First-Order Queries on Databases Embedded in an Infinite Structure

Martin Otto
RWTH Aachen[*]

Jan Van den Bussche[†]
University of Antwerp[‡]

Abstract

We consider "generic" (isomorphism-invariant) queries on relational databases embedded in an infinite background structure. Assume a generic query is expressible by a first-order formula over the embedded domain that may involve both the relations of the database and the relations and functions of the background structure. Then this query is already expressible by a first-order formula involving just an auxiliary linear ordering as background structure. We present an elementary proof of this fact.

Keywords: relational database, query, first-order logic, model theory

One of the leading themes of research in database theory is that of queries and query languages [AHV94]. Usually, one works in the relational model where a database is viewed as a finite structure over some fixed relational vocabulary (called the *database schema*). A *query* then is a mapping which associates with each database a finite relation on its domain, of some fixed arity. Not any such mapping makes sense as a query, however: a basic consistency criterion required of queries is that they are compatible with isomorphisms. Indeed, two isomorphic databases are meant to represent the

[*] Address: Mathematische Grundlagen der Informatik, RWTH Aachen, Ahornstr. 55, D-52074 Aachen, Germany. E-mail: otto@informatik.rwth-aachen.de

[†] Post-doctoral research fellow of the Belgian National Fund for Scientific Research.

[‡] Address: UIA, Informatica, Universiteitsplein 1, B-2610 Antwerp, Belgium. E-mail: vdbuss@uia.ac.be.

same information content and should not be distinguished [AU79, CH80]. This consistency criterion is called the *genericity* of queries.

A fundamental way of expressing a query is by means of a first-order formula over the database schema, which uniformly defines on each database the answer relation of the query as applied to that database. The class of queries thus obtained is the class FO of all *first-order queries*. Note that first-order queries are indeed generic, since logical formulae cannot distinguish between isomorphic structures.

One can extend the class of first-order queries by allowing the defining formula to use extra information which is not properly part of the database. One basic example of this is to use formulae over the database schema extended with the binary relation symbol $<$ for a linear order. One then evaluates such a formula on a database by first extending the database with a linear ordering on its domain. Of course, in order not to violate genericity, the formula must satisfy the consistency criterion that its result is independent of the particular ordering chosen. We call such formulae *order-invariant*. Although order-invariance is recursively undecidable, we can still consider the class FO[$<$] of queries defined by order-invariant formulae.

FO is trivially included in FO[$<$], and it is known that this inclusion is strict; see for instance [AHV94, Excercise 17.27]. One may ask whether there are other kinds of extra background structure, besides linear order, which further increase the expressive power of FO in this manner. In this note, we show that the answer, in a precise and rather general sense, is negative.

Specifically, we formalize the idea of "providing extra information" by fixing an arbitrary infinite structure $\mathcal{A}$ over some finite vocabulary τ (with τ disjoint from the database schema), and using formulae over the database schema extended with the symbols in τ. One then evaluates such a formula by embedding the database in $\mathcal{A}$. Again, in order not to violate genericity, we restrict to formulae which – in spite of the external auxiliary structure – define a result that is independent of the particular embedding chosen. We call such formulae $\mathcal{A}$-*invariant*. We thus obtain the class FO[$\mathcal{A}$] of queries defined by $\mathcal{A}$-invariant formulae.

The framework of FO[$<$] is slightly different from that of FO[$\mathcal{A}$], since in the former case the database is *extended* with extra information while in the latter case the databases is *embedded* into it. But this difference is only formal; if $\mathcal{A}$ is simply an infinite linearly ordered set, then FO[$\mathcal{A}$] coincides with FO[$<$].

Our result can now be stated as follows:

Theorem 1 *For any finite vocabulary τ and any infinite τ-structure $\mathcal{A}$,* FO$[\mathcal{A}]$ *is included in* FO$[<]$.

To finish this introduction we mention that similar results with interesting ramifications have meanwhile been obtained independently by Benedikt et al. [BDLW96]; see also the remarks at the end. In the remainder of this note we define the notions discussed above more formally and prove the theorem.

