Towards Provenance in Heterogeneous Knowledge Bases

Matthias Knorr(✉), Carlos Viegas Damásio, Ricardo Gonçalves, and João Leite

Departamento de Informática, NOVA LINCS, FCT NOVA, Caparica, Portugal
{mkn,cd,rjrg,jleite}@fct.unl.pt

Abstract. A rapidly increasing amount of data, information and knowledge is becoming available on the Web, often written in different formats and languages, adhering to standardizations driven by the World Wide Web Consortium initiative. Taking advantage of all this heterogeneous knowledge requires its integration for more sophisticated reasoning services and applications. To fully leverage the potential of such systems, their inferences should be accompanied by justifications that allow a user to understand a proposed decision/recommendation, in particular for critical systems (healthcare, law, finances, etc.). However, determining such justifications has commonly only been considered for a single formalism, such as relational databases, description logic ontologies, or declarative rule languages. In this paper, we present the first approach for providing provenance for heterogeneous knowledge bases building on the general framework of multi-context systems, as an abstract, but very expressive formalism to represent knowledge bases written in different formalisms and the flow of information between them. We also show under which conditions and how provenance information in this formalism can be computed.

Keywords: Provenance · Heterogeneous knowledge bases · Multi-context systems

1 Introduction

A rapidly increasing amount of data, information and knowledge is becoming available on the Web, driven by the Semantic Web initiative led by the World Wide Web Consortium (W3C).[1] A number of language standards have been established in this initiative and to take advantage of all this available knowledge often requires their integration. This is particularly true for (but not limited to) integrations of rule languages, e.g., under answer set semantics [4] and ontology languages based on description logics [1], that are both highly expressive, but with orthogonal/complementary characteristics and modelling features (see, e.g., [12, 21, 23, 24] and references therein).

[1] https://www.w3.org/.

G. Gottlob et al. (Eds.): LPNMR 2022, LNAI 13416, pp. 287–300, 2022.
https://doi.org/10.1007/978-3-031-15707-3_22

However, in the course of the integration of such heterogeneous knowledge, it becomes increasingly difficult to trace the causes for a certain inference, or find the justification for some proposed decision, in particular, if the pieces of knowledge originate from different authors. It would therefore be important to provide methods that accompany inferences/decisions with explanations/justifications in a way a user can understand to allow for the validation of reasoning results, in particular for critical systems (healthcare, law, finances, etc.).

This has been recognized in different areas of Artificial Intelligence, and for several Knowledge Representation and Reasoning formalisms, the problem of finding justifications has been considered. In particular, a lot of work has focussed on tracing the origins of derivations, commonly under the name of provenance [5], e.g., in relational databases and Datalog [18,19], Logic Programming [8], Answer Set Programming [13], Description Logics ontology languages [2,6,20], as well as in SPARQL [7] and data streams [16]. Yet, provenance for heterogeneous knowledge bases has mostly been ignored, with the exception of [10], though limited to two very restricted settings.

In this paper, we investigate provenance for heterogeneous knowledge bases, utilising multi-context systems (MCSs) [3] as our formalism of choice. MCSs allow for the integration of a large variety of logic-based formalisms, and model the flow of information between them. They cover very general approaches for integrating ontologies and rules [22], thus allowing to study provenance in a more general manner, which then paves the way towards provenance in related approaches in the literature. We focus on providing justifications of inferences (the only question that has been handled in the literature are explanations of inconsistencies when repairing inconsistent multi-context systems [11], which is inherently different). Our contributions can be summarized as follows:

– We develop the first general approach for provenance in heterogeneous knowledge bases, and in multi-context systems in particular, annotating inferences with their justifying provenance information.
– We provide means to compute this provenance information annotating models, so-called equilibria, in MCSs.
– We establish under which conditions this provenance information can indeed be computed, showing its applicability to a wide class of formalisms.

The remainder of the paper is structured as follows. We recall notions on provenance semirings in Sect. 2. Then, in Sect. 3, we introduce provenance multi-context systems as a non-trivial extension of MCSs. In Sect. 4, we show how and when model notions for such provenance MCSs can be computed and provide considerations on complexity, before we conclude in Sect. 5.

2 Provenance Semirings

In the context of databases, commutative semirings have been introduced as a means of representing provenance information [18,19], such as providing information regarding what combination of tuples in a database certain query results were obtained from. Subsequently, commutative semirings have been adopted for representing provenance information in a wide variety of different logic formalisms and it has been shown that they cover other related approaches in the literature, such as finding minimal explanations/justifications [15]. They are thus well-suited to capture provenance information in formalisms composed of knowledge bases written in different (knowledge representation) languages, and we recall the main notions here.

A *commutative semiring* is an algebraic structure $\mathcal{K} = (K, \oplus_{\mathcal{K}}, \otimes_{\mathcal{K}}, 0_{\mathcal{K}}, 1_{\mathcal{K}})$ where $\oplus_{\mathcal{K}}$ and $\otimes_{\mathcal{K}}$ are commutative and associative binary operators over a set K, called the annotation domain of $\mathcal{K}$. The operators $\oplus_{\mathcal{K}}$ and $\otimes_{\mathcal{K}}$ have neutral elements $0_{\mathcal{K}}$ and $1_{\mathcal{K}}$, respectively, where $\otimes_{\mathcal{K}}$ distributes over $\oplus_{\mathcal{K}}$, and $0_{\mathcal{K}}$ is an annihilating element of $\otimes_{\mathcal{K}}$. This allows the definition of functions (so-called $\mathcal{K}$-relations) that map tuples (in the case of databases) to annotations over K such that only finitely many tuples are annotated with a value different from $0_{\mathcal{K}}$.

