Interactive Pay as you go Relational-to-Ontology Mapping

Christoph Pinkel[1]

fluid Operations AG, D-69190 Walldorf, Germany,
{firstname.lastname}@fluidops.com

Abstract. Ontology Based Data Access (OBDA) enables access to relational data with a complex structure through ontologies as conceptual domain models. To this end, mappings are required. A key aim of OBDA is to facilitate access to data with a complex structure. Ironically, though, in today's existing OBDA systems mappings typically need to be compiled by hand, which is a complex and labor intensive task.
Additionally, existing semi-automatic mapping approaches suffer from high human effort for cleaning up results. Fully automatic approaches, on the other side, suffer from a lack of precision and/or recall. In setups where the correctness of query results is crucial but the initial human effort must still be kept be small as possible, neither approach is acceptable. This situation calls for a guided, *pay as you go* feedback process for human mapping validation.
We envision a comprehensive suite of methods and techniques that work well with one another in a seamless mapping process and support mapping construction in the context of OBDA. This suite will in part consist of a recombination and adaptation of various existing methods, but will also comprise newly devised algorithms and techniques.

1 Problem Statement

In recent years it has become increasingly important for companies throughout the industry to analyze their data. This raises a number of technical problems, as the amount of available data is growing heavily not only in size but also in complexity. Ontology-based data access (OBDA) [1] is an approach that has recently emerged to provide semantic access to complex structured relational data. Using the ontology and the mappings, domain experts can access the data directly by formulating queries in terms that reflect their vocabulary and conceptualization. Using query rewriting techniques, the end-user queries are then translated into queries over the underlying data sources. A key requirement for OBDA is a set of declarative *mappings*, relating the ontological schema elements (e.g., classes and properties) with the relational schema elements (e.g., tables and attributes) of the underlying data sources.

Today, most approaches for ontology-based data access focus on the definition of mapping languages and the efficient translation of high-level user queries over an ontology into executable queries over relational data [1,2]. These approaches assume that a declarative mapping of the schema elements of the ontology to the relational elements is already given. So far, in real-world systems [3,4] that follow the ontology-based data access principle, the mappings have to be created manually. The costs for the manual creation of mappings constitute a significant entry barrier for applying OBDA in practice.

Though many research efforts have been made on automatic and semi-automatic mapping construction, so far none have been specifically fit to OBDA and its specific needs: there is a significant *impedance mismatch* between the relational and ontology models. Detecting similarities between them (and thus potential matches) therefore requires to take into account the different design patterns and fundamental properties of either side. While lexical similarities can be used to detect matches cross-model, the same is not so straight-forward for structural or semantic similarity aspects. For example, there can be very specific correspondences between certain *structural* aspects in a relational schema and certain *semantic* aspects in an ontology that do not seem to correspond at first sight.

Also, most existing semi-automatic mapping approaches suffer from high human effort for cleaning up results, while fully automatic approaches suffer from a lack of precision and/or recall. Neither is typically acceptable for data analysis scenarios in the industry. This situation calls for a guided, pay as you go feedback process including mapping validation by humans. And while pay as you go mapping construction has also been researched, existing approaches assume a process that is very different from the one that we consider best for complex OBDA scenarios. In particular, we assume that there are expert users with the ability to precisely formulate their information need in application domain terms. Such expertise could drive an informed query-by-query mapping process with highly detailed user feedback. To date, this potential is poorly used.

In practice, this often leads to a tedious process involving coordination in a team of various domain and IT experts.

2 Relevancy: Why is the problem relevant?

Effective understanding of complex data is a crucial task for enterprises to support decision making and retain competitiveness on the market. This task is not trivial especially since the data volume and complexity keep growing fast in the light of Big Data [5]. While there are many techniques and tools for scalable data analytics today (e.g., [6]), there is little known on how to find the *right* data.

