Automatic Strengthening of Graph-Structured Knowledge Bases

Vinay Chaudhri, Nikhil Dinesh, Stijn Heymans, and Michael Wessel

{ vinay.chaudhri | nikhil.dinesh | stijn.heymans | michael.wessel}@sri.com

SRI International

Artificial Intelligence Center

333 Ravenswood Avenue

Menlo Park, California 94025-3493, USA

Abstract

We address two problems in underspecified graph-structured knowledge bases (GSKBs): the *co-reference* and the *provenance problem*. The former asks *"Which existentially quantified variables in different but related axioms of a GSKB possibly denote identical individuals?"*, and the latter *"From which axioms in a GSKB is a piece of knowledge getting derived?"* In an *underspecified* GSKB, the desired co-reference problem cannot be addressed by means of sound inference, i.e., the co-reference information does not follow logically from the GSKB. We present an algorithm which rewrites an underspecified GSKB into a strengthened GSKB by Skolemization and addition of equality atoms such that the co-reference information can be drawn from it. This enlarges the logical theory (the deductive closure) of the GSKB, and hence strengthens its inferential power. Both problems are important for a variety of reasons, e.g., to reduce modeling effort and to keep the GSKB small by identification of entailed and hence redundant atoms. We are identifying a class of desirable logical models which we call preferred models – our approach is model-theoretic in nature. We prove that the strengthened GSKBs enforces those preferred models. The prefered models capture the desired co-references. The presented framework is a logical reconstruction of an algorithm which we successfully applied on a large-scale biological knowledge base, in which it identified more that 22,000 equality atoms.

1 Introduction

Underspecified knowledge bases occur naturally, for example, when modeling biological knowledge of the form:

S1 Every $Cell$ has part a $Ribosome$ and a $Chromosome$.

S2 Every $EukaryoticCell$ is a $Cell$.

S3 Every $EukaryoticCell$ has part a $Eukaryotic\text{-}Ribosome$, a $EukaryoticChromosome$, a $Nucleus$, such that the $EukaryoticChromosome$ is inside the $Nucleus$.

S4 Every $EukaryoticRibosome$ is a $Ribosome$.

S5 Every $EukaryoticChromsome$ is a $Chromsome$.

Here the question arises - is the $EukayoticChromosome$ that S3 is talking about actually the $Chromsome$ from S1? Such assumptions are often reasonable. We are calling a knowledge base which does not answer this question definitely *underspecified.* These kinds of question are studied to some extent in the field of computational linguistics, where it is called *anaphora resolution* [Carpenter, 1994], [Cohen, 2007]. We will use the term *co-reference resolution* in the following.

From a logical point of view, these sentences correspond to the following FOPL formulas; we are using the comma in consequents to denote conjunction, and combine $S2, S3$ into $S23$:

S1 $\forall x : Cell(x) \Rightarrow \exists x_1, x_2 :$
$hasPart(x, x_1), Ribosome(x_1),$
$hasPart(x, x_2), Chromosome(x_2)$

S23 $\forall x : EukaryoticCell(x) \Rightarrow \exists x_3, x_4, x_5 : Cell(x),$
$hasPart(x, x_3), Euk.Ribosome(x_3),$
$hasPart(x, x_4), Euk.Chromosome(x_4),$
$hasPart(x, x_5), Nucleus(x_5), inside(x_4, x_5)$

S4 $\forall x : Euk.Ribosome(x) \Rightarrow Ribosome(x)$

S5 $\forall x : Euk.Chromosome(x) \Rightarrow Chromosome(x)$

In the following, unary predicates are called *concepts*, and binary predicates *relations*. The concept D is called a *superconcept* of C if $\forall x : C(x) \Rightarrow D(x), \ldots$, and C is a *subconcept* of D.

Co-references are, in general, tedious to specify at knowledge authoring time, or impossible if the input is underspecified in the first place (e.g., if natural language or incomplete visual graphical formalisms are used for GSKB authoring). Hence, an automatic co-reference resolution algorithm is desired which will necessarily have to rely on some sort of guessing. Related motivation and mechanisms can be found in the literature. For example, the KM reasoning system [Clark and Porter, 1997] employs a so-called unification operator for this purpose. Dealing with underspecified object-oriented knowledge bases has also been studied in the context of answer set programming (ASP) in [Chaudhri and Tran, 2012] where a unification operator is presented. These and related approaches (e.g., based on Description Logics) are discussed in more detail in Section 5.

We consider co-reference resolution in GSKBs an important problem to solve because of its potential to reduce modeling effort and to maximize the deductive power of a GSKB (get more entailments from it). For example, if we extend $Ribosome(x_1)$ in $S1$ by saying that it is *inside* $Cytosol(x_6)$, resulting in

S1b $\forall x : Cell(x) \Rightarrow \exists x_1, x_2, x_6 :$
$hasPart(x, x_1), Ribosome(x_1),$
$hasPart(x, x_2), Chromosome(x_2),$
$inside(x_1, x_6), Cytosol(x_6)$

then we would like to derive that the same holds for $Euk.Ribosome(x_3)$ in $EukaryoticCell$, assuming that the $Ribosome$ which got inherited from $Cell$ got specialized. Thus, there is only *one* $Euk.Ribosome$ in $EukaryoticCell$, and no additional $Ribosome$. However, this only holds if there is a co-reference between x_3 and x_1. We can ensure that $x_3 = x_1$ holds in $S1b$ if we "strengthen" the GSKB by addition of equalities between Skolem function values representing the original existentials. Such a hypothesized strengthened GSKB might look as follows:

S1b' $\forall x : Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x))$

S23' $\forall x : EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x))),$
$f_3(x) = f_1(x), f_4(x) = f_2(x)$

Note that $\forall x : EukaroyticCell(x) \models \exists x_1, x_3 : hasPart(x, x_1), Euk.Ribosome(x_1), inside(x_1, x_3), Cytosol(x_3)$ in $\{S1b', S23', S4, S5\}$, due to $f_3(x) = f_1(x)$ in $S23'$. Hence, this piece of knowledge does not have to be remodelled in $S23'$.

The provenance of an atom is important, as it enables us to identify and remove implied and hence *redundant* atoms. This reduces the size of the GSKB and keeps it manageable. For example, in $S23$, it is not necessary to re-state that every $EukaryoticCell$ has part a $Ribosome$, because this can be derived from $\{S1, S2\}$. Often, established co-references affect the provenance information. An example is again given in $\{S1b, S23, S4, S5\}$ by $\forall x : EukaryoticCell(x) \not\models \exists x_1, x_3, x_9 : hasPart(x, x_3), Euk.Ribosome(x_3), inside(x_3, x_9), Cytosol(x_9)$, and hence, $inside(x_3, x_9)$ would *not* be redundant if added to $S23$. However, this entailment *does* hold in $\{S1b', S23', S4, S5\}$, because of $f_3(x) = f_1(x)$ co-reference in $S23'$.

