A modest proposal for data interlinking evaluation

Jérôme Euzenat

INRIA & LIG
Jerome.Euzenat@inria.fr

Abstract. Data interlinking is a very important topic nowadays. It is sufficiently similar to ontology matching that comparable evaluation can be overtaken. However, it has enough differences, so that specific evaluations may be designed. We discuss such variations and design.

1 Motivations

Data interlinking [4], i.e., finding links between different linked data sets, is an important topic in linked data [7]. There are many activities related to data interlinking, such as record linkage, deduplication, entity recognition, or named entily identification. Some comparison can be found in [6]. Data interlinking is similar to ontology matching in various respects and would benefit from widely, and not necessarily universally, accepted evaluations such as the Ontology Alignment Evaluation Initiative [2]. There has been isolated evaluation of link quality [5] and Instance matching evaluation has been performed as part of OAEI since 2009 [1]. Yet, there has been so far few participants to IM@OAEI and there is still no largely acknowledged benchmark for data interlinking activites.

In order to secure more participation, we analyse the specificities of data interlinking and propose diverse modalities for evaluating data interlinking.

2 The problem

The data interlinking problem could be described in the same way as the ontology matching problem was:

Given two linked data sets, usually tied to their vocabularies (or ontologies),
Generate a link set, i.e., a set of sameAs links between entities of the two data sets.

This is summarised in Figure 1.

We concentrate on sameAs links because these are by far the most important links to be retrieved when interlinking. In particular, the linked data injunction "use your own URI space" contribute to first identify locally the resources and then find the corresponding resources in other sources. Finding sameAs links, may be considered as a required first step, other types of relations may eventually be considered.

As in ontology matching, it is also possible to input the process with different external resources and parameters. More notably, a partial set of links that may be used for seeding the process. This was motivated in ontology matching by software engineering

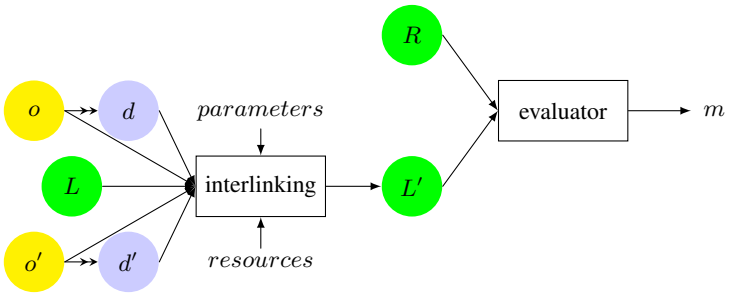

Fig. 1. Evaluating data interlinking (inspired from [3]). From two linked data sets (d and d') eventually related to their vocabularies (o and o') and possibly a training link set (L), the data interlinking process provides a link set (L') between the two data sets. This link set can be evaluated by comparison to a reference link set (R) to determine a particular measure (m).

concerns, but in data interlinking, such a partial link set may be considered a training set (see §3.3). In fact, an ontology alignment may also be provided to the interlinking process as will be considered later (§3.4).

In terms of evaluation, the same kind of procedure as in ontology matching may be used by comparing the provided link set to a reference link set. Measures such as precision, recall or time and memory consumption may be used.

Given the size of the data, it is difficult to provide correct and more specifically complete reference link sets. The OAEI 2011 IM task solved the problem by using links already provided by data providers. This is valuable when these links are curated manually. However, the quality of these links may still be questioned.

3 Specific interlinking features

We present some features of data interlinking, or of the way data interlinking is performed nowadays, that distinguish it from ontology matching. This suggests some hints to perform data interlinking evaluation.

3.1 Blocking vs. matching

Ontology matching, when confronted to large ontologies, cannot start upfront comparing instances with a similarity measure. The same applies to data interlinking. Hence many data interlinking procedures are designed in two steps:

blocking divides the sets of pairs of resources into subsets called blocks in which matching resources should be part;

matching compares entities in the same block in order to decide if they are the same or not.

Since these two steps are clearly different, well identified, and it is possible to use different matching methods with different blocking techniques, it is useful to evaluate independently the capacities of these two techniques.

