Towards Matching of Domain-Specific Schemas Using General-Purpose External Background Knowledge

Jan Philipp Portisch[1,2] [0000−0001−5420−0663]

[1] Data and Web Science Group, University of Mannheim, Germany
jan@informatik.uni-mannheim.de
[2] SAP SE Product Engineering Financial Services, Walldorf, Germany
jan.portisch@sap.com

Abstract. Schema matching is an important and time consuming part within the data integration process. Yet, it is rarely automatized – particularly in the business world. In recent years, the amount of freely available structured knowledge has grown exponentially. Large knowledge graphs such as BabelNet, DBnary (Wiktionary in RDF format), DBpedia, or Wikidata are available. However, these knowledge bases are hardly exploited for automated matching. One exception is the biomedical domain: Here domain-specific background knowledge is broadly available and heavily used with a focus on reusing existing alignments and on exploiting larger, domain-specific mediation ontologies. Nonetheless, outside the life sciences domain such specialized structured resources are rare. In terms of general knowledge, few background knowledge sources are exploited except for WordNet. In this paper, we present our research idea towards further exploiting general-purpose background knowledge within the schema matching process. An overview of the state of the art is given and we outline how our proposed research approach fits in. Potentials and limitations are discussed and we summarize our intermediate findings.

Keywords: data integration · schema matching · ontology matching · background knowledge · knowledge graphs · financial services industry.

Category: Early Stage Ph.D.

1 Introduction

1.1 Motivation

Data integration describes the effort to allow for a unified access across multiple autonomous and heterogeneous sources of data [5]. Up to date, the data integration process is manual and requires technical experts as well as domain specialists for most systems. As a consequence, data integration is slow and expensive. Within the data integration process for two given schemas (depicted

in Figure 1), schema matching is the first step and, therefore, of main interest for this research project. It is typically very complex and not automatized. One reason is that schemas are often defined with deep background knowledge that is not itself present within the schemas [7]. Schema matching is a problem for Open Data (e.g. matching publicly available domain ontologies or interlinking concepts in the linked open data cloud) as well as for private companies which need to integrate disparate data stores. The overall research goal is to improve

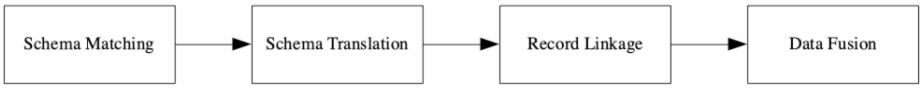

Fig. 1. Process for integrating two schemas, compiled from [34].

the data integration process by exploiting general-purpose knowledge graphs for schema matching. In terms of a business scenario, a favorable outcome would be the reduction of time that needs to be invested by human domain experts in order to accelerate data integration projects. Even though usability studies are not the main research interest of this project, an improvement can likely be achieved by providing users with a matching proposal that can be reviewed or used for human refinement. The focus of this work will be fully automatized schema matching but findings are also relevant for semi-automatic schema matching.

1.2 Industry Use Case: Matching Data Models From the Financial Services Industry

The software landscape of enterprises often resembles a heterogeneous patchwork of various systems by different vendors. Sometimes there are even multiple systems for the same task (e.g. after an acquisition). Different software components use their own data models with a large amount of overlapping parts. For a holistic understanding of the company, data has to be federated into one view. This problem is particularly pronounced in the financial services sector: Here, an understanding of a company's financial standing as well as its risk exposure is crucial for sustainable business decisions. Hence, there is an endogenous motivation to federate data. Additionally, regulators emerge to be an exogenous driver for this process by obligating financial institutions to report risk KPIs in a timely manner and even by regulating the IT infrastructure (like BCBS 239 [2]). The costs caused by regulation in the banking sector are considerable [11]. To handle the need of data federation and reporting, all individual data models of different software components have to be reconciled into one holistic view. The large size of corporate data models further complicates this process. SAP SE is developing such a data model for the financial services industry. Many applications and data stores need to be mapped into the defined data model. The

company recognizes the stated problem of schema matching and is, therefore, sponsoring this PhD project.

