A Survey of OpenRefine Reconciliation Services

Antonin Delpeuch[1][0000−0002−8612−8827]

Department of Computer Science, University of Oxford, UK
antonin.delpeuch@cs.ox.ac.uk

Abstract. We give an overview of the OpenRefine reconciliation API, a web protocol for tabular data matching. We suggest that such a protocol could be useful to the ontology matching community to evaluate systems more easily, following the success of the NIF ontology in natural language processing. This would make it easier for linked open data practitioners to build on the systems developed for evaluation campaigns. The OAEI task formats suggest some changes to the protocol specifications.

Keywords: record linkage · entity matching · reconciliation service · deduplication · web standards

1 Introduction

Integrating data from sources which do not share common unique identifiers often requires matching (or *reconciling*, *merging*) records which refer to the same entities. This problem has been extensively studied and many heuristics have been proposed to tackle it [1]. The Ontology Alignment Evaluation Initiative runs a yearly competition on this topic, offering a variety of task formats.

The OpenRefine reconciliation API[1] is a web protocol designed for this task. While most software packages for record linkage assume that the entire data is available locally and can be indexed and queried at will, this protocol proposes a workflow for the case where one of the data souces to be matched is held in an online database. By implementing such an interface, the online database lets users match their own datasets to the identifiers it holds. The W3C Entity Reconciliation Community Group[2], has been formed to improve and promote this protocol.

In this article, we survey the existing uses of the protocol and propose an architecture based on it to run evaluation campaigns in ontology matching.

[1] https://reconciliation-api.github.io/specs/latest/

[2] https://www.w3.org/community/reconciliation/

```
{
  "query": "Cesaria Evora",
  "type": "DifferentiatedPerson",
  "properties": [
    {
      "pid": "dateOfBirth",
       "v": "1941-08-27"
    }
  ]
}
```

(a) A reconciliation query

```
[
 {
  "id": "121291081",
  "name": "Évora, Cesária",
  "score": 92.627655,
  "match": true,
  "type":[
     {"id": "AuthorityResource"},
     {"id": "DifferentiatedPerson"}]
 },
 ...
]
```

(b) Response with candidates entities

Fig. 1: Example of a reconciliation workflow

2 Overview of the reconciliation protocol

The reconciliation API is essentially a search protocol tailored to the reconciliation problem. This protocol is implemented by many servers[3] and clients[4]. Consider the query in Figure 1. It contains the following components:

- The name of the entity to search for;
- An optional type to which the search should be restricted. The possible types are defined by the reconciliation service itself;
- An optional array of property values to refine the matching. The ontology is also defined by the reconciliation service.

We can submit this query to the reconciliation endpoint `https://lobid.org/gnd/reconcile`, which exposes the authority file of the German National Library (GND). As a response, we get a list of candidates ranked by score and a matching decision, predicting whether the entity matches the query.

The canonical client for this API is OpenRefine[5] [4], a data cleaning tool which can be used to transform raw tabular data into linked data. The tool proposes a semi-automatic approach to reconciliation, making it possible for the user to review the quality of the reconciliation candidates returned by the service. To that end, the reconcilation API lets services expose auto-complete endpoints and HTML previews for the entities they store, easing integration in the user interface of the client.

[3] A list of publicly available endpoints can be found at `https://reconciliation-api.github.io/testbench/`

[4] `https://reconciliation-api.github.io/census/clients/`

[5] `http://openrefine.org/`

3 Potential use in OAEI evaluation campaigns

In this section we turn our attention to the Ontology Alignment Evaluation Initiative, whose tasks cover among others the alignment of tabular data to knowledge bases. In these campaigns, reconciliation heuristics are evaluated on datasets covering various topics. Participants submit their systems which are run by evaluation platforms on test datasets, and their results are compared to reference alignments provided by the organizers. We argue that a web-based API such as the reconciliation API would be useful in OAEI campaigns, for multiple reasons.

The evaluation of candidate systems in OAEI events is carried out using various platforms. SEALS [8] is a Java-based tool to evaluate matching systems which has been used in OAEI campaigns for about 10 years. To be compatible with SEALS, matching systems must implement a Java interface which offers an API for ontology alignment. Participants who want to develop their systems in other programming languages have to write a Java wrapper around them, in order to be compatible with the evaluator. More recently, the HOBBIT [6] platform proposed a similar approach, where systems are submitted as Docker images and communicate with the evaluator in a similar way. Finally, the MELT platform [3] was proposed this year as a Java framework to develop systems compatible with both HOBBIT and SEALS. The newly launched SemTab challenge has been using the AIcrowd[6] platform so far. This platform does not evaluate systems directly, as participants submit the alignments produced by their systems on their own.

The complexity of this ecosystem is daunting for new participants. It also unlikely that systems packaged for the OAEI challenges are reused as such outside academia, for instance by an investigative journalist who would like to match company names to records in company registers or by a linked data enthusiast who would like to import a dataset in Wikidata.

