Fitting Ontologies and Constraints to Relational Structures (Extended Abstract)

Simon Hosemann[1], Jean Christoph Jung[2], Carsten Lutz[1,3] and Sebastian Rudolph[3,4]

[1]Leipzig University
[2]TU Dortmund University
[3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)
[4]TU Dresden

Abstract

We study the problem of fitting ontologies and constraints to positive and negative examples that take the form of a finite relational structure. As ontology and constraint languages, we consider the description logics $\mathcal{EL}$ and $\mathcal{ELI}$ as well as several classes of tuple-generating dependencies (TGDs): full, guarded, frontier-guarded, frontier-one, and unrestricted TGDs as well as inclusion dependencies. We pinpoint the exact computational complexity, design algorithms, and analyze the size of fitting ontologies and TGDs. We also investigate the related problem of constructing a finite basis of concept inclusions / TGDs for a given set of finite structures.

Keywords

Fitting Problems, Database Constraints, Ontologies, Tuple-Generating Dependencies, Description Logics

In a *fitting problem*, one is given a set of positive and negative examples, each of which takes the form of a logical structure, and the task is to produce a logical formula that is satisfied by every positive example and refuted by every negative example. Problems of this form play a fundamental role in several applications. A prime example is the classic paradigm of query by example, also known as query reverse engineering [1, 2, 3]. In that case, the positive and negative examples are database instances and the formula to be constructed is a database query. In concept learning in description logics (DLs) [4, 5, 6], the examples are ABoxes and the formula sought is a DL concept to be used as a building block in an ontology. We remark that fitting problems are intimately connected to PAC learning by the fundamental theorem of computational learning theory. A third example application is entity comparison [7, 8] where the examples are knowledge graphs and one wants to find a formula that takes the form of a SPARQL query.

This extended abstract is a summary of our recent work, in which we study fitting problems that aim to support the construction of ontologies and database integrity constraints [9]. We investigate (i) ontologies formulated in the DLs $\mathcal{EL}$, $\mathcal{ELI}$, or an existential-rule language, and (ii) database constraints taking the form of tuple-generating dependencies (TGDs). In $\mathcal{EL}$ and $\mathcal{ELI}$, an ontology is a set of concept inclusions (CIs), each of which can be translated into an equivalent TGD. Moreover, 'existential rule' and 'TGD' refer to the same thing, so from

DL 2025: 38th International Workshop on Description Logics, September 3–6, 2025, Opole, Poland

simon.hosemann@uni-leipzig.de (S. Hosemann); jean.jung@tu-dortmund.de (J. C. Jung); carsten.lutz@uni-leipzig.de (C. Lutz); sebastian.rudolph@tu-dresden.de (S. Rudolph)

CEUR Workshop Proceedings (CEUR-WS.org)

now on we speak of TGDs also in the context of ontologies. From our perspective there is in fact no difference between an ontology and a set of constraints: any set of TGDs can be used as an ontology when an open world semantics is adopted and as a set of constraints under a closed world semantics. As constraint / ontology languages we consider $\mathcal{EL}$- and $\mathcal{ELI}$-CIs, their extensions with $\bot$, unrestricted TGDs, and the following restricted classes of TGDs: full (FullTGD), guarded (GTGD), frontier-guarded (FGTGD), and frontier-one (F1TGD), as well as inclusion dependencies (IND).

Let us be more precise about the fitting problems that we study. In our setting the examples are finite relational structures that we refer to as instances. An *instance* I is a finite set of facts, where a fact $R(a_1, \ldots, a_n)$ consists of an n-ary relation symbol R and values $a_1, \ldots, a_n$. The active domain of I is the set of all values that occur in any fact of I. A pointed instance is a pair $(I, \bar{a})$, consisting of an instance I and a finite tuple of values $\bar{a}$. In the DL case, the considered instances may only contain facts using unary and binary relation symbols. Let $\mathcal{L}$ be one of the TGD classes mentioned above (including $\mathcal{EL}(\mathcal{I})$-CIs). Further let (P, N) be a pair of finite sets of instances, henceforth called a *fitting instance*. We say that an $\mathcal{L}$-ontology $\mathcal{O}$ *fits* (P, N) if $P \models \mathcal{O}$ for all $P \in \mathsf{P}$ and $N \not\models \mathcal{O}$ for all $N \in \mathsf{N}$. For a single $\mathcal{L}$-TGD ρ, fitting (P, N) is defined in exactly the same way. The induced decision problems of *fitting $\mathcal{L}$-ontology existence* and *fitting $\mathcal{L}$-TGD existence* ask whether a given (P, N) admits a fitting $\mathcal{L}$-ontology or a fitting $\mathcal{L}$-TGD. We also consider the corresponding construction problems, where the goal is to construct a fitting $\mathcal{L}$-ontology or a fitting $\mathcal{L}$-TGD for (P, N), if one exists.

