Towards Characterising Data Exchange Solutions in Open and Closed Words (Extended Abstract)

Henrik Forssell[1] Evgeny Kharlamov[2] Evgenij Thorstensen[1]

[1] University of Oslo [2] University of Oxford

Data exchange is the problem of translating information structured under a source schema into a target schema, given a source data set and a set of declarative mappings between the source and target schemata. The study of data exchange has recently received significant attention from both database theory and systems communities and we refer the reader to PODS and SIGMOD keynotes [2, 10] for overviews. Moreover, major database systems have adapted existing data exchange implementations [4, 13].

In data exchange, a set of schema mapping M is defined as a set of source-to-target tuple generating dependences [1]. In general such mappings only partially specify how to populate attributes of the target schema with data from the source instance S. Thus, a data exchange *solution* is in general an incomplete target data instance V that contains labeled nulls. Such V *represents* a set of possible complete target data instances denoted $\mathsf{Rep}(V)$. Several Rep functions were considered in the context of data exchange [5–7, 12], and they correspond to different data exchange semantics. Fagin et al [5, 6] proposed an open world (OWA) semantics based on the classical Rep of [8], which we denote Rep_O, Hernich et al [7] proposed a closed world (CWA) semantics, which was further extended by Libkin et al [12] to a semantics of mixed open-and-closed (OCWA) worlds and they both are based on a different notion of Rep, called Rep_A defined for *annotated* incomplete instances.

The *canonical solution* that is obtained by *chasing* [1] the source instance with mappings is considered in all semantics as a good data exchange solution for materialisation. However, the canonical solution may not be optimal for storing: it may contain redundant information and in general there might be another 'smaller' solution V that represents the same target instances. Thus, from the practical point of view, a *minimal* such V according to some order would be the best for materialisation. As the consequence, deciding whether two incomplete instances V_1 and V_2 represent the same set of complete ones is a fundamental problem underlying data exchange. For OWA semantics this decision problem can be *characterised* in terms of homomorphisms [5]: $\mathsf{Rep}_O(V_1) = \mathsf{Rep}_O(V_2)$ iff V_1 and V_2 are homomorphically equivalent. Moreover, the minimality problem has a unique solution, called the *core* [6]. The situation changes when we turn our attention to OCWA semantics: $\mathsf{Rep}_A(V_1) = \mathsf{Rep}_A(V_2)$ now cannot be *characterised* in terms of homomorphisms as before. Moreover, to the best of our knowledge, the problem of characterisation and minimality has not been studied in the context of OCWA.

The goal of this work is to address both the characterisation and minimality problem in the setting of OCWA semantics. As a first step we study the case of a restricted but natural class of OCWA mappings where all nulls are open while occurrences of constants can be either open or closed.

Our contributions are as follows. We propose an alternate definition of Rep, which we call Rep_C, that is based on homomorphic *covers*, and a new data exchange semantics

based on Rep_C. A homomorphic cover from an instance V' to V'' is a finite set of homomorphisms from V' to V'' such that the union of their images is all of V''. This allows us to characterise when $\mathsf{Rep}_C(V_1) = \mathsf{Rep}_C(V_2)$ in terms of homomorphisms and thus opens doors for the study of minimality. In particular, we show that $\mathsf{Rep}_C(V_1) = \mathsf{Rep}_C(V_2)$ iff V_1 and V_2 are *cover-equivalent* (they homomorphically cover each other). We then show that our definitions naturally extend the OCWA semantics [12], in the sense that each their data exchange solution can be translated into our that represents the same set of complete instances, but not the other way around. Finally for the problem of minimisation we introduce several natural orders on incomplete instances, show that for all of them there is in general no unique minimal element. At the same time we identify one, which we called *cover-core*, or *c-core* that has desirable semantic properties.

The notion of homomorphic cover has been used elsewhere (e.g. [3, 9, 11]). In our opinion several more data management scenarios can benefit from it. For instance, two conjunctive queries whose relational structures cover each other retrieve the same tuples from every relation of any database instance, a fact of potential relevance in e.g. data privacy settings. For another example, treating one conjunctive query as a view, it can be used to completely rewrite another if there exists a cover from the view. Thus in this setting, cover-equivalence corresponds to mutual complete rewritability.

Acknowledgements This was was partially supported by: the Norwegian Research Council grant no. 230525; SIRIUS SFI[1]; the EPSRC projects MaSI^3, DBOnto, and ED^3.

References

1. S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
2. P. A. Bernstein and S. Melnik. Model management 2.0: manipulating richer mappings. In *SIGMOD*, pages 1–12, 2007.
3. S. Chaudhuri and M. Y. Vardi. Optimization of real conjunctive queries. In *PODS*, 1993.
4. R. Fagin, L. M. Haas, M. A. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis. Clio: Schema mapping creation and data exchange. In *Conceptual Modeling: Foundations and Applications*, pages 198–236, 2009.
5. R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theor. Comput. Sci.*, 336(1):89–124, 2005.
6. R. Fagin, P. G. Kolaitis, and L. Popa. Data exchange: getting to the core. *ACM Trans. Database Syst.*, 30(1):174–210, 2005.
7. A. Hernich, L. Libkin, and N. Schweikardt. Closed world data exchange. *ACM Trans. Database Syst.*, 36(2):14:1–14:40, 2011.
8. T. Imielinski and W. L. Jr. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, 1984.
9. K. Knauer and T. Ueckerdt. Three ways to cover a graph. *Discrete Mathematics*, 339(2):745 – 758, 2016.
10. P. G. Kolaitis. Schema mappings, data exchange, and metadata management. In *PODS*, 2005.
11. E. V. Kostylev, J. L. Reutter, and A. Z. Salamon. Classification of annotation semirings over containment of conjunctive queries. *ACM Trans. Database Syst.*, 39(1):1, 2014.
12. L. Libkin and C. Sirangelo. Data exchange and schema mappings in open and closed worlds. *J. Comput. Syst. Sci.*, 77(3):542–571, 2011.
13. L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernández, and R. Fagin. Translating web data. In *VLDB*, 2002.

[1] `http://sirius-labs.no`