FO queries. Fix some finite relational vocabulary σ. We identify the class of all *databases* over the database schema σ with the class fin$[\sigma]$ of all finite σ-structures. A k-ary query Q then is a mapping

$$\begin{array}{rcl} \mathrm{fin}[\sigma] & \longrightarrow & \mathrm{fin}[\sigma\dot\cup\{Q\}] \\ \mathcal{B} & \longrightarrow & (\mathcal{B}, Q[\mathcal{B}]) \end{array}$$

where $Q[\mathcal{B}]$ is a k-ary relation on the domain of $\mathcal{B}$. This mapping has to satisfy the following condition (genericity): if $f\colon \mathcal{B} \xrightarrow{\simeq} \mathcal{C}$ is an isomorphism, then f maps $Q[\mathcal{B}]$ to $Q[\mathcal{C}]$, i.e. f also is an isomorphism $f\colon (\mathcal{B}, Q[\mathcal{B}]) \xrightarrow{\simeq} (\mathcal{C}, Q[\mathcal{C}])$.

Let $\varphi(x_1, \ldots, x_k)$ be a first-order formula over σ. On each $\mathcal{B}$, φ defines a relation

$$\phi[\mathcal{B}] := \{(a_1, \ldots, a_k) \in \mathrm{dom}(\mathcal{B})^k \mid \mathcal{B} \models \varphi[a_1, \ldots, a_k]\},$$

which obviously defines a query. The set of queries defined by first-order formulae in this way is denoted by FO.

Order-invariant formulae. Assume, without loss of generality, that the binary relation symbol $<$ is not in σ. Let $\psi(\bar{x})$ be a formula over the extended vocabulary $\sigma \,\dot\cup\, \{<\}$. If we extend $\mathcal{B} \in \mathrm{fin}[\sigma]$ with a linear ordering on its domain, we obtain an extended database $(\mathcal{B}, <^{\mathcal{B}}) \in \mathrm{fin}[\sigma\dot\cup\{<\}]$, on which ψ defines a relation $\psi[\mathcal{B}, <^{\mathcal{B}}]$.

If $\psi[\mathcal{B}, <^{\mathcal{B}}]$ happens to be the same no matter which ordering $<^{\mathcal{B}}$ we choose to extend $\mathcal{B}$ with, and this holds true for each $\mathcal{B}$, we call ψ *order-invariant.* In this case ψ defines a query on the original, non-extended databases given by $\psi(\mathcal{B}) := \psi[\mathcal{B}, <^{\mathcal{B}}]$, for each $\mathcal{B}$ and some (any) extension $(\mathcal{B}, <^{\mathcal{B}})$ of $\mathcal{B}$. The set of queries defined by order-invariant formulae in this way is denoted by FO[<].

$\mathcal{A}$-invariant formulae. Let τ be a finite vocabulary disjoint from σ. We do not require τ to be relational, τ may contain functions and constants. Let $\mathcal{A}$ be any fixed infinite τ-structure. An *embedding* of $\mathcal{B} \in \text{fin}[\sigma]$ into $\mathcal{A}$ is given by an injection $\mu\colon \text{dom}(\mathcal{B}) \to \text{dom}(\mathcal{A})$. We expand $\mathcal{A}$ with the isomorphic image $\mu(\mathcal{B})$ of $\mathcal{B}$, and also introduce a new unary relation U to denote the image of $\text{dom}(\mathcal{B})$ under μ as a subset of $\text{dom}(\mathcal{A})$. We thus obtain a structure $(\mathcal{A}, \mu(\mathcal{B}))$ over the combined vocabulary $\sigma \mathbin{\dot\cup} \tau \mathbin{\dot\cup} \{U\}$.