The contribution of this paper is the presentation of a novel GSKB rewriting algorithm. It rewrites a GSKB such as $\{S1b, S23, S4, S5\}$ into a strengthened GSKB, similar to $\{S1b', S23', S4, S5\}$. From the strengthened GSKB we compute the provenance of atoms and co-references. We will use a model theoretic notion of *preferred models* to characterize the additional desirable inferences that we wish to get from the underspecified GSKB. The information in the preferred model is used to rewrite a Skolemized version of the GSKB into a strengthened version of the GSKB, whose models are exactly the preferred ones. Obviously, deciding entailment of atoms, and hence the provenance problem, are in general undecidable problems in FOPL. In our fragment of FOPL(=) these problems are decidable.

The paper is structured as follows: We first formally define the GSKB framework and required notions of GSKB and strengthened GSKB, and the semantic notion of preferred models. We then present the algorithm and show that the strengthened GSKB (produced by the algorithm) has models which are preferred which hence gain the required additional conclusions in order to decide the provenance and co-reference problems. Next we evaluate the algorithm on a large-scale biological graph-structured GSKB from the AURA project [Gunning, D. and Chaudhri, V. K. et al., 2010]. Finally we conclude and discuss related and future work.

2 Graph Structured Knowledge Bases

In the following, we denote an atom or a conjunction of atoms with free variables $\{x, x_1, \ldots, x_n\}$ as $\varphi(x, \vec{x})$, with $\vec{x} = (x_1, \ldots, x_n)$. *Graph-structured knowledge bases* (GSKBs) are formulated in first order-logic with equality (FOPL(=)). We assume that there is a function $terms$ which returns the terms in a formula (or atom), e.g. $terms(R(t_1, t_2)) \triangleq \{t_1, t_2\}$:

Definition 1. *Basic Definitions. Let $\mathcal{C}$ be a countably infinite set of unary predicate names, $\mathcal{R}$ be a countably infinite set of binary predicate names, and $\mathcal{F} = \{f_1, f_2, \ldots\}$ be a countably infinite set of unary function names. We have no constants. Hence, $(\mathcal{C} \cup \mathcal{R}, \mathcal{F})$ constitutes the signature of the exploited fragment of first-order logic with equality. Elements in $\mathcal{C}$ are called concepts, and elements in $\mathcal{R}$ relations. Moreover, let $\mathcal{X} = \{x, x_1, x_2, \ldots\}$ be a countably infinite set of variables. We only allow function nesting of max. depth 2 - t is a* GSKB term *iff either $t \in \mathcal{X}$, or $t = f_i(x)$, or $t = f_i(f_j(x))$. In the following, assume that t, t_1, t_2 are GSKB terms:*

GSKB atoms: *Let $\{C, D\} \subseteq \mathcal{C}$, $R \in \mathcal{R}$, $\{v, w\} \subseteq \mathcal{X}$. Then, $C(v)$ and $C(f_i(x))$ are concept atoms, and $R(v, w)$, $R(x, f_i(x))$ are relation atoms. Moreover, there are equality and in-equality atoms of the following form: $f_i(x) = f_j(x), f_i(x) = f_j(f_k(x)), f_j(f_k(x)) = f_i(x)$, and $f_i(x) \neq f_j(x)$, with i, j, k pairwise unequal.*

GSKB rule: *For a concept C, a formula of the form $\forall x : C(x) \Rightarrow \exists! \vec{x} : \varphi(x, \vec{x})$, where $\varphi(x, \vec{x}) = \bigwedge_{i \in 1 \ldots m} \alpha_i$ is finite conjunction of GSKB atoms. This is shorthand for $\forall x : C(x) \Rightarrow \exists \vec{x} : pairwise\_disjoint(x, \vec{x}) \wedge \varphi(x, \vec{x})$, $\vec{x} = (x_1, \ldots, x_n)$, with the macro $pairwise\_disjoint(x, \vec{x}) \triangleq \bigwedge_{1 \leq i < j \leq n} x_i \neq x_j \wedge \bigwedge_{1 \leq i \leq n} x_i \neq x$.*

For a concept C with $\rho_C = \forall x : C(x) \Rightarrow \exists! \vec{x} : \varphi(x, \vec{x})$, denote $\varphi(x, \vec{x}) = \bigwedge_{i \in 1 \ldots m} \alpha_i$ as a set by $\tau_C = \{\alpha_1, \ldots, \alpha_m\}$.

We require that the terms in $terms(C)$ are connected to *x: for all $t \in terms(C)$, $connected(x, t)$ holds, where*

connected is defined as follows: $connected(t_1, t_2)$ *holds if* $\{R(t_1, t_2), R(t_2, t_1)\} \cap \tau_C \neq \emptyset$*, or there is some* t *s.t.* $connected(t_1, t)$ *and* $connected(t, t_2)$ *holds.*

GSKB: *A finite set of GSKB rules* Σ *in which there is at most one rule per concept.*

Input GSKB: *A GSKB which is function-free and without equality atoms.*

Auxiliary notions: *Given a GSKB* Σ*, we refer to the set of concepts used in* Σ *as* $concepts(\Sigma)$*, and* $\tau_{C,\Sigma}$ *to refer to the consequent of* $\rho_C \in \Sigma$*. We extend the other definitions to accept a* Σ *argument as well, e.g.,* $terms(C, \Sigma)$*, etc.*

For example, $\{S1b, S23, S4, S5\}$ is an ("underspecified") input GSKB, and $\{S1b', S23', S4, S5\}$ is a *strengthened* (output) GSKB; however, we need to replace the $\exists$ quantifier with $\exists!$. Strengthened GSKB is defined below. Note that somtimes the strengthening algorithm will not add anything, and the input with equal the ouput, e.g. for $\{S4, S5\}$.

We require that a input GSKB must be *coherent*:

Definition 2. *Coherent GSKB and coherent model. A GSKB* Σ *is coherent if there is standard first-order model* $\mathcal{I} = (\Delta, \cdot^{\mathcal{I}})$*,* $\mathcal{I} \models \Sigma$*, in which every concept* C *mentioned in* Σ *is interpreted in a non-empty way:* $C^{\mathcal{I}} \neq \emptyset$*. Such a model is called a* coherent model.