Today, enterprise information systems of large companies store petabytes of data distributed across multiple – typically relational – databases, each with hundreds or sometimes even thousands of tables (e.g., [7]). For example, an installation of an SAP ERP system comes with tens of thousands of tables [8]. Due to the complexity of data a typical scenario for data analyses today involves a domain expert who formulates an analytical request and an IT expert who has to understand the request, find the data relevant to it, and then translate the request into an executable query. In large enterprises this process may iterate several times between the domain and IT experts, the complexity of data and other factors, and may take up to several weeks.

In this light OBDA [1] has emerged as a useful technique to *facilitate* access to large databases with complex schemata. The overall aim is to save time and effort for formulating queries. In practice, however, this aim gets thwarted by the high effort to produce and maintain the required mappings for complex data.

Finding a holistic solution for reducing the effort of mapping construction is therefore key to enabling OBDA to solve significant real-world problems.

3 Related Work

A lot of research efforts have been made in the field of semi-automatic mapping approaches in general. Often, they employ lexical similarity of terms together with structural similarity ([9,10,11,12] or [13,14] for surveys). None of those, however, are designed to specifically consider the impedance mismatch between two data models as different as relational schemata and ontologies. In Yam++ [15] the authors exploit both sub-class and sub-property semantics as well as structural graph information in a multi-strategy approach. While this approach resembles our idea to leverage lexical, semantic and structural information to bridge the impedance mismatch of different data models, Yam++ still simply uses those different strategies within the task of ontology alignment.

There are also some approaches for mapping relational schemata to ontologies. However, no techniques have so far been designed for the specific needs of OBDA. In fact, most approaches instead try to transform the OBDA mapping generation problem into a better understood, yet not equivalent problem (e.g., ontology alignment [13]). For example, [16] transforms relational schemata and ontologies into directed labeled graphs respectively and reuse COMA [17] for essentially syntactic graph matching.

The few approaches for directly matching aspects from relational schemata to corresponding aspects in ontologies to date are not are not being used with OBDA and have been written with a different motivation and under vastly different preliminaries. Ronto [18] uses a combination of syntactic strategies to discover mappings by distinguishing the types of entities in relational schemata. The authors of [19] exploit structure of ontologies and relational schemata by calculating the confidence measures between virtual documents corresponding to them via the TF/IDF model. The authors support that any purely manual approach to constructing mappings would be tedious and therefore improbable. Finally, [20] describes an approach to derive complex correspondences for a relational schema to ontology mapping using simple correspondences as input. The paper mentions the problem of different design patterns used in ontologies and relational databases, but stops short of addressing the issue in general. Instead, the authors focus on a special case to follow their primary aim of deriving complex mappings.

None of these approaches do support user feedback or incremental mapping construction in a pay as you go fashion.

Approaches that do involve basic user interaction include (e.g., [21,22,23,24]). Typically, these pay as you go approaches assume a classical information retrieval scenario with a large number of users, massive but simple *end user* feedback, a lot of noise and *statistical* methods to harvest feedback. This is in contrast to our plans of harvesting very explicit feedback from a small number of expert users. Similarly, classical human computer interaction and, more recently, crowd sourcing have been looked into (e.g., [25]) but they remain just as limited in perspective of seeing users as a large sample to be observed statistically.

4 Research Questions

We plan to provide a comprehensive suite of methods and techniques for a seamless, interactive OBDA mapping process incorporating the feedback of expert users. The overall aim is to reduce human effort in the process.

Research topics therefore primarily include to:

- identify relevant matching aspects between relational schemata and ontologies and find a suitable model to express them;
- adapt existing match discovery methods to work with this model and evaluate them to identify the most suitable ones; develop novel algorithms where existing methods fail to meet expectations;
- devise an interactive, incremental pay as you go process to suggest mappings based on those matches in a query-driven fashion with the aim of minimizing human effort; advance pay as you go techniques for schema mappings to effectively incorporate expert feedback;
- and to leverage partial mappings to enhance the quality of subsequent mapping suggestions.

