Adventures with Datalog: Walking the Thin Line Between Theory and Practice (Extended Abstract)

Georg Gottlob*

Department of Computer Science, University of Oxford
7 Parks Rd, Oxford OX1 3QG, UK

Datalog. Datalog (see [8]) is a logic programming language whose rules (in the basic version) syntactically coincide with universally quantified function-free Horn clauses. A Datalog program uses as input an *extensional database (EDB)* and computes from it an *intensional database (EDB)*. Datalog has been extended by negation in rule bodies (e.g. *stratified negation*) and various other features. Datalog with stratified negation can be evaluated in polynomial time (PTIME) in data complexity [19], and captures PTIME over linearly ordered structures.

Disjunctive Datalog, DLV, and the DLVSystem Company. Disjunctive logic programming (see [18]) and, in particular, disjunctive Datalog, extend the basic formalism by the possibility of using disjunctions in rule heads. In [9], the complexity of disjunctive Datalog was studied and it was shown that reasoning with this language is Π_2^p complete and that disjunctive Datalog with (most forms of) negation captures Π_2^p. Based on these results, a first version of the DLV disjunctive Datalog system [17], was implemented in the late 1990s at TU Vienna in an effort led by Nicola Leone, and was continued in Calabria in Rende (Cosenza), where it is still ongoing. In 2005, the DLVSystem company (dlvsystem.com) was founded by Leone with several other DLV contributors. The company has since worked with many corporate customers and has very successfully solved important industrial problems, see [1].

Monadic Datalog over Trees, Web Data Extraction, and Lixto. HTML or XML Web documents constitute tree-structured data. Data extraction from tree-structured documents is a task, which can be essentially described by monadic second-order logic (MSO). Since MSO is not a practical language and has a high complexity, we identified *monadic Datalog* as a candidate extraction language. We proved in [13, 14, 15] that reasoning with monadic Datalog over finite trees is $O(size(P) \times size(D))$ for program P and EDB D, and that, over trees, monadic Datalog has the same expressive power as MSO. Based on these results, we developed the *Lixto* system [2, 12] for visual data extraction which gave rise to the Lixto spin-out company, which was acquired in 2013 by McKinsey to be part of their *Periscope* solution.

Datalog$^\pm$ and Fully Automated Web Data Extraction. Several applications need data from thousands of websites (real estate, used cars, etc). The idea thus arose to create *knowledge-based* fully automated wrapper (=extractor) generators for such applications. We thus investigated suitable Datalog variants for knowledge representation and reasoning that would also be able to perform *ontological reasoning tasks*, such as those featured by the DL-Lite[10] description logic family. This can be done in Datalog[$\exists$] which extends Datalog by the possible use of existential quantification in rule heads. However, this language is undecidable, hence we studied decidable versions that restrict Datalog[$\exists$] syntactically [6, 7, 11]. This gave rise to the Datalog$^\pm$ family, which contains among other languages *guarded* and *weakly guarded* (w.g.) Datalog[$\exists$]. We showed that reasoning is PTIME-complete with the former and EXPTIME-complete with the latter in data complexity, and is 2EXPTIME-complete in combined complexity for both languages, and that w.g. Datalog[$\exists$] captures EXPTIME. Based on Datalog$^\pm$ rules, in the

*Supported by the Royal Society *Raison Data* project RP\R1\201074 and by the Alan Turing Institute.

context of the DIADEM ERC Advanced Grant (2010-15) at Oxford, we developed the DIADEM system, a fully automated rule-based wrapper generator, which is adaptable to disparate application areas, and founded in 2015 the *Wrapidity* startup, which was acquired in 2016 by Meltwater, a media intelligence firm, where the DIADEM technology has since been intensively used for extracting massive amounts of news articles and company data from the Web.

Warded Datalog[∃] and Vadalog for Reasoning in Knowledge Graphs. We proposed the Datalog$^{\pm}$ language *warded Datalog*[∃] [16, 5], which (i) can express DL-Lite and similar languages and is thus suited for ontological reasoning, (ii) extends plain Datalog, and (iii) has tractable data complexity for the relevant reasoning tasks. This language is the core of the Vadalog language and knowledge graph management system [3, 4], which gave rise to the *DeepReason.ai* spin-out, funded in 2018 and acquired by Meltwater in 2021.

References

[1] Weronika Adrian et al. The ASP system DLV: Advancements and Applications. *KI-Künstliche Intelligenz*, 32(2):177–179, 2018.

[2] R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with lixto. In *VLDB - Intl. Conference on Very Large Data Bases* , pages 119–128. Morgan Kaufmann, 2001.

[3] Luigi Bellomarini, Georg Gottlob, Andreas Pieris, and Emanuel Sallinger. Swift logic for big data and knowledge graphs. In *Intl. Conf. on Artificial Intelligence, IJCAI'17*, pages 2–10, 2017.

[4] Luigi Bellomarini, Emanuel Sallinger, and Georg Gottlob. The vadalog system: datalog-based reasoning for knowledge graphs. *Proceedings of the VLDB Endowment*, 11(9):975–987, 2018.

[5] Gerald Berger, Georg Gottlob, Andreas Pieris, and Emanuel Sallinger. The Space-Efficient Core of Vadalog. In *ACM Symp. on Principles of Database Systems*, pages 270–284. ACM, 2019.

[6] Andrea Calì, Georg Gottlob, and Michael Kifer. Taming the infinite chase: Query answering under expressive relational constraints. *Journal of Artificial Intelligence Research*, 48:115–174, 2013.

[7] Andrea Calì, Georg Gottlob, and Thomas Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. *Journal of Web Semantics*, 14:57–83, 2012.

[8] S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases*. Springer, 1990.

[9] Thomas Eiter, Georg Gottlob, and Heikki Mannila. Disjunctive Datalog. *ACM Transactions on Database Systems (TODS)*, 22(3):364–418, 1997.

[10] Diego Calvanese et al. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. of Automated reasoning*, 39(3):385–429, 2007.

[11] G. Gottlob et al. Expressiveness of guarded existential rule languages. In *Proceedings of the 33rd ACM Symp. on Principles of Database Systems*, pages 27–38. ACM, 2014.

[12] Georg Gottlob et al. The Lixto Data Extraction Project - Back and Forth between Theory and Practice. In *ACM Symp. on Principles of Database Systems*, pages 1–12. ACM, 2004.

[13] Georg Gottlob and Christoph Koch. Monadic queries over tree-structured data. In *LICS 2002*, pages 189–202. IEEE Computer Society, 2002.

[14] Georg Gottlob and Christoph Koch. Monadic datalog and the expressive power of languages for web information extraction. *Journal of the ACM (JACM)*, 51(1):74–113, 2004.

[15] Georg Gottlob, Reinhard Pichler, and Fang Wei. Monadic datalog over finite structures of bounded treewidth. *ACM Transactions of Computational Logic (TOCL)*, 12(1):3:1–3:48, 2010.

[16] Georg Gottlob and Andreas Pieris. Beyond sparql under owl 2 ql entailment regime: Rules to the rescue. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[17] Nicola Leone et al. . The DLV system for knowledge representation and reasoning. *ACM Transactions of Computational Logic (TOCL)*, 7(3):499–562, 2006.

[18] T. Przymusinski. Stable semantics for disjunctive programs. *New Gen. Comput.*, 9(3), 1991.

[19] Moshe Y. Vardi. The complexity of relational query languages. In *STOC'82*, pages 137–146, 1982.