Moreover, we define standard notions such as $superconcepts$ as follows:

Definition 3. *Auxiliary Definitions. Let* C *be a concept,* Σ *be a GSKB. We then define the following functions and predicates w.r.t.* Σ*:*

- $asserted\_types(C, \Sigma) \triangleq \{D \mid D(t) \in \tau_{C,\Sigma}\}$
- $has\_asserted\_type_\Sigma(C, D)$ *iff* $D \in asserted\_types(C, \Sigma)$
- $asserted\_superconcepts(C, \Sigma) \triangleq \{D \mid D(x) \in \tau_{C,\Sigma}\}$
- $has\_asserted\_superconcept_\Sigma(C, D)$ *iff* $D \in asserted\_superconcepts(C, \Sigma)$
- $superconcepts(D, \Sigma) \triangleq \{E \mid has\_asserted\_superconcept^+_\Sigma(D, E)\ \}$
- $has\_superconcept_\Sigma(C, D)$ *iff* $D \in superconcepts(C, \Sigma)$
- $all\_types_\Sigma(C) \triangleq \{E \mid D \in asserted\_types(C, \Sigma), E \in superconcepts(D, \Sigma)\} \cup superconcepts(C, \Sigma)$
- $has\_type_\Sigma(C, D)$ *iff* $D \in all\_types(C, \Sigma)$

where R^+ *denotes the transitive closure of relation* R*.*

We require that the relations $has\_superconcept_\Sigma$ and has_type_Σ are irreflexive and define:

Definition 4. *Admissible GSKB. An input GSKB* Σ *is called admissible if* Σ *is coherent,* $has\_superconcept_\Sigma$ *and* has_type_Σ *are irreflexive, and if there are no implied concept atoms in the rules: for all* $C \in concepts(\Sigma)$*, if* $D(t) \in \tau_{C,\Sigma}$*, then for all* $E \in superconcepts(D, \Sigma)$*:* $E(t) \notin \tau_{C,\Sigma}$*,*

The following is straightforward:

Proposition 1. *Every admissible GSKB* Σ *has a coherent, finite model.*

Proof. Given that we do not support negation of concepts or relations, and given that inequality atoms are only introduced by the $\exists!$ quantor, inconsistencies such as $x \neq x$ cannot occur. Moreover, since GSKB $has\_superconcept_\Sigma$ and has_type_Σ are irreflexive, the GSKB is acyclic, and the consequent of every rule can be "unfolded", analog to the unfolding of an acyclic TBox in description logics [Baader *et al.*, 2003]. This produces a finite consequent for every rule. Next, for every $\rho_C \in \Sigma$, C can be instantiated s.t. $i_C \in C^{\mathcal{I}}$ holds, and we can easily satisfy the existentials in the consequent by creating one instance per variable. The process terminates and produces a model of Σ which is coherent and finite. □

We need a notion of connectedness on models:

Definition 5. *Predicate* $connected$ *on models. Let* $\mathcal{I} = (\Delta, \cdot^{\mathcal{I}})$ *be a model of* Σ*. For* $i, j \in \Delta^{\mathcal{I}}$*, we define* $connected_{\mathcal{I}}(i, j)$ *if, for some* $R \in \mathcal{R}$*,* $\{(i, j), (j, i)\} \cap R^{\mathcal{I}} \neq \emptyset$*, or there is some* $k \in \Delta_{\mathcal{I}}$ *s.t.* $connected_{\mathcal{I}}(i, k)$ *and* $connected_{\mathcal{I}}(k, j)$*.*

In the following we are considering admissible GSKBs only, and we are interested in their *preferred models*. The intuition behind the notion of a preferred model is the following: for every concept C, there should be a *prototypical model* of C which is *not* also a model of some non-superconcept of C, in the form of a connected graph that "mirrors" the atoms in $\tau_{C,\Sigma}$ – due to the disjointness axioms there will be at least one individual per variable in ρ_C in this model. Moreover, the prototypical model for C also contains inherited "graphs" from concepts in $all\_types_\Sigma(C)$. Hence, the graph satisfying the atoms $\tau_{C,\Sigma}$ is only a subgraph of the full model for C. Most importantly, the notion of a preferred model captures the intuition that inherited content can be specialized, and hence should give rise to co-references: in the prototypical model for $EukaryoticCell$, the $Chromsome$ inherited from its superclass $Cell$ will be represented by the same individual as its own local $Euk.Chromsome$. Note that this minimizes the extension of $Chromsome$. The same argument applies to arbitrary conjunctions: we will *not* identify the inherited $Chromsome$ with the local $Euk.Ribosome$, as this would result in a model in which $Chromsome \wedge Euk.Ribosome$ is interpreted non-empty, and there are models in which this conjunction is interpretated by the empty set. These intuitions are formalized as follows:

Definition 6. *Preferred model of admissible GSKB* Σ*. Let* Σ *be an admissible GSKB, and* $\mathcal{I} \models \Sigma$ *be a coherent finite model. Then,* $\mathcal{I}$ *is called* preferred *if the following holds:*

1. *for every concept* $C \in concepts(\Sigma)$*, there is (at least) one* $i \in C^{\mathcal{I}}$ *s.t. for all* D*, if* $has\_superconcept(D, C)$*, then* $i \notin D^{\mathcal{I}}$ *– hence, there is at least one element which is "unique" to* C*, and denoted by* i_C*.*
2. *for* $C \in concepts(\Sigma)$*, define* $participants_{\mathcal{I}}(C) \triangleq \{j \mid connected_{\mathcal{I}}(i_C, j)\}$*. Then, for all* $C, D \in concepts(\Sigma)$*, with* $C \neq D$*, the following holds:* $participants(C) \cap participants(D) = \emptyset$*.*

3. *for every non-empty subset $\mathcal{CS} \subseteq concepts(\Sigma)$, there is no preferred model $\mathcal{I} \neq \mathcal{I}'$, with $\Delta_{\mathcal{I}'} \subseteq \Delta_{\mathcal{I}}$ s.t. $\bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'} \subset \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}}$.*

Consider the preferred models of $\{S1, S23, S4, S5\}$. We are forced to have at least one "unique" $Cell$ which is not a $EukaryoticCell$, due to 1. Otherwise, every $Cell$ would acquire the properties of $EukaryoticCell$s, which is not desirable. Moreover, none of the individuals connected to that unique $Cell$ are shared by another concept, due to 2. Hence, the concept models have the forms of "non-overlapping graphs", and inherited content is "mapped in". We are forced to minimize the extension of every concept, as well as of every conjunction of concepts. This prevents models in which, for example, $Ribosome^{\mathcal{I}} \cap Euk.Chromosome^{\mathcal{I}} \neq \emptyset$ holds, as there are smaller models in which they are interpreted disjointly: $Ribosome^{\mathcal{I}} \cap Euk.Chromosome^{\mathcal{I}} = \emptyset$. Note that the inequality atoms in Σ only prevent "merging" of variables within the same formula, but the individual for $Chromosome(y_2)$ inherited from $Cell$ could in principle be made co-referential with the local $Euk.Ribosome(y_3)$ in $EukaryoticCell$. *This is prevented in an admissible model.* Also, looking at the model of $EukaryoticCell$, the co-reference between the from $Cell$ inherited $Chromosome(y_2)$ and its own local $Euk.Chromosome(y_4)$ is made explicit, since this will result in the smallest (extension of) $Chromosome^{\mathcal{I}}$. A model in which a $EukaryoticCell$ would have two different $Chromosome$s would be larger and in violation to 3. So, we only make those conjunction true in an preferred model that we *have to make true* - for example, $Cell^{\mathcal{I}} \cap EukaryoticCell^{\mathcal{I}} \neq \emptyset$, due to $S23$, and there is no model in which this conjunction is interpreted by a smaller set.