2 State of the Art

2.1 Background Knowledge in Ontology Matching

Schema matching can be interpreted as ontology matching task because techniques for ontology matching can also be applied to other schema matching tasks such as database schema matching [7]. In addition, approaches exist to transform other data schemas, such as entity relationship (ER) models, into ontologies [8]. Ontology and schema matching systems are evaluated by the *Ontology Alignment Evaluation Initiative (OAEI)* every year since 2005 [6]. In terms of background knowledge, many systems[3] use *WordNet* as a general language resource. Besides the latter one, few other general-purpose resources are exploited: Lin and Krizhanovsky [20] employ *Wiktionary* for translation look-ups within a larger matching system [21]. The *WikiMatch* [14] system exploits the *Wikipedia* search API by determining concept similarity through the overlap of returned *Wikipedia* articles for a search term. *WeSeE Match* [24] queries search APIs and determines similarity based on TF-IDF scores on the returned Web site titles and excerpts. Background knowledge sources are also used for multilingual matching tasks. Here, translation APIs are typically called, for example *Microsoft Bing Translator* by *KEPLER* [18] or *Google Translator* by *LogMap* [17].

In the biomedical and life science domain, specialized external background knowledge is broadly available and heavily exploited for ontology matching. Chen et al. [3] extend the LogMap matching system to use *BioPortal*, a portal containing multiple ontologies, alignments, and synonyms, by (i) applying an overlap based approach as well as by (ii) selecting a suitable ontology automatically and using it as mediating ontology. As mappings between biomedical ontologies are available, those are used as well: Groß et al. [12] exploit existing mappings to third ontologies, so called *intermediate ontologies*, to derive mappings. This approach is extended by Annane et al. [1] who use the BioPortal by exploiting existing alignments between the ontologies found there for matching through a path-based approach: By linking source and target concepts into the global mapping graph, the paths that connect the concepts in that graph are used to derive new mappings. In the same domain, research has also been conducted on background knowledge selection. Faria et al. [9] propose the usage of a metric, called *Mapping Gain (MG)*, which is based on the number of additional alignments found given a baseline alignment. Quinx et al. [31] use a keyword-based vector similarity approach to identify suitable background knowledge sources. Similarly, Hartung et al. [13] introduce a metric, called *effectiveness*, that is based on the mapping overlap between the ontologies to be matched. While in the biomedical

[3] In 2013, Euzenat and Shvaiko [7] counted more than 80 schema matching systems that exploit *WordNet*.

domain, many specialized resources are available and data schemas are heavily interlinked, this is not the case for other domains. As a consequence, such methods cannot be easily translated and applied.

In terms of the exploitation strategies, i.e. methods to use background knowledge to derive mappings, that are applied, it is notable that embedding-based approaches, such as *RESCAL* [23] or *RDF2Vec* [32], are largely underexplored.

2.2 Tooling

In order to evaluate and compare existing as well as new matching approaches, sufficient tooling is required. The Alignment API [4] defines an interface for matchers as well as alignments. It has been gradually extended and also contains evaluation capabilities. The API is used by the main evaluation platforms presented below and defines the alignment output format that is in use by the OAEI today. Two well-known evaluation platforms are employed in the ontology matching community: The *Semantic Evaluation at Large Scale (SEALS)* and the more recent *Holistic Benchmarking of Big Linked Data (HOBBIT)*. Both platforms define a matcher interface as well as a packaging pattern. Packaged matchers can be run on the platforms on evaluation data sets and evaluation scores such as *precision*, *recall*, and F_1 can be calculated. Both platforms are used in OAEI campaigns.