Let $\chi(\bar{x})$ be a first-order formula over $\sigma \mathbin{\dot\cup} \tau$. We use χ to define a relation on $\text{dom}(\mathcal{B})$, denoted by $\chi[\mathcal{B}, \mu]$, as follows:

$$\chi[\mathcal{B}, \mu] := \{\bar{a} \in \text{dom}(\mathcal{B}) \mid (\mathcal{A}, \mu(\mathcal{B})) \models \chi^U[\mu(\bar{a})]\}.$$

The superscript U in χ^U denotes *relativization* to the embedded domain of $\mathcal{B}$, $U = \mu(\text{dom}(\mathcal{B}))$, which means that the quantifiers in χ are restricted to range only over this embedded domain. This is an essential restriction in the setup. We indicate in an example below that without this restriction the main theorem is no longer valid. Note, however, that even this restricted use of the background structure, though it does not give access to all its elements via direct quantification, still gives access to outside elements exactly in as far as these are parameterized through terms (of vocabulary τ) from within the embedded domain.

If $\chi[\mathcal{B}, \mu]$ is the same no matter which embedding μ we choose, and this for each $\mathcal{B}$, then we call χ *$\mathcal{A}$-invariant.* In this case χ defines a query on databases over σ given by $\chi[\mathcal{B}] := \chi[\mathcal{B}, \mu]$, for some (any) embedding μ of $\mathcal{B}$ into $\mathcal{A}$. The set of queries defined by $\mathcal{A}$-invariant formulae in this way is here denoted by FO[$\mathcal{A}$].

Towards our proof, we next present two little lemmas.

Lemma 1 *Let $\tau = \{<\}$ and let $\mathcal{A}$ be a τ-structure such that $<^{\mathcal{A}}$ is a linear ordering of* $\text{dom}(\mathcal{A})$*. Then* FO[$\mathcal{A}$] *equals* FO[<].

Proof. Let $\mathcal{B}$ be a database. The orderings of $\mathcal{B}$ are precisely the inverse images of $<^{\mathcal{A}}$ under embeddings of $\mathcal{B}$ into $\mathcal{A}$. Hence, a formula ψ over $\sigma \cup \{<\}$

is order-invariant iff it is $\mathcal{A}$-invariant. Furthermore, if ψ is order-invariant and μ is an embedding of $\mathcal{B}$ into $\mathcal{A}$, then for any $\bar{a} \in \mathrm{dom}(\mathcal{B})$,

$$(\mathcal{A}, \mu(\mathcal{B})) \models \psi^U[\mu(\bar{a})] \quad \Leftrightarrow \quad (\mathcal{B}, \mu^{-1}(<^{\mathcal{A}})) \models \psi[\bar{a}],$$

so that ψ defines the same query regardless of whether it is considered as an order-invariant formula or as an $\mathcal{A}$-invariant one. Hence, FO[$\mathcal{A}$] equals FO[$<$]. ■

Lemma 2 *Let $\mathcal{A}$ and $\mathcal{A}'$ be two elementarily equivalent τ-structures. Then* FO[$\mathcal{A}$] *equals* FO[$\mathcal{A}'$].

Proof. Let $\chi(\bar{x})$, $\bar{x} = (x_1, \ldots, x_k)$, be a formula over $\sigma \cup \tau$. Let $\mathcal{B} \in \mathrm{fin}[\sigma]$ and let $\bar{a} \in \mathrm{dom}(\mathcal{B})$, $\bar{a} = (a_1, \ldots, a_k)$. Associate to each $d \in \mathrm{dom}(\mathcal{B})$ a different variable y_d not already occurring in χ and think of $d \mapsto y_d$ as an intended interpretation that associates each element with "its" variable. We define two sentences $\chi_{\forall}^{\mathcal{B},\bar{a}}$ and $\chi_{\exists}^{\mathcal{B},\bar{a}}$ over τ in several steps as follows:

1. First, we define a quantifier-free formula $\chi^{\mathcal{B}}(\bar{x}, (y_d)_{d \in \mathrm{dom}(\mathcal{B})})$ in vocabulary τ as follows. Replace each atom of the form $R\,\bar{t}$ in χ, with $R \in \sigma$, by
$$\bigvee_{\bar{d} \in R^{\mathcal{B}}} \bar{t} = \bar{y}_{\bar{d}},$$
where $\bar{y}_{\bar{d}}$ is shorthand for the tuple $(y_{d_1}, \ldots, y_{d_k})$ if $\bar{d} = (d_1, \ldots, d_k)$. Then any existential quantifier $\exists x$ is replaced by the disjunction over all instantiations y_d for x in the quantified formula. Correspondingly, any universal quantification $\forall x$ is replaced by the conjunction over all instantiations y_d for x.