Note that an preferred model is not a "minimal" model in the classical sense. Consider $\forall x : C(x) \Rightarrow \exists!x1 : R(x, x1), D(x1)$, $\forall x : SubC(x) \Rightarrow \exists!x2 : C(x), R(x, x2), E(x2)$. In the classical minimal model $\mathcal{I}$, we would have $\#\Delta = 2$, and it would satisfy $D \wedge E$. Also, $C^{\mathcal{I}} = SubC^{\mathcal{I}}$. But this is not what we want. It violates 1, 2, as well as 3. The preferred model will need at least 5 nodes.

In principle, there can be more than one preferred model and hence, more than one strengthened version of the GSKB. For example, consider the GSKB

$$\begin{aligned} &C(x) \Rightarrow \exists!x_1 : R(x, x_1), E(x_1) \\ &SubC(x) \Rightarrow \exists!x_2, x_3 : \ C(x), \\ &\qquad\qquad\qquad\qquad R(x, x_2), E(x_2), F(x_2), \\ &\qquad\qquad\qquad\qquad R(x, x_3), E(x_3), G(x_3). \end{aligned}$$

Here, x_1 in C can be co-referential with either x_2 in $SubC$, or with x_3.

In the next section, we will show the following constructively, by specifying an algorithm which constructs an preferred model for a given admissible GSKB Σ:

Proposition 2. *Every admissible GSKB has an preferred model.*

We can now state the purpose of the GSKB strengthening algorithm more clearly. Given an admissible GSKB Σ (note that this is an input GSKB), we are interested in finding a strengthened version of Σ:

Definition 7. *Strengthened version of Σ. Given an admissible (input) GSKB Σ, we are calling Σ' a* strengthened version *of Σ if the following holds:*

1. *for every rule $\rho_C \in \Sigma$, there is a rule $\rho'_C \in \Sigma'$ that uses only the variable x: $\mathcal{T}(\rho'_C) \cap \mathcal{X} = \{x\}$.*
2. *if $\mathcal{I}' \models \Sigma'$ is a standard first-order model of Σ' which is coherent, then $\mathcal{I}' \models \Sigma$, and $\mathcal{I}'$ is an preferred model for Σ. Hence, $\Sigma' \models \Sigma$.*

From a strengthened GSKB, we can decide provenance and co-reference as follows:

Definition 8. *Provenance and co-reference determination. Let C be a concept, Σ' be a strengthened GSKB, and $\mathcal{P} \subseteq \tau_{C,Sigma'}$. With $\beta = \bigwedge_{\alpha \in \mathcal{P}} \alpha$, we then say that β (and hence all the atoms in $\mathcal{P}$) are*

- *local (or asserted) in C if*
  $\Sigma' \setminus \{\rho_C\} \cup \{\forall x : C(x) \Rightarrow \bigwedge_{\alpha \in \tau_{C,\Sigma'} \setminus \mathcal{P}} \alpha\}$
  $\not\models \forall x : C(x) \Rightarrow \beta$,
- *and inherited otherwise. More specifically, β (and $\mathcal{P}$) is inherited from D, iff $D(t) \in \tau_{C,\Sigma'}$, and $\beta' = \bigwedge_{\alpha \in \mathcal{P}'} \alpha$ with $\mathcal{P}' = \{\alpha_{[f_i(t) \Rightarrow f_i(x)]} \mid \alpha \in \mathcal{P}\}$ is local in D, and there is no more general $SupD$ with $has\_superconcept_{\Sigma'}(D, SupD)$ such that β (and $\mathcal{P}$) is inherited from $SupD$.*

Moreover, given concepts C, D, two GSKB terms $t_1 \in terms(C)$, $t_2 \in terms(D)$ are said to be co-referential in Σ' iff either $t_1 = t_2 = x$, $t_1 = f_i(x)$, $t_2 = f_j(f_k(x))$, or $t_2 = f_i(x)$, $t_1 = f_j(f_k(x))$, and $\Sigma' \models (\forall x : C(x) \Rightarrow f_i(x) = f_j(x) \vee (\forall x : D(x) \Rightarrow f_i(x) = f_j(x))$.

Note that a conjunction β is local as soon as *some* atom is already local. Hence, if a complex conjunction β (resp. $\mathcal{P}$) is local, this does *not* mean that *all* its atoms have to be local – some atoms may be inherited.

Proposition 3. *Provenance and co-reference are decidable in a strengthened GSKB Σ'.*

The proof is given in the next Section.

3 Constructing a Strengthened GSKB

The algorithm works by performing the following steps:

1. Produce the skolemized version of Σ, Σ_S, by bringing every rule in Σ into Skolem normal form. The Skolemized axioms contain no nested function terms, only terms of the form $f_i(x)$ and x. Let $\mathcal{O} \triangleq \{o_C \mid C \in concepts(\Sigma)\}$ be a set of constants, and also add $\{C(o_C) \mid C \in concepts(\Sigma)\}$ to Σ_S.
2. Construct the *minimal Herbrand model* $\mathcal{I}_{\mathcal{H}} = (\Delta_{\mathcal{H}}, \cdot^{\mathcal{I}_{\mathcal{H}}})$ of Σ_S. The minimal Herbrand model is unique and finite, given that Σ is admissible (and does not contain disjunctions in the consequents). Note that the minimal Herbrand model will automatically satisfy the inequality atoms, and it will also satisfy points 1 and 2 from Definition 6, due to the set of constants $\mathcal{O}$ which are instantiated as $\{C(o_C) \mid C \in concepts(\Sigma)\} \subseteq \Sigma_S$, and with the exception of x, there are no shared terms

in the rules of Σ_S, as Skolemization creates fresh function symbols for every variable. Thus, o_C represents the root individual of the unique model for concept C, with $o_C^{\mathcal{I}_\mathcal{H}} = i_C, i_C \in C^{\mathcal{I}_\mathcal{H}}$.

3. Transform $\mathcal{I}_\mathcal{H}$ into an preferred model of Σ, $\mathcal{I}_\mathcal{A} = (\Delta_\mathcal{A}, \cdot^{\mathcal{I}_\mathcal{A}})$. $\Delta_\mathcal{A}$ is the quotient set of $\Delta_\mathcal{H}$ under the $=$ equivalence relation, $\Delta_\mathcal{A} = \Delta_\mathcal{H} \backslash =$. Hence, the elements of $\Delta_\mathcal{A}$ represent the equivalence classes of equated Skolem ground terms from the Herbrand universe $\Delta_\mathcal{H}$. This step is non-deterministic, as there may be more than one preferred model for Σ.
4. Use $\mathcal{I}_\mathcal{A}$ to construct a strengthened GSKB Σ' from Σ_S which is satisfied by that model. Use the equivalent clusters in $\Delta_\mathcal{A}$ to generate equality atoms.
5. From Σ' it is possible to decide the provenance and the co-reference problem, on a syntactic basis.