As an example in the case of databases, we may consider a commutative semiring where each tuple in any given table is annotated with an annotation name. Then, the annotations of query results correspond to the combinations of these annotations names, i.e., those corresponding to tuples, using $\oplus_{\mathcal{K}}$ to represent alternatives and $\otimes_{\mathcal{K}}$ to represent the join of tuples.

This idea is captured in general in the provenance polynomials semiring $\mathbb{N}[X] = (\mathbb{N}[X], +, \times, 0, 1)$ where polynomials over annotation variables X are used with natural number coefficients and exponents over these variables. Other relevant semirings in the literature can be obtained from it by introducing additional properties on the operations such as idempotence on $+$ and/or $\times$, or absorption, giving rise to a hierarchy of semirings [19]. Several such semirings have been used for notions of provenance in different logic-based formalisms (e.g., [2,6], and also [15] for further references). Among these semirings, $\mathbb{N}[X]$ is the most general one and universal, in the sense that for any other commutative semiring $\mathcal{K}$, a semiring homomorphism can be defined, allowing the computations for $\mathcal{K}$ to be done in $\mathbb{N}[X]$.

Monus semirings or *m-semirings* [14] extend such commutative semirings by adding natural orders $\preceq_{\mathcal{K}}$, which are partial orders that order elements of the annotation domain based on the $\oplus_{\mathcal{K}}$ operation, namely, $k_1 \preceq_{\mathcal{K}} k_2$ if there exists k_3 such that $k_1 \oplus_{\mathcal{K}} k_3 = k_2$. The monus operation $k_1 \ominus_{\mathcal{K}} k_2$ then refers to the unique smallest element k_3 in such a partial order such that $k_2 \oplus_{\mathcal{K}} k_3 \preceq_{\mathcal{K}} k_1$. This allows capturing negation and has been generalized to recursive Datalog queries and logic programs under different semantics [8] using a semiring that utilizes boolean formulas over two sets of variables – positive facts and their negations. Similar ideas have been applied to handle semiring provenance for First-Order Logic [17,25], utilizing first-order formulas in negation normal form.

3 Provenance Multi-context Systems

Multi-context systems (MCSs) [3] are defined as a collection of components, so-called contexts, each of which allows one to represent knowledge in some logic-based formalism. Each such logic is associated with a set of well-formed knowledge bases in the logic (its admitted syntax), possible belief sets, indicating how models are defined in this logic (its admitted semantics), and a function assigning to each possible knowledge base a set of acceptable such belief sets. MCSs use so-called bridge rules that allow one to model the flow of information between these contexts, in the sense that they admit the incorporation of knowledge in one context based on the beliefs considered true in other contexts. The semantics of MCSs is then assigned using equilibria that take the acceptable belief sets and the interaction between contexts into account.

In this section, we introduce provenance multi-context systems that extend MCSs with the means to explain inferences obtained from the modular integration of its contexts. Here, rather than recalling first MCSs and then introduce their extension, to not unnecessarily burden the presentation with partially repetitive technical definitions, we introduce provenance multi-context systems right away clarifying in the course of this introduction how and where our notion extends the previous one.

The first important question is how provenance should be represented in such a modular framework. Given that different notions of provenance have been introduced for different logical formalisms, with varying granularity of the provided provenance information, our objective is to maintain the modular character of MCSs, and admit that possibly different notions of provenance be used in each of the contexts, and provide provenance annotations for inferences in the contexts taking into account provenance information from other contexts via bridge rules.

We start by defining the set of variables allowed to be used as annotations. To account for possibly varying algebras in different contexts with differing binary operators, we introduce a number of different annotation languages V_i, each intended to correspond to one of the contexts, that can be interleaved by means of one particular language V_*, which is meant to correspond to the integration of information in bridge rules between contexts.

Definition 1. *Let* $N = N_* \cup \bigcup N_i$ *be a countably infinite set of names and* $\Sigma = \Sigma_* \cup \bigcup \Sigma_i$ *be a countably infinite set of binary operators such that, for all* i*, all* N_i *are mutually distinct,* $N_* \cap N_i = \emptyset$*, and* $N \cap \Sigma = \emptyset$*. The* set of annotation variables $V = V_* \cup \bigcup V_i$ *is defined inductively for all* i *with* $1 \leq i \leq n$ *:*

(1) $N_i \subseteq V_i$*;*
(2) $(v_1 \circ v_2) \in V_i$ *for* $v_1, v_2 \in V_i$ *and* $\circ \in \Sigma_i$*;*
(3) $(r \circ v_1 \circ \cdots \circ v_m) \in V_*$ *for* $r \in N_*$*, for all* k*,* $1 \leq k \leq m$*,* $v_k \in V_i$ *for some* i*,* $\circ \in \Sigma_*$*, and* $m \geq 0$*;*
(4) $V_* \subseteq V_i$*.*

Condition (1) specifies that names intended to annotate formulas are valid annotation variables within the respective sublanguage. Then, condition (2) indicates how to obtain complex annotations within each of the defined sublanguages. Condition (3) defines that annotation variables in the language V_* are composed of one name from N_* and 0 or more annotations from the other languages, which is intended to represent the composition of annotations v_k within a bridge rule (as defined in the following). Finally, condition (4) admits that annotations from V_* be used in the other annotation languages.