Example 1. *Consider the instances $P = \{R(a, b), R(b, a)\}$, $N = \{R(a, b), R(b, c), R(c, a)\}$. Then $(\{P\}, \{N\})$ has no fitting $\mathcal{ELI}$-CI, but it has fitting GTGDs such as*

$$R(x, y) \to R(y, x).$$

Now let $N' = N \cup \{R(b, a), R(c, b), R(a, c)\}$. Then $(\{P\}, \{N'\})$ has no fitting GTGD. But it has fitting FGTGDs such as

$$R(x, y) \land R(y, z) \land R(z, x) \to R(x, x).$$

Example 2. *Having $\bot$ or not makes a difference. Let $P = \{R(a, b)\}$, $N = \{R(a, a)\}$. Then $\exists R.\exists R.\top \sqsubseteq \bot$ fits $(\{P\}, \{N\})$, but $(\{P\}, \{N\})$ has no fitting $\mathcal{ELI}$-ontology.*

All negative claims in Examples 1 and 2 are a consequence of the semantic characterizations for fitting $\mathcal{L}$-TGD existence established in [9]. The characterization for $\mathcal{EL}_\bot$ and $\mathcal{ELI}_\bot$ is explicitly stated in Theorem 1 below.

How are fitting ontologies and fitting TGDs related? The following is an immediate consequence of the definition of fitting and the semantics of ontologies and TGDs.

Lemma 1. *Let (P, N) be a fitting instance. Then there is an $\mathcal{L}$-ontology that fits (P, N) if and only if for every $N \in \mathsf{N}$, there is an $\mathcal{L}$-TGD that fits $(\mathsf{P}, \{N\})$.*

Hence, if (P, N) admits a fitting $\mathcal{L}$-ontology, it admits one with at most $|\mathsf{N}|$ TGDs.

The problem of fitting an ontology to a given set of examples turns out to be closely related to a problem that has been studied in the area of description logic and is known as finite basis construction [10, 11, 12]. There, one fixes an ontology language $\mathcal{L}$ and is given as input a finite instance I and the task is to produce an $\mathcal{L}$-ontology $\mathcal{O}$ such that $I \models \rho$ if and only if $\mathcal{O} \models \rho$, for all $\mathcal{L}$-TGDs ρ. We generalize this problem to a finite set H of input instances. The following lemma connects finite basis construction with fitting $\mathcal{L}$-ontology existence. Informally, it states that a finite basis of the positive examples is a canonical candidate for a fitting $\mathcal{L}$-ontology.

Lemma 2. *Let (P, N) be a fitting instance and let $\mathcal{O}_\mathsf{P}$ be a finite $\mathcal{L}$-basis of P. Then $\mathcal{O}_\mathsf{P}$ fits (P, N) if and only if (P, N) has a fitting $\mathcal{L}$-ontology.*

If finite $\mathcal{L}$-bases always exists, we can thus solve the $\mathcal{L}$-ontology fitting problem for any (P, N) by constructing $\mathcal{O}_\mathsf{P}$ and checking whether it fits the input examples. This approach in fact often yields decidability and tight upper complexity bounds.

We first consider the DLs $\mathcal{EL}$ and $\mathcal{ELI}$ as well as their extensions with the $\bot$ concept. We reprove the existence of finite bases for $\mathcal{EL}$, already known from [13, 10], and simultaneously prove that finite bases exist also for $\mathcal{ELI}$ which to the best of our knowledge is a new result. In contrast to the proofs from [13, 10], our proofs are direct in that they do not rely on the machinery of formal concept analysis. The constructed bases are of double exponential size, but can be succinctly represented in single exponential size by structure sharing. We also show that these size bounds are tight, both for $\mathcal{EL}$ and for $\mathcal{ELI}$. We obtain from this an ExpTime upper bound for the fitting existence problem for $\mathcal{EL}$- and $\mathcal{ELI}$-ontologies.