3 Problem Statement and Contributions

3.1 Research Questions

Up to date, publicly available knowledge graphs and resources are rarely exploited outside the biomedical domain despite their continuous growth. In particular when it comes to general background knowledge, few other resources than WordNet are used. Therefore, we see a great potential for general external knowledge sources within the matching process in the public and also in the private domain. By now, even general knowledge sources such as Wikidata contain many tail-entities and facts that might be valuable for domain specific matching tasks. While the exploitation of domain-specific knowledge sources may be more desirable, this is very often not feasible due to missing availability of such resources. We strive to explore and answer the following research questions:
RQ 1: How can general-purpose background knowledge be integrated into the schema matching process to provide value?
RQ 2: Which general-purpose external resources are valuable for data integration and what are determining factors?
RQ 3: Which background-knowledge-focused exploitation strategies are valuable in the schema matching process and are applicable to general-purpose resources?
RQ 4: Which combination of background knowledge source and exploitation strategy is most helpful in schema matching?

3.2 Further Contributions

As the stated research problem is very relevant for businesses, in particular in the financial services sector, a further contribution of this PhD will be the evaluation and application of findings in concrete business applications. Thereby, Semantic Web technologies may also be integrated into SAP standard products.

4 Research Methodology and Approach

The schema matching problem is interpreted as ontology matching problem. This allows implementing matchers against a predefined API and reduces the technical heterogeneity problem that occurs with different data schemas. In addition, existing alignments and matching tasks of the ontology matching community can be easily reused due to the same technical setting.

Naturally, schemas are not always available as ontologies – particularly in the enterprise sector. However, due to the versatility of ontologies, schemas (and their semantic definitions such as ER models) can be translated into ontologies without any loss of information. Here, the OWL format is exploited as mediating technical format.[4] After the transformation process, the ontologies can be fed into a matching system. Here, we intend to develop a unitized system that allows using different sources of background knowledge as well as different matching strategies. Unlike many other matching systems, the focus of the system here is not limited to only 1:1 correspondences but also to 1:N ones which makes it more applicable for matching relational database schemas. The resulting alignments are then parsed by an evaluation platform that allows comparing different matching systems (see Section 5). This approach has been piloted for five SAP integration scenarios and proven as technically feasible [26]. An overview of the approach is depicted in Figure 2.

5 Evaluation Plan

In terms of comparison and evaluation metrics, the most common approach is to compare the *precision*, *recall*, and F_1 scores of different approaches. We also plan to consider runtime performance aspects whereas memory consumption is regarded of lower importance given that matching itself does not have to be performed on a consumer PC. An additional suitable evaluation metric is mapping gain (although introduced in a different context). Lastly, statistical significance testing can also be applied: Recently, McNemar's test has been used to determine whether matching results are significantly different in a statistical sense [22]. We

[4] Note that the semantic expressiveness or quality of the generated technical ontologies is only as good as the inputs for the transformation and influences the results of automated matching methods. However, the outlined approach is also used for semantically richer models such as conceptual data models that are frequently used in the financial services industry, for instance.

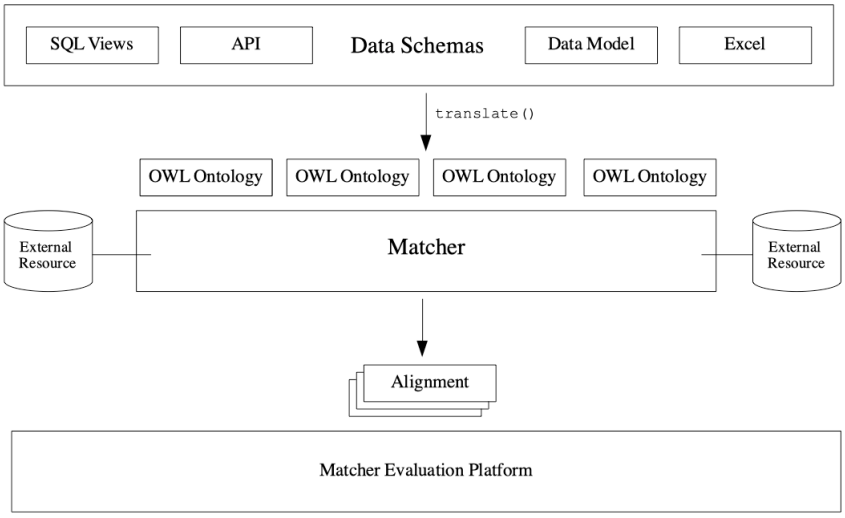

Fig. 2. High-level approach to evaluate matchers with different sources of background knowledge and strategies on existing ontologies as well as proprietary industry data models.