Since steps 1 and 2 are standard and well-know [Hedman, 2004], let us define the algorithm for step 3. We need two more utility notions before we can proceed:

Definition 9. *Relations $\mathcal{E}$ and $\mathcal{U}$, and equivalence classes. Let $\mathcal{I}_\mathcal{H} = (\Delta_\mathcal{H}, \cdot^{\mathcal{I}_\mathcal{H}})$ be the minimal unique Herbrand model after step 2 of Σ_S above. Let $\mathcal{E}$ be a binary relation over terms from the Herbrand universe $\Delta_\mathcal{H}$, and define*

$$closure(\mathcal{E}) \triangleq \bigcup_{C \in concepts(\Sigma), k \in \Delta_\mathcal{H}} \{(f_1(k), f_2(k)) \mid (f_1(o_C), f_2(o_C)) \in \mathcal{E}^{\circledast}\} \cup \{(f_1(f_2(k)), f_3(k)) \mid (f_1(f_2(o_C)), f_3(o_C)) \in \mathcal{E}^{\circledast}\} \cup \{(f_1(f_2(k)), f_3(f_4(k))) \mid (f_1(f_2(o_C)), f_3(f_4(o_C))) \in \mathcal{E}^{\circledast}\}$$

and $\cdot^{\circledast}$ denotes the reflexive, symmetric, and transitive closure *of a relation. Let $[i]^\mathcal{E} \triangleq \{j \mid (i,j) \in closure(\mathcal{E})\}$. Moreover, let $\mathcal{U} \triangleq \{[i]^\mathcal{E} \neq [j]^\mathcal{E} \mid i_1 \in [i]^\mathcal{E}, j_1 \in [j]^\mathcal{E}, C \in concepts(\Sigma), (i_1 \neq j_1) \in \tau_{C,\Sigma_S}\}$ be the set of inequality atoms.*

Intuitively, $(i,j) \in \mathcal{E}$ represents $i = j$, and $[i]^\mathcal{E}$ represents the equivalence class of i. The relation $\mathcal{E}$ (and hence the equivalence classes) will grow as pairs of equated individuals / terms are added by the algorithm given below. Intuitively, the closure operator makes sure that whenever two terms starting from the same root node o_C are equated in the unique model of C, that then this equality will also hold for all its C instantiations in other parts of the model. Note that also $\mathcal{U}$ will grow, representing inferences such as $i \neq j, k \neq l, j = k \Rightarrow i \neq l$.

The algorithm can now be stated as follows:

Algorithm 1. *Construction of an preferred model for Σ. Let $\mathcal{I}_\mathcal{H} = (\Delta_\mathcal{H}, \cdot^{\mathcal{I}_\mathcal{H}})$ be the minimal unique Herbrand model of Σ_S after step 2 above.*

1. *define $hasRoot(i) \triangleq o_C$ iff $connected_{\mathcal{I}_\mathcal{H}}(o_C, i)$ holds, for every $C \in concepts(\Sigma)$.*
2. *then, non-deterministically apply the following* merging rule *on the model as long as it is applicable:*

   ***if** there are individuals $i, j \in \Delta_\mathcal{H}$, $i \neq j$, with $hasRoot(i) = hasRoot(j) = o_C$ and $ind\_types(i) \subseteq ind\_types(j)$, $i \notin [j]^\mathcal{E}$, $[i]^\mathcal{E} \neq [j]^\mathcal{E} \notin \mathcal{U}$, **then** $\mathcal{E} \triangleq \mathcal{E} \cup \{(i,j)\}$.*

   Assume the rule application stops with a global maximum of inequality atoms s.t. $\#\mathcal{U}$ is maximized. Since this is a non-deterministic algorithm, such a run exists, and we can assume that the non-deterministic algorithm will produce it.
3. *define $\mathcal{I}_\mathcal{A} = (\Delta_\mathcal{A}, \cdot^{\mathcal{I}_\mathcal{A}})$ as follows:*
   $\Delta_\mathcal{A} \triangleq \{[i]^\mathcal{E} \mid i \in \Delta_\mathcal{H}\}$*, and for all $C \in concepts(\Sigma)$:* $C^{\mathcal{I}_\mathcal{A}} \triangleq \{[i]^\mathcal{E} \mid i \in C^{\mathcal{I}_\mathcal{H}}\}$*, for all $R \in \mathcal{R}$:* $R^{\mathcal{I}_\mathcal{A}} \triangleq \{([i]^\mathcal{E}, [j]^\mathcal{E}) \mid (i,j) \in R^{\mathcal{I}_\mathcal{H}}\}$.

The algorithm terminates, since $\mathcal{I}_\mathcal{H}$ is finite, so there is a finite set of merging possibilities in the rule. The solution which maximizes $\#\mathcal{U}$ can obviously be found by search in a deterministic version.

Lemma 1. *$\mathcal{I}_\mathcal{A} = (\Delta_\mathcal{A}, \cdot^{\mathcal{I}_\mathcal{A}})$ is an admissible model for Σ.*

Proof. Obviously, $\mathcal{I}_\mathcal{A}$ is finite and coherent, as it was constructed by the algorithm based on the unique finite Herbrand model. Assume that $\mathcal{I}_\mathcal{A}$ is not an preferred model for Σ. By construction, $\mathcal{I}_\mathcal{A}$ is a model of Σ_S, as the merging rule preserves the model character of $\mathcal{I}_\mathcal{H}$. Since $\mathcal{I}_\mathcal{H}$ is a model of the Skolemized version, it is also a model of Σ, since $\Sigma_S \models \Sigma$ for the Skolemized GSKB [Hedman, 2004]. Hence, $\mathcal{I}_\mathcal{A}$ is a model of Σ, also.

It remains to show that it is admissible. Assume that it is not. Since points 1 and 2 from Definition 6 are already satisfied by construction, only 3 can be violated. Then, there must be some other model $\mathcal{I}'$ and some $\mathcal{CS} \subseteq concepts(\Sigma)$ such that $\bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'} \subset \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}}$, witnessed by $[i]^\mathcal{E} \in \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}}$ with $[i]^\mathcal{E} \notin \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'}$.

1. If $\bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'} = \emptyset$, then this violates the assumption that $\Sigma_\mathcal{H}$ was a minimal Herbrand model (which does not make things true without need). Hence, $\bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}} = \emptyset$ as well, which contradicts $[i]^\mathcal{E} \in \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}}$.
2. Assume $\mathcal{CS} = \{D\}$ is a single concept name. As $\mathcal{I}_\mathcal{H}$ was a minimal model, the existence of i, with $i \in [i]^\mathcal{E}$, is somehow enforced by Σ_S, hence there is some term $t_i \in terms(C, \Sigma_S)$ with $D \in ind\_types(t_i)$. Moreover, for the same reason, $D^{\mathcal{I}'} \neq \emptyset$, as otherwise it wouldn't be a model, but $i \notin D^{\mathcal{I}'}$. Consequently, there is some $j \in D^{\mathcal{I}'}$ with $i \neq j$. Then, there must also be some $t_j \in terms(C, \Sigma_S)$ with $D \in ind\_types(t_j)$, with $t_i \neq t_j$.
   There are a couple of cases:
   (a) Assume $ind\_type(t_i) \subseteq ind\_types(t_j)$
      i. if $C' = C$ and hence $hasRoot(i) = hasRoot(j) = C$, then $(t_i \neq t_j) \notin \tau_{C,\Sigma_S}$ and $[i]^\mathcal{E} \neq [j]^\mathcal{E} \notin \mathcal{U}$, as otherwise $\mathcal{I}$ would not be a model. But then, the merging rule would have been applied and merged i and j, such that $[i]^\mathcal{E} = [j]^\mathcal{E} = \{i, j\}$. Rule application could not have been blocked by the precondition $[i]^\mathcal{E} \neq [j]^\mathcal{E} \notin \mathcal{U}$, because $\mathcal{I}_\mathcal{A}$ was produced by a run in which $\#\mathcal{U}$ was maximized. This means that the rule will be applicable and equate i and j, contradicting the assumption that the algorithm has terminated.
      ii. otherwise, $C \neq C'$, then we don't have to worry: as stated in Definition 6, $participants(C) \cap participants(C') = \emptyset$.