2. Let $\chi^{\mathcal{B},\bar{a}}$ be the result of substituting $\bar{y}_{\bar{a}} = (y_{a_1}, \ldots, y_{a_k})$ for the free variables $\bar{x} = (x_1, \ldots, x_k)$ in $\chi^{\mathcal{B}}(\bar{x}, (y_d)_{d \in \mathrm{dom}(\mathcal{B})})$.

3. Put
$$\begin{aligned} \chi_{\forall}^{\mathcal{B},\bar{a}} &:= (\forall y_d)_{d \in \mathrm{dom}(\mathcal{B})} \Big(\bigwedge_{d \neq d'} y_d \neq y_{d'} \longrightarrow \chi^{\mathcal{B},\bar{a}} \Big) \\ \chi_{\exists}^{\mathcal{B},\bar{a}} &:= (\exists y_d)_{d \in \mathrm{dom}(\mathcal{B})} \Big(\bigwedge_{d \neq d'} y_d \neq y_{d'} \wedge \chi^{\mathcal{B},\bar{a}} \Big) \end{aligned}$$

The following are readily verified, for any τ-structure $\mathcal{C}$:

- $\mathcal{C} \models \chi_{\forall}^{\mathcal{B},\bar{a}}$ if and only if $\bar{a} \in \chi[\mathcal{B},\mu]$ for *all* embeddings μ of $\mathcal{B}$ into $\mathcal{C}$;
- $\mathcal{C} \models \chi_{\exists}^{\mathcal{B},\bar{a}}$ if and only if $\bar{a} \in \chi[\mathcal{B},\mu]$ for *some* embedding μ of $\mathcal{B}$ into $\mathcal{C}$.

The lemma now follows: if χ is $\mathcal{A}$-invariant then it is also $\mathcal{A}'$-invariant. Indeed,

$$\begin{array}{rcll} \chi \text{ is } \mathcal{A}\text{-invariant} & \Leftrightarrow & \mathcal{A} \models (\chi_{\exists}^{\mathcal{B},\bar{a}} \rightarrow \chi_{\forall}^{\mathcal{B},\bar{a}}) & \text{for each } (\mathcal{B},\bar{a}) \\ & \Leftrightarrow & \mathcal{A}' \models (\chi_{\exists}^{\mathcal{B},\bar{a}} \rightarrow \chi_{\forall}^{\mathcal{B},\bar{a}}) & \text{for each } (\mathcal{B},\bar{a}) \\ & \Leftrightarrow & \chi \text{ is } \mathcal{A}'\text{-invariant.} & \end{array}$$

The second equivalence holds because $\mathcal{A}$ and $\mathcal{A}'$ are elementarily equivalent, the first and third by the two above-stated properties.

For $\mathcal{A}$-invariant χ, if $\bar{a} \in \chi[\mathcal{B}]$ on $\mathcal{A}$ then $\bar{a} \in \chi[\mathcal{B}]$ on $\mathcal{A}'$. Indeed,

$$\begin{array}{rcl} \bar{a} \in \chi[\mathcal{B}] \quad \text{on } \mathcal{A} & \Leftrightarrow & \mathcal{A} \models \chi_{\forall}^{\mathcal{B},\bar{a}} \\ & \Leftrightarrow & \mathcal{A}' \models \chi_{\forall}^{\mathcal{B},\bar{a}} \\ & \Leftrightarrow & \bar{a} \in \chi[\mathcal{B}] \quad \text{on } \mathcal{A}'. \end{array}$$

■

We also make use of the Ehrenfeucht-Mostowski Theorem on first-order indiscernibles, which is an important consequence of Ramsey's Theorem and compactness, much used in classical model theory, cf. [H93, CK90].