plan to benchmark different background knowledge sources combined with different knowledge exploitation strategies on publicly available (OAEI) data sets as well as on industry specific data sets provided by SAP SE. First preliminary results are outlined in Subsection 6.4. We further plan to explore and include novel exploitation approaches, i.e. embedding-based ones, into our evaluation.

6 Preliminary or Intermediate Results

6.1 Evaluation Runtime

In order to evaluate and compare matching systems, the *Matching EvaLuation Toolkit (MELT)* [16] has been developed. MELT allows to develop, package, and evaluate various matching systems and is integrated with the existing tooling that is used within the ontology matching community, i.e., it is compatible with SEALS and HOBBIT. Compared to existing evaluation frameworks, MELT is superior in terms of the granularity of the evaluation that can be performed and the provided functionality to evaluate multiple matchers on multiple tasks. MELT is also capable of generating an interactive dashboard which alows consume matcher results through a Web interface [25].

6.2 Creation and Evaluation of Data Model Mappings from the Financial Services Domain

In order to evaluate the matcher performance on real-world data models, five preliminary SAP data model alignments that have been created by domain ex-

perts have been translated into ontologies using a set of predefined translation rules. The translations were inspired by the ones suggested in [8] and have been extended. The existing mappings were translated into the alignment format as defined by the Alignment API. After all data was translated into publicly known data formats, current OAEI ontology matchers have been run on the data. First results [26] indicated that even top-notch OAEI matchers performed comparatively bad on real-world financial services data models.

6.3 Training of Embedded Background Knowledge

So far, embedding strategies have rarely been exploited when it comes to external knowledge in schema matching. First experiments have been conducted with the *WebIsALOD* [15] data set, a large hypernymy knowledge graph extracted from the Web, and showed positive results [30,29] for schema matching. For a deeper exploration of these strategies in schema matching, knowledge graph embeddings have been trained on four large knowledge graphs: *DBpedia* [19], *WebIsALOD* [15], *Wiktionary* [33], and *WordNet* [10]. In order to obtain the concept vectors from the knowledge graph, the *RDF2Vec* approach has been applied. The embedding models as well as the code have been published together with Web APIs[5] [28]. The models have been evaluated on three semantic similarity gold standards. First results indicate that the embeddings rather represent relatedness than similarity. As a consequence, they are likely capable of generating mappings that cannot be found by other methods but are less precise. In their current form, they could be used to improve current matching methods but perform badly when used as the only similarity function (see Subsection 6.4). In the evaluation, it could furthermore be shown that combining different graph models can outperform the single best model.

6.4 A Comparison of Sources of General Knowledge: Strategy vs. Data Source

In a larger study, three different exploitation strategies (synonymy-based, hypernymy-based, embedding-based) have been evaluated on four different knowledge graphs (DBpedia, WebIsALOD, DBnary, WordNet) with the objective to determine whether the strategy or the choice of the knowledge graph is a more dominant factor for ontology matcher performance. The results showed that – given the evaluation setting – the synonymy-based strategy performs best on all knowledge graphs. In addition, no superior general-purpose knowledge graph could be identified. This study is yet to be published.

6.5 Further Findings

Two OAEI matchers have been submitted to the OAEI: (i) The *Alod2Vec Matcher* [29] showed that it is possible to train embeddings for a background knowledge

[5] `http://kgvec2go.org/`

set and to exploit them, albeit the contribution of the background data set was low in this case. (ii) The *Wiktionary Matcher* [27] exploits multiple recent Wiktionary graphs in different language versions. It could be shown that Wiktionary can be used as background source with reasonable matching and run time performance. An additional finding was that the publicly built knowledge source is capable of handling multilingual matching tasks.