Note that names, which are meant to be used to identify formulas in individual contexts, are distinct, while operators may overlap in between different Σ_i, and between Σ_* and different Σ_i, to account for the possibility that different contexts may use the same provenance formalism.

Based on this, we can introduce provenance logics as a means to capture a large variety of formalisms that allow tracing the reasons for inferences, thus generalizing the logics of MCSs as described in the beginning of this section.

Definition 2. *A* provenance logic L *is a tuple* $(\mathcal{K}, \mathbf{KB}, \mathbf{BS}, \mathbf{ACC})$ *where*

(1) $\mathcal{K}$ *is a commutative semiring over polynomials over some* V_i *with* $\oplus_{\mathcal{K}}, \otimes_{\mathcal{K}} \in \Sigma_i$*, and a natural order* $\preceq_{\mathcal{K}}$*;*
(2) $\mathbf{KB}$ *is the set of well-formed knowledge bases of* L *such that each* $kb \in \mathbf{KB}$ *is a set composed of formulas distinctly annotated with elements from* V_i*;*
(3) $\mathbf{BS}$ *is the set of possible annotated belief sets, i.e., functions that map beliefs from the set of possible beliefs* B_L *of* L *to* V_i*, such that false beliefs are mapped to* $0_{\mathcal{K}}$*;*
(4) $\mathbf{ACC} : \mathbf{KB} \rightarrow 2^{\mathbf{BS}}$ *is a function describing the semantics of* L *by assigning to each knowledge base a set of acceptable annotated belief sets.*

In comparison to logics for MCSs [3], provenance logics include a commutative semiring $\mathcal{K}$ and formulas in knowledge bases are in addition annotated according to one of the languages V_i (see Definition 1). Also, the idea of possible belief sets from MCSs is extended in that sets of annotated beliefs are used. I.e., rather than using sets of beliefs that are meant to be true, sets of beliefs with their corresponding annotations are considered. The function **ACC** then assigns semantics to knowledge bases by associating them with acceptable annotated belief sets.

Note that some approaches in the literature assign polynomials to beliefs that are not true, e.g., to account for possible changes so that something becomes true, but here, for the sake of generality and in the spirit of MCSs, we omit this, and focus on determining the provenance of true elements.

Example 1. We present some example provenance logics.

- L_{db} – Databases with provenance under bag semantics [18]:
  - $\mathcal{K}_{db}$: $\mathbb{N}[X]$;
  - $\mathbf{KB}_{db}$: the set of annotated databases together with queries expressed in an appropriate query language, such as Datalog;
  - $\mathbf{BS}_{db}$: the set of sets of atoms with annotations;
  - $\mathbf{ACC}_{db}(kb)$: the set of tuples in kb and query results over db with their annotation according to $\mathcal{K}_{db}$;
- L_{dl} – Description Logic $\mathcal{ELH}^r$ [2]:
  - $\mathcal{K}_{dl}$: $Trio[X]$, i.e., $\mathbb{N}[X]$ with idempotent $\times$;
  - $\mathbf{KB}_{dl}$: set of well-formed annotated $\mathcal{ELH}^r$ ontologies;
  - $\mathbf{BS}_{dl}$: the set of sets of annotated atomic inferences;
  - $\mathbf{ACC}_{dl}(kb)$: the set of atomic inferences from kb with their annotation according to $\mathcal{K}_{dl}$;

– L_{lp} – Normal logic programs under answer set semantics (adapted from [8]):
  - $\mathcal{K}_{lp}$: $PosBool[X]$, i.e., $\mathbb{N}[X]$ with idempotent $+$ and $\times$ and absorption on $+$, over positive atoms;
  - $\mathbf{KB}_{lp}$: the set of annotated normal logic programs;
  - $\mathbf{BS}_{lp}$: the set of sets of atoms with annotations;
  - $\mathbf{ACC}_{lp}(kb)$: the answer sets of kb with annotations according to $\mathcal{K}_{lp}$;

Similar to MCSs, bridge rules are used to specify how knowledge is transferred between the different components, but here we also have to take provenance information into account.

Definition 3. *Given a collection of provenance logics $L = \langle L_1, \ldots, L_n\rangle$, an L_i-*bridge rule *over L, $1 \leq i \leq n$, is of the form:*

$$\pi@s \leftarrow (r_1 : p_1), \ldots, (r_j : p_j), \quad \mathbf{not}\,(r_{j+1} : p_{j+1}), \ldots, \mathbf{not}\,(r_m : p_m) \tag{1}$$

where $\pi \in N_*$ and, for $1 \leq k \leq m$, $1 \leq r_k \leq n$ and $p_k \in B_{L_{r_k}}$, and, for each $kb \in \mathbf{KB}_i$, $kb \cup \{v@s\} \in \mathbf{KB}_i$ for every $v \in V_*$. We refer with $H(\pi)$ and $B(\pi)$ to the head and the body of the bridge rule, respectively. A bridge rule is called monotonic *if it does not contain elements of the form $\mathbf{not}\,(r : p)$, and* non-monotonic *otherwise.*

Note that each of the r_k refer to one of the logics and the beliefs p_k belong to the corresponding set of possible beliefs $B_{L_{r_k}}$ of logic L_{r_k} (cf. (3) of Definition 1). Also note that π is the annotation name of the bridge rule itself, whereas v is an annotation variable associated to the bridge rule head s, intended to be incorporated into the knowledge base kb_i, such that v takes the annotations of the bridge rule elements into account (as made precise when we define the semantics).