Additionally, a number of side topics may be touched, including issues around mapping validation, advancing models for evaluation of human effort and human computer-interfacing aspects in the interactive process.

5 Hypotheses

In accordance with the general motivation given before, we assume that the need for large and complex relational to ontology mappings exists and will even grow further. We also assume that manual mappings are unfeasible in this context as they require too much human effort.

Hypothetically, we assume that the current (semi-)automatic approaches are insufficient to tackle this problem and that these insufficiencies are, to a large part, to the ignorance of existing approaches towards the impedance mismatch between relational schemata and ontologies, as well as to the predominant *all-at-once approach*. We use the term of all-at-once approaches to refer to the typical mapping suggestion systems where large mapping problems are being solved in one step, resulting in a huge number of suggestions including many false positives.

We hypothesize that an interactive, incremental approach can levy many limitations of (semi-)automatic approaches, but that repeated feedback at *several intermediate stages* of the mapping process is therefore required.

As a last hypothesis central we assume that the overall human effort involved to complete an OBDA mapping will be greatly reduced in such a process versus any all-at-once approach.

6 Approach and Preliminary Results

We propose to build a system to facilitate mapping generation specifically for OBDA, heavily using interaction with expert users in an incremental, query-driven fashion. In contrast to earlier approaches we assume that the integration process will be assisted by a small number of domain experts who have the time and knowledge to precisely formulate their information needs and give feedback when asked for. We also assume that those users require correct and complete results, i.e., maximum precision and full recall. To get users started as quick as possible and deliver results as early as possible in the mapping process, we embark on a pay as you go approach with advanced user interaction. We then combine semi-automated schema matching techniques based on query input with highly informed feedback from explicit user interactions.

Whenever a query is not yet supported by sufficient existing mapping information to deliver precise results the system engages in any number of interactions necessary to complete the missing information. Similarly, the user could always interrupt the process to specify information that he or she deems helpful to complete the mapping sooner.

Interaction methods could comprise (a) explicit on-the-fly matching of some of the variables or properties to some tables or columns in the input data base (if the user has the necessary knowledge for doing so), (b) sample result oriented mapping by providing some example values, or (c) by sorting out possible partial interpretations to reduce the number of possibilities. A series of additional measures could be imagined as well, a lot could probably be gained to reverse-applying provenance and lineage analysis techniques (such as [26,27]).

Eventually, the system would learn a mapping from successful queries, interaction and feedback.

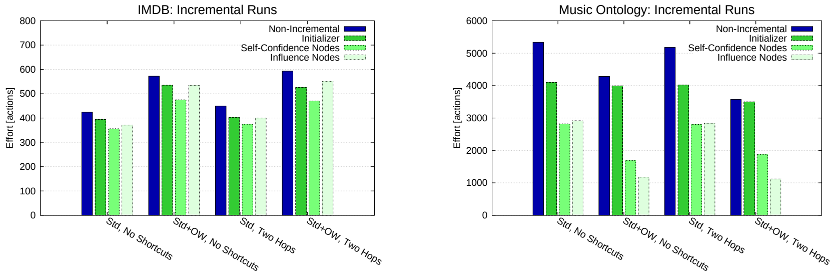
6.1 First Results

As a first step we have implemented a prototype called *IncMap*, which semi-automatically suggests simple matches (one-to-one correspondences) between an ontology and a relational schema. The prototype works query driven, producing matches only for basic schema elements (i.e., concepts and properties) required by the query at hand. For those elements the user is presented with a ranked series of suggestions, which he or she can either accept or reject. After a basic element is confirmed we re-rank suggestions for following elements.

The matching approach of *IncMap* is inspired by the Similarity Flooding algorithm of Melnik et al. [28]. However, applying the Similarity Flooding algorithm naively for matching schema elements of a relational schema to an OWL ontology results in rather poor quality of the suggested correspondences as we show in our experiments. A major reason is the impedance mismatch between ontologies and relational schemata. We therefore adapted the internal graph model to account for having an ontology on one side and a relational schema on the other.

