Evaluating Ontology Matchers on Real-World Financial Services Data Models

Jan Portisch[1,2][0000−0001−5420−0663], Michael Hladik[2][0000−0002−2204−3138], and Heiko Paulheim[1][0000−0003−4386−8195]

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

Abstract. Financial data in enterprises is often stored using different data models, yet, it needs to be integrated in order to foster comprehensive evaluations. Conceptually, each of those data models can be understood as an ontology, and automated ontology matching can be applied as a first step towards data integration. In this paper, we analyze the performance of existing ontology matching tools for matching financial data models. The data has been provided by SAP SE and consists of real data schemas that are used in the financial services area and mappings between them. We have created five data sets by translating enterprise data schemas to ontologies and expert mappings to ontology alignment gold standards. We evaluate state of the art ontology matchers on our newly created data set. Our experiments show that current matching systems struggle to handle enterprise data sets and achieve significantly lower scores compared to data sets of other evaluation initiatives.

Keywords: Ontology Matching · Ontology Alignment · Data Integration · Data Management · Financial Services

1 Motivation

For financial services enterprises, an understanding of the company's financial standing as well as its risk exposure is crucial for business decisions. Naturally, 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 2392 [1]). To handle the need of data federation and reporting, all individual data schemas of different software components have to be reconciled into one holistic view of the company. The required mappings between the data models require a high amount of manual work to be carried out by well-paid domain experts. Automatic or semiautomatic support during this process can help businesses in tackling these challenges in an efficient way.

Studer et al. define an ontology as "a formal, explicit specification of a shared conceptualization" [11]. Ontology matching or ontology alignment is the non-trivial task of finding correspondences between entities of a set of given ontologies

[4]. The matching can be performed manually or through the use of an automated matching system. For systematically evaluating the quality of such matchers, the Ontology Alignment Evaluation Initiative (OAEI) has been running campaigns [3] every year since 2005.

Ontologies have already been used in enterprise settings before [9] – but despite advances in ontology matching, research in this area has not yet been applied in the corporate world where it could be of use for instance for data integration.

2 Approach

2.1 Ontologies as Data Structure Descriptor

Concepts and data structures can be described using various notations and syntaxes. At SAP Financial Services, for example, data sources and data consumption layers are described, among others, by conceptual data models, physical data models, API documentation, or simply by SQL DDL statements. Depending on the notation abstraction chosen, the expressiveness varies. Ontologies can be used to describe data structures, since they are more expressive than the aforementioned notations.

In a first step, the available data was collected, and data structures were translated into ontologies using the Web Ontology Language (OWL). In a second step, the known mappings were transformed into the alignment format as defined by the *Alignmnet API* [2] which is also used by the OAEI. This process is described in the following subsection. The data is further explained in subsection 2.3. The resulting data sets follow an open format and can be processed by regular ontology matchers.

2.2 Transformation of Data Schemas

To address the problem of heterogeneity of notations, all schemas were transformed into ontologies by schema-specific adapters. We have adopted the approach for translating entity relationship models to ontologies introduced in [5], and extended it to account for model-specific idiosyncrasies. The semantically richest data structures used here are conceptual data models. Generally, entities are translated to classes, attributes are translated to datatype properties with a maximal cardinality of 1 and with the corresponding class as domain, relationships are translated to object properties, and inheritances are directly taken into account using `rdfs:subClassOf`. In addition, mandatory attributes were assigned the restriction of a minimal cardinality of 1 and key fields were marked using `owl:hasKey` which was introduced in OWL 2. Similarly, the cardinalities of relationships were translated into the ontology by using restrictions. Labels and definitions can also be found in the resulting ontology whenever they are available in the original source structure. This process was likewise applied in a similar fashion to the other data structures evaluated here where applicable.

2.3 Data

The data has been provided for research by SAP SE. The *SAP Financial Services Data Platform (FSDP)* is a solution with the purpose to help financial institutions with their data management. It includes a semantically rich conceptual data model (CDM) and a performance optimized physical data model (PDM) that can be deployed on a column-based database. As analytical (OLAP) and transactional (OLTP) applications run on the platform, inbound and outbound mappings are required. Data sources and consumers are mapped to the CDM. All mappings used here were manually created by multiple experts from the banking and insurance domain within SAP and map to the FSDP CDM.

The first data set (D_1) is derived from the mapping between the conceptual and the physical data model of FSDP. This is the largest data set. Because of performance improvements and implementation adaptions, the entities of the models are different. The second data set (D_2) consists of a mapping between the FSDP CDM and a regulatory reporting application which brings its own data model. The third data set (D_3) maps between the FSDP CDM and an SAP accounting solution. The fourth data set (D_4) is a mapping between the FSDP business partner and the business partner of SAP ERP. The last data set (D_5) maps between a loans management system and FSDP. Data sets D_2, D_3, and D_5 are work in process and only the mapped structures were kept in the corresponding ontology. Data sets D_1 and D_4 are complete. Table 1 gives an overview over the data sets used.

Data Set	**Source**			**Target**			**# of Corr.**	**Arity**
	C	P_D	P_O	**C**	P_D	P_O		
D_1	760	3373	687	438	6878	0	4645	n:n
D_2	760	4355	687	1	70	0	251	n:n
D_3	760	4355	687	11	100	0	131	n:n
D_4	760	4355	687	12	43	0	60	n:n
D_5	760	4355	687	6	19	0	31	n:n

Table 1. The derived data sets consisting of two ontologies each and an alignment. C refers to the number of classes, P_D to the number of datatype properties, and P_O to the number of object properties.