With this in place, we can introduce provenance multi-context systems.

Definition 4. *A* provenance multi-context system (pMCS) *is a collection of contexts $M = \langle C_1, \ldots, C_n\rangle$ where $C_i = (L_i, kb_i, br_i)$, $L_i = (\mathcal{K}_i, \mathbf{KB}_i, \mathbf{BS}_i, \mathbf{ACC}_i)$ is a provenance logic, $kb_i \in \mathbf{KB}_i$ a knowledge base, and br_i is a set of L_i-bridge rules over $\langle L_1, \ldots, L_n\rangle$.*

We assume that the annotations used for the elements occurring in the individual kb_i are unique elements from N_i, and that each context uses a different set of annotations V_i. Also, while different forms of specifying the annotations of formulas can be found in the literature, here we use uniformly the notation introduced for bridge rules, i.e., the annotation is given in front of a formula with @ as separator.

Example 2. Consider $M = \langle C_1, C_2, C_3\rangle$ such that:

– C_1 is a database context with L_{db}, $kb = \{d_1@p(1,1), d_2@p(1,2)\}$ with a single relation $p/2$ with two tuples, $br_1 = \emptyset$, and query q defined by $q(x,y) \leftarrow p(e,x), p(e,y)$;

- C_2 a DL context with L_{dl}, $kb_2 = \{o_1@A \sqsubseteq B\}$, and $br_2 = \{b_1@A(w) \leftarrow \mathbf{not}\,(3 : l)\}$;
- C_3 an ASP context with L_{lp}, $kb_3 = \{r_1@l \leftarrow \mathit{not}\ m, n\}$, and $br_3 = \{c_1@n \leftarrow (1 : q(1, 1)), c_2@m \leftarrow (2 : B(w))\}$.

As C_1 has no bridge rules, we obtain $\mathbf{ACC}_{db}(kb_1) = \{S_1\}$ with $S_1(p(1, 1)) = d_1$, $S_1(p(1, 2)) = d_2$, $S_1(q(1, 1)) = d_1^2$, $S_1(q(1, 2)) = S_1(q(2, 1)) = d_1 \times d_2$, and $S_1(q(2, 2)) = d_2^2$. For both kb_2 and kb_3, $\mathbf{ACC}_i(kb_i) = \{S_i\}$ with S_i mapping every atomic inference/atom to 0 (as the bridge rules are not considered for the semantics of individual contexts).

We now turn to the semantics of pMCSs. We first introduce belief states that contain one possible annotated belief set for each context and serve as suitable model candidates.

Definition 5. *Let $M = \langle C_1, \ldots, C_n\rangle$ be a pMCS. A* belief state *of M is a collection $S = \langle S_1, \ldots, S_n\rangle$ such that each S_i is an element of $\mathbf{BS}_i$.*

We next identify specific belief states, called equilibria, that take bridge rules into account for determining acceptable belief states, similar to MCSs. We adapt this with annotations building on the algebraic approach for non-monotonic rules [8] to pass annotation information via bridge rules. The main idea is to use annotations from V_* assuming the existence of distinct negative names (using **not**) in the respective N_i, one per negated p_k with $j+1 \leq k \leq m$ for bridge rules of the form (1). This is necessary as we assume that false beliefs are annotated with $0_{\mathcal{K}}$, thus no annotations exist for such negations.

We first fix the commutative semiring for bridge rules.

Definition 6. *The* commutative semiring for bridge rules *$\mathcal{BR}$ is defined as PosBool[V_*], for $\wedge, \vee \in \Sigma_*$, with idempotent meet ($\wedge$) and join ($\vee$), absorption on $\vee$, and logical consequence as natural order, i.e., $k_1 \preceq_{\mathcal{BR}} k_2$ iff $k_1 \models k_2$.*

We can now define when a bridge rule is applicable in a belief state, namely when the beliefs in the rule body hold true for positive elements and false for negative elements.

Definition 7. *Let $M = \langle C_1, \ldots, C_n\rangle$ be a pMCS and π an L_i-bridge rule over L of form (1). Then π* is applicable in a belief state *S, denoted $S \models B(\pi)$, iff*

(1) for $1 \leq k \leq j$, $S_{r_k}(p_k) = n$ for some annotation $n \neq 0$;
(2) for $j + 1 \leq k \leq m$, $S_{r_k}(p_k) = 0$.

As false elements are annotated with 0, we can use the annotations corresponding to beliefs being false in the annotations of the inferred/added bridge rule heads. This allows us to define equilibria.