A linearly ordered subset I of $\mathrm{dom}(\mathcal{A})$ – think of it as a linear order $\mathcal{I} = (I, \prec)$ embedded into $\mathcal{A}$ – is called a *chain of indiscernibles in* $\mathcal{A}$ if for any formula $\phi(x_1,\ldots,x_k)$ over τ and any two increasing sequences $c_1 \prec \cdots \prec c_k$ and $d_1 \prec \cdots \prec d_k$ from I it is true that

$$\mathcal{A} \models \phi[c_1,\ldots,c_k] \quad \Leftrightarrow \quad \mathcal{A} \models \phi[d_1,\ldots,d_k].$$

In other words, truth in $\mathcal{A}$ of a formula $\phi(x_1,\ldots,x_k)$ on a tuple $a_1,\ldots,a_k$ of elements in I depends only on the way these elements are ordered by $\prec$.

Ehrenfeucht-Mostowski Theorem *For each infinite $\mathcal{A}$ there is an elementarily equivalent structure $\mathcal{A}'$ which has an infinite chain of indiscernibles (of any prescribed order type, in fact).*

We shall see below that actually we could employ Ramsey's Theorem directly and avoid the passage to an elementarily equivalent structure $\mathcal{A}'$. The application of the Ehrenfeucht-Mostowski Theorem, on the other hand, leads to a neat and uniform translation. We are now ready for a very simple proof of the theorem. For an $\mathcal{A}$-invariant formula, we provide an $<$-invariant formula that is equivalent in the sense of defining the same query.

Proof of Theorem 1. By Lemma 2, we may assume without loss of generality that $\mathcal{A}$ itself has a chain $\mathcal{I} = (I, \prec)$ of indiscernibles.

Let $\chi(\bar{x})$ be $\mathcal{A}$-invariant. We may assume without loss of generality that in each atomic subformula of χ of the form $Rt_1 \ldots t_m$ with $R \in \sigma$, every term t_i is a variable. Indeed, we can always replace this subformula by the formula

$$(\exists z_1) \ldots (\exists z_m) \Big(Rz_1 \ldots z_m \wedge \bigwedge_{i=1}^{m} z_i = t_i \Big).$$

Note that this replacement is correct even though quantifiers range only over the domain of the embedded structure. Note also that this replacement serves to guarantee that no atomic subformula of χ contains symbols from both σ and τ.

Consider now an evaluation of χ over a finite subdomain included in $I \subset \mathrm{dom}(\mathcal{A})$. The truth of any atomic subformula $\phi(y_1, \ldots, y_m)$ over τ in χ, when evaluated in a tuple $d_1, \ldots, d_m$ of elements of I, depends only on the way $d_1, \ldots, d_m$ are ordered with respect to $\prec$. In other words, the truth depends only on the order type of $d_1, \ldots, d_m$.[1] Hence, on $\mathcal{I}$, each atomic subformula over τ is equivalent to a formula over $\{<\}$, namely, a disjunction of order types. Denote by ψ the formula obtained from χ by replacing each atomic τ-subformula by its equivalent formula over $\{<\}$ in this manner. Note that ψ is a formula over $\sigma \cup \{<\}$.

Since χ is $\mathcal{A}$-invariant, and since any embedding of a database $\mathcal{B}$ into $\mathcal{I}$ is also an embedding of $\mathcal{B}$ into $\mathcal{A}$, ψ is $\mathcal{I}$-invariant. Moreover, for any database $\mathcal{B}$ and tuple $\bar{a}$ on $\mathrm{dom}(\mathcal{B})$, $\bar{a} \in \chi[\mathcal{B}]$ on $\mathcal{A}$ if and only if $\bar{a} \in \psi[\mathcal{B}]$ on $\mathcal{I}$. We thus conclude that the query expressed by χ is in FO[$\mathcal{I}$] (by ψ). Finally, by Lemma 1, this query then is in FO[$<$]. ■

[1] An order type on the variables $y_1, \ldots, y_m$ is a maximal consistent conjunction of atoms of the form $y_i < y_j$ and negations thereof.