(b) Assume $ind\_type(t_j) \subseteq ind\_types(t_i)$: analog to the previous case.

(c) Assume $ind\_type(t_i) \not\subseteq ind\_types(t_j)$. Then there is some $E \in ind\_type(t_i), E \notin ind\_types(t_j)$. As $\mathcal{I}_\mathcal{A}$ was a minimal Herbrand model, and there is no way for $[i]^\mathcal{E}$ to "vanish" from $E^{\mathcal{I}_\mathcal{A}}$, there must be $[i]^\mathcal{E} \in E^{\mathcal{I}_\mathcal{A}}$ and hence $[i]^\mathcal{E} \in \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'}$. Contradiction.

3. If $\mathcal{CS} = \{D_1, \ldots, D_n\}$, then there must already be some $\mathcal{CS}' = \{D_m, D_n\}, \mathcal{CS}' \subseteq \mathcal{CS}$ for which we have such an i. If $has\_superconcept(D_m, D_n)$ or vice versa, then there is already some $\mathcal{CS}' = \{D_m\}$, and this is handled by 2. Otherwise, D_m, D_n are not in a sub/superconcept relationship, and corresponding instances are not getting merged by the merging rule. But similar to 2c), this will lead us to conclude that $[i]^\mathcal{E} \in \bigcap_{C \in \mathcal{CS}} C^{\mathcal{I}'}$, contradicting the assumption.

Hence, $\mathcal{I}_\mathcal{A}$ is an preferred model. Note that this proves Proposition 2. □

For what remains to be shown is how we can compute a strengthened GSKB from Σ_S and $\mathcal{I}_\mathcal{A}$.

Definition 10. *Construction of strengthened GSKB Σ'. Let Σ_S be the skolemized version of the admissible GSKB, and $\mathcal{I}_\mathcal{A}$ be an preferred model of Σ. We then rewrite the rules in Σ_S as follows; note that $\alpha_{[t_1 \Rightarrow t_2]}$ means "in α, substitute all occurrences of t_1 with t_2":*

$$\Sigma' \triangleq \{rewrite(\rho_C, terms(C, \Sigma_S)) \mid \rho_C \in \Sigma_S\}, \text{with}$$
$$rewrite(\rho_C, terms) \triangleq C(x) \Rightarrow$$
$$\bigwedge_{\alpha \in \tau_{C,\Sigma_S}} \alpha \wedge$$
$$\bigwedge_{t \in terms, t \neq o_C} hasRoot(t, x)_{[o_C \to x]} \wedge$$
$$\bigwedge_{t_1, t_2 \in terms, t_1 \neq t_2} t_1 \neq t_{2[o_C \to x]} \wedge$$
$$\bigwedge_{t_1 \in terms, t_2 \in [t_1]} t_1 = t_{2[o_C \to x]}$$

In addition, we need the following axioms:

1. $\Sigma' \triangleq \Sigma' \cup \{C(o_C) \mid C \in concepts(\Sigma)\}$
2. $\Sigma' \triangleq \Sigma' \cup \{o_C \neq o_D \mid C, D \in concepts(\Sigma), C \neq D\}$
3. $\Sigma' \triangleq \Sigma' \cup \{\forall x, y, z : hasRoot(x, y), hasRoot(y, z) \Rightarrow hasRoot(x, z)\}$
4. $\Sigma' \triangleq \Sigma' \cup \{\forall x, y : hasRoot(x, o_C), hasRoot(y, o_D) \Rightarrow x \neq y\}$,

for all $C, D \in concepts(\Sigma), C \neq D$.

Lemma 2. *If $\mathcal{I} \models \Sigma'$, then $\mathcal{I}$ is an preferred model for Σ.*

Proof. As Σ' has been constructed from Σ_S by adding equality atoms to explicitly represent the co-references with inherited Skolem function successors, which have been identified by the merging rule from an preferred model of Σ, it is clear that any model of Σ' will force the same co-references, and hence, satisfy point 3 in Definition 6. Moreover, point 1 in Definition 10 makes sure that we have non-empty concept models for every concept by requiring an instance, hence satisfying condition 1 in Definition 6. Point 2 in Definition 10 enforces disjointness between those constants, and point 3 declares $hasRoot$ as a transitively closed relation. In combination with the added $hasRoot$ atoms in Σ', and with the axioms in point 4 of Definition 10, this ensures that condition 2 in Definition 6 is satisfied, requiring that the unique concept models do not overlap (no sharing of participants). □

Let us return to our example. For $\Sigma = \{S1b, S23, S4, S5\}$ we will get Σ_S as follows:

$Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$inside(f_1(x), f_0(x)), Cytosol(f_0(x)),$
$pairwise\_disjoint(\{x, f_0(x), f_1(x), f_2(x)\})$
$EukaryoticCell(x) \Rightarrow Cell(x)$
$hasPart(x, f_3(x)), Euk.Ribosome(f_3(x)),$
$hasPart(x, f_4(x)), Euk.Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x)),$
$pairwise\_disjoint(\{x, f_3(x), f_4(x), f_5(x)\})$
$Cell(o_{Cell}), EukaryoticCell(o_{EukarioticCell})$
$Ribosome(o_{Ribosome}) \ldots$

If we look at the minimal Herbrand model of Σ_S, we find that the following atoms are satisfied for $o_{Euk.Cell}$:

$hasPart(o_{Euk.Cell}, f_1(o_{Euk.Cell})),$
$hasPart(o_{Euk.Cell}, f_2(o_{Euk.Cell})),$
$inside(f_1(o_{Euk.Cell}), f_0(o_{Euk.Cell})),$
$Ribosome(f_1(o_{Euk.Cell})),$
$Chromosome(f_2(o_{Euk.Cell})),$
$Cytosol(f_0(o_{Euk.Cell})),$
$hasPart(o_{Euk.Cell}, f_3(o_{Euk.Cell})),$
$hasPart(o_{Euk.Cell}, f_4(o_{Euk.Cell})),$
$hasPart(o_{Euk.Cell}, f_5(o_{Euk.Cell})),$
$inside(f_4(o_{Euk.Cell}), f_5(o_{Euk.Cell})),$
$Euk.Ribosome(f_3(o_{Euk.Cell})),$
$Euk.Chromosome(f_4(o_{Euk.Cell})),$
$Nucleus(f_5(o_{Euk.Cell})),$

Moreover, there are pairwise inequality atoms between $o_{Euk.Cell}, f_3(o_{Euk.Cell}), f_4(o_{Euk.Cell}), f_5(o_{Euk.Cell})$ and between $o_{Euk.Cell}, f_0(o_{Euk.Cell}), f_1(o_{Euk.Cell}), f_2(o_{Euk.Cell})$.