Definition 8. *Let $M = \langle C_1, \ldots, C_n\rangle$ be a pMCS. A belief state $S = \langle S_1, \ldots, S_n\rangle$ of M is an* equilibrium *if, for all i with $1 \leq i \leq n$, the following condition holds:*

$$S_i \in \mathbf{ACC}_i(kb_i \cup \{\, v@H(\pi) \mid \pi \in br_i \text{ and } S \models B(\pi)\,\})$$

where the annotation of $H(\pi)$*,* v*, is defined over all* $\pi_l \in br_i$ *with* $H(\pi_l) = H(\pi)$ *such that* $S \models B(\pi_l)$ *as:*

$$v = \bigvee \pi_l \wedge v_1 \wedge \ldots \wedge v_j \wedge \textbf{not } p_{j+1} \wedge \ldots \textbf{not } p_m$$

such that, for $1 \leq k \leq j$*,* $S_{r_k}(p_k) = v_k$*.*

Hence, a belief state S is an equilibrium if, for each context, the corresponding annotated belief set is acceptable for the knowledge base of the context enhanced with the heads of those bridge rules of the context that are admissible in S. The corresponding annotations for bridge rule heads are constructed as representations of the alternative provenance information (via disjunction) resulting from different bridge rules with the same head. Before we explain the reason for that, we consider an example without bridge rules with the same head.

Example 3. Consider M from Example 2. Since C_1 does not contain bridge rules, S_1 is fully determined in Example 2. Then, by the first rule in br_3, we have that $S_3(n) = c_1 \wedge d_1^2$. If the other rule in br_3 is not applicable, then $S_3(l) = r_1 \times_3 (c_1 \wedge d_1^2)$ holds. In this case, the only bridge rule in br_2 is not applicable, thus $B(w)$ cannot be inferred from C_2 which ensures that the second rule in br_3 is not applicable. In fact, together with S_2 mapping every atomic inference to 0, we obtain an equilibrium.

We next provide an example showing that absorption on disjunctions is necessary to ensure the existence of equilibria.

Example 4. Consider $M = \langle C_1 \rangle$ with $L_1 = L_{lp}$, $kb_1 = \{n@q\}$, $br_1 = \{r_1@p \leftarrow (1 : p), r_2@p \leftarrow (1 : q)\}$. Then, $S = \langle S_1 \rangle$ with $S_1(q) = n$ and $S_1(p) = r_2 \times_1 n$ is an equilibrium of M. Note that both bridge rules are applicable in S, so without absorption, v would be $(r_2 \times_1 n) \vee (r_1 \times_1 r_2 \times_1 n)$ and no equilibrium would exist.

One could argue that this problem can be avoided by not joining disjunctively the provenance information for bridge rules with the same head and by passing several bridge rule heads with differing provenance information. But this requires that absorption be present in the target context's provenance notion, and since this cannot be guaranteed, nor do we want to impose it as an additional restriction, we deem the solution where absorption is resolved on the level of bridge rules the most suitable one.

We also illustrate that idempotency of $\wedge$ is required.

Example 5. Take $M = \langle C_1 \rangle$ with $L_1 = L_{lp}$, $kb_1 = \{\}$, $br_1 = \{r@p \leftarrow (1 : p)\}$. Then, $S = \langle S_1 \rangle$ with $S_1(p) = 0$ and $S' = \langle S_1 \rangle$ with $S_1(p) = r \times_1 p$ are equilibria of M. Note that r is applicable in S' and without idempotency of $\wedge$, v would be $r \times_1 r \times_1 p$, i.e., S' would not be an equilibrium.

One may wonder whether S' indeed should be an equilibrium, since the truth of p relies on self-support using the bridge rule, i.e., we would be interested in minimal equilibria. Formally, an equilibrium S is *minimal* if there is no equilibrium $S' = \langle S'_1, \ldots, S'_n \rangle$ such that $S'_i \preceq_{\mathcal{K}_i} S_i$ for all i with $1 \leq i \leq n$ and $S'_j \prec_{\mathcal{K}_j} S_j$ for some j with $1 \leq j \leq n$. However, it has been shown [3] that equilibria in MCSs are not necessarily minimal due to cyclic dependencies as in Example 5, and this carries over to pMCSs.

Still, we can show that equilibria of pMCSs are faithful with equilibria of MCSs as well as the provenance annotations in individual formalisms in the following sense.

Proposition 1. *Let $M = \langle C_1, \ldots, C_n \rangle$ be a pMCS and M' the MCS obtained from M by omitting all semirings $\mathcal{K}$ and all annotations from M, and, in each logic, replace the annotated belief sets by the set of beliefs not mapped to* 0.

- *For context C_i in M with $br_i = \emptyset$ and equilibrium S of M, $S_i \in \mathbf{ACC}_i(kb_i)$.*
- *If S is an equilibrium of M, then S' is an equilibrium of M' with $S'_i = \{b \mid S_i(b) \neq 0\}$.*

The converse of 2. does not hold in general because we require a method to determine the provenance annotations corresponding to equilibria of MCSs. In the next section, we investigate how and under which conditions such equilibria together with their provenance information can be effectively determined.

4 Grounded Equilibria

Building on material developed for MCSs [3], we introduce a restriction of pMCSs, called reducible pMCSs, for which minimal equilibria can be computed. To this end, we first consider definite pMCSs, a further restriction similar in spirit to definite logic programs, where reasoning is monotonic and where a unique minimal equilibrium exists.

