Property;
        edoal:and rdfs:label.
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain dbpedia:City.
];
      align:entity2 [ a align:Property;
        edoal:and gn:name;
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain gn:P. ];
      align:relation align:equivalent.
    ].
```

Even if such an alignment would provide information to data interlinking tools, this is still not sufficient. Of course, it tells which properties should be equivalent and thus can be used for identifying entities. But it does not tell how to take them into account. So, this alignement would be sufficient to link entities if the values of `rdfs:label` were exactly the same as those of `gn:name` and the values of `populationTotal` were exactly the same as those of `population`, but not otherwise.

EDOAL provides more features for transforming this information. This could be helpful but the problem is deeper: data interlinking is a decision problem rather that just a transformation. It is the role of the data linking specification to tell when a `dbpedia:City` and a `gn:P` should be considered the same. This is why we propose to use data interlinking specifications together with alignments.

Indeed, using an explicit alignment, provided that it is expressive enough, can serve two functions:

1. narrowing the search space through pointing to equivalent concepts, and
2. providing the properties that can be used for identifying concepts.

A link specification like Silk-LSL (Bizer et al., 2009) fulfills two roles: - it is an alignment: it specifies the classes in which entities to link can be found, it specifies how to link entities. It could be possible to refer to an external alignment between the two underlying ontologies instead of specifying it in the linking specification. This approach would present obvious reuse advantages when other datasets requiring the same alignment, i.e., using the same ontologies, need to be interlinked.

Given that the alignment is available, it is possible to simplify the Silk specification and refer to the alignment, by introducing three types of information: which alignments to use (`UseAlignment`), entities of which correspondences must be linked (`LinkCell`) and which matched properties can be compared for identifying entities (`CellParam`).

```
<UseAlignment rdf:resource="#dbp-geo" />
<Interlink id="cities">
  <LinkType>owl:sameAs</LinkType>
  <LinkCell rdf:resource="#map1" />
  <LinkCondition>
    <AVG>
      <Compare metric="jaroSimilarity">
        <CellParam rdf:resource="#map2" />
      </Compare>
      <Compare metric="numSimilarity">
        <CellParam rdf:resource="#map3" />
      </Compare>
    </AVG>
  </LinkCondition>
  <Thresholds accept="0.9" verify="0.7" />
  <Output acceptedLinks="accepted_links.n3"
    verifyLinks="verify_links.n3"
    mode="truncate" />
</Interlink>
```

The specifics of the data interlinking task remain in this specification: how to compare values, how to aggregate their results and when to issue the link or not.

This approach presents several advantages:

1. The link specification is simplified, reducing the manual input;
2. The alignment can be reused for linking any two datasets described according to these two ontologies;
3. There is a clear separation between links, linking specification, and ontology alignments.

5 CONCLUSIONS

We have proposed an architecture based on three different languages having each its own precise purpose: expressing links, expressing linking specifications, and expressing ontology alignments. This architecture can be used in order to organize a better collaboration between ontology matchers and data interlinking tools. This can be achieved with only minimal extensions to existing languages. In particular, we have illustrated the ontology alignment part with EDOAL, an expressive ontology alignment language that offers the necessary concepts for being used in data interlinking. On the data interlinking side, we have focussed on the Silk-LSL language which seems to be at once declarative and powerful enough to express a wide range of constraints on data interlinking. Extending it with the capacity to benefit from ontology alignments would allow tools using it to benefit from the wide range of ontology alignment techniques and tools.

ACKNOWLEDGEMENTS

This work was conducted in the context of the Datalift project funded by the French ANR (Agence Nationale de la Recherche) under grant ANR-10-CORD-009.

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. (2009). Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'. In *Linked Data on the Web Workshop (LDOW09), Workshop at 18th International World Wide Web Conference (WWW09)*, Madrid, Spain.

Bizer, C., Volz, J., Kobilarov, G., and Gaedke, M. (2009). Silk - a link discovery framework for the web of data. In *18th International World Wide Web Conference*.

Bouquet, P., Stoermer, H., and Bazzanella, B. (2008). An Entity Naming System for the Semantic Web. In *Proceedings of the 5th European Semantic Web Conference (ESWC2008)*, LNCS.

David, J., Euzenat, J., Scharffe, F., and dos Santos, C. T. (2011). The alignment api 4.0. *Semantic Web*, 2(1):3–10.

Elmagarmid, A., Ipeirotis, P., and Verykios, V. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Euzenat, J., Scharffe, F., and Zimmermann, A. (2007). D2.2.10: Expressive alignment language and implementation. Project deliverable 2.2.10, Knowledge Web NoE (FP6-507482).

Euzenat, J. and Shvaiko, P. (2007). *Ontology matching*. Springer-Verlag, Heidelberg (DE).

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Raimond, Y., Abdallah, S., Sandler, M., and Giasson, F. (2007). The music ontology. In *Proceedings of the International Conference on Music Information Retrieval*.

Scharffe, F. (2009). *Correspondence Patterns Representation*. PhD thesis, University of Innsbruck.

Scharffe, F. and Euzenat, J. (2011). MeLinDa: an interlinking framework for the web of data. Rapport de recherche 7691, INRIA.

Scharffe, F. and Fensel, D. (2008). Correspondence patterns for ontology mediation. In Springer, editor, *Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008)*, pages 83–92.

Volz, J., Bizer, C., and Gaedke, M. (2009). Web of data link maintenance protocol. Protocol specification, Frei Universität Berlin.

Winkler, W. (2006). Overview of record linkage and current research directions. Technical Report 2006-2, Statistical Research Division. U.S. Census Bureau.