3 Preliminary Results

For a first analysis, all OAEI 2018 matchers were ran on the data set. In addition, a simple string matcher[3] has been used as baseline. The individual matcher performance is given in Table 2. For an overall statistic, macro average was chosen due to the different size and difficulty of the data sets. Macro averages can

[3] `BaselineStringMatcher` of the MELT framework [6].

be found in Table 3. All statistics were calculated using the MELT framework[4] [6].

Out of the matchers evaluated, only 5 matchers returned non-empty alignments. Out of those, ALOD2Vec [10], LogMap Light [7], and Kepler [8] were the only matchers to find a non-empty alignment for all data sets.[5] Kepler performs best in terms of F_1. It is outperformed by LogMap Light when using macro recall as benchmark.

Early experiments indicate that current matchers struggle to match real world industry data schemas. Likely explanations are missing background knowledge, shallow and weakly structured ontologies, and potential overfitting to publicly available benchmarks. In addition, most OAEI data sets and matchers focus on a 1-1 alignment arity while the data sets evaluated here are more complex.

		Alod2Vec	**AML**	**LogMap**	**LogMap Lt**	**Kepler**	**Baseline**
D_1	**Precision**	0.3596	0.6016	0.9628	0.3432	0.6950	0.7210
	Recall	0.7991	0.6129	0.0893	0.7929	0.5681	0.5414
	F_1	0.4960	0.6072	0.1635	0.4790	**0.6252**	0.6185
D_2	**Precision**	0.5555	-	-	0.4000	0.7143	0.6667
	Recall	0.0199	-	-	0.0239	0.0398	0.0159
	F_1	0.0385	-	-	0.0451	**0.0754**	0.0311
D_3	**Precision**	0.2333	-	-	0.0769	0.2714	0.2667
	Recall	0.0534	-	-	0.1603	0.1450	0.0612
	F_1	0.087	-	-	0.1040	**0.1891**	0.0994
D_4	**Precision**	0.8571	-	-	0.8571	0.8889	0.8571
	Recall	0.100	-	-	0.100	0.1333	0.100
	F_1	0.1791	-	-	0.1791	**0.2319**	0.1791
D_5	**Precision**	0.0909	-	1.0000	0.1176	0.5000	0.1000
	Recall	0.0323	-	0.0323	0.0645	0.1290	0.0322
	F_1	0.0476	-	0.0625	0.0833	**0.2051**	0.0488

Table 2. Individual Performance Results of the Matchers. The best F_1 score is printed in bold.

4 Challenges and Future Work

While the current prototypical set-up shows how ontology matching can be applied in a real enterprise setting, there are still many challenges that need to be addressed. The current data sets presented in this paper give a first indication of the performance of current state of the art matchers on real financial services data models. However, the data sets are yet small and incomplete. We plan to extend the current data base by increasing the amount of data and to improve

[4] https://github.com/dwslab/melt/

[5] Note that all matchers in Table 2 could process each data set – i.e., the problems are rather semantic than technical.

System	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F_1
Alod2Vec	0.4193	0.2010	0.2717
AML	0.1203	0.1226	0.1214
LogMap	0.3926	0.0245	0.0458
LogMap Lt	0.3590	**0.2283**	0.2791
Kepler	**0.6139**	0.1973	**0.3052**
Baseline	0.5223	0.1501	0.2332

Table 3. Macro Averages of the 5 Test Cases. The best macro precision, recall, and F_1 are printed in bold.

its quality. When the data base is grown to a more significant size and a high level of quality can be ensured, we consider offering a blind alignment track at the OAEI. Since the results show that financial services data models cannot be matched without background knowledge, future work will also focus on evaluating suitable sources of background knowledge, and on developing robust matchers that can handle loosely structured data schemas.

Acknowledgements. Acknowledgements go to Gaurav Sharma and Stephan Schub for helping compiling the mappings and overcoming technical obstacles.

References

1. Basel Committee on Banking Supervision: Principles for Effective Risk Data Aggregation and Risk Reporting. Bank for Internat. Settlements, Basel (2013), http://www.bis.org/publ/bcbs239.htm
2. David, J., Euzenat, J., Scharffe, F., Trojahn dos Santos, C.: The Alignment API 4.0. Semantic Web Journal **2**(1), 3–10 (2011)
3. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: six years of experience. In: Journal on data semantics XV, pp. 158–192. Springer (2011)
4. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, New York, 2nd edn. (2013)
5. Fahad, M.: ER2OWL: generating OWL ontology from ER diagram. In: Intelligent Information Processing. IFIP Advances in Information and Communication Technology, vol. 288, pp. 28–37. Springer (2008)
6. Hertling, S., Portisch, J., Paulheim, H.: MELT - Matching EvaLuation Toolkit. In: Semantics 2019 SEM2019 Proceedings. Karlsruhe (2019, to appear)
7. Jiménez-Ruiz, E., Grau, B.C., Cross, V.: Logmap family participation in the OAEI 2018. In: OM@ISWC. CEUR Workshop Proceedings, vol. 2288, pp. 187–191. CEUR-WS.org (2018)
8. Kachroudi, M., Diallo, G., Yahia, S.B.: KEPLER at OAEI 2018. In: OM@ISWC. CEUR Workshop Proceedings, vol. 2288, pp. 173–178. CEUR-WS.org (2018)
9. Oberle, D.: How ontologies benefit enterprise applications. Semantic Web **5**(6), 473–491 (2014)
10. Portisch, J., Paulheim, H.: Alod2vec matcher. In: OM@ISWC. CEUR Workshop Proceedings, vol. 2288, pp. 132–137. CEUR-WS.org (2018)
11. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data Knowledge Engineering **25**(1-2), 161–197 (1998)