We start with monotonic logics, where **ACC** is deterministic and monotonic. Formally, a logic $L = (\mathcal{K}, \mathbf{KB}, \mathbf{BS}, \mathbf{ACC})$ is *monotonic* if (1) $\mathbf{ACC}(kb)$ is a singleton set for each $kb \in \mathbf{KB}$, and (2) $S \preceq_{\mathcal{K}} S'$ whenever $kb \subseteq kb'$, $\mathbf{ACC}(kb) = \{ S \}$, and $\mathbf{ACC}(kb') = \{ S' \}$.

This excludes non-monotonic logics, but many of them are *reducible*, namely, if for some $\mathbf{KB}^* \subseteq \mathbf{KB}$ and some reduction function $red : \mathbf{KB} \times \mathbf{BS} \rightarrow \mathbf{KB}^*$:

(1) the restriction of L to $\mathbf{KB}^*$ is monotonic, and
(2) for each $kb \in \mathbf{KB}$, and all $S, S' \in \mathbf{BS}$:
  (a) $red(kb, S) = kb$ whenever $kb \in \mathbf{KB}^*$,
  (b) $red(kb, S) \subseteq red(kb, S')$ whenever $S' \preceq S$, and
  (c) $S \in \mathbf{ACC}(kb)$ iff $\mathbf{ACC}(red(kb, S)) = \{ S \}$.

This is adapted from MCSs, and inspired by the Gelfond-Lifschitz reduction for logic programs, indicating that (a) reduced *kb*s do not need to be reduced any further, (b) *red* is antitonic, and (c) acceptable annotated belief sets can be checked based on *red*. Note that the latter condition implies that the annotations are determined based on the reduced knowledge base.

This is generalized to contexts that are reducible, namely, if their logic is reducible, and if the reduction function is not affected by the addition of annotated bridge rule heads. Formally, a context $C_i = (L_i, kb_i, br_i)$ is *reducible* if

- its logic L_i is reducible and,
- for all belief sets S_i and all $H \subseteq \{v@H(\pi) \mid \pi \in br_i, v \in V_i\}$: $red(kb_i \cup H, S_i) = red(kb_i, S_i) \cup H$.

A pMCS is *reducible* if all of its contexts are.

It has been argued that a wide variety of logics is reducible [3]. Thus, reducible pMCSs admit the integration of a wide variety of logical formalisms provided the provenance notion fits or can be adjusted to the semiring requirements in Definition 2 and annotations can be determined based on reduced kbs, which arguably is the case for many KR formalisms.

We can now determine definite pMCSs as follows. A reducible pMCS $M = \langle C_1, \ldots, C_n \rangle$ is *definite* if

1. all bridge rules in all contexts are monotonic,
2. for all i and all $S \in \mathbf{BS}_i$, $kb_i = red_i(kb_i, S)$.

Thus, in definite pMCSs, bridge rules are monotonic and knowledge bases are already in reduced form. Therefore, its logics are monotonic, $\mathbf{ACC}_i(kb_i)$ is a singleton set, and the $\mathbf{ACC}_i$ themselves are monotonic. Hence, reasoning is montonic, and a unique minimal equilibrium exists.

Definition 9. *Let M be a definite pMCS. A belief state S of M is the* grounded equilibrium of M, *denoted by* $\mathbf{GE}(M)$, *if S is the unique minimal equilibrium of M.*

This unique equilibrium of a pMCS $M = \langle C_1, \ldots, C_n \rangle$ can be computed as follows. For $1 \leq i \leq n$, $kb_i^0 = kb_i$ and for each sucessor ordinal $\alpha + 1$,

$$kb_i^{\alpha+1} = kb_i^{\alpha} \cup \{v@H(\pi) | \pi \in br_i \wedge E^{\alpha} \models B(\pi)\}$$

where $E^{\alpha} = (E_1^{\alpha}, \ldots, E_n^{\alpha})$, $\mathbf{ACC}_i(kb_i^{\alpha}) = \{E_i^{\alpha}\}$; and v defined as in Definition 8, and for limit ordinal α, $kb_i^{\alpha} = \bigcup_{\beta \leq \alpha} kb_i^{\beta}$. Furthermore, let $kb_i^{\infty} = \bigcup_{\alpha > 0} kb_i^{\alpha}$.

Essentially, we start with the set of given knowledge bases and E^0 corresponds to the belief state resulting from M without the bridge rules. The iteration then stepwise checks based on the current belief state which bridge rules are applicable, enhancing the knowledge bases which in turn increases the annotated belief sets in the iteration of E^{α}, based on which further bridge rules become applicable, until a fixpoint is reached. This indeed yields the unique grounded equilibrium.

Proposition 2. *Let M be a definite pMCS. Then, belief state $S = (S_1, \ldots, S_n)$ is the grounded equilibrium of M iff $\mathbf{ACC}_i(kb_i^{\infty}) = \{S_i\}$, for $1 \leq i \leq n$.*

Note that this construction not only allows us to iteratively determine what is true (with annotation $\neq 0$)/can be inferred as in MCSs, it also allows us to simultaneously calculate what are the actual corresponding annotations.