In particular, from a reference link set, it is possible to determine how many pairs in the link set a particular blocking technique misses (blocking recall) and how many non necessary pairs a blocking technique imposes to compare (blocking precision). Similarly, a matching technique could be evaluated with a given block structure if this is necessary: the evaluation can be achieved by comparing the part of the reference alignments which can be found from the given blocks.

3.2 Scalability

Linked data has to deal with large amounts of data. Even if this is also the case in ontology matching, automatic data interlinking is really useful with large amounts of data. So, besides qualitative evaluation, it is critical to assess the behaviour of interlinking tools when data sizes get larger.

3.3 Learning

Another effect of the size of linked data is that learning is more relevant, mainly for two reasons:

- The size of the data makes it difficult to study it for choosing the best approach and after extracting a training sample, much work remains to be done;
- The regularity of the data facilitates machine learning efficiency.

So, it is not surprising that learning methods are successful in data interlinking [9, 8]. This provides incentive to evaluate data interlinking techniques using machine learning. For that purpose, it is necessary to provide tests in which a part of the reference link set is provided as training set to the systems which have then to deliver a complete link set.

3.4 Instance and ontology matching

Data interlinking is dependent on the vocabularies used in data sets. This vocabulary may be described by a schema or an ontology or not described explicitly. When the vocabulary is explicit, various situations may occur:

- both data sets use the same vocabularies,
- they use different vocabularies but an alignment between these vocabularies exists,
- they use different vocabularies and no alignment is available.

Some matchers may be specialised for some of these situations and it may be useful to recognise this by providing different evaluation tasks. In particular, it seems useful to test these configurations:

- without published vocabulary (or ontology),
- with the same vocabulary or, alternatively, with aligned vocabularies,
- with different vocabularies, without alignment (as in the IIMB data set of IM@OAEI).

4 Conclusions

We considered the specifics of data interlinking that it may be useful to take into account in order to design data interlinking benchmarks that may be more widely adopted by people working in the field. In conclusion, it seems that in addition to or in combination with existing tasks provided in IM@OAEI, such benchmarks should consider including:

- scalability tests (retaining one tenth, one hundredth, one thousandth of the data);
- training sets for tools using machine learning;
- separating the evaluation of blocking and matching for users who specifically consider one of these aspects only;
- tests with no ontology, the same ontology and different ontologies.

Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) under grant ANR-10-CORD-009 (Datalift).

References

1. Jérôme Euzenat, Alfio Ferrara, Laura Hollink, Antoine Isaac, Cliff Joslyn, Véronique Malaisé, Christian Meilicke, Andriy Nikolov, Juan Pane, Marta Sabou, François Scharffe, Pavel Shvaiko, Vassilis Spiliopoulos, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Cássia Trojahn dos Santos, George A. Vouros, and Shenghui Wang. Results of the ontology alignment evaluation initiative 2009. In *Proc. 4th International Workshop on Ontology Matching (OM) at the International Semantic Web Conference (ISWC)*, pages 73–126, Chantilly (VA US), 2009.
2. Jérôme Euzenat, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, and Cássia Trojahn dos Santos. Ontology alignment evaluation initiative: six years of experience. *Journal on Data Semantics*, XV:158–192, 2011.
3. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.
4. Alfio Ferrara, Andriy Nikolov, and François Scharffe. Data linking for the semantic web. *International journal of semantic web and information systems*, 7(3):46–76, 2011.
5. Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti, editors, *Proc. 9th ESWC, Heraklion (GR)*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102, 2012.
6. Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James Curran. Evaluating entity linking with wikipedia. *Artificial intelligence*, 2012.
7. Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.
8. Robert Isele and Christian Bizer. Learning linkage rules using genetic programming. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel Cruz, editors, *Proc. 6th ontology matching workshop (OM), Bonn (DE)*, volume 814 of *CEUR Workshop Proceedings*, 2011.
9. Axel-Cyrille Ngonga Ngomo. A time-efficient hybrid approach to link discovery. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel Cruz, editors, *Proc. 6th ontology matching workshop (OM), Bonn (DE)*, volume 814 of *CEUR Workshop Proceedings*, 2011.