In order to obtain lower complexity bounds, we provide a semantic characterization of fitting $\mathcal{EL}$- and $\mathcal{ELI}$-CI existence in terms of simulations and direct products. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. For unary pointed instances (I, a) and (J, b) we write $(I, a) \preceq_\mathcal{L} (J, b)$ iff there exists an $\mathcal{L}$-simulation from I to J that contains the pair (a, b). Recall that an $\mathcal{EL}$-simulation preserves concept names and the existence of role-successors, whereas an $\mathcal{ELI}$-simulation must in addition preserve role-predecessors, reflecting inverse roles. For a non-empty finite set H of instances with pairwise disjoint active domains, we use $\biguplus \mathsf{H}$ to denote the instance $\bigcup \mathsf{H}$. When the domains of the instances in H are not pairwise disjoint, we assume that renaming is used to achieve disjointness before forming $\biguplus \mathsf{H}$. We next present the characterization for fitting $\mathcal{EL}_\bot$- and $\mathcal{ELI}_\bot$-CI existence.

Theorem 1. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_k\}$ and let $P = \biguplus \mathsf{P}$. Then no $\mathcal{L}_\bot$-concept inclusion fits (P, N) if and only if for all $\bar{a} = (a_1, \ldots, a_k) \in \Delta^{\prod \mathsf{N}}$, the following condition is satisfied:*

$$S_{\bar{a}} = \{(P, b) \mid (\prod \mathsf{N}, \bar{a}) \preceq_\mathcal{L} (P, b)\} \text{ is non-empty} \quad \text{and} \quad \prod S_{\bar{a}} \preceq_\mathcal{L} (N_i, a_i) \text{ for some } i \in [k].$$

An extended version of Theorem 1, also covering the cases of $\mathcal{EL}$ and $\mathcal{ELI}$ without $\bot$ is provided in [9]. The semantic characterization gives rise to an algorithm for fitting $\mathcal{EL}(\mathcal{I})$-CI existence and opens up an alternative path to algorithms for fitting $\mathcal{EL}(\mathcal{I})$-ontology existence. It also enables us to prove lower complexity bounds and we in fact show that all four problems are ExpTime-complete. The proof of the theorem is constructive in the sense that it also yields an algorithm for fitting CI and fitting ontology construction. Regarding fitting ontology existence and construction, Lemma 1 yields a simple reduction to the CI fitting case that gives the desired results. We also prove tight bounds on the sizes of fitting CIs and fitting ontologies, which are identical to the size bounds on finite bases described above.

We next turn to TGDs. For guarded TGDs, we implement exactly the same program described above for $\mathcal{EL}(\mathcal{I})$, but obtain different complexities. We show that finite GTGD-bases always exist and establish a tight single exponential bound on their size. Succinct representation does not help to reduce the size. We give a characterization of fitting GTGD existence and fitting GTGD-ontology existence in terms of products and homomorphisms, show that fitting GTGD existence and fitting GTGD-ontology existence is coNExpTime-complete, and give a tight single exponential bound on the size of fitting GTGDs and GTGD-ontologies. The coNExpTime upper bound may be obtained either via finite bases or via the semantic characterization.

For the remaining classes of TGDs, the approach via finite bases fails: for the frontier-guarded, frontier-one, and full case, we prove that finite bases need not exist. For inclusion dependencies, finite bases trivially exist but approaching fitting via this route does not result in an optimal upper complexity bound. For unrestricted TGDs, the existence of finite bases is left open.

Theorem 2. *For $\mathcal{L} \in \{FGTGD, F1TGD, FullTGD\}$, there exist instances that have no finite $\mathcal{L}$-basis.*

Example 3. *Consider the instance $I = \{R(a,b), R(b,a)\}$. It has no finite FGTGD- and no finite F1TGD-basis. For every $n \geq 1$, consider the frontier-one TGD*

$$\rho_n = \bigwedge_{i\in[n-1]} R(x_i, x_{i+1}) \wedge R(x_n, x_1) \rightarrow R(x_1, x_1).$$

The TGD ρ_n expresses that if x_1 lies on a cycle of length n, then x_1 has a reflexive loop. We have $I \models \rho_n$ for all odd n because (i) a cycle homomorphically maps to I if and only if it is of even length and (ii) I contains no reflexive loops. Note that the rule bodies of the TGDs ρ_n with n odd get larger with increasing n. Intuitively, this means that also the rule bodies of any finite FGTGD-basis of I must be of unbounded size, which means that there is no finite FGTGD-basis.