We argue here that the communication between the evaluator and participating systems could be done via a web protocol such as the reconciliation API. This architecture is already been used in other domains. For instance, in natural language processing, it is used for *entity linking* (annotating text with mentions of named entities aligned to a knowledge base). The GERBIL platform [7] evaluates systems for this task using a web API based on NIF [2], an ontology to represent text annotation tasks. Experiments can be configured from a web interface, letting the user choose systems, datasets and evaluation metrics. Experiment results are then archived publicly.

The use of a web-based architecture has three main benefits. First, academics can evaluate their entity linking system simply by submitting to GERBIL the URL of their service. They can easily compare their systems to other services available online. Debugging services on some input data can be done easily with

[6] `https://www.aicrowd.com/challenges/semtab-2020-cell-entity-annotation-cea-challenge`

a web browser.[7] Second, systems can be used outside academia easily, as users only need to interact with a simple web API without installing anything. In turn, this use of the systems by practitioners can help source new datasets for evaluation campaigns. For instance, the Wikidata reconciliation service serves millions of queries each month. These queries can be logged, analyzed and turned into new datasets which match real-world use cases closely.

4 Adapting the protocol to the OAEI tasks

The protocol specifications are actively being discussed and improved with feedback from users, service providers and other stakeholders. Therefore, if we identify aspects of the protocol which do not fit well with the use case sketched above, it is possible to address them in a new version of the specifications.

In the SemTab challenge, the task is to match table cells to entities of a knowledge graph, without any information about the relations between columns or the domain of the dataset: these must be inferred by the service too. In contrast, reconciliation queries already identify the role of each data field using the service's ontology. One could therefore wonder whether the reconciliation protocol should be adapted not to require this information.

The anonymous reviewers have also been helpful in pointing out points that we have then forwarded to the Community Group. For instance, in some tasks a given cell can be matched to multiple entities[8]. Another useful comment was made about the absence of multilingual support in the API,[9] which had also been brought up in a different context.

5 Conclusion

We have surveyed a range of services which conform to the reconciliation API. The use of a web API such as the reconciliation API could well benefit academic initiatives such as OAEI, especially for the newly-lauched challenge on alignment of tabular data to knowledge bases [5]. Therefore, we hope to see fruitful interactions between these two communities in the future. We encourage all interested parties to join the W3C Entity Reconciliation Community Group[10].

6 Acknowledgements

We thank the anonymous reviewers, the OpenCorporates team, Vladimir Alexiev and the W3C Entity Reconciliation Community Group for their feedback on this project. This work was supported by OpenCorporates as part of the

[7] The reconciliation testbench can be used to submit queries to services: https://reconciliation-api.github.io/testbench/

[8] https://github.com/reconciliation-api/specs/issues/51

[9] https://github.com/reconciliation-api/specs/issues/52

[10] https://www.w3.org/community/reconciliation/

“TheyBuyForYou” project on EU procurement data. This project has received funding from the European Commission’s Horizon 2020 research and innovation programme (grant agreement n 780247).

References

1. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media (2012)
2. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP Using Linked Data. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Salinesi, C., Norrie, M.C., Pastor, Ó. (eds.) Advanced Information Systems Engineering, vol. 7908, pp. 98–113. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41338-4$_7$
3. Hertling, S., Portisch, J., Paulheim, H.: MELT - Matching EvaLuation Toolkit. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) Semantic Systems. The Power of AI and Knowledge Graphs, vol. 11702, pp. 231–245. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-33220-4$_1$7
4. Huynh, D., Morris, T., Mazzocchi, S., Sproat, I., Magdinier, M., Guidry, T., Castagnetto, J.M., Home, J., Johnson-Roberson, C., Moffat, W., Moyano, P., Leoni, D., Peilonghui, Alvarez, R., Vishal Talwar, Wiedemann, S., Verlic, M., Delpeuch, A., Shixiong Zhu, Pritchard, C., Sardesai, A., Thomas, G., Berthereau, D., Kohn, A.: OpenRefine (2019). https://doi.org/10.5281/zenodo.595996
5. Jimenez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K.: SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In: The Semantic Web. pp. 514–530. Springer, Cham (May 2020). https://doi.org/10.1007/978-3-030-49461-2$_3$0
6. Ngomo, A.C.N., Röder, M.: HOBBIT: Holistic Benchmarking for Big Linked Data p. 2
7. Usbeck, R., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Röder, M., Waitelonis, J., Wesemann, L., Ngonga Ngomo, A.C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D.: GERBIL: General Entity Annotator Benchmarking Framework. In: Proceedings of the 24th International Conference on World Wide Web - WWW ’15. pp. 1133–1143. ACM Press, Florence, Italy (2015). https://doi.org/10.1145/2736277.2741626
8. Wrigley, S.N., García-Castro, R., Nixon, L.: Semantic evaluation at large scale (SEALS). In: Proceedings of the 21st International Conference Companion on World Wide Web - WWW ’12 Companion. p. 299. ACM Press, Lyon, France (2012). https://doi.org/10.1145/2187980.2188033