We measured the effort for a user to complete the mapping as far as required by the reference queries by simply counting the number of required validations, i.e., the number of “reject” or “accept” clicks.

The most significant finding in this simplified scenario so far is that the overall effort can be significantly reduced in an incremental setting (i.e., by considering previous user feedback and by re-ranking suggestions accordingly) over a non-incremental setting.



(a) IMDB Schema/Movie Ontology (b) MusicBrainz Schema/Music Ontol.

Fig. 1. *IncMap* Experimental Evaluation

Figure 1 shows the measured effort on two different schema/ontology pairs for different parametric settings, comparing the non-incremental case vs. three different flavors of incremental harvesting of user feedback. Results clearly show that incremental approaches generally outperform the non-incremental approach, in some cases even by more than 50%.

The different parametric settings (bar groups) refer to different internal graph representations that we have tried out. Incremental methods differ in how user feedback is incorporated for re-ranking. Schemata and ontologies used are IMDB[1], the Movie Ontology[2], MusicBrainz[3] and the Music Ontology[4].

7 Reflections

So far, no holistic approaches for semi-automatic mapping generation have been published that focus specifically on OBDA. However, as argued before, there are a number of very specific properties in OBDA that could be harvested to produce mapping suggestions. It should therefore be generally possible to improve over existing approaches by devising algorithms that are tuned to the specific requirements.

[1] http://www.imdb.com

[2] http://www.movieontology.org

[3] http://musicbrainz.org/doc/MusicBrainz_Database

[4] http://www.musicontology.com

More than that, however, we believe to succeed by considering two aspects as mutually dependent, that are usually designed separately: the mapping process during which users repeatedly interact with the mapping system on the one hand, and the actual mapping and matching algorithms used in this system on the other. By proceeding incrementally and in multiple steps we can interactively request all kind of expert user feedback that we consider useful and harvest this feedback to suggest better mappings.

8 Evaluation Plan

Automatically generated mappings are usually being evaluated on the basis of precision and recall. In our motivating scenario, however, mappings are typically required to be eventually perfect w.r.t. user expectations.

Instead, the mainly relevant measure is the amount of *human effort* that it takes to complete the process, i.e., to reach the perfect mapping. Besides the possibility to perform user studies there are no generally accepted benchmarks to model human effort for our purpose, though.

Our plan to evaluate our suite of methods is therefore threefold: First, we use simple models to estimate user effort, e.g., by counting user actions. This is what we also did for our first results that we briefly discussed in Section 6. Second, we plan to extend these models to become more accurate and possibly contribute to efforts of creating suitable benchmark models for human effort in interactive mapping systems. And third, we plan to perform actual user studies of our approach at different stages of the project.

9 Acknowledgements

This work was supported by the Seventh Framework Program (FP7) of the European Commission under Grant Agreement 318338, the Optique project.

References

1. Poggi, A., Lembo, D., Calvanese, D., Giacomo, G.D., Lenzerini, M., Rosati, R.: Linking Data to Ontologies. J. Data Semantics **10** (2008) 133–173
2. Kontchakov, R., Lutz, C., Toman, D., Wolter, F., Zakharyaschev, M.: The Combined Approach to Ontology-Based Data Access. In: IJCAI. (2011) 2656–2661
3. Hepp, M., Wechselberger, A.: OntoNaviERP: Ontology-Supported Navigation in ERP Software Documentation. In: International Semantic Web Conference. (2008)
4. Blunschi, L., Jossen, C., Kossmann, D., Mori, M., Stockinger, K.: SODA: Generating SQL for Business Users. PVLDB **5**(10) (2012) 932–943
5. Beyer, M.A., Lapkin, A., Gall, N., Feinberg, D., Sribar, V.T.: 'Big Data' is Only the Beginning of Extreme Information Management. Gartner rep. G00211490 (2011)
6. Apache Hadoop: http://hadoop.apache.org/ (2013)
7. Crompton, J.: Keynote talk at the W3C Workshop on Sem. Web in Oil & Gas Industry (2008) http://www.w3.org/2008/12/ogws-slides/Crompton.pdf.
8. SAP HANA Help: http://help.sap.com/hana/html/sql_export.html (2013)