7 Conclusions and Lessons Learned

The presented approach has potential because it explores further sources of general background knowledge that can easily be integrated in any matcher and at the same time is compatible with existing exploitation strategies on domain-specific data sets. In addition, a specific business use case for the financial services domain is explored that may push the usage of Semantic Web technologies in the business world. Lastly, new exploitation methods are explored and compared which may give guidance to practitioners.

The inclusion of real-world data schemas introduces additional complications such as many-to-one correspondences that are not well represented in most existing matching systems and still need to be addressed. One risk is that the proposed background knowledge sources are insufficient for domain specific matching tasks and do not contribute at all to solving the problem. However, first results indicate that there is a positive effect in introducing larger general knowledge graphs to domain-specific problems.

Preliminary findings showed that it is possible to translate existing schemas into ontologies. It could also be shown that existing matching systems perform comparatively bad on real-world financial services data schemas. In the current (preliminary) evaluation, it was found that embedding based strategies on background knowledge do not yet outperform explicit strategies. Additionally, it could be shown that collaboratively built, non-expert reviewed background data sets such as *BabelNet* or *Wiktionary* achieve similar or better results for the task of ontology matching compared to *WordNet*.

Acknowledgements. I would like to thank my supervisor, Prof. Heiko Paulheim, for his valuable feedback, guidance, and support in the realization of this work.

References

1. Annane, A., Bellahsene, Z., Azouaou, F., Jonquet, C.: Selection and combination of heterogeneous mappings to enhance biomedical ontology matching. In: Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings. Lecture Notes in Computer Science, vol. 10024, pp. 19–33 (2016)
2. Basel Committee on Banking Supervision: Principles for Effective Risk Data Aggregation and Risk Reporting. Bank for International Settlements, Basel (2013)

3. Chen, X., Xia, W., Jiménez-Ruiz, E., Cross, V.V.: Extending an ontology alignment system with bioportal: a preliminary analysis. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014. CEUR Workshop Proceedings, vol. 1272, pp. 313–316. CEUR-WS.org (2014)
4. David, J., Euzenat, J., Scharffe, F., dos Santos, C.T.: The alignment API 4.0. Semantic Web **2**(1), 3–10 (2011)
5. Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration, chap. 1, p. 6. Morgan Kaufmann (2012)
6. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., dos Santos, C.T.: Ontology alignment evaluation initiative: Six years of experience. J. Data Semantics **15**, 158–192 (2011)
7. Euzenat, J., Shvaiko, P.: Ontology Matching, chap. 13. Springer, New York, 2nd edn. (2013)
8. Fahad, M.: ER2OWL: Generating OWL Ontology from ER Diagram. In: Intelligent Information Processing IV, 5th IFIP International Conference on Intelligent Information Processing, October 19-22, 2008, Beijing, China. IFIP Advances in Information and Communication Technology, vol. 288, pp. 28–37. Springer (2008)
9. Faria, D., Pesquita, C., Santos, E., Cruz, I.F., Couto, F.M.: Automatic Background Knowledge Selection for Matching Biomedical Ontologies. PloS one **9**(11) (2014)
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts (1998)
11. Groenfeldt, T.: Taming the high costs of compliance with tech (2018), https://www.forbes.com/sites/tomgroenfeldt/2018/03/22/taming-the-high-costs-of-compliance-with-tech/
12. Groß, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping composition for matching large life science ontologies. In: Proceedings of the 2nd International Conference on Biomedical Ontology, Buffalo, NY, USA, July 26-30, 2011. CEUR Workshop Proceedings, vol. 833. CEUR-WS.org (2011)
13. Hartung, M., Groß, A., Kirsten, T., Rahm, E.: Effective composition of mappings for matching biomedical ontologies. In: The Semantic Web: ESWC 2012 Satellite Events - ESWC 2012 Satellite Events. pp. 176–190 (2012)
14. Hertling, S., Paulheim, H.: WikiMatch - using wikipedia for ontology matching. In: Shvaiko, P., Euzenat, J., Kementsietsidis, A., Mao, M., Noy, N., Stuckenschmidt, H. (eds.) OM-2012: Proceedings of the ISWC Workshop. vol. 946, pp. 37–48 (2012)
15. Hertling, S., Paulheim, H.: Webisalod: Providing hypernymy relations extracted from the web as linked open data. In: The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Proceedings, Part II. Lecture Notes in Computer Science, vol. 10588, pp. 111–119. Springer (2017)
16. Hertling, S., Portisch, J., Paulheim, H.: MELT - matching evaluation toolkit. In: Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings. pp. 231–245 (2019)
17. Jiménez-Ruiz, E., Grau, B.C.: Logmap: Logic-based and scalable ontology matching. In: The Semantic Web - ISWC 2011, Bonn, Germany, October 23-27, 2011, Proceedings, Part I. Lecture Notes in Computer Science, vol. 7031, pp. 273–288. Springer (2011)
18. Kachroudi, M., Diallo, G., Yahia, S.B.: KEPLER at OAEI 2018. In: Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference. CEUR Workshop Proceedings, vol. 2288, pp. 173–178. CEUR-WS.org (2018)

19. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015)
20. Lin, F., Krizhanovsky, A.: Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. CoRR **abs/1109.0732** (2011)
21. Lin, F., Sandkuhl, K., Xu, S.: Context-based ontology matching: Concept and application cases. J. UCS **18**(9), 1093–1111 (2012)
22. Mohammadi, M., Atashin, A.A., Hofman, W., Tan, Y.: Comparison of ontology alignment systems across single matching task via the McNemar's test. TKDD **12**(4), 51:1–51:18 (2018)
23. Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011. pp. 809–816. Omnipress (2011)
24. Paulheim, H.: Wesee-match results for OEAI 2012. In: Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012. CEUR Workshop Proceedings, vol. 946. CEUR-WS.org (2012)
25. Portisch, J., Hertling, S., Paulheim, H.: Visual analysis of ontology matching results with the MELT dashboard. In: The Semantic Web: ESWC 2020 Satellite Events (2020), (to appear)
26. Portisch, J., Hladik, M., Paulheim, H.: Evaluating ontology matchers on real-world financial services data models. In: Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems (SEMANTiCS 2019), Karlsruhe, Germany, September 9th - to - 12th, 2019. CEUR Workshop Proceedings, vol. 2451. CEUR-WS.org (2019)
27. Portisch, J., Hladik, M., Paulheim, H.: Wiktionary matcher. In: Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019. CEUR Workshop Proceedings, vol. 2536, pp. 181–188. CEUR-WS.org (2019)
28. Portisch, J., Hladik, M., Paulheim, H.: KGvec2go - knowledge graph embeddings as a service. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC). Marseille, France (2020), (to appear)
29. Portisch, J., Paulheim, H.: ALOD2Vec matcher. In: Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference. pp. 132–137 (2018)
30. Portisch, J.P.: Automatic schema matching utilizing hypernymy relations extracted from the web (2018), `https://madoc.bib.uni-mannheim.de/52029/`
31. Quix, C., Roy, P., Kensche, D.: Automatic selection of background knowledge for ontology matching. In: Proceedings of the International Workshop on Semantic Web Information Management, SWIM 2011, Athens, Greece, June 12, 2011. p. 5. ACM (2011)
32. Ristoski, P., Rosati, J., Noia, T.D., Leone, R.D., Paulheim, H.: Rdf2vec: RDF graph embeddings and their applications. Semantic Web **10**(4), 721–752 (2019)
33. Sérasset, G.: Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. Semantic Web **6**(4), 355–361 (2015)
34. Wang, X., Haas, L.M., Meliou, A.: Explaining data integration. IEEE Data Eng. Bull. **41**(2), 47–58 (2018)