If we next look at $\mathcal{I}_\mathcal{A}$, we will find that $[f_3(o_{Euk.Cell}])] = [f_1(o_{Euk.Cell})] = \{f_3(o_{Euk.Cell}), f_1(o_{Euk.Cell})\}$ holds, and likewise $[f_4(o_{Euk.Cell}])] = [f_2(o_{Euk.Cell})] = \{f_2(o_{Euk.Cell}), f_4(o_{Euk.Cell})\}$. Hence, the desired co-references have been established, e.g., the from $Cell$ inherited $Ribosome(f_1(o_{Euk.Cell}))$ is identified as being co-referential with the "local" $Euk.Ribosome(f_3(o_{Euk.Cell}))$.

The abridged strengthened GSKB Σ' then looks as follows:

$Cell(x) \Rightarrow$
$hasPart(x, f_1(x)), Ribosome(f_1(x)),$
$hasPart(x, f_2(x)), Chromosome(f_2(x)),$
$hasRoot(f_1(x), x), hasRoot(f_2(x), x),$
$pairwise\_disjoint(\{x, f_1(x), f_2(x)\})$
$EukaryoticCell(x) \Rightarrow Cell(x),$
$hasPart(x, f_3(x)), Ribosome(f_3(x)),$
$f_3(x) = f_1(x), f_4(x) = f_2(x),$
$hasPart(x, f_4(x)), Chromosome(f_4(x)),$
$hasPart(x, f_5(x)), Nucleus(f_5(x)),$
$inside(f_4(x), f_5(x)),$
$hasRoot(f_3(x), x), hasRoot(f_4(x), x),$
$hasRoot(f_5(x), x),$
$pairwise\_disjoint(\{x, f_3(x), f_4(x), f_5(x)\})$

$Ribosome(o_{Ribosome}), Chromosome(o_{Chromosome})$
$\ldots o_{Cell} \neq o_{EukaryoticCell} \ldots$ (axiom sets 2–4 from Def. 10)

We claim that we can decide the provenance problem for the strengthened GSKB Σ' syntactically as follows; also recall that in an admissible KB, the consequents do not contain redundant concept atoms:

Definition 11. *Syntactic provenance of atoms in Σ'. In a strengthened GSKB Σ', for $C \in concepts(\Sigma)$, let $\alpha \in \tau_{C,\Sigma'}$ be an atom:*

- *$\alpha = C(f(x))$ is inherited from D if $D(f_s(x)) \in \tau_{C,\Sigma'}$ with $D \in \{C\} \cup all\_superclasses(C, \Sigma')$ and $f'(f_s(x)) = f(x) \in \tau_{C,\Sigma'}$ with $C(f'(x)) \in \tau_{D,\Sigma'}$, and there is no more general class $SupD$ with $has\_superconcept(D, SupD)$ for which this is also the case.*
- *$\alpha = R(f_1, f_2)$ is inherited from D if $D(f_s(x)) \in \tau_{C,\Sigma'}$ with $D \in \{C\} \cup all\_superclasses(C, \Sigma')$ and $\{f'_1(f_s(x)) = f_1(x), f'_2(f_s(x)) = f_2(x)\} \subseteq \tau_{C,\Sigma'}$ with $R(f'_1, f'_2) \in \tau_{D,\Sigma'}$, and there is no more general class $SupD$ with $has\_superconcept(D, SupD)$ for which this is also the case.*

If α is not inherited from some concept, it is called local *to C.*

Looking at the example GSKB Σ', we see that the atoms $hasPart(x, f_3(x))$ are inherited from $Cell$, due to $f_3(x) = f_1(x)$, and $hasPart(x, f_4(x))$, due to $f_4(x) = f_2(x)$. Consequently, $hasPart(x, f_5(x)), Nucleus(f_5(x)), inside(f_4(x), f_5(x))$ are local to $EukaryoticCell$. Hence, for the original GSKB Σ, $hasPart(x, y_3)$ and $hasPart(x, y_4)$ were inherited from $Cell$, and $hasPart(x, y_5), Nucleus(y_5), inside(y_4, y_5)$ are local to $EukaryoticCell$.

We claim that we can decide the co-reference problem for the strengthened GSKB Σ' syntactically as follows:

Definition 12. *Syntactic co-reference of terms in Σ'. Two terms with $t_1 \in terms(C, \Sigma'), t_2 \in terms(D, \Sigma')$ are co-referential, if $t_1 = t_2 = x$, or $t_1(x) = t_2(t) \in \tau_{C,\Sigma'}$, or $t_2(x) = t_1(t) \in \tau_{D,\Sigma'}$ (note that $t = x$, or $t = f_s(x)$).*

Looking at the example GSKB Σ', we see that $f_3(x) = f_1(x)$ are co-referential and hence the $Ribosome$ in $Cell$ is the same as the $Euk.Ribosome$ in $EukaryoticCell$, and likewise for the $Chromosome$ due to $f_4(x) = f_2(x)$.

Lemma 3. *Syntactic provenance according to Def. 11 is sound and complete for deciding semantic provenance according to Def. 8. Syntactic co-reference according to Def. 12 is sound and complete for deciding semantic co-reference according to Def. 8.*

Proof. Soundness is immediate. Completeness is a straightforward too, as Skolem functions are not shared by different consequents in Σ', and Σ was admissible. Moreover, for two different Skolem functions f_i, f_j, with $i \neq j$, $f_i(t) = f_j(t)$ will hold for a certain term t in *all* models of Σ' if and only if this was explicitly enforced by an equality atom. Note that this also proves Proposition 3. □

We can generalize these results to the original GSKB Σ as follows. To check the provenance of $\tau_{C,\Sigma}$ we need to keep track during Skolemization which atom $\alpha' \in \tau_{C,\Sigma'}$ corresponds to α, and check the provenance of α' in Σ'. Likewise, to check to co-referentiality of two variables, let t_1 and t_2 be its corresponding (Skolem function) terms in the Skolemized versions. Now, y_1 and y_2 are co-referential in Σ iff t_1 and t_2 are co-referential in Σ'. Looking at the example GSKB Σ, we see that y_1 from $S1$ is co-referential with y_3 in $S23$ since $f_3(x) = f_1(x)$ in Σ', and y_2 from $S1$ is co-referential with y_4 in $S23$ due to $f_4(x) = f_2(x)$ in Σ'.