We may, however, still approach fitting existence in a direct way or via a semantic characterization. For inclusion dependencies (IND), we use direct arguments to show that fitting IND existence and fitting IND-ontology existence is NP-complete, and that the size of fitting IND-ontologies is polynomial. For all remaining cases, we establish semantic characterizations in terms of products and homomorphisms and then use them to approach fitting existence. In this way, we prove the following. Fitting ontology existence and fitting TGD existence are coNExpTime-complete for TGDs that are frontier-guarded or frontier-one. For full TGDs, fitting TGD existence is coNExpTime-complete and fitting ontology existence is in Σ_2^p and DP-hard. In the case of unrestricted TGDs, both problems are coNExpTime-hard and we prove a co2NExpTime upper bound for fitting ontology existence and a co3NExpTime upper bound for fitting TGD existence. We also show tight single exponential size bounds for fitting TGDs and ontologies in the case of frontier-guarded and frontier-one TGDs. We do the same for fitting full TGDs while if there is a fitting FullTGD-ontology, then there is always one of polynomial size. For unrestricted TGD and TGD-ontology fittings, we give a single exponential lower bound and a triple (for TGDs) and double (for ontologies) exponential upper bound on the size.

Acknowledgments

This work is partly supported by BMFTR in DAAD project 57616814 (SECAI). The second and third author were supported by DFG project JU 3197/1-1.

References

[1] H. Li, C. Chan, D. Maier, Query from examples: An iterative, data-driven approach to query construction, Proc. VLDB Endow. 8 (2015) 2158–2169. doi:10.14778/2831360.2831369.

[2] P. Barceló, M. Romero, The complexity of reverse engineering problems for conjunctive queries, in: Proc. of ICDT, volume 68 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, pp. 7:1–7:17. doi:10.4230/LIPICS.ICDT.2017.7.

[3] B. ten Cate, V. Dalmau, M. Funk, C. Lutz, Extremal fitting problems for conjunctive queries, in: Proc. of PODS, ACM, 2023, pp. 89–98.

[4] J. Lehmann, P. Hitzler, Concept learning in description logics using refinement operators, Mach. Learn. 78 (2010) 203–250. doi:10.1007/S10994-009-5146-2.

[5] J. C. Jung, C. Lutz, H. Pulcini, F. Wolter, Separating data examples by description logic concepts with restricted signatures, in: Proc. of KR, 2021, pp. 390–399. doi:10.24963/KR.2021/37.

[6] B. ten Cate, M. Funk, J. C. Jung, C. Lutz, SAT-based PAC learning of description logic concepts, in: Proc. of IJCAI, ijcai.org, 2023, pp. 3347–3355. doi:10.24963/IJCAI.2023/373.

[7] A. Petrova, E. Sherkhonov, B. C. Grau, I. Horrocks, Entity comparison in RDF graphs, in: Proc. of ISWC, volume 10587 of *LNCS*, Springer, 2017, pp. 526–541. doi:10.1007/978-3-319-68288-4_31.

[8] A. Petrova, E. V. Kostylev, B. C. Grau, I. Horrocks, Query-based entity comparison in knowledge graphs revisited, in: Proc. of ISWC, volume 11778 of *LNCS*, Springer, 2019, pp. 558–575. doi:10.1007/978-3-030-30793-6_32.

[9] S. Hosemann, J. C. Jung, C. Lutz, S. Rudolph, Fitting ontologies and constraints to relational structures, in: Proc. of KR, 2025. Accepted; to appear.

[10] F. Distel, Learning description logic knowledge bases from data using methods from formal concept analysis, Ph.D. thesis, Dresden University of Technology, 2011. URL: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa-70199.

[11] R. Guimarães, A. Ozaki, C. Persia, B. Sertkaya, Mining $\mathcal{EL}\bot$ bases with adaptable role depth, J. Artif. Intell. Res. 76 (2023) 883–924. doi:10.1613/JAIR.1.13777.

[12] F. Kriegel, Efficient axiomatization of OWL 2 EL ontologies from data by means of formal concept analysis, Proc. of AAAI 38 (2024).

[13] F. Baader, F. Distel, A finite basis for the set of EL-implications holding in a finite model, in: Proc. of ICFCA, volume 4933 of *LNCS*, Springer, 2008, pp. 46–61. doi:10.1007/978-3-540-78137-0_4.