Example 6. Consider M from Example 2 with $br_2 = \emptyset$ and kb reduced to $\{r_1@l \leftarrow n\}$, i.e., such M is definite. We can verify that the computed grounded equilibrium is $S = \langle S_1, S_2, S_3 \rangle$ with S_1 and S_2 defined as in Example 2 and S_3 s.t. $S_3(l) = r_1 \times_3 (c_1 \wedge d_1^2)$ and $S_3(n) = c_1 \wedge d_1^2$. Here, S_1 is determined based on kb_1^0, whereas $S_3(n)$ results from the applicable bridge rule, and $S_3(l)$ from that and rule r_1.

In the more general case of reducible pMCSs, we introduce a reduct taking into account the provenance information.

Definition 10. *Let $M = \langle C_1, \dots, C_n \rangle$ be a reducible pMCS s.t. $C_i = (L_i, kb_i, br_i)$, and $S = \langle S_1, \dots, S_n \rangle$ a belief state of M. The* S-reduct *of M is defined as $M^S = \langle C_1^S, \dots, C_n^S \rangle$, where, for $1 \leq i \leq n$, $C_i^S = (L_i, red_i(kb_i, S_i), br_i^S)$, and*

$$br_i^S = \{v@s \leftarrow (r_1 : p_1), \dots, (r_j : p_j) \mid \pi \in br_i \text{ of the form (1) such that } S_{r_i}(p_i) = 0 \text{ for all } j+1 \leq i \leq m, \text{ and } v = \pi \wedge \mathbf{not}\ p_{j+1} \wedge \dots \mathbf{not}\ p_m\}.$$

Thus, in the reduct, knowledge bases are reduced w.r.t. the considered belief state, and bridge rules are either omitted if there is a **not** $(r_k : p_k)$ in the bridge rule such that p_k is true in S_{r_i}, i.e., with annotation different from 0, or maintained in the reduct with the negated elements, just adapting the annotation v to take these negated elements into account in the annotation.

The resulting S-reduct of M is definite and we can check whether S is a grounded equilibrium in the usual manner.

Definition 11. *Let M be a reducible pMCS. A belief state S of M is a* grounded equilibrium *of M if $S = \mathbf{GE}(M^S)$.*

We can show that such grounded equilibria are minimal equilibria.

Proposition 3. *Every grounded equilibrium of a reducible pMCS M is a minimal equilibrium of M.*

Thus, for grounded equilibria, the converse of 2. (Prop. 1) can be obtained, and pMCSs indeed provide provenance annotations for grounded equilibria in MCSs.

Example 7. Consider M from Example 2. Note that M is reducible using the usual Gelfond-Lifschitz reduct for C_3 and since C_1 and C_2 are monotonic. We obtain one grounded equilibrium S as specified in Example 6 because b_1 in br_2 and c_2 in br_3 are not applicable in S. There is a second grounded equilibrium $S' = \langle S_1', S_2', S_3' \rangle$ with $S_1' = S_1$ for S_1 from Example 6, S_2' with $S_2'(A(w)) = b_1 \wedge \mathbf{not}\ l$ and $S_2'(B(w)) = o_1 \times_2 (b_1 \wedge \mathbf{not}\ l)$, and S_3' with $S_3'(m) = c_2 \wedge o_1 \times_2 (b_1 \wedge \mathbf{not}\ l)$ and $S_3'(n) = c_1 \wedge d_1^2$.

We observe that the resulting annotations concisely represent the formulas required to obtain an inference, and that this information modularily preserves the characteristics of the semirings used in individual contexts, e.g., the annotation of n in S'_3 contains that d_1 is used twice (in the database context), even though in C_3 such repetition would be omitted due to idempotent $\times$.

We close the section with considerations on the computational complexity where we assume familiarity with basic notions including the polynomial hierarchy. First, we recall that output-projected equilibria have been considered in the context of MCSs [11] as a means to facilitate consistency checking, by restricting the focus to the beliefs that occur in the bridge rules, showing in particular that for each output-projected equilibrium there exists a corresponding equilibrium. Then, consistency of an MCS, whose size, for fixed logics, is measured in the size of the knowledge bases and the size of the bridge rules, can be determined by guessing an output-projected belief state S and checking for each context whether it accepts the guessed S w.r.t. the active bridge rule heads. The complexity of the latter step, called context complexity, influences the complexity of consistency checking and has been determined for a number of logics [11]. Then, the context complexity of an MCS M, $\mathcal{CC}(M)$, can be determined based on upper and lower context complexities, which allows one to study the complexity of problems such as the existence of equilibria w.r.t. context complexity of MCSs.

Now, for pMCSs, we have to take the anotations into account, and it turns out that this increases the computational complexity in general. As argued in the case of DL $\mathcal{ELH}^r$, where standard reasoning problems are polynomial, annotations may be exponential in size [2], and similar problems can be observed for other logics. Still, for definite pMCSs, we can take advantage of the deterministic way to compute the unique grounded equilibrium and show that we can avoid this exponential size of annotations when solving the problem of whether there is a grounded equilibrium such that the annotation of belief p is n, i.e., $S_i(p) = n$. The essential idea is to take advantage of the bound on the size of annotations imposed by n, and limit the computation to the relevant part of the equilibrium, and adapt at the same time the notion of $\mathcal{CC}(M)$ from MCSs to take size of the monomials in the individual context into account.