Remark 1 As the proof really only requires the indiscernibility condition on $\mathcal{I}$ for a finite set of (atomic) formulae, it is not even necessary to invoke the power of the Ehrenfeucht-Mostowski Theorem. By Ramsey's Theorem one may obtain an infinite chain satisfying indiscernibility for a finite collection of formulae through suitable choice within the given $\mathcal{A}$ (this is the route taken in [BDLW96]).

We conclude with an example showing that the usual first-order semantics with unrestricted quantification over the entire background structure behaves completely differently. Consider the countably infinite random graph $\mathcal{R}$ for the background structure – with vocabulary consisting of a binary edge relation E –, into which we embed the finite structure $\mathcal{B}$. It is easily checked that any monadic second-order formula in the vocabulary of $\mathcal{B}$ can in this setting be captured by an $\mathcal{R}$-invariant first-order formula. We merely replace any quantification $\exists X\phi(X)$ by a quantification $\exists y\phi(\{x|Eyx\})$. Here $\phi(\{x|Eyx\})$ is shorthand for the result of replacing each atom Xu that may occur in ϕ by the atom Eyu. This replacement is semantically appropriate in the proposed setting, since for any finite subset $U \subset \mathrm{dom}(\mathcal{R})$ and any $X \subset U$, there is an outside vertex y such that Eyx for $x \in U$ if and only if $x \in X$. This is just an instance, in fact, of the extension axioms that characterize the random graph [BH79, EF95, H93]. But monadic second-order logic over finite structures is known to be strictly more expressive than first-order logic even in the presence of a linear ordering. Consider for instance structures of monadic vocabulary plus order, so-called word models. Monadic second-order logic exactly defines those classes of word models that correspond to regular languages (a theorem of Büchi, Elgot and Trakhtenbrot), while first-order logic only defines those that correspond to star-free regular languages (McNaughton and Papert), see for instance [T82].

This shows that the inclusion claim of Theorem 1 does not hold in general, if unrestricted first-order quantification over the background structure is admitted. It also immediately suggests the question under which model theoretic requirements on the background structure (or rather, according to Lemma 2, on its first-order theory) the inclusion does go through in this stronger sense after all. Note that the above example immediately suggests sparseness conditions on definable sets. One known positive case was that of the additive arithmetic of the reals investigated in [PVV95]. Several people, including the present authors, have conjectured that among linearly ordered background structures o-minimality [PS86] might give a sufficient condition. This conjecture has meanwhile been proved by Benedikt et al. [BDLW96] in a slightly different setting.

References

[AHV94] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases.* Addison-Wesley, 1994.

[AU79] A.V. Aho and J.D. Ullman. Universality of data retrieval languages. In *Proceedings of the 6th ACM Symposium on Principles of Programming Languages*, pages 110–120, 1979.

[BDLW96] M. Benedikt, G. Dong, L. Libkin, and L. Wong. Relational expressive power of constraint query languages. In *Proceedings 15th ACM Symposium on Principles of Database Systems*, pages 5–16, 1996.

[BH79] A. Blass and F. Harary. Properties of almost all graphs and complexes. *Journal of Graph Theory*, 3:225–240, 1979.

[CH80] A. Chandra and D. Harel. Computable queries for relational database systems. *Journal of Computer and System Sciences*, 21(2):156–178, 1980.

[CK90] C.C. Chang and H.J. Keisler. *Model Theory.* Third edition. North-Holland, 1990.

[EF95] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory.* Springer-Verlag, 1995.

[H93] W. Hodges. *Model Theory.* Cambridge University Press, 1993.

[PVV95] J. Paredaens, J. Van den Bussche, and D. Van Gucht. First-order queries on finite structures over the reals. In *Proceedings of the 10th IEEE Symposium on Logic in Computer Science*, pages 79–89, 1995.

[PS86] A. Pillay and C. Steinhorn. Definable sets in ordered structures, I. *Transactions of the AMS*, 295(2):565–592, 1986.

[T82] W. Thomas. Classifying regular events in symbolic logic. *Journal of Computer and System Sciences*, 25:360–376, 1982.