However, given that a GSKB may have more than one strengthened version, "to decide" should be understood in a *credulous* way here. Only in case a provenance information or co-reference holds in *all* strengthened GSKBs, this would be a *skeptical* conclusion; it is clear that all strengthened GSKBs can in principle be constructed, due to finiteness of $\mathcal{I}_{\mathcal{H}}$. We can hence present the main result of this paper as follows:

Corollary 1. *Given a strengthened GSKB Σ', we can decide the provenance and co-reference problem on a syntactic basis. For an admissible (input) GSKB Σ, we can decide the* credulous provenance and credulous co-reference problem *by constructing* a *strengthened GSKB Σ' first, and check there. The* skeptical provenance and skeptical co-reference problem *can be decided by constructing* all *strengthened GSKBs, and checking if a positive answer holds in all of them.*

4 Implementation & Evaluation

We have applied the algorithm on the AURA GSKB [Gunning, D. and Chaudhri, V. K. et al., 2010], which currently contains 5662 graph-structured concepts, with a number of 141909 atoms. The GSKB strengthening algorithm identifies 116442 of these triples as inherited, which amounts to $\approx$ 82 %. Moreover, 22667 equality atoms were hypothesized, and 2858 cases of variable specialization in a subconcept were found (e.g., the $Chromosome$ in $Cell$ got specialized to $Euk.Chromosome$ in $EukaryoticCell$). The required runtime is $\approx$ 15 hours on a Intel Xeon E5607 2.2 GHz PC with 8 GB of RAM running Windows 7 64 Bit with a 64 Bit Common Lisp implementation.

For the sake of implementability, the implemented algorithm does not really construct a Skolemized version of Σ and neither its Herbrand model. Rather, it works directly on the level of formulas, by considering graph morphisms between concept graphs. The operations performed on these graphs can be understood as operations on the minimal Herbrand model. The actual deterministic implementation of the non-deterministically specified algorithm requires search in order to maximize the number of inequality assertions. The search space (number of possible individual mergings) is tremendously large, and can only be tackled by employing a large number of search heuristics, including time outs and pruning of the search space. The achieved progress is good enough from a practical point of view (i.e., 82 % of all AURA atoms in the GSKB are identified as being inherited).

We have given the constructed strengthened GSKBs to the subject matter experts in the knowledge factory [Chaudhri, V. and Dinesh, N., et al, 2011] and they have evaluated the quality of the rewritten GSKB by spot-checking the provenance

of atoms. So far, the results are very encouraging, and only a small percentage of atoms is falsely given a local provenance (they should be inherited). This is caused by the heuristics which prune the search space. The identified co-references were correct to a very large extent.

5 Related Work, Conclusions, and Outlook

The problem of anaphora resolution has been studied in the NLG literature to some extent; for example, [Carpenter, 1994; Cohen, 2007] use default rules to hypothesize equality assertions between variables (NLG referents).

The work of [Chaudhri and Tran, 2012] uses answer set programming (ASP) to add so-called UMap atoms to the GSKB specified as an ASP program. This is similar to our approach here, but we are using a standard first-order semantics and equality atoms. Note that in ASP programs, all constants are distinct by definition. Also, they rely on a more complicated axiomatic system encoded as ASP rules to prevent unwanted unifications, whereas our approach is in principle model-theoretic and hence does not require additional axiomatic framework. Scalability of the ASP approach has not been demonstrated yet.

As mentioned, the reasoning system KM uses a so-called unification mechanism in order to tackle the same problem [Clark and Porter, 1997], but a lack of a formal framework makes it difficult to understand and to debug. One problem with the approach is that unifications are not represented explicitly as (equality or Umap) atoms in the GSKB. Hence, retraction and comprehensibility is very difficult and time consuming, especially in case heuristically generated bad unifications have to be revised manually later. Dealing with the effects of such destructive heuristic unifications ha time consuming in the AURA project.

It is well-known that the modeling of graph structures is challenging in description logic (DL), as derivations from the tree-model property usually result in decidability problems [Vardi, 1996] which can often only be regained by imposing severe artificial modeling restrictions. Although some progress has been made on modeling graph structures with DLs [Motik *et al.*, 2009], those extensions are still too restricted to be employed in a large-scale modeling effort such as AURA. Our experience is that graph structures are central to biology, and approximating them by trees results in coarse models. Our framework allows us to express the graph structures truthfully, but comes with other restrictions, too. The AURA system and the actual implementation of the algorithm covers additional expressive means that we have not formally reconstructed yet (transitive, functional, hierarchical relations, number restrictions, and disjointness axioms, cyclical GSKBs, etc.) This is future work. To the best of our knowledge, there is no body of work in the DL community that provides answers to the problems addressed in this paper, namely, how to construct a strengthened GSKB by hypothesizing equality atoms. To the best of our knowledge, no abduction or hypothesization algorithm has every successfully been applied to such a large-scale GSKB.

The strengthened GSKB is also the basis for a couple of AURA knowledge base exports in SILK, ASP, FOPL, and TPTP FOF syntax. We also have an OWL2 translation, in which the graphs are approximated by trees (using unraveling). The equality atoms are omitted. These GSKBs can be downloaded on request from `http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html` (a password and a license agreement are required). It is interesting to note that our second smallest OWL2 export (we have several variants, in increasing complexity) cannot be checked for consistency by contemporary DL reasoners. However, our algorithm has identified 116442 inherited atoms, hypothesized 22667 equality atoms, and found 2858 cases of variable specialization in 18 hours. Compared to the performance of DL reasoners on the AURA KB, that is a success story.

References

[Baader *et al.*, 2003] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[Carpenter, 1994] Bob Carpenter. Skeptical and credulous default unification with applications to templates and inheritance. In *Inheritance, Defaults and the Lexicon*, pages 13–37. Cambridge University Press, 1994.

[Chaudhri and Tran, 2012] Vinay K. Chaudhri and Son Cao Tran. Specifying and reasoning with under-specified knowledge base. In *International Conference on Knowledge Representation and Reasoning*, 2012.

[Chaudhri, V. and Dinesh, N., et al, 2011] Chaudhri, V. and Dinesh, N., et al. Preliminary steps towards a knowledge factory process. In *The Sixth International Conference on Knowledge Capture*, 2011.

[Clark and Porter, 1997] Peter Clark and Bruce Porter. Building concept representations from reusable components. In *Proceedings of AAAI*. AAAI Press, 1997.

[Cohen, 2007] Ariel Cohen. Anaphora resolution as equality by default. In Antnio Horta Branco, editor, *Anaphora: Analysis, Algorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*.

[Gunning, D. and Chaudhri, V. K. et al., 2010] Gunning, D. and Chaudhri, V. K. et al. Project halo update progress toward digital aristotle. *AI Magazine*, Fall 2010.

[Hedman, 2004] Shawn Hedman. *A First Course in Logic: An Introduction to Model Theory, Proof Theory, Computability, and Complexity*. Oxford Texts in Logic, 2004.

[Motik *et al.*, 2009] Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, and Ulrike Sattler. Representing ontologies using description logics, description graphs, and rules. *Artif. Intell.*, 173(14):1275–1309, September 2009.

[Vardi, 1996] VARDI, M. Y. 1996. Why is modal logic so robustly decidable? In *Descriptive Complexity and Finite Models, Proceedings of a DIMACS Workshop*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 31, 1996.