9. Jiménez-Ruiz, E., Grau, B.C.: LogMap: Logic-Based and Scalable Ontology Matching. In: International Semantic Web Conference (1). (2011) 273–288
10. Lambrix, P., Tan, H.: SAMBO – A system for aligning and merging biomedical ontologies. J. Web Sem. **4**(3) (2006) 196–206
11. Fagin, R., Haas, L.M., Hernández, M.A., Miller, R.J., Popa, L., Velegrakis, Y.: Clio: Schema Mapping Creation and Data Exchange. In: Conceptual Modeling: Foundations and Applications. (2009) 198–236
12. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Trans. Knowl. Data Eng. (2009) 1218–1232
13. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. **25**(1) (2013) 158–176
14. Rahm, E., Bernstein, P.A.: A Survey of Approaches to Automatic Schema Matching. In: VLDB J. (2001) 334–350
15. Ngo, D., Bellahsene, Z.: YAM++: A Multi-strategy Based Approach for Ontology Matching Task. In: EKAW. (2012) 421–425
16. Dragut, E.C., Lawrence, R.: Composing Mappings Between Schemas Using a Reference Ontology. In: CoopIS/DOA/ODBASE (1). (2004) 783–800
17. Do, H.H., Rahm, E.: COMA – A System for Flexible Combination of Schema Matching Approaches. In: VLDB. (2002) 610–621
18. Papapanagiotou, P., Katsiouli, P., Tsetsos, V., Anagnostopoulos, C., Hadjiefthymiades, S.: Ronto: Relational to Ontology Schema Matching. In: AIS SIGSEMIS BULLETIN. (2006) 32–34
19. Hu, W., Qu, Y.: Discovering Simple Mappings Between Relational Database Schemas and Ontologies. In: ISWC/ASWC. (2007) 225–238
20. An, Y., Borgida, A., Mylopoulos, J.: Inferring Complex Semantic Mappings Between Relational Tables and Ontologies from Simple Correspondences. In: OTM Conferences (2). (2005)
21. Tran, T., Wang, H., Haase, P.: Hermes: Data Web Search on a Pay-as-you-go Integration Infrastructure. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (2009) 189 – 203 The Web of Data.
22. Jeffery, S.R., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go User Feedback for Dataspace Systems. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. SIGMOD '08, New York, NY, USA, ACM (2008) 847–860
23. Hentschel, M., Haas, L., Miller, R.J.: Just-in-time data integration in action. Proc. VLDB Endow. **3** (September 2010) 1621–1624
24. Madhavan, J., Jeffery, S.R., Cohen, S., (luna Dong, X., Ko, D., Yu, C., Halevy, A., Inc, G.: Web-scale Data Integration: You Can Only Afford to Pay As You Go. In: In Proc. of CIDR-07. (2007)
25. Parameswaran, A., Polyzotis, N.: Answering Queries using Humans, Algorithms and Databases. In: Conference on Inovative Data Systems Research (CIDR 2011), Stanford InfoLab (January 2011)
26. Chapman, A., Jagadish, H.V.: Why not? In: Proceedings of the 35th SIGMOD international conference on Management of data. SIGMOD '09, New York, NY, USA, ACM (2009) 523–534
27. Glavic, B., Alonso, G., Miller, R.J., Haas, L.M.: TRAMP: Understanding the Behavior of Schema Mappings Through Provenance. Proc. VLDB Endow. **3** (September 2010) 1314–1325
28. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: ICDE, IEEE Computer Society (2002)