Proposition 4. *Let M be a definite pMCS with context complexity $\mathcal{CC}(M)$ and S its unique grounded equilibrium. The problem of determining whether $S_i(p) = n$ for $p \in B_{L_i}$ and $n \in V_i$, is in $\mathcal{C}$ if $\mathcal{CC}(M) = \mathcal{C}$ for $\mathcal{C} \supseteq \mathbf{P}$.*

It turns out that this result does not hold in the general case, because verifying whether some belief state is an equilibrium requires the computation of the entire belief state, thus subject to the exponential size of annotations in general.

5 Conclusions

We have introduced provenance multi-context systems as the first approach for provenance in heterogeneous knowledge bases, allowing us to obtain annotations for the inferences in the models of the integrating formalism. We have shown how these models with annotations, equilibria, can be computed and, given the generality of the approach, under which conditions this is possible, showing that the approach is viable for the integration of a wide variety of formalisms.

For future work, we will investigate the usage of power series for dealing with provenance approaches that use infinite semirings [18], as well as consider the application of semiring provenance for fixed-point logic [9].

Acknowledgment. We thank the anonymous reviewers for their comments and FCT projects RIVER (PTDC/CCI-COM/30952/2017), FORGET (PTDC/CCI-INF/32219/2017), and NOVA LINCS (UIDB/04516/2020) for partial support.

References

1. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook. Cambridge University Press (2007)
2. Bourgaux, C., Ozaki, A., Peñaloza, R., Predoiu, L.: Provenance for the description logic $\mathcal{ELH}^r$. In: IJCAI, pp. 1862–1869. ijcai.org (2020)
3. Brewka, G., Eiter, T.: Equilibria in heterogeneous nonmonotonic multi-context systems. In: Proceedings of AAAI, pp. 385–390. AAAI Press (2007)
4. Brewka, G., Eiter, T., Truszczynski, M.: Answer set programming at a glance. Commun. ACM **54**(12), 92–103 (2011)
5. Buneman, P.: The providence of provenance. In: Gottlob, G., Grasso, G., Olteanu, D., Schallhart, C. (eds.) BNCOD 2013. LNCS, vol. 7968, pp. 7–12. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39467-6_3
6. Calvanese, D., Lanti, D., Ozaki, A., Peñaloza, R., Xiao, G.: Enriching ontology-based data access with provenance. In: IJCAI, pp. 1616–1623. ijcai.org (2019)
7. Damásio, C.V., Analyti, A., Antoniou, G.: Provenance for SPARQL queries. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012. LNCS, vol. 7649, pp. 625–640. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35176-1_39
8. Viegas Damásio, C., Analyti, A., Antoniou, G.: Justifications for logic programming. In: Cabalar, P., Son, T.C. (eds.) LPNMR 2013. LNCS (LNAI), vol. 8148, pp. 530–542. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40564-8_53
9. Dannert, K.M., Grädel, E., Naaf, M., Tannen, V.: Semiring provenance for fixed-point logic. In: CSL, LIPIcs, vol. 183, pp. 17:1–17:22 (2021)
10. Dividino, R.Q., Schenk, S., Sizov, S., Staab, S.: Provenance, trust, explanations - and all that other meta knowledge. KI **23**(2), 24–30 (2009)
11. Eiter, T., Fink, M., Schüller, P., Weinzierl, A.: Finding explanations of inconsistency in multi-context systems. Artif. Intell. **216**, 233–274 (2014)
12. Eiter, T., Ianni, G., Lukasiewicz, T., Schindlauer, R., Tompits, H.: Combining answer set programming with description logics for the semantic web. Artif. Intell. **172**(12–13), 1495–1539 (2008)
13. Fandinno, J., Schulz, C.: Answering the "why" in answer set programming - a survey of explanation approaches. TPLP **19**(2), 114–203 (2019)
14. Geerts, F., Poggi, A.: On database query languages for k-relations. J. Appl. Log. **8**(2), 173–185 (2010)
15. Glavic, B.: Data provenance. Found. Trends Datab. **9**(3–4), 209–441 (2021)
16. Glavic, B., Esmaili, K.S., Fischer, P.M., Tatbul, N.: Ariadne: managing fine-grained provenance on data streams. In: Proceedings of DEBS, pp. 39–50. ACM (2013)

17. Grädel, E., Tannen, V.: Semiring provenance for first-order model checking. arXiv preprint arXiv:1712.01980 (2017)
18. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: Proceedings of PODS, pp. 31–40. ACM (2007)
19. Green, T.J., Tannen, V.: The semiring framework for database provenance. In: Proceedings of PODS, pp. 93–99. ACM (2017)
20. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: Sheth, A., et al. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 323–338. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88564-1_21
21. Knorr, M., Alferes, J.J., Hitzler, P.: Local closed world reasoning with description logics under the well-founded semantics. Artif. Intell. **175**(9–10), 1528–1554 (2011)
22. Knorr, M., Slota, M., Leite, J., Homola, M.: What if no hybrid reasoner is available? hybrid MKNF in multi-context systems. J. Log. Comput. **24**(6), 1279–1311 (2014)
23. Lukumbuzya, S., Ortiz, M., Simkus, M.: Resilient logic programs: answer set programs challenged by ontologies. In: AAAI, pp. 2917–2924. AAAI Press (2020)
24. Motik, B., Rosati, R.: Reconciling description logics and rules. J. ACM **57**(5), 93–154 (2010)
25. Tannen, V.: Provenance analysis for FOL model checking. ACM SIGLOG News **4**(1), 24–36 (2017)